

Multivariate Biomedical Signal Processing

Iead Rezek

Department of Engineering Science; University of Oxford,
Parks Road, Oxford, OX1 3PJ, UK

December 7, 2005

1 Introduction

Suppose one is interested in modelling the statistical properties not of a single variable but of an entire set of variables. To focus on the statistics of every variable separately without considering the other variables would lead to incorrect results because interdependencies between the variables have been ignored. Consider, for example modelling heart rate and respiration separately. These two physiological control systems are clearly highly interdependent due to the nervous system, to name just one of many coupling factors.

Statistically, the interdependence between variables can be expressed through their probability distributions. Given two parameters, say μ_1 and μ_2 , their joint probability distribution is denoted by $p(\mu_1, \mu_2)$. If and only if the two parameters are independent (e.g. there is no hidden factor affecting their values) will their joint distribution factorise

$$p(\mu_1, \mu_2) = p(\mu_1)p(\mu_2) \quad . \quad (1)$$

If independence is not guaranteed, the only split of the joint distribution allowed by the laws of probability is one consisting of a univariate and a conditional distribution, e.g.

$$p(\mu_1, \mu_2) = p(\mu_1)p(\mu_2|\mu_1) \quad . \quad (2)$$

Thus, while I am allowed to study the statistics of one parameter in isolation, $p(\mu_1)$, I must consider the parameter's affect on the statistics of the second, $p(\mu_2|\mu_1)$ for accurate description of the parameters' joint statistics. However, even if the independence assumption is valid, statistical modelling becomes much simpler if all parameters can be estimated with one single multivariate distribution.

The distribution on the left-hand side of equation (2) has a two parameter argument, i.e. it is a multivariate distribution. What is evident from equation (2) is that rather than studying a set of distributions one can simplify the task somewhat by investigating only one multivariate distribution. From this multivariate distribution all univariate densities can then be extracted¹.

2 Multivariate Linear Models

Multivariate statistics arise in the most simple modelling tasks, such as a linear model. When faced with a new signal and the data generating model is completely unknown it is natural to begin the analysis with this simplest of possible models.

Suppose a set of noisy signal observations of length T is described by the linear model

$$d_t = \sum_{k=1}^M b_k g_{t,k} + e_t \quad \forall t = 1, 2, \dots, T. \quad (3)$$

¹In fact, $p(\mu_2|\mu_1)$ in equation (2) is also a multivariate distribution but the issue of simplifying modelling multiple distributions still remains.

The model has a set of M basis functions, $g_{t,k}$, which are a function of time but their form is assumed to be known or fixed, e.g. harmonic functions evaluated at time t . The basis functions are linearly weighted by their M coefficients, b_k , which are to be estimated. Last, there is also noise, e_t , which corrupts the observations and necessitates the statistical analysis.

To simplify notation from this point on, I stack all parameters b_1, \dots, b_M to form a vector, \mathbf{b} . Thus, the joint distribution is now denoted by $p(\mathbf{b})$. The same stacking can be done for all data points d_1, \dots, d_T to give \mathbf{d} and for the noise samples, $\mathbf{e} = \text{vec}(e_1, \dots, e_T)$. For each observation sample, t , there is a row of M basis functions $g_{t,1}, \dots, g_{t,M}$ which can be stacked into a matrix \mathbf{G} to form

$$\mathbf{G} = \begin{pmatrix} g_{1,1} & \cdots & g_{1,M} \\ & \ddots & \\ g_{T,1} & \cdots & g_{T,M} \end{pmatrix} \quad (4)$$

In vector notation, the linear model of equation (3) can now be written compactly in a multivariate form

$$\mathbf{d} = \mathbf{G}\mathbf{b} + \mathbf{e} \quad (5)$$

Having established the model by which the data is generated the next step is to state the statistical properties or assumptions. I will begin by making the common assumption that the noise samples, \mathbf{e} , are identically and independently drawn from a Gaussian distribution with zero mean and variance σ^2

$$p(e_t|\sigma) = \mathcal{N}(0, \sigma^2) \quad \forall t = 1, \dots, T \quad (6)$$

Thus, \mathbf{e} follows a T -variate Gaussian distribution which has a zero mean vector and a covariance matrix that has σ^2 as its only elements along the diagonal

$$p(\mathbf{e}|\sigma) = (2\pi\sigma^2)^{-\frac{T}{2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{e}^\top\mathbf{e}\right\} \quad (7)$$

where \top is the transpose operator. By inserting the model

$$\mathbf{e} = \mathbf{d} - \mathbf{G}\mathbf{b} \quad (8)$$

into equation (7) I obtain the data likelihood, that is probability of the data, \mathbf{d} , conditioned on the parameters for the noise, σ , and the model, \mathbf{b} ,

$$p(\mathbf{d}|\mathbf{b}, \sigma) = (2\pi\sigma^2)^{-\frac{T}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{d} - \mathbf{G}\mathbf{b})^\top(\mathbf{d} - \mathbf{G}\mathbf{b})\right\} \quad (9)$$

By assuming the noise to be identically and independently distributed yet parameterising everything using vector notation I made use of the mathematical convenience of multivariate densities.

3 Multivariate Maximum Likelihood Inference

The task of statistical inference is to learn the values of the parameters for the noise, σ , and the model, \mathbf{b} . The choice will fall on the values that "best" describe the observations and the data likelihood provides one very obvious definition of "best": find the values, $\hat{\mathbf{b}}, \hat{\sigma}$ for \mathbf{b} and σ that maximise the conditional probability, $p(\mathbf{d}|\mathbf{b}, \sigma)$, of the observations. For mathematical convenience and without any loss of generality, the maximisation of the likelihood is often performed on its logarithm

$$\hat{\mathbf{b}}, \hat{\sigma} = \arg \max \log p(\mathbf{d}|\mathbf{b}, \sigma) \quad (10)$$

For the multivariate linear model, the log-likelihood to be maximised is simply

$$\log(p(\mathbf{d}|\mathbf{b}, \sigma)) = -\frac{T}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{d} - \mathbf{G}\mathbf{b})^\top(\mathbf{d} - \mathbf{G}\mathbf{b}). \quad (11)$$

Maximisation of the multivariate log-likelihood naturally requires matrix differential calculus, which has its specifically defined operations and are described in some excellent textbooks and

reference manuals [1, 2, 3]. Rather than differentiating T terms with respect to the set of scalar values b_1, \dots, b_M and σ , in the multivariate case (11) only one term must be differentiated with respect to the vector \mathbf{b} and the scalar σ . Thus, setting

$$\begin{aligned} 0 &= \frac{\partial}{\partial \mathbf{b}} \log(p(\mathbf{d}|\mathbf{b}, \sigma)) \\ &= \frac{\partial}{\partial \mathbf{b}} \left(-\frac{1}{2\sigma^2} (\mathbf{d} - \mathbf{G}\mathbf{b})^\top (\mathbf{d} - \mathbf{G}\mathbf{b}) \right) \\ &= -\frac{1}{2\sigma^2} \left(2\mathbf{G}^\top (\mathbf{d} - \mathbf{G}\mathbf{b}) \right) \end{aligned} \tag{12}$$

and solving for \mathbf{b} one obtains

$$\mathbf{b} = (\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{G}^\top \mathbf{d} \tag{13}$$

Maximisation with respect to σ follows the standard differential calculus rules. However, if the model included a full noise covariance matrix, Σ

$$(\mathbf{d} - \mathbf{G}\mathbf{b})^\top \Sigma^{-1} (\mathbf{d} - \mathbf{G}\mathbf{b}),$$

maximisation with respect to Σ gives as the maximum likelihood solution for Σ

$$\Sigma = (\mathbf{d} - \mathbf{G}\mathbf{b})(\mathbf{d} - \mathbf{G}\mathbf{b})^\top.$$

4 Multivariate Bayesian Inference

Unlike the maximum likelihood approach Bayesian inference takes a different approach to parameter estimation. Rather than maximising, say $p(D|\theta)$, the probability of the observations given some parameters, θ , the Bayesian approach also weights the likelihood by the probability of the parameter values occurring *a priori*, $p(\theta)$. Thus, the question Bayesian inference answers is: What is the revised or posterior distribution of the parameters given the newly obtained observations, $p(\theta|D)$. Thus, a Bayesian applies Bayes' rule

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}. \tag{14}$$

I will not delve in the pros and cons of either approach but refer the interested reader to the literature [4, 5, 6]. As will become apparent later in this exposition, compared to the maximum likelihood method which primarily required differential calculus Bayesian inference requires mostly integration.

The posterior distribution, left-hand side of equation (14), is more often than not a multivariate distribution describing the joint statistics of all the model parameters. For the earlier linear model example, the posterior will have \mathbf{b} and σ as its arguments. Not only are such high dimensional distributions impossible to visualise but it is very likely that of interest is only a subset of the arguments of the posterior distribution. To obtain these, the full joint must be marginalised, i.e. integrated with respect to the irrelevant parameters (known as nuisance parameters in Bayesian terminology). In the case of the linear model above a possible choice of marginal distributions includes the marginal of \mathbf{b}

$$p(\mathbf{b}|\mathbf{d}) = \int p(\mathbf{b}, \sigma|\mathbf{d}) d\sigma \tag{15}$$

and the marginal noise distribution

$$p(\sigma|\mathbf{d}) = \int p(\mathbf{b}, \sigma|\mathbf{d}) d\mathbf{b}. \tag{16}$$

To compute the joint distribution in the integrand of equations (15) and (16) we use Bayes rule (14) which requires a likelihood function, equation (9), and a prior distribution. The two

priors that need to be defined for the linear model are the noise prior, $p(\sigma)$, and the coefficient prior, $p(\mathbf{b})$. For the noise I will use assume a prior of the form $\frac{1}{\sigma}$. The prior is called Jeffrey's prior, can be found in all the relevant literature, and effectively simply encodes a uniform distribution assumption in the positive real space. For the coefficients, \mathbf{b} , I will also assume that they are uniformly distributed within some range, i.e. there is an equally valued probability for all values of \mathbf{b} , say $p(\mathbf{b}) = q$. Thus, the full posterior distribution is given as

$$p(\sigma, \mathbf{b}|\mathbf{d}) \propto p(\mathbf{d}|\sigma, \mathbf{b}) \frac{1}{\sigma}, \quad (17)$$

where q and the normalisation constant of $p(\sigma) \propto \frac{1}{\sigma}$ have been absorbed in the posterior's normalisation constant. More explicitly the posterior reads

$$p(\sigma, \mathbf{b}|\mathbf{d}) \propto (2\pi\sigma^2)^{-\frac{T}{2}} \exp \left[-\frac{(\mathbf{d} - \mathbf{G}\mathbf{b})^\top (\mathbf{d} - \mathbf{G}\mathbf{b})}{2\sigma^2} \right] \frac{1}{\sigma}. \quad (18)$$

As reasoned earlier, the full posterior distribution (18) has a high dimensionality and most modelling tasks require marginalisation of the distribution over subsets of its dimensions. Before proceeding with the marginalisation, some term collection within the exponent of the posterior distribution is useful and is detailed in appendix A. The modified posterior of equation (18) then takes the following form

$$p(\sigma, \mathbf{b}|\mathbf{d}) \propto (2\pi\sigma^2)^{-\frac{T}{2}} \exp \left[\frac{(\mathbf{b} - \hat{\mathbf{b}})^\top \mathbf{G}^\top \mathbf{G} (\mathbf{b} - \hat{\mathbf{b}}) - \mathbf{d}^\top \mathbf{G} \mathbf{G}^\dagger \mathbf{d} + \mathbf{d}^\top \mathbf{d}}{-2\sigma^2} \right] \frac{1}{\sigma}, \quad (19)$$

where $\hat{\mathbf{b}}$ was substituted for $\mathbf{G}^\dagger \mathbf{d}$ and \mathbf{G}^\dagger for $(\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{G}^\top$. The reason for this form is that a part of the posterior now has the form of a Gaussian distribution in \mathbf{b} with some additional terms which do not involve \mathbf{b} and thus can be absorbed into the normalising constant.

4.0.1 The Conditional Coefficient Distribution

Before proceeding with the computation of the marginal distributions, it is instructive to study the posterior distribution for the case that the noise standard deviation, σ , is been known, say it has the value $\hat{\sigma}$. If it is known it is fixed and thus can be socked up in the the normalising constant. Then it becomes clearer to see that the the posterior is a Gaussian distribution in the parameters \mathbf{b} . Thus, the expectation of \mathbf{b} is given by

$$\begin{aligned} \langle \mathbf{b} \rangle &= \frac{1}{\int p(\mathbf{b}, \hat{\sigma}^2|\mathbf{d}) d\mathbf{b}} \int \mathbf{b} p(\mathbf{b}, \hat{\sigma}^2|\mathbf{d}) d\mathbf{b} \\ &= \frac{1}{\int p(\mathbf{b}, \hat{\sigma}^2|\mathbf{d}) d\mathbf{b}} \int \mathbf{b} \exp \left[\frac{(\mathbf{b} - \hat{\mathbf{b}})^\top \mathbf{G}^\top \mathbf{G} (\mathbf{b} - \hat{\mathbf{b}})}{-2\hat{\sigma}^2} \right] d\mathbf{b} \\ &= (\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{G}^\top \mathbf{d}. \end{aligned} \quad (20)$$

The second central moment is given by

$$\langle \mathbf{b}\mathbf{b}^\top \rangle_c = \hat{\sigma}^2 (\mathbf{G}^\top \mathbf{G})^{-1}, \quad (21)$$

and the non-central second central moment by

$$\langle \mathbf{b}\mathbf{b}^\top \rangle = \langle \mathbf{b}\mathbf{b}^\top \rangle_c + \langle \mathbf{b} \rangle \langle \mathbf{b} \rangle^\top. \quad (22)$$

4.0.2 The Marginal Coefficient Distribution

To obtain the marginal distribution of the model coefficients, \mathbf{b} , I need to integrate the posterior distribution (19) with respect to σ

$$p(\sigma, \mathbf{b}|\mathbf{d}) \propto \int_0^\infty (2\pi\sigma^2)^{-\frac{T}{2}} \exp \left[\frac{(\mathbf{b} - \hat{\mathbf{b}})^\top \mathbf{G}^\top \mathbf{G} (\mathbf{b} - \hat{\mathbf{b}}) + S}{-2\sigma^2} \right] \frac{1}{\sigma} d\sigma,$$

where S substitutes for the term $\mathbf{d}^\top \mathbf{d} - \mathbf{d}^\top \mathbf{G} \mathbf{G}^\dagger \mathbf{d}$, which are all independent of \mathbf{b} . Thus, the actual integration steps are

$$\begin{aligned} p(\mathbf{b}|\mathbf{d}, \sigma) &\propto \int_0^\infty (2\pi\sigma^2)^{-\frac{T}{2}} \sigma^{-1} \exp\left[\frac{(\mathbf{b} - \hat{\mathbf{b}})^\top \mathbf{G}^\top \mathbf{G} (\mathbf{b} - \hat{\mathbf{b}}) + S}{2\sigma^2}\right] d\sigma \\ &\propto (2\pi)^{-\frac{T}{2}} \int_0^\infty \sigma^{-T+2} \exp\left[\frac{(\mathbf{b} - \hat{\mathbf{b}})^\top S^{-1} \mathbf{G}^\top \mathbf{G} (\mathbf{b} - \hat{\mathbf{b}}) + 1}{2\sigma^2}\right] d\sigma \\ &\propto (2\pi)^{-\frac{T}{2}} \int_0^\infty x^{-(\frac{T-2}{2})} \exp\left[\left((\mathbf{b} - \hat{\mathbf{b}})^\top S^{-1} \mathbf{G}^\top \mathbf{G} (\mathbf{b} - \hat{\mathbf{b}}) + 1\right) x\right] dx \end{aligned}$$

where $\sigma^2 = \frac{1}{x}$ and $d\sigma = -1/2\sigma^3 dx$. Thus, the marginal in the coefficient is given by

$$p(\mathbf{b}|\mathbf{d}, \sigma) \propto \frac{\Gamma(\frac{T}{2})}{\left((\mathbf{b} - \hat{\mathbf{b}})^\top S^{-1} \mathbf{G}^\top \mathbf{G} (\mathbf{b} - \hat{\mathbf{b}}) + 1\right)^{\frac{T}{2}}}. \quad (23)$$

This is a multivariate Student-t distribution with $T - 1$ degrees of freedom, mean $\hat{\mathbf{b}}$ and a scaling of $S(\mathbf{G}^\top \mathbf{G})^{-1}$.

4.0.3 The Marginal Noise Distribution

Apart from the model coefficients the other important marginal distribution of the model describes the noise properties. This requires integrating the posterior distribution (19) with respect to \mathbf{b}

$$p(\sigma, \mathbf{b}|\mathbf{d}) \propto \int_0^\infty (2\pi\sigma^2)^{-\frac{T}{2}} \exp\left[\frac{(\mathbf{b} - \hat{\mathbf{b}})^\top \mathbf{G}^\top \mathbf{G} (\mathbf{b} - \hat{\mathbf{b}}) + S}{-2\sigma^2}\right] \frac{1}{\sigma} d\mathbf{b}, \quad (24)$$

which is a multiple integral in all elements of the vector \mathbf{b} . The integration given the from above is difficult to perform for all cross terms of $\mathbf{b}^\top \mathbf{G}^\top \mathbf{G} \mathbf{b}$ in the integrand need to to be considered. However, integration will be simplified if all these cross terms can be eliminated and this can be achieved by transforming $\mathbf{G}^\top \mathbf{G}$, using an eigen-decomposition,

$$\mathbf{G}^\top \mathbf{G} = \mathbf{P} \mathbf{\Lambda} \mathbf{P}^\top \quad (25)$$

into two matrices, one of which, $\mathbf{\Lambda}$, is diagonal with elements λ_i $i = 1, \dots, M$. The second matrix is an orthogonal matrix (for the purposes of this exposition it can be considered a rotation matrix), which implies that the norm of \mathbf{P} , $|\mathbf{P}| = 1$ and that its inverse $\mathbf{P}^{-1} = \mathbf{P}^\top$. Substituting the decomposition (25) for $\mathbf{G}^\top \mathbf{G}$, the first exponent term in the integrand of (24) can be rewritten to

$$\begin{aligned} (\mathbf{b} - \hat{\mathbf{b}})^\top \mathbf{G}^\top \mathbf{G} (\mathbf{b} - \hat{\mathbf{b}}) &= (\mathbf{a} - \hat{\mathbf{a}})^\top \mathbf{\Lambda} (\mathbf{a} - \hat{\mathbf{a}}) \\ &= \sum_i^M \lambda_i (a_i - \hat{a}_i)^2, \end{aligned}$$

where $\mathbf{a} = \mathbf{P}\mathbf{b}$ and $\hat{\mathbf{a}} = \mathbf{P}\hat{\mathbf{b}}$. Thus, the exponent in the integrand has been transformed and now involves only independent terms of a_i . Before proceeding with the integral, the change of variables needs to be completed by setting $d\mathbf{b} = |\mathbf{P}|d\mathbf{a} = d\mathbf{a}$. Now integration is greatly simplified as the

following steps illustrate

$$\begin{aligned}
\int \exp \left[-\frac{1}{2\sigma^2} \sum_i^M \lambda_i (a_i - \hat{a}_i)^2 \right] d\mathbf{a} &= \prod_i^M \int \exp \left[-\frac{1}{2\sigma^2} \lambda_i (a_i - \hat{a}_i)^2 \right] da_i \\
&= \prod_i^M \sqrt{\frac{2\pi\sigma^2}{\lambda_i}} \\
&= \frac{(2\pi\sigma^2)^{\frac{M}{2}}}{\sqrt{\prod_i \lambda_i}} \\
&= \frac{(2\pi\sigma^2)^{\frac{M}{2}}}{\sqrt{|\mathbf{G}^T \mathbf{G}|}}.
\end{aligned}$$

The last step follows directly from $|\mathbf{G}^T \mathbf{G}| = |\mathbf{P} \mathbf{\Lambda} \mathbf{P}^T| = |\mathbf{P}| |\mathbf{\Lambda}| |\mathbf{P}| = |\mathbf{\Lambda}| = \prod_i^M \lambda_i$.

Finally, we arrive at the full marginal over σ which is given by

$$p(\sigma|\mathbf{d}) = \frac{(2\pi\sigma^2)^{-\frac{T-M}{2}}}{\sqrt{|\mathbf{G}^T \mathbf{G}|}} \exp \left[-\frac{S}{2\sigma^2} \right], \quad (26)$$

were $S = \mathbf{d}^T \mathbf{d} - \mathbf{d}^T \mathbf{G} \mathbf{G}^\dagger \mathbf{d}$. This is an inverse Chi-Square distribution in σ^2 , scaled by S and having $T - M$ degrees for freedom

$$p(\sigma|\mathbf{d}) \propto (\mathbf{d}^T \mathbf{d} - \mathbf{d}^T \mathbf{G} \mathbf{G}^\dagger \mathbf{d}) \chi_{T-M}^{-2}. \quad (27)$$

Using standard formulae for the moments of this distribution [7], we obtain an estimate of the expected noise variance

$$\langle \sigma^2 \rangle = \frac{(\mathbf{d}^T \mathbf{d} - \mathbf{d}^T \mathbf{G} \mathbf{G}^\dagger \mathbf{d})}{T - M - 2}, \quad (28)$$

for $T - M > 2$.

5 Examples

The simplicity of linear models means that their application is ubiquitous. The same can be said about their multivariate counterpart. The harmonic model and the autoregressive models are probably the most popular types of models. To give the reader a feel for the workings of the model I will illustrate the harmonic model's use and Bayesian inference on simulated data. As a real-world signal analysis example I will apply the AR model to the analysis of an 8 hour sleep electroencephalogram (EEG) recording.

5.1 Harmonic Model

Suppose the basis functions g of the linear model (3) are chosen to be sine and cosine functions

$$d_t = \sum_{k=1}^M b_{2k-1} \cos(\omega t) + b_{2k} \sin(\omega t) + e_t. \quad (29)$$

To obtain the multivariate form we stack, as describe earlier. Thus, for example, the matrix \mathbf{G} also called the design matrix consists of two columns of sine and cosine functions of a particular fixed frequency ω and evaluated at T time points

$$\mathbf{G} = \begin{pmatrix} \cos(\omega 1) & \sin(\omega 1) \\ \vdots & \vdots \\ \cos(\omega T) & \sin(\omega T) \end{pmatrix}. \quad (30)$$

It should come as no surprise that, due to the orthogonality property of the harmonic functions, the matrix product $\mathbf{G}^\top \mathbf{G}$ is diagonal. More precisely described in [8] for large sample sizes,

$$\mathbf{G}^\top \mathbf{G} = \frac{T}{2} \mathbf{I}_M$$

where \mathbf{I}_M is the $M \times M$ identity matrix.

Thus the expectation of the coefficients, using the earlier Bayesian analysis, is given by

$$\langle \mathbf{b} \rangle \approx \frac{2}{T} \mathbf{G}^\top \mathbf{d} = \frac{2}{T} \begin{pmatrix} \sum_{t=1}^T d_t \cos(\omega t) \\ \sum_{t=1}^T d_t \sin(\omega t) \end{pmatrix} \stackrel{def}{=} \begin{pmatrix} R(\omega) \\ I(\omega) \end{pmatrix}. \quad (31)$$

The last equation is only a substitution for notational convenience. The second central moment of the model coefficients is given by

$$\langle \mathbf{b} \mathbf{b}^\top \rangle_c = \sigma^2 (\mathbf{G}^\top \mathbf{G})^{-1}, \quad (32)$$

and the (non-central) second central moment by

$$\langle \mathbf{b} \mathbf{b}^\top \rangle = \langle \mathbf{b} \mathbf{b}^\top \rangle_c + \langle \mathbf{b} \rangle \langle \mathbf{b} \rangle^\top, \quad (33)$$

which for \mathbf{G} given by (30) yields

$$\langle \mathbf{b} \mathbf{b}^\top \rangle = \sigma^2 \frac{2}{T} + \left(\frac{2}{T} \right)^2 \begin{pmatrix} R^2(\omega) & R(\omega)I(\omega) \\ R(\omega)I(\omega) & I^2(\omega) \end{pmatrix}. \quad (34)$$

If we are only interested in the power spectrum, which is given by the sum of the squared coefficients \mathbf{b} ,

$$\langle b_i^2 + b_j^2 \rangle = \text{tr} \langle \mathbf{b} \mathbf{b}^\top \rangle = \frac{4}{T} (\sigma^2 + C(\omega)), \quad (35)$$

where $C(\omega) = \frac{1}{N} (R^2(\omega) + I^2(\omega))$ is known as the Schuster periodogram and closely related to the discrete Fourier transform.

An example simulated single sine wave is shown in figure 1. The harmonic, which had unit amplitude and a frequency of 1Hz, was sampled at 10Hz for 10 seconds. The added noise consisted of samples drawn from a zero mean and unit variance Gaussian distribution. The marginal Student-t distribution of the coefficient is depicted below the time series trace in figure 1. It peaks at roughly 0.83, where the loss in magnitude, $b < 1$, is due to noise. Although Student-t, the distribution has an almost Gaussian shape which is a known characteristic of Student-t distribution in cases of large sample sizes.

The numerator in Bayes' rule (14) can be used as an indicator for model fit (see [9] for instructive explanation). Roughly, the larger its value, the higher the probability of the observations and thus model fit. To evaluate the numerator, sometimes referred to as evidence, requires integrating out *all* model parameters and an illustration of this is given in [10]. What will affect the value of the evidence, in the harmonic model, is the frequency of the sine function: the basis function frequency matching the true signal frequency gives the highest model fit compared to any other basis function frequency. An example of this is shown in figure 2, together with a discrete Fourier transform using conventional FFT of the sine wave. Contrasting both spectra it is clear that the conventional Fourier model has multiple spurious spikes which result from modelling the noise in the signal. By construction, the Fourier transform is deterministic and thus considers noise, too, to be information worth modelling.

5.2 Autoregressive Model for Sleep EEG

As a real world example of modelling using multivariate linear models consider inference on electroencephalographic recordings taken from a subject during 8 hours of sleep. An example section is shown in figure 3 One of the features of sleep EEG is that it changes its properties depending on the sleep's depth. For instance, deep sleep correlates with slow wave forms in the EEG, whilst

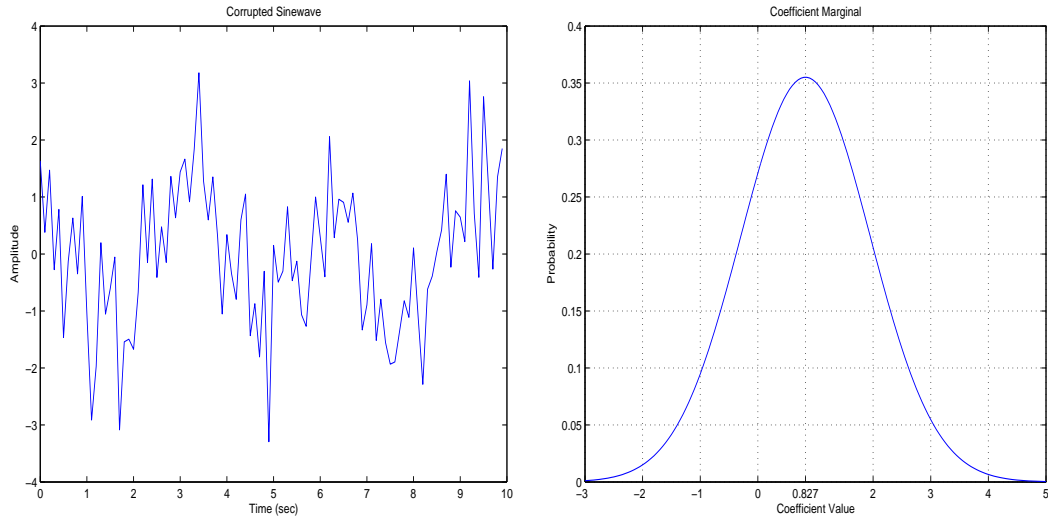


Figure 1: Simulated noisy sine wave (top) and its corresponding marginal coefficient distribution

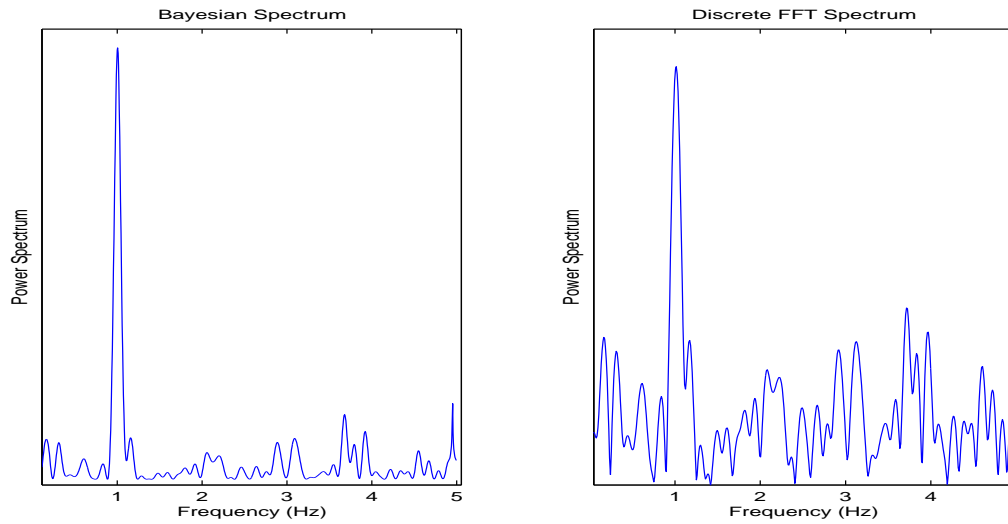


Figure 2: Left: Evidence $P(\mathbf{d})$ of the sine wave time series for different basis function frequencies; Right: Periodogram

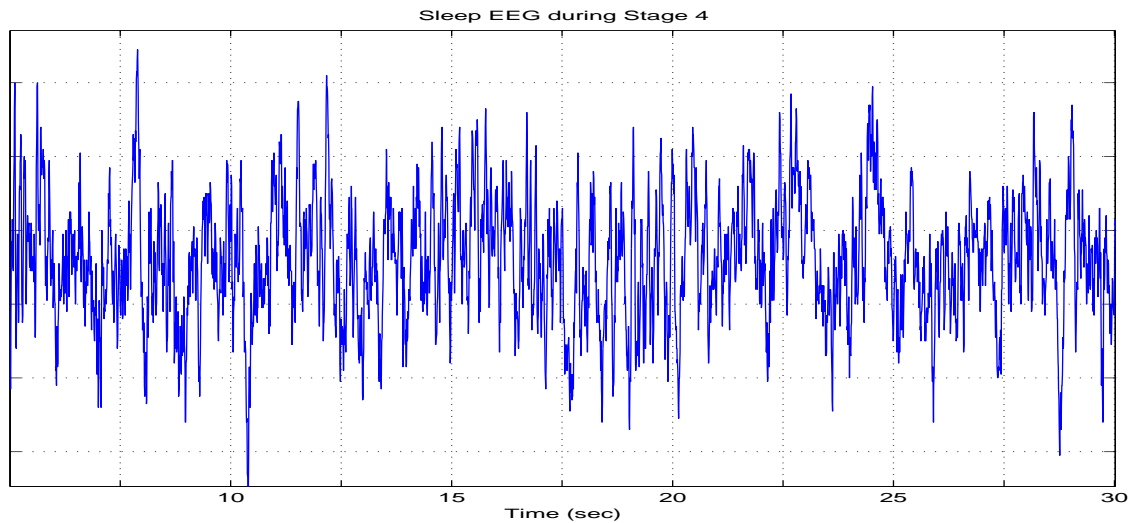


Figure 3: Thirty second section of an EEG recording during Sleep

REM sleep is peppered with fast signal dynamics. It is therefore of interest to related the EEG to sleep stage and thus obtain a full sleep profile of a patient, which can then be used for further diagnostic purposes.

As this time series cannot be assumed stationary for its full duration, I have segmented the EEG recording into 5sec long segments, within each I fit an 6^{th} -order Bayesian autoregressive model. The AR model's "basis functions", g , (they are not usually named as such) of the linear model (3) are a set of past observations

$$d_t = \sum_{k=1}^M b_k d_{t-k} + e_t. \quad (36)$$

In its multivariate form the design matrix \mathbf{G} consists merely stacked past observations

$$\mathbf{G} = \begin{pmatrix} d_M & \cdots & d_1 \\ \vdots & \ddots & \vdots \\ d_{T-1} & \cdots & d_{T-M} \end{pmatrix}. \quad (37)$$

and $\mathbf{d} = [d_{M+1}, \dots, d_T]^T$.

The expected values of AR coefficients for each segment is shown in figure 4. The patterns in the coefficient value plot show a correlation with the underneath plotted hypnogram. The Hypnogram is the sleep stage as classified by a human observer. How strong the correlation is can be seen after clustering all expected coefficient values and comparing the cluster allocation of each to the hypnogram, figure 5. The overall pattern has been captured but estimated cluster labels correspond to multiple manual classification labels.

6 Extensions

6.1 Extensions to Multivariate Models

The model in the preceding sections has, for illustrative purposes, focused on modelling a simple time series using multivariate description. Thus, discrete time was the dimension of the multivariate model. More typical is the use of multivariate models for simultaneous modelling multiple

observation vectors or signals, such as multi-channel EEG recordings. Such a linear model could take the form

$$\mathbf{d}_t = \sum_{k=1}^M \mathbf{b}_k \mathbf{G}_{k,t} + \mathbf{e}_t \quad \forall t = 1, 2, \dots, T. \quad (38)$$

Thus, the number of channels corresponds to the dimension of the multivariate model, leaving time to be treated separately. By suitably stacking multiple channels and time, a new multivariate model can be constructed that is identical to the model (5) - The structure of the design matrix would naturally be dependent on the model and the stacking. If one is to keep the form of the model then modelling requires inference with matrix-variate densities [11]. This can be desirable in order to retain the structure of the matrix [12]. For example, one may want to model evoked potentials from several EEG recording sites in their entire length. Thus, the commonalities between sites are modelled by one set of parameters whilst the time course is modelled by another set of parameters, and both sets have different prior assumptions.

The next step up in complexity is to assume that each observation is governed by its own model (and set of parameters) but that parameters of the models are subject to joint control by a latent model. These kinds of models, known as coupled latent space models, are far less common, but have been used to model relationships, for instance between heart-rate and respiration [13, 14].

6.2 Approximate Bayesian Inference

The Bayesian framework provides a set of a very powerful and very popular mechanisms[15]. From a biomedical point of view its chief advantage is that it leads to distributions for all parameters. The implication is that error-bars can be drawn to indicate the degree of uncertainty in the model. Armed with this information a clinician can make substantially more balanced decisions.

Unfortunately, multivariate integrals (univariate integrals, too, for that matter) are not always analytic so that explicit solutions cannot be found. In fact, analyticity is guaranteed only in a limited class of models known as conjugate models [16]. When the integral is no longer analytic, the only alternative is to use numerical integration methods, among them the Markov Chain Monte Carlo (MCMC) samplers, such as Gibbs and Metropolis Hastings [17]. The principle of these samplers is to replace the integral, e.g. in

$$F(x) = \int f(x)p(x)dx$$

by a sum

$$F(x) \approx \sum_{i=1}^N f(x_i)$$

in which the evaluation points x_i are chosen with probability $p(x)$ to ensure the correctness of the approximation in the large sample limit, $N \rightarrow \infty$.

Often, the full joint distribution is too highly dimensional and marginalisation is computationally costly. For such circumstances and when there are loops in the graphical representation of the model's distribution, see figure 6, approximate methods are used. One such method is the Variational method [18, 19] which seeks to approximate the full joint distribution by a set of smaller distributions, e.g. $Q(A)Q(B)Q(C)Q(D) \leftrightarrow P(A, B, C, D)$ of figure 6. The approximate set of distributions are adapted such that the set minimises the Kullback-Leibler divergence

$$KL(Q\|P) = \int Q(A)Q(B)Q(C)Q(D) \log \frac{Q(A)Q(B)Q(C)Q(D)}{P(A, B, C, D)} dA dB dC dD$$

between them. This gives substantially faster algorithms at the cost of optimality that MCMC samplers possess. For the details of the method I refer the reader to [18] for a good tutorial introduction.

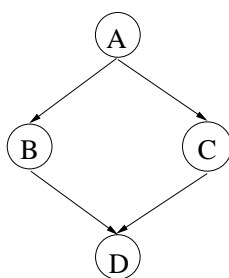


Figure 6: Graphical Representation of the joint probability $P(A, B, C, D) = P(A)P(B|A)P(C|A)P(D|C, B)$.

7 Conclusion

The purpose of this tutorial is to provide the reader with an initial feel for the mathematical operations typically encountered in multivariate statistics. The linear model allows all inference to be analytic and, thus, is a more instructive. However, non-linear models, which require numerical tools, still share a large part of the mathematical manipulations seen for the linear case. The statistical paradigms of maximum likelihood and Bayesian inference each require a different type of calculus, multivariate differential and integral calculus respectively. Because the latter is usually less well described in the literature it received more emphasis in this text. The apparently added complexity of Bayesian methods is easily offset if one considers that additional statements about error bars of the model parameters, crucial for medical applications, are part and parcel of the approach.

References

- [1] S.R. Searle. *Matrix Algebra Useful for Statistics*. John Wiley & Sons, 1982.
- [2] J.R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, 1988.
- [3] H. Lütkepohl. *Handbook of Matrices*. John Wiley, 1996.
- [4] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. John Wiley and Sons, 1994.
- [5] G.E.P. Box and G.C. Tiao. *Bayesian Inference in Statistical Analysis*. John Wiley and Sons, 1992.
- [6] A. Stuart and J.K. Ord. *Distribution Theory*, volume 1 of *Kendall's Advanced Theory of Statistics*. Edward Arnold, 1994.
- [7] N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1 & 2. New York: John Wiley & Sons, 1995.
- [8] G.L. Bretthorst. *Bayesian Spectrum Analysis and Parameter Estimation*. New-York: Springer Verlag, 1989.
- [9] D.J.C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.
- [10] J.J.K. O'Ruanidh and W.J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer, Berlin, 1996.
- [11] A.K. Gupta and D.K. Nagar. *Matrix Variate Distributions*. Number 104 in Monographs and Surveys in Pure and Applied Mathematics. Chapman & Hall/CRC, 2000.

- [12] I. Rezek, S.J. Roberts, and P. Sykacek. Ensemble coupled hidden markov models for joint characterisation of dynamic signals. In C.M. Bishop and B.J. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Key West, FL, Jan 3-6 2003.
- [13] I. Rezek, M. Gibbs, and S.J. Roberts. Maximum *a-posteriori* estimation of coupled hidden markov models. *Journal of VLSI Signal Processing Systems*, 32:55–66, 2002. invited paper.
- [14] I. Rezek, Peter Sykacek, and S.J. Roberts. Learning interaction dynamics with coupled hidden markov models. *IEE Proceedings - Science, Measurement and Technology.*, 147(6):345–350, November 2000. <http://www.robots.ox.ac.uk/~irezek>.
- [15] D. Husmeier, R. Dybowski, and S. Roberts, editors. *Probabilistic Modeling in Bioinformatics and Medical Informatics*. Advanced Information and Knowledge Processing. Springer Verlag, 2004.
- [16] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2000.
- [17] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman Hall, 1996.
- [18] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An Introduction to Variational Methods for Graphical Models. In M.I. Jordan, editor, *Learning in Graphical Models*. Kluwer Academic Press, 1997.
- [19] T.S. Jaakkola and M.I. Jordan. Bayesian parameter estimation through variational methods. *Statistics and Computing*, 1997.
- [20] G.H. Golub and C.F. van Loan. *Matrix Computations*. John Hopkins University Press, 3 edition, 1983.

A Expanding a Linear Model with uninformative Prior

A.1 Completing Squares

In this section we describe a way to derive the posterior functional form of the exponent by means of completing the squares.

We are given the term in the exponent

$$\begin{aligned}
 (\mathbf{d} - \mathbf{G}\mathbf{b})^\top(\mathbf{d} - \mathbf{G}\mathbf{b}) &= \mathbf{d}^\top\mathbf{d} - \mathbf{d}^\top\mathbf{G}\mathbf{b} - \mathbf{b}^\top\mathbf{G}^\top\mathbf{d} + \mathbf{b}^\top\mathbf{G}^\top\mathbf{G}\mathbf{b} \\
 &= \mathbf{d}^\top\mathbf{d} - 2\mathbf{d}^\top\mathbf{G}\mathbf{b} + \mathbf{b}^\top\mathbf{G}^\top\mathbf{G}\mathbf{b}
 \end{aligned}$$

where we could manage to sum the cross-terms because they are scalars

(i.e. $\text{scalar}^\top = \text{scalar}$). Introducing $\mathbf{C} = \mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1}$, we now expand the to obtain a square term

$$\begin{aligned}
 &= \mathbf{b}^\top(\mathbf{G}^\top\mathbf{G}\mathbf{b} - \mathbf{G}^\top\mathbf{G}\mathbf{C}^\top\mathbf{d}) - \mathbf{d}^\top\mathbf{C}\mathbf{G}^\top\mathbf{G}\mathbf{b} + \mathbf{d}^\top\mathbf{d} \\
 &= \mathbf{b}^\top(\mathbf{G}^\top\mathbf{G}\mathbf{b} - \mathbf{G}^\top\mathbf{G}\mathbf{C}^\top\mathbf{d}) - \mathbf{d}^\top\mathbf{C}(\mathbf{G}^\top\mathbf{G}\mathbf{b} \\
 &\quad - \mathbf{G}^\top\mathbf{G}\mathbf{C}^\top\mathbf{d}) - \mathbf{d}^\top\mathbf{C}\mathbf{G}^\top\mathbf{G}\mathbf{C}^\top\mathbf{d} + \mathbf{d}^\top\mathbf{d} \\
 &= (\mathbf{b} - \mathbf{d}^\top\mathbf{C})^\top\mathbf{G}^\top\mathbf{G}(\mathbf{b} - \mathbf{d}^\top\mathbf{C}) - \mathbf{d}^\top\mathbf{C}\mathbf{G}^\top\mathbf{G}\mathbf{C}^\top\mathbf{d} + \mathbf{d}^\top\mathbf{d},
 \end{aligned}$$

which allows us to write the full posterior (18) as

$$p(\sigma, \mathbf{b}|\mathbf{d}) \propto (2\pi\sigma^2)^{-\frac{T}{2}} \exp \left[\frac{(\mathbf{b} - \hat{\mathbf{b}})^\top\mathbf{G}^\top\mathbf{G}(\mathbf{b} - \hat{\mathbf{b}}) - \mathbf{d}^\top\mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top\mathbf{d} + \mathbf{d}^\top\mathbf{d}}{-2\sigma^2} \right] \frac{1}{\sigma},$$

where we have substituted $\hat{\mathbf{b}}$ for $(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top\mathbf{d}$.

A.2 Expanding the Model matrix (Pseudoinverse)

The completion of squares, as shown in Section (A.1), is quite cumbersome and not very obvious. An easier approach is by expressing the inverse of the model matrix in terms of the inverse of the matrix product of model matrices. In the term in the exponent we can extract the model matrix \mathbf{G} as follows

$$(\mathbf{d} - \mathbf{G}\mathbf{b})^\top(\mathbf{d} - \mathbf{G}\mathbf{b}) = (\mathbf{G}^{-1}\mathbf{d} - \mathbf{b})^\top \mathbf{G}^\top \mathbf{G} (\mathbf{G}^{-1}\mathbf{d} - \mathbf{b}).$$

However, the model matrix \mathbf{G} is a rectangular matrix, and thus the inverse is not unique. We can, however, express (\mathbf{G}^{-1}) in terms of $\mathbf{G}^\top \mathbf{G}$:

$$\begin{aligned} \mathbf{G} &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \\ \Rightarrow \mathbf{G}^\top \mathbf{G} &= (\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top)(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top) \\ &= (\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top), \end{aligned}$$

where $\mathbf{\Sigma}$ is a diagonal matrix and \mathbf{V} and \mathbf{U} are orthonormal matrices [20] (i.e. $\mathbf{V}^{-1} = \mathbf{V}^\top \Rightarrow \mathbf{V}^\top \mathbf{V} = \mathbf{I}$, the identity matrix). Also

$$\begin{aligned} \mathbf{G}^{-1}\mathbf{G} = \mathbf{I} &= (\mathbf{G}^\top \mathbf{G})^{-1}(\mathbf{G}^\top \mathbf{G}) \\ \Rightarrow \mathbf{G}^{-1} &= (\mathbf{G}^\top \mathbf{G})^{-1}(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top)(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)(\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^\top) \\ &= (\mathbf{G}^\top \mathbf{G})^{-1}(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top) \\ &= (\mathbf{G}^\top \mathbf{G})^{-1}\mathbf{G}^\top. \end{aligned}$$

Thus, we arrive again at

$$(\mathbf{d} - \mathbf{G}\mathbf{b})^\top(\mathbf{d} - \mathbf{G}\mathbf{b}) = (\mathbf{b} - \hat{\mathbf{b}})^\top \mathbf{G}^\top \mathbf{G} (\mathbf{b} - \hat{\mathbf{b}}),$$

where $\hat{\mathbf{b}} = (\mathbf{G}^\top \mathbf{G})^{-1}\mathbf{G}^\top \mathbf{d}$.