A P P E N D I X

1

Linear models and inference

K. Friston

INTRODUCTION

In this appendix, we gather together different perspectives on inference with multivariate linear models. In brief, we will see that all inference, whether it is based on likelihood ratios (i.e. classical statistics), canonical variates analysis, linear discriminant analysis, Bayesian analysis or information theory, can be regarded as tests of the same thing, namely, the statistical dependence between one set of variables and another. This is useful to appreciate because it means there is no difference between inference using generative models (i.e. functions of causes that generate data) and inference based on recognition models (i.e. functions of data that recognize causes). This equivalence rests on the invertability of linear models, which means that recognition or classification models are simply the inverse of generative or forward models. In other words, there is no difference between reverse-correlation methods (e.g. Hansen *et al.*, 2004; Hasson *et al.*, 2004) or brain-reading (Cox and Savoy, 2003) and conventional analyses (e.g. Friston *et al.*, 1996; Worsley *et al.*, 1997; Kherif *et al.*, 2002), provided the models are linear.

In what follows, we look at the problem of establishing statistical dependencies from an information theory point of view and then revisit the same issue from a classical, a Bayesian and finally a multivariate perspective.

INFORMATION THEORY AND DEPENDENCY

The aim of classification is to find some function of data y that can be used to classify or predict their causes x. Conversely, the aim of hypothesis testing is to show that hypothetical causes predict data. For simplicity, we will assume both x and y represent s independent samples drawn from multivariate distributions. The ability to predict one, given the other, rests on the statistical dependencies between x and y that are quantified by their mutual information. Irrespective of the form of these dependencies, the mutual information is given by the difference between both their entropies and the entropy of both:

$$\begin{aligned} I(x,y) &= H(y) + H(y) - H(x,y) \\ &= H(y) - H(y|x) \end{aligned} \qquad \textbf{A1.1}$$

This is the difference between the entropy, or information, of one minus the information given the other. Under Gaussian assumptions, we have only to consider moments of the densities to second-order (i.e. covariances). Their Gaussian form allows us to express the densities in terms of $\Sigma_y \otimes I_s$, the covariance of *vec(y)* and its conditional covariance $\Sigma_{y|x} \otimes I_s$.

$$\begin{aligned} H(y) &= \langle -\ln p(y) \rangle = \frac{1}{2} \ln |\Sigma_y \otimes I_s| \\ H(y|x) &= \langle -\ln p(y|x) \rangle = \frac{1}{2} \ln |\Sigma_{y|x} \otimes I_s| \\ I(x,y) &= \frac{s}{2} (\ln |\Sigma_y| - \ln |\Sigma_{y|x}|) \\ &= \frac{s}{2} \ln |\Sigma_{y|x}^{-1} \Sigma_y| \end{aligned} \qquad \textbf{A1.2}$$

Here and throughout, constant terms have been ignored. The mutual information can be estimated using sample covariances. From Eqn. **A1.1**, and using standard results for the determinant of block matrices:

$$\begin{aligned} I(x,y) &= \frac{s}{2} (\ln |y^T y| - \ln |y^T y - y^T x (x^T x)^{-1} x^T y|) \\ H(x) &= \frac{s}{2} \ln |\Sigma_x| = \frac{s}{2} \ln |x^T x| \\ H(y) &= \frac{s}{2} \ln |\Sigma_y| = \frac{s}{2} \ln |y^T y| \\ H(x,y) &= \frac{s}{2} \ln \left| \begin{bmatrix} x^T x & y^T x \\ x^T y & y^T y \end{bmatrix} \right| \end{aligned} \qquad \textbf{A1.3}$$

Statistical Parametric Mapping, by Karl Friston et al.
ISBN–13: 978-0-12-372560-8 ISBN–10: 0-12-372560-7

Comparison of Eqn. **A1.2** and Eqn. **A1.3** shows:

$$\begin{aligned}\Sigma_y &= y^T y \\ \Sigma_{y|x} &= y^T y - y^T x(x^T x)^{-1} x^T y\end{aligned} \quad \textbf{A1.4}$$

In fact, we will see below that these are the maximum likelihood estimates of the covariances. It is sometimes useful to express the mutual information or predictability in terms of linearly separable components using the generalized eigenvector solution, with a leading diagonal matrix of eigenvalues υ:

$$\begin{aligned}\Sigma_y c &= \Sigma_{y|x} c \upsilon \\ c^T c &= I \\ I(x, y) &= \frac{s}{2} \ln |\Sigma_{y|x}^{-1} \Sigma_y| \\ &= \frac{s}{2} \ln |\upsilon| \\ &= \frac{s}{2} \ln \upsilon_1 + \frac{s}{2} \ln \upsilon_2 + \cdots\end{aligned} \quad \textbf{A1.5}$$

where c_i are the generalized eigenvectors, which define orthogonal mixtures of y that express the greatest mutual information with x. This information is simply $\frac{s}{2} \ln \upsilon_i$. We will see below that c_i are canonical vectors. Practically speaking, one could use these vectors to predict x, using one or more canonical variates $v_i = y c_i$.

OTHER PERSPECTIVES

Classical inference

In a classical setting, we test a null hypothesis. This calls for a model comparison, usually of a null model against an alternate model. We will assume a linear mapping between the causes and data:

$$y = x\theta + \varepsilon \quad \textbf{A1.6}$$

where ε is some well-behaved error term. The null hypothesis is $\theta = 0$. Following the Neyman-Pearson Lemma, classical statistics uses the maximum likelihood ratio, which, in this context, is:

$$\begin{aligned}L &= \frac{p(y|x, \hat{\theta})}{p(y)} \Rightarrow \\ \ln L &= \ln p(y|x, \hat{\theta}) - \ln p(y) \\ \hat{\theta} &= \max_{\theta} p(y|x, \theta) = (x^T x)^{-1} x^T y\end{aligned} \quad \textbf{A1.7}$$

The maximum likelihood value of the parameters is the usual least squares estimator (this is because we assumed the errors are IID (independent and identically distributed) and can be derived simply by solving $\partial \ln p(y|x, \theta)/\partial\theta = 0$. The maximum log-likelihoods, under the null and alternate hypotheses are:

$$\begin{aligned}\ln p(y) &= -\frac{s}{2} \ln |R_0| - \frac{1}{2} vec(y)^T (R_0^{-1} \otimes I_s) vec(y) \\ R_0 &= \max_{R_0} \ln p(y) = y^T y \\ \ln p\left(y|x, \hat{\theta}\right) &= -\frac{s}{2} \ln |R| - \frac{1}{2} vec(r)^T (R^{-1} \otimes I_s) vec(r) \\ r &= y - x\hat{\theta} \\ R &= \max_{R} \ln p(y|x, \hat{\theta}) = r^T r \\ &= y^T y - y^T x(x^T x)^{-1} x^T y\end{aligned} \quad \textbf{A1.8}$$

$R_0 = \Sigma_y$ and $R = \Sigma_{y|x}$ are the sum of squares and products (SSQP) of the residuals under the null and alternate hypotheses respectively. The maximum likelihood expression for R_0, like the parameters, is obtained easily by solving $\partial \ln p(y)/\partial R_0^{-1} = 0$. Similarly for R, substituting Eqn. **A1.8** into Eqn. **A1.7** shows that the log-likelihood ratio statistic is simply the mutual information:

$$\ln L = \frac{s}{2}(\ln |R_0| - \ln |R|) = \frac{s}{2}(\ln |\Sigma_y| - \ln |\Sigma_{y|x}|) = I(x, y) \quad \textbf{A1.9}$$

In this context, the likelihood ratio is known as Wilk's Lambda $\Lambda = L^{-1}$. When the dimensionality of y is one, this statistic is the basis of the F-ratio. When the dimensionality of x is also one, the square root of the F-ratio is the t-statistic. Classical inference uses the null distribution of the log-likelihood ratio to reject the null hypothesis that $\theta = 0$ to infer that $I(x, y) > 0$.

A Bayesian perspective

A Bayesian perspective on the predictability issue would call for a comparison of two models, with and without x as a predictor. This would proceed using the differences in log-evidence or marginal likelihoods between the alternative and null models:

$$\ln p(y|x) - \ln p(y) = -\frac{s}{2} \ln |R| + \frac{s}{2} \ln |R_0| = I(x, y) \quad \textbf{A1.10}$$

In the context of linear models, this is simply the mutual information.

A multivariate perspective

In linear multivariate models, such as canonical variates analysis, canonical correlation analysis, and linear discriminant function analysis, one is trying to find a mixture of y that affords the best discrimination, in relation

to x. This proceeds by maximizing the length of a vector projected onto the subspace of y, which can be explained by x, relative to its length in the subspace that cannot. More formally, subject to the constraint $c^T c = I$, we want:

$$c = \max_c \frac{c^T T c}{c^T R c} \qquad \textbf{A1.11}$$
$$T = \theta^T x^T x \theta$$

where T is referred to as the SSQP due to treatments. This is the null space of the residual SSQP. This means the total SSQP of y decomposes into the orthogonal covariance components:

$$R_0 = T + R \qquad \textbf{A1.12}$$

The canonical vectors c are the principal generalized eigenvectors:

$$Tc = Rc\lambda$$
$$c^T c = I$$

However, from Eqn. **A1.12**:

$$(R_0 - R)c = Rc\lambda$$
$$R_0 c - Rc = Rc\lambda \qquad \textbf{A1.13}$$
$$R_0 c = Rc(\lambda + I) =$$
$$\Sigma_y c = \Sigma_{y|x} c(\lambda + I)$$

which has exactly the same form as Eqn. **A1.5**. In other words, the canonical vectors are simply the mixtures that express the greatest mutual information with x; the amount of information is $\frac{s}{2}\ln(\lambda_i + 1)$, where $\lambda_i = \upsilon_i - 1$ is the i-th canonical value. From Eqn. **A1.5**, Eqn. **A1.9** and Eqn. **A1.13** we get:

$$\begin{aligned} -\ln \Lambda &= I(x, y) \\ &= \frac{s}{2}\ln \upsilon_1 + \frac{s}{2}\ln \upsilon_2 + \cdots \\ &= \frac{s}{2}\ln(\lambda_1 + 1) + \frac{s}{2}\ln(\lambda_2 + 1) + \cdots \end{aligned} \qquad \textbf{A1.14}$$

Tests for the dimensionality of the subspace are based on the canonical values $\ln(\lambda_i + 1)$ (see Chapter 37).

SUMMARY

In the context of linear mappings under Gaussian assumptions, the heart of inference lies in the generalized eigenvalue solution. This solution finds pairs of generalized eigenvectors that show the greatest statistical dependence between two sets of multivariate data. The generalized eigenvalues encode the mutual information between the i-th variate and its corresponding vector. The total information is the log-likelihood ratio, or log-Bayes factor, comparing models with and without a linear mapping. Special cases of this quantity are Wilk's Lambda, Hotelling's T-square, the F-ratio and the t-statistic, upon which classical inference is based.

REFERENCES

Cox DD, Savoy RL (2003) Functional magnetic resonance imaging (fMRI) 'brain reading': detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* **192**: 61–70

Friston KJ, Stephan KM, Heather JD *et al.* (1996) A multivariate analysis of evoked responses in EEG and MEG data. *NeuroImage* **3**: 167–74

Hansen KA, David SV, Gallant JL (2004) Parametric reverse correlation reveals spatial linearity of retinotopic human V1 BOLD response. *NeuroImage* **23**: 233–41

Hasson U, Nir Y, Levy I *et al.* (2004) Intersubject synchronization of cortical activity during natural vision. *Science* **303**: 1634–40

Kherif F, Poline JB, Flandin G *et al.* (2002) Multivariate model specification for fMRI data. *NeuroImage* **16**: 1068–83

Worsley KJ, Poline JB, Friston KJ *et al.* (1997) Characterizing the response of PET and fMRI data using multivariate linear models. *NeuroImage* **6**: 305–19