Dimensionality Tests for Canonical Variates Analysis

W.D. Penny

Wellcome Trust Centre for Neuroimaging,
University College, London WC1N 3BG, UK.

April 24, 2013

1 Log-Likelihood Ratio

We consider the linear relationship between a $[1 \times d_1]$ variable y_n and a $[1 \times d_2]$ variable x_n where

$$y_n = x_n \beta + e_n \tag{1}$$

The matrix of regression coefficients β is $[d_2 \times d_1]$ and the Gaussian error e_n is $[1 \times d_1]$. We have $n = 1..N$ independent data points giving rise to the N rows in the matrices Y, X and E such that

$$Y = X\beta + E \tag{2}$$

If there is no relation between the variables then the log likelihood of the data is

$$\log p(Y) = -\frac{N}{2} \log |\Sigma_y| \tag{3}$$

where Σ_y is the sample covariance. If there is a relation between the variables then the log likelihood of the data under the model having maximum likelihood coefficients $\beta_{ML} = (X^T X)^{-1} X^T Y$ is

$$\log p(Y|\beta_{ML}) = -\frac{N}{2} \log |\Sigma_{y|x}| \tag{4}$$

where

$$\Sigma_{y|x} = \Sigma_y - \Sigma_{xy}^T \Sigma_x^{-1} \Sigma_{xy} \tag{5}$$

and Σ_{xy} is the covariance between x and y, and Σ_x is the covariance of x. The log-likelihood ratio, Λ, is therefore

$$\begin{aligned} \Lambda &= \log \frac{p(Y|\beta_{ML})}{p(Y)} \\ &= \frac{N}{2} \log |\Sigma_{y|x}^{-1} \Sigma_y| \end{aligned} \tag{6}$$

If s_i is the ith eigenvalue of $\Sigma_{y|x}^{-1}\Sigma_y$ we can write

$$\Lambda = \frac{N}{2}\sum_{i=1}^{h}\log s_i \tag{7}$$

where $h = \min(d_1, d_2)$. This is also known as Wilk's Lambda. We also define the quantity

$$\Lambda_{j,k} = \frac{N}{2}\sum_{i=j}^{k}\log s_i \tag{8}$$

$\Lambda_{1,m}$ is the log-likelihood ratio for a CVA model with m canonical variates.

1.1 Equivalent Expressions

The variability in the data can be expressed as

$$\Sigma_y = \Sigma_{\hat{y}} + \Sigma_{y|x} \tag{9}$$

where $\Sigma_{\hat{y}}$ is the covariance explained by the model and $\Sigma_{y|x}$ is the covariance not explained by the model.

If λ_i are eigenvalues of $\Sigma_{y|x}^{-1}\Sigma_{\hat{y}}$ then the above relationship can be used to show that $s_i = \lambda_i + 1$ (see Appendxi A1 of SPM book). Hence an alternative expression for the log likelihood ratio is

$$\Lambda = \frac{N}{2}\sum_{i=1}^{h}\log(1+\lambda_i) \tag{10}$$

Here, $\Sigma_{\hat{y}}$ can be formed directly from model predictions

$$\begin{aligned} \hat{Y} &= X\beta_{ML} \\ \Sigma_{\hat{y}} &= \hat{Y}^T\hat{Y} \end{aligned} \tag{11}$$

and $\Sigma_{y|x}$ from the residuals

$$\begin{aligned} R &= Y - \hat{Y} \\ \Sigma_{y|x} &= R^T R \end{aligned} \tag{12}$$

The ith canonical correlation can be expressed as

$$r_i = \sqrt{\frac{\lambda_i}{\lambda_i + 1}} \tag{13}$$

Hence, a third equivalent form for the log likelihood ratio is

$$\Lambda = -\frac{N}{2}\sum_{i=1}^{h}\log(1-r_i^2) \tag{14}$$

The function `spm_cva.m` uses equation 10. Similarly, we can write

$$\Lambda_{j,k} = \frac{N}{2}\sum_{i=j}^{k}\log s_i \tag{15}$$

$$
\begin{aligned}
&= \frac{N}{2}\sum_{i=j}^{k}\log(1+\lambda_i) \\
&= -\frac{N}{2}\sum_{i=j}^{k}\log(1-r_i^2)
\end{aligned}
$$

2 Bartlett's Test

Bartlett's Test for the dimension of a CVA model is based on classical inference. It tests the null hypothesis that canonical correlations for dimensions m to h are all zero. Strength of evidence against the null is assessed using

$$\Lambda_{m,h} \approx \chi^2(df) \tag{16}$$

where $df = (d_1 - m)(d_2 - m)$. We denote the corresponding p-value as p_m. The estimated model order is the largest value of m for which $p_m < 0.05$.

3 Bayes Factors

The log evidence for a model with no parameters (null model) is simply the log likelihood of the data, $L_0 = \log p(Y)$. The log evidence for model m with parameters β is given by

$$L_m = \log \int p(Y|\beta)p(\beta)d\beta \tag{17}$$

This can be approximated by BIC as

$$BIC = \log p(Y|\beta_{ML}) - \frac{k}{2}\log N \tag{18}$$

or

$$AIC = \log p(Y|\beta_{ML}) - k \tag{19}$$

where k is the number of parameters in the model. For a CVA model of dimension m we have $k = m(d_1+d_2)$. Log Bayes factors can therefore be approximated as differences in BIC/AIC scores. Hence, under BIC, the log Bayes factor for a CVA model of dimension m versus a model with dimension zero (null model) is given by

$$LogBF(m)_{BIC} = \Lambda_{1,m} - \frac{k}{2}\log N \tag{20}$$

and under AIC as

$$LogBF(m)_{AIC} = \Lambda_{1,m} - k \tag{21}$$

The estimated model order is the one which has the largest LogBF. Negative values of $LogBF(m)$ express evidence in favour of the null model.

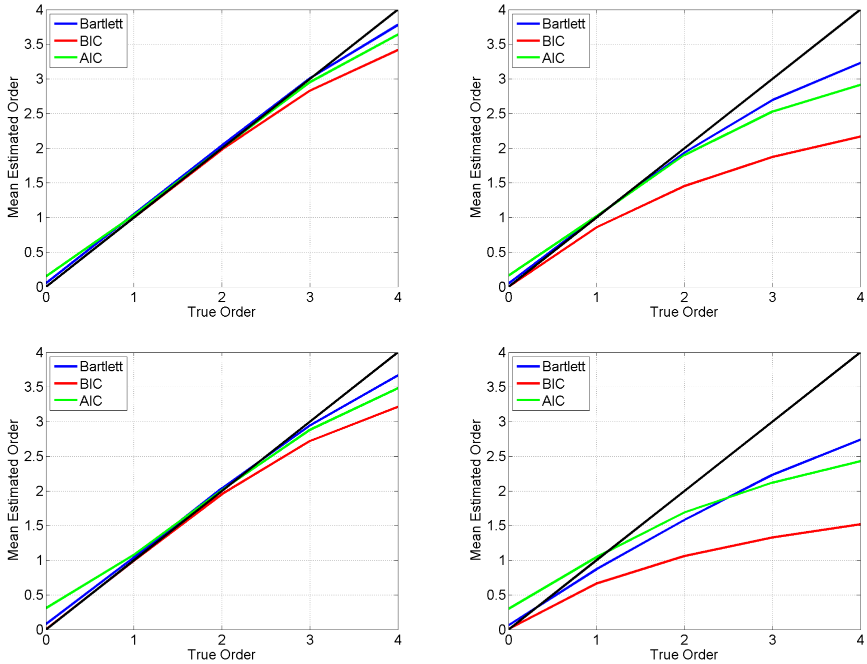

Figure 1: **Top Left** $N = 64$ data points and observation noise variance $\sigma^2 = 0.1$. The mean estimated canonical correlations at the true model order were 0, 0.97, 0.93, 0.85 and 0.64. **Top Right** $N = 64$ data points and observation noise variance $\sigma^2 = 1$. The mean estimated canonical correlations at the true model order were 0, 0.83, 0.71, 0.55 and 0.37. **Bottom Left** $N = 32$ data points and observation noise variance $\sigma^2 = 0.1$. The mean estimated canonical correlations at the true model order were 0, 0.98, 0.95, 0.88 and 0.67. **Bottom Right** $N = 32$ data points and observation noise variance $\sigma^2 = 1$. The mean estimated canonical correlations at the true model order were 0, 0.87, 0.77, 0.63 and 0.44.

4 Simulations

Here we generated data from a latent variable model corresponding to probabilistic CVA (Wong, 2006)

$$\begin{aligned} y_n &= w_y z_n + e_n \\ x_n &= w_x z_n + r_n \end{aligned} \tag{22}$$

where z_n is of dimension m, and $d_1 = dim(y_n)$, $d_2 = dim(x_n)$. We set $d_1 = 4$ and $d_2 = 8$.

We generate w_y and w_x as standard Gaussian variates. We then produce the nth data sample by drawing z_n as a standard Gaussian variate and e_n and r_n as zero mean Gaussian variates with variance σ^2. This produces Y and X. We then estimated the model order using Bartlett's test and Bayes factors based on BIC and AIC. This whole process is repeated $Nrep = 1000$ times and we record the mean estimated order.