33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Bayesian model evidence for order selection and correlation testing

Leigh A. Johnston, Iven M. Y. Mareels and Gary F. Egan

Abstract— **Model selection is a critical component of data analysis procedures, and is particularly difficult for small numbers of observations such as is typical of functional MRI datasets. In this paper we derive two Bayesian evidence-based model selection procedures that exploit the existence of an analytic form for the linear Gaussian model class. Firstly, an evidence information criterion is proposed as a model order selection procedure for auto-regressive models, outperforming the commonly employed Akaike and Bayesian information criteria in simulated data. Secondly, an evidence-based method for testing change in linear correlation between datasets is proposed, which is demonstrated to outperform both the traditional statistical test of the null hypothesis of no correlation change and the likelihood ratio test.**

I. Introduction

Model comparison and model selection procedures are of fundamental importance to system identification techniques; implicit in the inference of a system's states and parameters is the understanding that estimates are meaningful only in the context of the chosen system model. The assessment of model fit according to an objective criterion is therefore a critical step in data analysis. Bayesian model evidence (BME) quantifies the fit of a model to a given set of data. The calculation of BME can however be non-trivial, often requiring approximation via Variational Bayes techniques or Markov chain Monte Carlo methods in order to render the calculation tractable [1].

We investigate the closed form expression for BME of the linear Gaussian model class that exploits the ability to marginalise over the model class parameters [2]. We present two identification procedures arising from the analytic BME expression: (1) An evidence information criterion, similar in form to the Akaike information criterion [3] and Bayesian or Schwarz information criterion [4], is proposed, in which a penalty dependent on the model order is applied to the maximised likelihood function. We apply the evidence information criterion to estimation of model order for autoregressive models of moderate length time-series, as is commonplace in fMRI analyses [5]. (2) The statistical testing of correlation changes between brain regions, indicating connectivity differences [6], has traditionally been achieved using the normalising transformation suggested by Fisher [7], [8]. Following an evidence-based reasoning, we propose a test for change in linear relationship between two datasets that is both simple to implement and is shown to be more sensitive and specific than Fisher's method.

This work was supported by NICTA Victoria Research Laboratory

L. Johnston is with the Melbourne School of Engineering, University of Melbourne, the NICTA Victoria Research Laboratory, and the Florey Neuroscience Institutes, l.johnston@unimelb.edu.au.

I. Mareels is with the Melbourne School of Engineering, University of Melbourne, i.mareels@unimelb.edu.au.

G. Egan is with Monash Biomedical Imaging, Monash University, and the Centre for Neuroscience, University of Melbourne, gary.egan@monash.edu.

II. Theory

Let $\mathbf{y}$ be a $(n \times 1)$ vector and $\mathbf{X}$ a $(n \times p)$ matrix of observed real-valued data. The linear Gaussian model class, $\mathcal{M}_G$, considered here-in is defined by parameters $\Theta_G = \{\mathbf{a}, \sigma\}$, according to the relationship

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{e}, \qquad \mathbf{e} \sim N(\mathbf{0}_n, \sigma^2 I_n), \tag{1}$$

where $\mathbf{a} \in \mathbb{R}^p$ is a $(p \times 1)$ vector of unknown parameters, and $\sigma > 0$ is an unknown standard deviation. Here $\mathbf{0}_n$ and I_n denote a length n zero vector and an nxn identity matrix, respectively.

The Bayesian evidence of model $\mathcal{M}_G$ for observed data $\{\mathbf{y}, \mathbf{X}\}$ is given by the denominator of the Bayes rule expansion for model $\mathcal{M}_G$. For notational simplicity, we define the term 'evidence' to denote the Naperian log evidence:

$$\mathcal{E}(\mathbf{y}, \mathbf{X}, \mathcal{M}_G) = \ln \int_\sigma \int_{\mathbf{a}} P(\mathbf{y}, \mathbf{X} | \mathbf{a}, \sigma)\, P(\mathbf{a})\, P(\sigma)\, d\mathbf{a}\, d\sigma, \tag{2}$$

which, for the linear Gaussian model class, evaluates to the closed form expression [2]

$$\begin{aligned}\mathcal{E}(\mathbf{y}, \mathbf{X}, \mathcal{M}_G) = \ln \Gamma\left(\frac{n-p}{2}\right) - \frac{1}{2}\ln|\mathbf{X}'\mathbf{X}| - \frac{n-p}{2}\ln(2\pi) \\ + \left(\frac{n-p}{2} - 1\right)\ln 2 - \left(\frac{n-p}{2}\right)\ln S + \ln f(p).\end{aligned} \tag{3}$$

Here Γ is the gamma function, f is a known parametric function and S is the residual sum of squares,

$$S = \mathbf{y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}. \tag{4}$$

The choice of priors is critical to the evidence calculation. Noninformative, improper priors for the parameters of the Gaussian distribution are the flat prior for $\mathbf{a}$ and the scale invariant $P(\sigma) = 1/\sigma$ for the standard deviation [1]. The flat prior for $P(\mathbf{a})$ is problematic, however, as comparison of models with different dimensions can cause an *a priori* determined outcome regardless of the data, known as the Barlett or Lindley paradox [9]. The predominant approach proposed to avoid this problem is to use proper priors. This is the approach we follow here, with the proposal that the prior over $\mathbf{a}$ is a parametric function of the dimension of $\mathbf{a}$ only: $P(\mathbf{a}) = f(p) > 0$.

III. Evidence Information Criterion

The primary use of BME is in model selection, where the largest evidence computed over a set of competing models is chosen as the best model of a particular dataset. The analytic expression for the linear Gaussian Bayesian model evidence enables derivation of an evidence information criterion (EIC), similar in form to the popular Akaike (AIC) and Bayesian (BIC) information criteria. Akaike developed the AIC in the early 1970's as an information theoretic criterion grounded in model prediction rather than traditional hypothesis testing that he perceived to be subjective [3]. The BIC, proposed by Schwarz in 1978, is derived from an asymptotic Bayesian expansion that is therefore optimal at large sample sizes. In contrast, the EIC proceeds directly from the closed form BME expression for the linear Gaussian model class, as follows.

The maximum likelihood estimates of the parameters of the linear Gaussian signal model are

$$\mathbf{a}_{ML} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \quad \sigma_{ML} = \sqrt{\frac{S}{n}}. \tag{5}$$

The value of the log likelihood function evaluated at the ML estimates is

$$\ln P\Big(\mathbf{y},\mathbf{X}\Big|\mathbf{a}_{ML},\sigma_{ML}\Big) = \frac{n}{2}\ln n - \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln S - \frac{n}{2}. \tag{6}$$

The "Evidence Information Criterion", based on BME, results from direct comparison of (6) and (3):

$$\begin{aligned} EIC = &-2\ln P\Big(\mathbf{y},\mathbf{X}\Big|\mathbf{a}_{ML},\sigma_{ML}\Big) - p\ln(2\pi) - 2\ln\Gamma\left(\frac{n-p}{2}\right) \\ &+ \ln|\mathbf{X}'\mathbf{X}| - (n-p-2)\ln 2 + n(\ln n - 1) - p\ln S \\ &- 2\ln f(p). \end{aligned} \tag{7}$$

The scale factor of 2 ensures the same functional form as the Akaike and Bayesian information criteria:

$$AIC = -2\ln P\Big(\mathbf{y},\mathbf{X}\Big|\mathbf{a}_{ML},\sigma_{ML}\Big) + 2p \tag{8}$$

$$BIC = -2\ln P\Big(\mathbf{y},\mathbf{X}\Big|\mathbf{a}_{ML},\sigma_{ML}\Big) + p\ln n \tag{9}$$

Optimal model orders are selected by minimising the information criteria. The penalty terms in the AIC and BIC, that which are added to the loglikelihood, depend only on data length, n, and model order, p. The EIC penalty, on the other hand, contains data-dependent terms, $p\ln S$ and $\ln|\mathbf{X}'\mathbf{X}|$ and the parameterised prior function, $f(p)$. Despite the assertion by [10] of asymptotic equivalence of the BIC and a closed form expression similar to the EIC, we observe that the data-dependent penalty terms alter the performance of the criterion significantly for a wide range of model orders and data lengths. The component of the EIC penalty that is independent of data and prior choice is given by $-p\ln(2\pi) - 2\ln\Gamma\left(\frac{n-p}{2}\right) - (n-p-2)\ln 2 + n(\ln n - 1)$. The AIC and BIC penalty terms are compared with this data-independent EIC penalty across model order in Fig. 1 for $n = 100$ and $n = 10{,}000$. It is evident that the EIC penalises higher model orders more strongly than the AIC and even the more stringent BIC. The data-dependent terms in the EIC act to increase the penalty at high SNR, as $(\ln|\mathbf{X}'\mathbf{X}| - p\ln S) > 0$, while at low SNR, the EIC penalty for increased model orders is reduced by the data-dependent terms. Note that a similar approach to order selection is followed in [11], in which the eigenvalues of the data covariance matrix for a maximally high-order model are examined, with a magnitude reduction beyond the optimal model order.

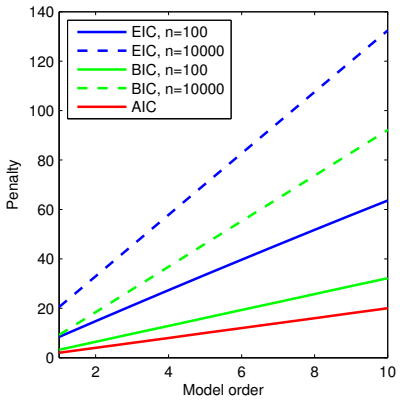

Fig. 1. The evidence information criterion: Comparison of AIC and BIC penalty terms with the component of the EIC penalty independent of data and prior choice, for $n = 100$ and $10{,}000$.

A. Model order selection for AR models

Consider the autoregressive model of order p, denoted $AR(p)$,

$$z_k = \sum_{i=1}^{p} a_i z_{k-i} + e_k, \qquad e_k \sim N(0,\sigma^2). \tag{10}$$

As the parameter vector, $\mathbf{a}$, defines p poles within the unit circle (stability assumed), the prior is assigned to be

$$f(p) = (4\pi)^{-p}, \tag{11}$$

expressing a uniform distribution over the surface of the unit circle. A tighter prior can be realised, given that poles appear in conjugate pairs, however (11) has been found to be sufficiently constrained. The prior considered in [10] enforces model stability, but loses a closed form posterior.

We compared the EIC, AIC and BIC, for estimating the optimal model order of simulated AR data. Data for each true model order, $p = 1,\ldots,8$, was generated for 1000 simulation runs, each with $\sigma_e^2 = 1$ and AR parameters drawn from a uniform distribution over a disc with inner radius $r = 0.950$ and outer radius $r = 0.999$. The restriction to the disc with poles of large magnitude is necessary for the generated time-series to take on the characteristics of the true model order. The results for two time-series lengths, $n = 100$ and $n = 500$ are displayed in Fig. 2 with log of the relative frequency (empirical probability) of the estimated model order against the true model order. It is evident that the EIC provides the most accurate model order estimates, particularly noticeable in the weights on the super-diagonal which for BIC are significant across all true model orders, but which for EIC are

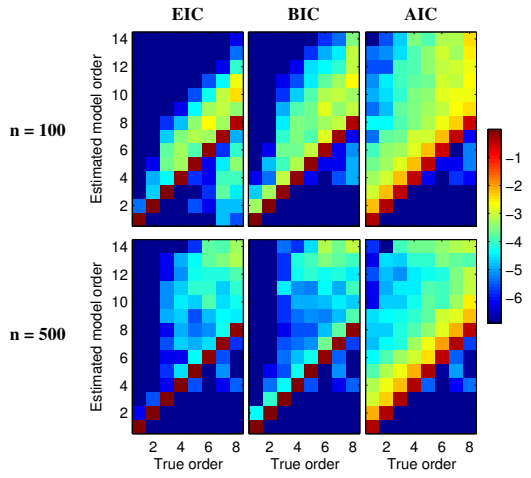

Fig. 2. The evidence information criterion: Results for EIC, BIC and AIC model order selection. True vs estimated model order for length $n = 100$ and $n = 500$ $AR(p)$ simulations. Colour denotes the logarithm of relative frequency.

extremely low at low true model orders. AIC, as expected, over-estimates the model order.

IV. Evidence of Correlation Change

The second application of the analytic form for BME that we present here is a test for a change in correlation between two sets of data. Consider two sets of data, each containing two time-series of length n, $\{\mathbf{y}_1, \mathbf{x}_1\}$ and $\{\mathbf{y}_2, \mathbf{x}_2\}$. Correlation is a measure of the strength of the linear Gaussian relationship between $\mathbf{y}_i$ and $\mathbf{x}_i$, traditionally summarised by the correlation coefficient, $\rho_{\mathbf{x}_i\mathbf{y}_i}$,

$$\rho_{\mathbf{x}_i\mathbf{y}_i} = \frac{\mathbf{x}_i'\mathbf{y}_i}{\sqrt{\mathbf{x}_i'\mathbf{x}_i\ \mathbf{y}_i'\mathbf{y}_i}}, \quad i = 1,2. \tag{12}$$

The standard method by which to test the null hypothesis that correlation has not changed across the datasets, $\rho_{\mathbf{x}_1\mathbf{y}_1} = \rho_{\mathbf{x}_2\mathbf{y}_2}$, is to use Fisher's variance stabilising transformation of the correlation coefficients [7], [8], to convert to z-scores. A statistically significant rejection of the null hypothesis is determined by a z-score exceeding a chosen significance level, α, of a zero-mean, unit variance Gaussian distribution [12].

An evidence-based statistical test proceeds differently. The analytic form for BME permits direct calculation of the difference between the evidence of two independent linear Gaussian relationships and the evidence of the same linear Gaussian relationship. A statistically significant result, suggesting that the linear Gaussian relationships differ between the two datasets, is determined by the change in evidence exceeding a given evidence threshold level.

Consider first the calculation of the evidence of two independent linear Gaussian relationships. For the timeseries data, the linear Gaussian model matrix, $\mathbf{X}$, is a vector, and therefore the model order is $p = 1$, and $\mathbf{a} \equiv a$. Using the additive property of the BME,

$$\begin{aligned}
&\mathcal{E}(\mathbf{y}_1, \mathbf{x}_1, \mathcal{M}_G) + \mathcal{E}(\mathbf{y}_2, \mathbf{x}_2, \mathcal{M}_G) \\
&\quad = \ln \int_\sigma \int_a P(\mathbf{y}_1, \mathbf{x}_1 | a, \sigma)\, da\, d\sigma \int_\sigma \int_a P(\mathbf{y}_2, \mathbf{x}_2 | a, \sigma)\, da\, d\sigma \\
&\quad = 2 \ln \Gamma\left(\frac{n-1}{2}\right) - (n-1)\ln(2\pi) + (n-3)\ln 2 - 2\log M \\
&\quad - \frac{1}{2} \sum_{i=1,2} \left(\ln(\mathbf{x}_i'\mathbf{x}_i) - (n-1)\ln(\mathbf{y}_i'\mathbf{y}_i) - (n-1)\ln(1-\rho^2_{\mathbf{x}_i\mathbf{y}_i}) \right).
\end{aligned} \tag{13}$$

Here we have assumed a uniform prior distribution, $P(a) = 1/M$, $M > 0$, where $M = a_{\max} - a_{\min}$ is the width of the uniform distribution. The size of M controls the sensitivity-specificity trade-off, as discussed below in relation to numerical examples.

Consider next the combined time-series vectors, $\mathbf{y}_c = [\mathbf{y}_1', \mathbf{y}_2']'$ and $\mathbf{x}_c = [\mathbf{x}_1', \mathbf{x}_2']'$, each of length $2n$. The evidence of the same linear Gaussian relationship is given by

$$\begin{aligned}
\mathcal{E}(\mathbf{y}_c, \mathbf{x}_c, \mathcal{M}_G) = \ln \Gamma\left(\frac{2n-1}{2}\right) - \frac{1}{2}\ln(\mathbf{x}_c'\mathbf{x}_c) - \frac{2n-1}{2}\ln(2\pi) \\
+ \left(\frac{2n-3}{2}\right)\ln 2 - \log M - \frac{2n-1}{2}\ln(\mathbf{y}_c'\mathbf{y}_c) \\
- \frac{2n-1}{2}\ln\left(1 - \rho^2_{\mathbf{x}_c\mathbf{y}_c}\right).
\end{aligned} \tag{14}$$

The evidence of a change in linear relationship is given by the difference between (14) and (13),

$$\begin{aligned}
&\Delta\mathcal{E}(\mathbf{y}_1, \mathbf{x}_1, \mathbf{y}_2, \mathbf{x}_2, \mathcal{M}_G) \\
&\quad = \mathcal{E}(\mathbf{y}_1, \mathbf{x}_1, \mathcal{M}_G) + \mathcal{E}(\mathbf{y}_2, \mathbf{x}_2, \mathcal{M}_G) - \mathcal{E}(\mathbf{y}_c, \mathbf{x}_c, \mathcal{M}_G).
\end{aligned} \tag{15}$$

A. Examples of evidence-based correlation test application

We tested the sensitivity and specificity of the evidence-based test, with $M = 1, 10, 100$, for detecting changes in correlation, compared with the Fisher z-transformation test and a likelihood ratio test. $N = 10{,}000$ datasets were generated under the null hypothesis via sampling of the parameters $a_1^{(i)} = a_2^{(i)} \sim \mathrm{U}(-1,1)$, $\sigma_1^{(i)} = \sigma_2^{(i)} \sim \mathrm{U}(0,3)$ and $\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)} \sim N(\mathbf{0}_{200}, I_{200})$, from which the $\mathbf{y}_1^{(i)}$ and $\mathbf{y}_2^{(i)}$ were generated, $i = 1, \ldots, N$. Similarly, $N = 10{,}000$ datasets were generated under the alternative hypothesis of two distinct linear Gaussian models through sampling $a_1^{(i)}, a_2^{(i)} \sim \mathrm{U}(-1,1), a_1^{(i)} \neq a_2^{(i)}$, $\sigma_1^{(i)}, \sigma_2^{(i)} \sim \mathrm{U}(0,3), \sigma_1^{(i)} \neq \sigma_2^{(i)}$, and $\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)} \sim N(\mathbf{0}_{200}, I_{200})$.

From these null- and alternative-hypothesis datasets, true positive and false positive rates were calculated across a range of thresholds for the evidence-based, likelihood ratio and z-transformation tests. The results are shown in the standard receiver-operating-characteristic (ROC) curve graphical form, which displays true positive rate against false positive rate (Fig. 3). It is evident from Fig. 3) plots that M controls the sensitivity-specificity (type I/II error) trade-off; the marked points correspond to an evidence threshold of $\Delta\mathcal{E} = 0$ for each of the three M values.

The evidence-based test clearly outperform the Fisher z-transformation test; at the marked point indicating an $\alpha = 0.05$ significance threshold corresponding in the two-tailed

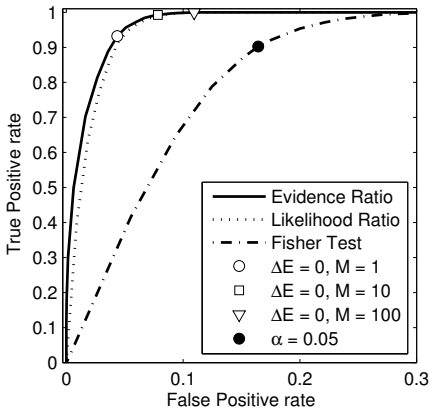

Fig. 3. Test of correlation change: ROC curve, false vs true positive rates for simulated data, $N = 10{,}000$ datasets generated under each of the null and alternative hypotheses.

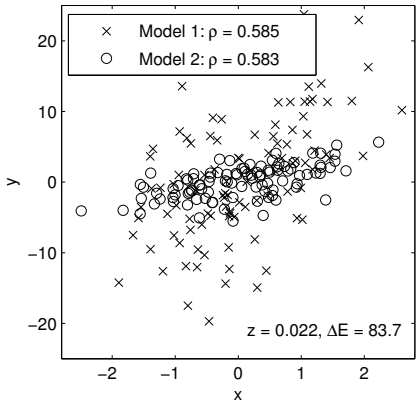

Fig. 4. Test of correlation change: Example demonstrating two datasets with near equal correlation that were drawn from two independent linear Gaussian models. Fisher test declares acceptance of the null hypothesis of no correlation change ($p = 0.509$), while evidence-based test strongly in favour of two independent linear relationships, with $\Delta\mathcal{E} = 83.7$.

Fisher test to a 90% true positive rate, the false positive rate of the Fisher test is more than double that of the evidence-based test. The evidence-based test is also shown in Fig. 3 to slightly outperform the likelihood ratio test. A further advantage of the evidence-based test over the likelihood ratio test is that $\Delta\mathcal{E} = 0$ provides a natural threshold for declaring a result in favour of a changed versus unchanged linear relationship, as $\Delta\mathcal{E}$ was seen to be predominantly negative when data generated under the null hypothesis and positive under the alternative. The (log) likelihood ratio, on the other hand, was positive for data generated under both the null and alternative hypotheses, and a subjective threshold must therefore be chosen to discriminate appropriately between the two.

The evidence-based test provides meaningful results in cases where the Fisher test fails. Two sets of correlated data, $\{\mathbf{x}_1, \mathbf{y}_1\}$ and $\{\mathbf{x}_2, \mathbf{y}_2\}$, were generated from two distinct linear Gaussian models, $\mathcal{M}_1 : (a_1 = 5, \sigma_1 = 8)$ and $\mathcal{M}_2 : (a_2 = 1, \sigma_2 = 2)$, with $\mathbf{x}_1, \mathbf{x}_2 \sim N(\mathbf{0}_{100}, I_{100})$. Fig. 4 displays the scatterplots of the two datasets, which clearly visually indicate two distinct linear relationships. The sample correlations, however, are $\rho_{\mathbf{x}_1\mathbf{y}_1} = 0.585$ and $\rho_{\mathbf{x}_2\mathbf{y}_2} = 0.583$. The corresponding Fisher z-score, $z = 0.022$ equivalently a p-value of 0.509, indicates no change in correlation between the two datasets. The change in evidence, calculated for $M = 1$, is $\Delta\mathcal{E} = 83.7$. A threshold of $\Delta\mathcal{E} = 7$ gave a significance level of 10^{-4} in Fig. 3, while 5% was achieved by a threshold of $\Delta\mathcal{E} = 1$. Therefore the result of $\Delta\mathcal{E} = 83.7$ is an overwhelmingly strong indication of the presence of two linear relationships, rather than one.

V. Conclusions

We have applied Bayesian model evidence-based reasoning to two fundamental problems in fMRI analysis, that of order selection for autoregressive models, and inference of correlation change. The proposed order selection procedure, based on the evidence information criterion, was shown to outperform AIC and BIC and is particularly well-suited to moderate sample sizes. Correlation testing, typically performed using Fisher's transform, was cast in the Bayesian model evidence framework, a strength of which over its likelihood ratio counterpart was highlighted in the objective choice of zero change in evidence as the decision threshold. Future work will involve the use of increased computational power to numerically evaluate Bayesian model evidence expressions outside the linear Gaussian model class that are currently left to approximation.

References

[1] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge University Press, 2003.

[2] S. J. Roberts and W. D. Penny, "Variational Bayes for generalized autoregressive models," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2245–2257, 2003.

[3] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, Dec. 1974.

[4] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[5] B. P. Rogers, S. B. Katwald, V. L. Morgana, C. L. Asplunde, and J. C. Gore, "Functional mri and multivariate autoregressive models," *Magnetic Resonance Imaging*, vol. 28, pp. 1058–1065, 2010.

[6] G. Marrelec, J. Daunizeau, M. Plgrini-Issac, J. Doyon, and H. Benali, "Conditional correlation as a measure of mediated interactivity in fMRI and MEG/EEG," *IEEE Transactions on Signal Processing*, vol. 53, pp. 3503–3516, Sept. 2005.

[7] R. A. Fisher, "Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large populatation," *Biometrika*, vol. 10, pp. 507–521, 1915.

[8] R. A. Fisher, "On the 'probable error' of a coefficient of correlation deduced from a small sample," *Metron*, vol. 1, pp. 3–21, 1921.

[9] R. W. Strachan and H. K. van Dijk, "The practical implementation of Bayesian model section," *Oxford Bulletin of Economics and Statistics*, vol. 65 (suppl.), pp. 863–876, 2003.

[10] F. Gustafsson and H. Hjalmarsson, "Twenty-one ML estimators for model selection," *Automatica*, vol. 31, no. 10, pp. 1377–1392, 1995.

[11] T. Cassar, K. P. Camilleri, and S. G. Fabri, "Order estimation of multivariate ARMA models," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 3, pp. 494–503, 2010.

[12] D. R. Cox, "The null distribution of the first serial correlation coefficient," *Biometrika*, vol. 53, no. 3-4, pp. 623–626, 1966.