# An Efficient Approach to Graphical Modeling of Time Series

R. J. Wolstenholme and A. T. Walden, *Senior Member, IEEE*

***Abstract*—A method for selecting a graphical model for $p$-vector-valued stationary Gaussian time series was recently proposed by Matsuda and uses the Kullback–Leibler divergence measure to define a test statistic. This statistic was used in a backward selection procedure, but the algorithm is prohibitively expensive for large $p$. A high degree of sparsity is not assumed. We show that reformulation in terms of a multiple hypothesis test reduces computation time by $O(p^2)$ and simulations support the assertion that power levels are attained at least as good as those achieved by Matsuda's much slower approach. Moreover, the new scheme is readily parallelizable for even greater speed gains.**

***Index Terms*—Undirected graph, Kullback–Leibler divergence, multiple hypothesis test, vector-valued time series.**

## I. Introduction

THERE has been much interest in recent years in the construction of graphical models from $p$-vector-valued (or multivariate) stationary time series $\{\boldsymbol{X}_t\}$ where $\boldsymbol{X}_t = [X_{1,t}, \ldots, X_{p,t}]^T \in \mathbb{R}^p$, $t \in \mathbb{Z}$, and $^T$ denotes transposition. The purpose of graphical models is to aid visualization of connections between multiple time series: each of the time series is represented by one vertex and it is wished to define connections via edges between the vertices of the graph. The lack of an edge indicates the lack of a connection between the corresponding series.

Formally, a graph $G = (V, E)$ consists of vertices $V$ and edges $E$, where $E \subset \{(j,k) \in V \times V : j \neq k\}$. (We are considering simple graphs where there are no loops from a vertex to itself, nor multiple edges between two vertices.) To represent $\{\boldsymbol{X}_t\}$ the vertices of the graph correspond to the $p$ individual series $\{X_{j,t}\}$, so $V = \{1, \ldots, p\}$. Edges connect ordered pairs of distinct vertices. Edges $(j,k) \in E$ for which both $(j,k) \in E$ and $(k,j) \in E$ are called undirected edges. An undirected graph is one with only undirected edges and it only represents interaction between the series. An edge $(j,k)$ is called directed if $(j,k) \in E$, with $(k,j) \notin E$. A directed graph is one in which all edges are directed and it typically encodes directions of influence or of causation between series.

In this paper we will consider the modelling only of *undirected graphs*. There are $p(p-1)/2$ unordered pairs of vertices for the graph and $2^{p(p-1)/2}$ possible distinct graph structures. A high degree of sparsity of edges is not assumed. We are interested in both moderate $p$ and large $p$; both are practically important and present a challenge to graphical modelling when a high degree of sparsity is not assumed.

The statistical framework for graphical modelling of vector-valued time series was begun by Brillinger [3] who considered both directed and undirected graphs. Two different nonparametric approaches were subsequently developed, by Dahlhaus [4] for undirected graphs, and by Bach and Jordan [1] for directed graphs. Recent work for directed graphs includes copula-Granger methodology [2].

In the approach of [4] the absence of an edge in the graphical model between series $j$ and $k$ is indicated by the corresponding partial coherence, being zero at all frequencies $f$. The partial coherence is a frequency domain version of the partial correlation coefficient and measures, at a frequency $f$, the correlation between series $j$ and $k$ when all other series involved are held constant. The partial coherence is denoted $\gamma^2_{jk\bullet\{\backslash jk\}}(f)$, where $\{\backslash jk\} = \{1 \leq i \leq p : i \neq j, k\}$, and the $_{\bullet\{\backslash jk\}}$ terminology indicates that these series are held constant. The assessment of the interaction between series $j$ and $k$ thus discounts the indirect effects of the other series. Estimated partial coherencies will include sampling variability and will never be exactly zero, so that hypothesis testing is required to test edge $(j,k)$ to see if it should be declared to be missing. The problem here is that the partial coherence for edge $(j,k)$ must be zero-tested for every frequency computed: Dahlhaus [4] suggested a test based simply on the maximum of the nonparametrically-estimated partial coherence over the frequency range, but the exact asymptotic null distribution of his test statistic is not known and only approximations have been used in practice. Nevertheless, this nonparametric approach has seen considerable use [10], [11], [26] and some variants have been developed and applied in the context of connectivity of brain signals [8], [18], [19].

The approach of [1] for *directed* graphs, while inapplicable here, had as a key component the use of the Kullback-Leibler (KL) divergence between stationary processes, formulated earlier by [13]. In this paper we use the KL divergence for determining *undirected* graphs.

As an alternative to [4], it was suggested in [7] and [23] to instead use parametric graphical models, known as 'graphical interaction models' which utilise vector autoregressive (VAR) processes to model $\{\boldsymbol{X}_t\}$. Here the VAR parameters are constrained by an associated graph; by then ranging over *all* the $2^{p(p-1)/2}$ possible graphs and (typically low) orders of the VAR

Manuscript received June 02, 2014; revised October 31, 2014 and March 12, 2015; accepted March 25, 2015. Date of publication April 13, 2015; date of current version May 15, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dmitry Malioutov. The work of R. J. Wolstenholme is supported in part by the EPSRC (U.K.).

The authors are with the Department of Mathematics, Imperial College London, London SW7 2BZ, U.K. (e-mail: rjw08@imperial.ac.uk; a.walden@imperial.ac.uk).

Digital Object Identifier 10.1109/TSP.2015.2422679

model, an information criterion (IC) can be used to select an appropriate model. However such an exhaustive search procedure is only suitable for small $p$.

Instead of using exhaustive searches, a topology selection scheme which uses a more efficient approach was given in [24]. It uses penalized maximum likelihood where the penalty term reflects sparsity constraints. For every pair of series, the resulting partial coherence is then subjected to thresholding to determine whether it can be considered to be everywhere null (for determining the missing edges). Having thus determined the missing edges, the graph is determined and constrained parameters can be estimated. By ranging over a small number of possible VAR orders and penalty weights, and computing an IC in each case, the graph giving the minimum value of the IC is selected. Unfortunately, the correct/optimum level to take for the critical thresholding step is unknown in practice.

The case where a sparse VAR model can be assumed (i.e., many parameters are zero) is considered in [5]. Partial coherence is used in selecting the group structure of the AR coefficients. For a common data set of dimension $p = 5$ results were close to those of [23]. The sparse VAR model was also used in the lasso-based method of [22].

A fully nonparametric approach to graphical modelling has the advantage of avoiding the possibility of model misspecification that can arise with parametric modelling when addressing real-world data. Indeed, Matsuda [16] proposed the identification of a graphical model for $\{\boldsymbol{X}_t\}$ based on the use of nonparametrically-estimated Kullback-Leibler (KL) divergence between two graphical models. Matsuda's test statistic is simple to compute and its asymptotic null distribution is standard Gaussian. It allows to test whether a particular nested subgraph is "correct"—in the sense that it contains the true graph—and thus to determine if restricting the set of edges poses a real constraint. Matsuda used an iterative procedure: at each step the null hypothesis that a subgraph with one edge less is correct is tested. At each such iteration the test therefore has to be carried out as many times as there are edges remaining in the graph; this is computationally very costly because of the number of test statistics needing to be computed, especially for large $p$. Moreover, for general non-decomposable graphs the computation of the test statistic employs a second iterative procedure to satisfy the constraints imposed by the currently selected graph.

While still based on Matsuda's test statistic, we introduce a much more efficient approach to identifying the model which is also well-suited to modern distributed processing:

1. We consider only tests that compare the fully connected or complete graph (alternative) with graphs that have exactly one missing edge (null hypothesis). The standard Gaussian test statistic can be calculated without resorting to any time-consuming iterations.
2. These tests are carried out using the well-known Holm method for multiple hypothesis testing. The method provides strong familywise error control which means that the type I error of rejecting any of the tested null hypotheses falsely does not exceed the specified significance level. This contrasts with Matsuda's procedure where it is unclear how the error rate used in the stepwise selection is related to the overall properties of the procedure.
3. The decreased number of tests required in our scheme, as well as the reduced computational burden for evaluating the test statistics themselves, (as iterative fitting algorithms are no longer required), produces large efficiency gains. Indeed, the number of computations for our approach is $O(p^4)$ compared to $O(p^6)$ for Matsuda's implementation.
4. In simulations our algorithm achieves power at least as good as that achieved by Matsuda's original and much slower approach.
5. In contrast to Matsuda's implementation there is no dependency between the calculation of each of the test statistics and so our algorithm can be scaled for higher dimensionality just by using more processors, i.e., it is readily parallelizable for even greater speed gains.

In Section II we review background ideas in time series graphical modelling (including the concept of a *correct* graph). Section III summarizes the construction of Matsuda's test statistic and gives a worked example showing how it is used in his backward stepwise selection procedure. In Section IV we describe our much more computationally efficient multiple hypothesis test (MHT) employing Matsuda's test statistic. The computational efficiencies of the two approaches are contrasted in Section V, justifying the $O(p^2)$ improvement for the MHT algorithm, empirically illustrated in Section VI.A. Statistical powers are compared for the two algorithms in Section VI.B, and the MHT algorithm is seen to do at least as well as Matsuda's algorithm. That the MHT algorithm performs well for higher-dimensional models (large $p$), and is readily parallelizable for even greater speed gains, is shown in Section VII. The methodology is satisfactorily applied to $p = 10$ EEG data in Section VIII. Concluding comments are provided in Section IX.

## II. Graphs and VAR Models

Throughout the paper, for a matrix $\boldsymbol{A}$, $A_{jk}$ refers to the $(j, k)$th element of $\boldsymbol{A}$ and $A^{jk}$ refers to the $(j, k)$th element of $\boldsymbol{A}^{-1}$, unless otherwise stated. Without loss of generality $\{\boldsymbol{X}_t\}$ is taken to have a mean of zero.

### A. Time Series Graphical Models

The edges between the vertices represent partial correlation between two series, i.e., there is no connection between nodes $j$ & $k$ if and only if $X_j$ and $X_k$ are partially uncorrelated given $X_{\{\backslash jk\}}$. To be precise, we remove the linear effects of $X_{\{\backslash jk\}}$ from $X_j$ to obtain the $j$th residual series defined as $\nu_{j,t} = X_{j,t} - \sum_{v\in\{\backslash jk\}} \sum_u a_{jv,u} X_{v,t-u}$, where the $p - 2$ filters $\{a_{jv,u}, u \in \mathbb{Z}\}$ give the minimum mean square prediction error. The $k$th residual series is defined likewise. The sequence $s_{\nu_j\nu_k,\tau} = \text{cov}\{\nu_{j,t+\tau}, \nu_{k,t}\}, \tau \in \mathbb{Z}$, is called the partial cross-covariance sequence and the two residual series are uncorrelated if it is everywhere zero. If $X_j$ and $X_k$ are partially uncorrelated we write $X_j \perp\!\!\!\perp X_k \mid X_{\{\backslash jk\}}$. Let $(j, k) \notin E \iff X_j \perp\!\!\!\perp X_k | X_{\{\backslash jk\}}$. Then $G$ is called a partial correlation graph. For Gaussian time series a null partial correlation equates to independence between the $j$th and $k$th conditioned series, and in this case we have a conditional independence graph.

The Fourier transform of the partial cross-covariance sequence is the partial cross-spectral density function, denoted $S_{jk\bullet(\backslash jk)}(f)$. The partial coherence is defined as

$$\gamma^2_{jk\bullet(\backslash jk)}(f) = \frac{|S_{jk\bullet(\backslash jk)}(f)|^2}{S_{jj\bullet(\backslash jk)}(f) S_{kk\bullet(\backslash jk)}(f)}, \quad -1/2 \le f < 1/2.$$

Since $S_{jk\bullet(\backslash jk)}(f) \equiv 0$ for all $-1/2 \le f < 1/2 \iff s_{\nu_j\nu_k,\tau} = 0$ for all $\tau \in \mathbb{Z}$ we see that

$$(j,k) \notin E \iff S_{jk\bullet(\backslash jk)}(\,\cdot\,) \equiv 0 \iff \gamma^2_{jk\bullet(\backslash jk)}(\,\cdot\,) \equiv 0.$$

Let $\boldsymbol{S}(f)$ denote the spectral matrix of $\{\boldsymbol{X}_t\}$ at frequency $f$, assumed to exist and be of full rank. Denoting the $(j,k)$th element of $\boldsymbol{S}^{-1}$ by $S^{jk}(f)$, the partial coherence can be expressed as, (e.g., [4]), $\gamma^2_{jk\bullet\{\backslash jk\}}(f) = |S^{jk}(f)|^2/[S^{jj}(f)S^{kk}(f)]$, and therefore

$$(j,k) \notin E \iff S^{jk}(f) = 0, \quad -1/2 \le f < 1/2,$$

i.e., if $X_j$ and $X_k$ are partially uncorrelated then there is a zero in the corresponding entry of the inverse spectral matrix [4]. (Partial correlation graphical models for time series are undirected as $(j,k) \notin E \iff (k,j) \notin E$.)

### B. Correct Graphs

An important concept in what follows is that of a *correct graph*. Such graphs can be used to identify the underlying graphical model for multivariate time series. The following definition is a slightly clarified version of that in [16].

*Definition 1:* If $(V,E)$ is the true graphical model for $\{\boldsymbol{X}_t\}$, then $(V,E')$ is correct for $(V,E)$, if

$$S^{jk}(f) = 0, \quad (j,k) \notin E' \quad \text{and} -1/2 \le f < 1/2. \tag{1}$$

Note that by this definition, if an edge is missing in $E'$ it must also be absent in $E$ for $(V,E')$ to be correct. A correct graph $(V,E')$, when imposed on top of $(V,E)$, will completely cover all its edges as $E \subseteq E'$. Also, the complete graph—containing all edges between vertices—is correct for any graphical model.

By way of an example, let $G = (V,E)$ in Fig. 1 be the true graphical model. Then the complete graph $G_0 = (V,E_0)$ completely covers $G$ and is correct for $G$. Likewise, $G_1 = (V,E')$ completely covers $G$ and is correct for $G$. However, when $G_2 = (V,E'')$ is imposed over $G$ the edge between $\{X_{2,t}\}$ and $\{X_{4,t}\}$ in $G$ is not covered. So $(2,4) \notin E''$ but $(2,4) \in E$. Therefore $E \not\subseteq E''$ and $G_2$ is not a correct graph for $G$.

It should be emphasized that we use the phrasing "$(V,E)$ is the *true* graphical model for $\{\boldsymbol{X}_t\}$" and reserve the use of the word *correct* for the special context of Definition 1.

### C. VAR Models

Here we give a very brief summary of some relevant results on VAR processes, useful for understanding ideas in our simulation examples such as "jointly influencing." We stress however that the methodology discussed in the paper is more widely applicable.

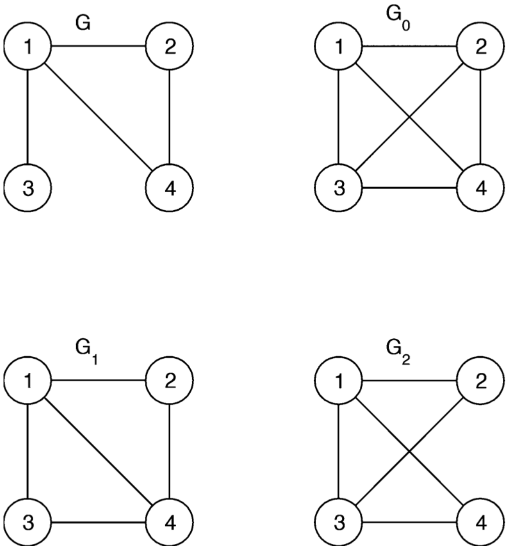

Fig. 1. Illustration of the concept of a correct graph. $G$ is the true graph. As explained in the text, $G_0$ and $G_1$ are correct for $G$ while $G_2$ is not.

$\{\boldsymbol{X}_t\}$ is a real-valued zero mean $p$-vector-valued autoregressive process of order $\ell$, or VAR$_p(\ell)$, if it is of the form $\boldsymbol{X}_t = \sum_{u=1}^{\ell} \boldsymbol{\Phi}_u \boldsymbol{X}_{t-u} + \boldsymbol{\epsilon}_t$, where the $\{\boldsymbol{\Phi}_u\}$ are $p \times p$ coefficient matrices, and $\boldsymbol{\epsilon}_t = [\epsilon_{1,t}, \ldots, \epsilon_{p,t}]^T$ is a $p$-vector-valued white noise process with a mean vector of zero and covariance matrix $\boldsymbol{\Sigma_\epsilon}$. If $\det\{\boldsymbol{I}_p - \sum_{u=1}^{\ell} \boldsymbol{\Phi}_u z^u\} \ne 0$ for all $|z| \le 1$, where $\boldsymbol{I}_p$ is a $p \times p$ identity matrix, then the process is stationary [15, p. 25]. We define $\boldsymbol{\Phi}(f) \equiv -\sum_{u=0}^{\ell} \boldsymbol{\Phi}_u \mathrm{e}^{-\mathrm{i}2\pi f u}$ and $\boldsymbol{\Phi}_0 \equiv -\boldsymbol{I}_p$.

Let $\Phi_{ij,u}$ be the $(i,j)$th element of $\boldsymbol{\Phi}_u$ where we are interested in the case $i \ne j$. Then $\Phi_{ij,u}$ is said to be the *influence* from $X_{j,t-u}$ on $X_{i,t}$ [4]. There is no influence from component $j$ on $i$ if $\Phi_{ij,u} = 0, u = 1, \ldots, \ell$, so that $\Phi_{ij}(\,\cdot\,) = 0$.

$\boldsymbol{S}(f) = \boldsymbol{\Phi}^{-1}(f)\boldsymbol{\Sigma_\epsilon}[\boldsymbol{\Phi}^{-1}(f)]^H, -1/2 \le f < 1/2$, is the spectral matrix for $\{\boldsymbol{X}_t\}$ where $^H$ denotes conjugate transpose. Then $\boldsymbol{S}^{-1}(f) = \boldsymbol{\Phi}^H(f)\,\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\,\boldsymbol{\Phi}(f)$, $-1/2 \le f < 1/2$. If $\boldsymbol{\Sigma_\epsilon} = \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}_p$ it follows [4] that if the $j$th and $k$th series do not jointly influence another series $i \ne j,k$ (i.e., $\Phi_{ij}(\,\cdot\,) = 0$ and/or $\Phi_{ik}(\,\cdot\,) = 0$), then the $j$th and $k$th series will be partially uncorrelated if and only if $\Phi_{jk}(\,\cdot\,) = 0$ and $\Phi_{kj}(\,\cdot\,) = 0$.

Later we will make use of the VAR$_5(1)$ model

$$\boldsymbol{X}_t = \boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t \tag{2}$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}_5(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon})$, the 5-dimensional Gaussian distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma_\epsilon}$.

For testing and illustration purposes we will make use of several models, named as follows:

Model A: Here $\boldsymbol{\Sigma_\epsilon} = \boldsymbol{I}_5$ and

$$\boldsymbol{\Phi}_1 = \begin{bmatrix} 0.2 & 0 & -0.1 & 0 & -0.5 \\ 0.4 & -0.2 & 0 & 0.2 & 0 \\ -0.2 & 0 & 0.3 & 0 & 0.1 \\ 0.3 & 0.1 & 0 & 0.3 & 0 \\ 0 & 0 & 0 & 0.5 & 0.2 \end{bmatrix}. \tag{3}$$

By inspection of $\boldsymbol{\Phi}_1$, we see that the set of missing edges is $\{(2,3), (2,5), (3,4)\}$.

Model B: (Matsuda [16]). Here $\boldsymbol{\Sigma_\epsilon} = \boldsymbol{I}_5$ and

$$\boldsymbol{\Phi}_1 = \begin{bmatrix} 0.2 & 0 & 0.3 & 0 & 0.3 \\ 0.3 & -0.2 & x & 0 & 0 \\ 0.2 & x & 0.3 & 0 & 0 \\ 0.2 & 0.3 & 0 & 0.3 & 0 \\ 0.2 & 0 & 0.2 & 0.2 & 0.2 \end{bmatrix}, \tag{4}$$

and we consider the cases $x = 0$ and 0.1, as used in [16]. For $x = 0$ the set of missing edges in our model is $\{(2,3),(2,5)\}$, where we note that although entries (3, 4) and (4, 3) are both zero, neither entry (5, 3) nor (5, 4) are zero so that series 3 and 4 jointly influence the 5th, and therefore edge $(3,4)$ is not missing. When $x = 0.1$, the set of missing edges is simply $\{(2,5)\}$.

Model C: This consists of $\boldsymbol{\Phi}_1$ of the form (4) with $x = 0$ but now with $\boldsymbol{\Sigma_\epsilon}^{-1} = \boldsymbol{I}_5$ except that entries (1, 2) and (2, 1) of $\boldsymbol{\Sigma}^{-1}$ are equal to 0.5. As a result of these two off-diagonal entries being non-zero, instead of missing edges $\{(2,3),(2,5)\}$ only (2, 3) is missing.

## III. Test Statistic

### A. Test for Missing Edges

Given $\{\boldsymbol{X}_t\}$ with graph $(V,E)$ and spectral matrix $\boldsymbol{S}(f)$, consider graph $(V,E')$ and matrix $\boldsymbol{T}(f)$ satisfying

$$T_{jk}(f) = S_{jk}(f),\ (j,k) \in E';\ T^{jk}(f) = 0,\ (j,k) \notin E'. \tag{5}$$

Unique existence of $\boldsymbol{T}(f)$ is shown in ([17], Lemma 7).

*Proposition 1 [16, p. 401]:* Given graph $(V,E')$, if $\boldsymbol{T}(f)$ satisfies the constraints in (5) then $(V,E')$ is correct for $(V,E)$ if and only if $\boldsymbol{S}(f) = \boldsymbol{T}(f)$.

The result in Proposition 1 can be used to determine whether graph $(V,E_2)$ is correct, given $(V,E_1)$ is correct, where $E_2 \subseteq E_1$. With $(V,E_1)$ assumed correct we have $\boldsymbol{T}_1(f) = \boldsymbol{S}(f)$. If we calculate estimators $\hat{\boldsymbol{T}}_1(f), \hat{\boldsymbol{T}}_2(f)$ using observed data, then intuitively a large difference between them suggests $\hat{\boldsymbol{T}}_2(f) \neq \hat{\boldsymbol{T}}_1(f) \approx \boldsymbol{S}(f)$ and by Proposition 1, $(V,E_2)$ would be deemed incorrect.

Assuming $(V,E_1)$ is correct, a test can be constructed between a null $(H_0)$ and alternative $(H_A)$ hypothesis:

$$H_0 : (V,E_2) \text{ is correct vs } H_A : (V,E_2) \text{ is incorrect}$$

where a measure of divergence between $\hat{\boldsymbol{T}}_1(f)$ and $\hat{\boldsymbol{T}}_2(f)$ is used to build the test statistic.

For example, suppose we want to determine whether two series are partially uncorrelated, or in fact simply uncorrelated in this case. Define

$$\hat{\boldsymbol{T}}_1(f) = \begin{bmatrix} \hat{S}_{11}(f) & \hat{S}_{12}(f) \\ \hat{S}^*_{12}(f) & \hat{S}_{22}(f) \end{bmatrix}, \tag{6}$$

the estimated spectral matrix. With $(V,E_1)$ being the complete model, with the two vertices connected, we can test against $(V,E_2)$, the model where the vertices aren't connected. The matrix satisfying (5) for $(V,E_2)$ is then

$$\hat{\boldsymbol{T}}_2(f) = \begin{bmatrix} \hat{S}_{11}(f) & 0 \\ 0 & \hat{S}_{22}(f) \end{bmatrix}. \tag{7}$$

### B. Spectral Estimator

Given vector observations $\boldsymbol{X}_0,\ldots,\boldsymbol{X}_{N-1}$, the matrix periodogram estimator $\hat{\boldsymbol{S}}^{(P)}(f)$ of $\boldsymbol{S}(f)$ takes the form $\hat{\boldsymbol{S}}^{(P)}(f) = \boldsymbol{W}(f)\boldsymbol{W}^H(f)$, where $\boldsymbol{W}(f) = \sum_{t=0}^{N-1}\boldsymbol{X}_t \mathrm{e}^{-\mathrm{i}2\pi ft}/\sqrt{N}$. $\hat{\boldsymbol{S}}^{(P)}(f)$ has unit periodicity. Let $f_j = j/N$, the $j$th Fourier frequency, then given a symmetric positive weight sequence $\{w_k\}$ for $k = -M,\ldots,M$, with $\sum w_k = 1$, the frequency-averaged periodogram is

$$\hat{\boldsymbol{S}}(f_j) = \sum_{k=-M}^{M} w_k \hat{\boldsymbol{S}}^{(P)}(f_{j-k}). \tag{8}$$

This estimator was used by Matsuda [16] in the derivation of his test statistic. It is necessary and sufficient for $\hat{\boldsymbol{S}}(f)$ to be non-singular that $2M+1 \geq p$, i.e., we have $p$ or more non-zero values in our weight sequence, e.g., [9, p. 3007]. For consistency of the spectral estimator we require $M,N \to \infty$ such that $M/N \to 0$; for the finite sample sizes used in practice we would expect $M \gg p$. $M$ can be chosen using, for example, the method of 'window closing' [20] or by cross-validation [16].

### C. Construction of Test Statistic

Estimators $\hat{\boldsymbol{T}}_1(f)$ and $\hat{\boldsymbol{T}}_2(f)$ can be found by applying the constraints in (5) to $\hat{\boldsymbol{S}}(f_j)$ in (8); the recursion of [27] is used for this purpose along with a result from [25] which justifies convergence—see [16, p. 403].

To measure the difference between $\hat{\boldsymbol{T}}_1(f)$ and $\hat{\boldsymbol{T}}_2(f)$ Matsuda [16] used the estimated Kullback-Leibler divergence, $eKL(\boldsymbol{T}_1,\boldsymbol{T}_2)$. With $N$ assumed even this is

$$\frac{1}{N}\sum_{j=1}^{N/2}\Big[\mathrm{tr}\left\{\hat{\boldsymbol{T}}_1(f_j)\hat{\boldsymbol{T}}_2^{-1}(f_j)\right\} - \log\det\left\{\hat{\boldsymbol{T}}_1(f_j)\hat{\boldsymbol{T}}_2^{-1}(f_j)\right\} - p\Big].$$

Under the following assumptions, Matsuda derived a statistic based on $eKL(\boldsymbol{T}_1,\boldsymbol{T}_2)$ which has an asymptotically standard normal, $\mathcal{N}(0,1)$, statistic:

1. $\{\boldsymbol{X}_t\}$ is a $p$-vector-valued Gaussian stationary process.
2. $\boldsymbol{S}(f)$ is positive definite for $|f| \leq 1/2$.
3. $S_{jk}(f)$ is twice continuously differentiable for $j,k = 1,\ldots,p$ and $-1/2 \leq f < 1/2$.
4. $M = O(N^\beta)$ ($M$ is at most of order $N^\beta$) for $1/2 < \beta < 3/4$ and the weight sequence $\{w_k\}$ is of the form $w_k = u\left(\frac{k}{2M}\right)$, $k = -M,\ldots,M$, where $u(\cdot)$ is a continuous even function on $[-1/2, 1/2]$.

Matsuda [16] defined the test statistic $Z_N(\boldsymbol{T}_1,\boldsymbol{T}_2)$ as

$$\left[\frac{2MN}{D_u(m_2-m_1)}\right]^{1/2}\left[eKL(\boldsymbol{T}_1,\boldsymbol{T}_2) - \frac{C_u(m_2-m_1)}{2M}\right] \tag{9}$$

where $m_i = \#\{(j,k) : (j,k) \notin E_i,\ j<k\}$, (the number of missing edges in the model), and $C_u, D_u$ are constants with values determined by $u(\cdot)$, see [16]. Given assumptions 1-4 it follows that [16]

- Under $H_0$,

$$Z_N(\boldsymbol{T}_1,\boldsymbol{T}_2) \to \mathcal{N}(0,1) \text{ as } N \to \infty \tag{10}$$

- Under $H_A, Z_N(\boldsymbol{T}_1, \boldsymbol{T}_2)$ takes the form

$$\left[\frac{2MN}{D_u(m_2 - m_1)}\right]^{1/2} KL(\boldsymbol{S}, \boldsymbol{T}_2) + o_p([MN]^{1/2}) \qquad (11)$$

  where $\boldsymbol{S}$ is the true spectral matrix, $KL(\cdot, \cdot)$ denotes the true Kullback-Leibler divergence, and $o_p([MN]^{1/2})$ denotes a term of smaller order in probability than $[MN]^{1/2}$.

Under $H_A$, the dominant term of the test statistic, the divergence, is positive and it therefore has a one-sided critical region. So for values of the statistic greater than a critical level, $H_0$ is rejected in favour of $H_A$. Also from (11) the statistic diverges to infinity at rate $[MN]^{1/2}$ under $H_A$, so that the test can be more powerful than other standard tests which diverge at the rate $N^{1/2}$ [16].

*Remark 1:* We draw attention to the fact that Matsuda's statistical results assume that the processes involved are Gaussian. He considered [16, p. 407] that this might not be a necessity, but presently this is an open question. Bach and Jordan [1] also assumed Gaussianity in their study for directed graphs.

### D. Matsuda's Algortihm

Matsuda [16] used the test statistic (9) and the recursion in [27] in a backward stepwise selection algorithm to identify the best graphical model for $\{\boldsymbol{X}_t\}$. Start by setting $(V, E_0)$ equal to the complete graph with no missing edges and choose significance level $\alpha$. Set $k = 0$ and begin:

1. Let $(V, E_{k+1}^1), (V, E_{k+1}^2), \ldots, (V, E_{k+1}^{L_k})$ be the $L_k$ distinct graphs with one more missing edge than $(V, E_k)$. Calculate the test statistics
$$Z_N^i = Z_N\left(\boldsymbol{T}_k, \boldsymbol{T}_{k+1}^i\right), \quad i = 1, \ldots, L_k,$$
with $\boldsymbol{T}_{k+1}^i$ the statistic corresponding to model $(V, E_{k+1}^i)$.
2. With $\Phi(\cdot)$ denoting the standard Gaussian distribution function, find $C_k(\alpha)$ satisfying
$$C_k(\alpha) = \Phi^{-1}((1 - \alpha)^{1/L_k}) \qquad (12)$$
and if for all $i$, $Z_N^i > C_k(\alpha)$, then stop the procedure and select $(V, E_k)$ as the graphical model for $\{\boldsymbol{X}_t\}$. Otherwise, set $(V, E_{k+1}) = (V, E_{k+1}^j)$ where $Z_N^j$ is the smallest statistic calculated.
3. Set $k = k + 1$ and loop back to step 1.

Under the assumption that all $Z_N^i$ are standard Gaussian—which they will be asymptotically if $(V, E_{k+1}^i)$ is a correct graph—the result

$$P\left\{\bigcap_{i=1}^{L_k} \left(Z_N^i \leq C_k(\alpha)\right)\right\} \geq \prod_{i=1}^{L_k} P\left\{Z_N^i \leq C_k(\alpha)\right\} = (1 - \alpha)$$

means that under the hypothesis that *all* $(V, E_{k+1}^i)$ are correct, the type I error rate is asymptotically less than $\alpha$ and the critical region is conservative [16, p. 404].

*Remark 2:* Perhaps a more intuitive definition for the type I error rate, which we use later, would be the probability of not removing an edge when $(V, E_{k+1}^i)$ is correct, i.e., it should have been removed. This is because we know the distribution of $Z_N^i$ when $(V, E_{k+1}^i)$ is correct, so this error rate can be calculated. The error rate used in the stepwise selection is only relevant in terms of the tests carried out at each step. It is unclear how it is related to the overall properties of the procedure [6, p. 158].

TABLE I
TEST STATISTICS $Z_N(T_k, T_{k+1}^i)$ AND CRITICAL LEVELS $C_k(0.05)$ FOR MATSUDA'S ALGORITHM

| Edge | $k=0$ | $k=1$ | $k=2$ | $k=3$ |
|---|---|---|---|---|
| (1,2) | 53.71 | 54.03 | 57.02 | 67.63 |
| (1,3) | 12.72 | 14.62 | 17.55 | 17.54 |
| (1,4) | 22.25 | 24.14 | 24.14 | 23.71 |
| (1,5) | 67.92 | 68.96 | 70.12 | 79.62 |
| (2,3) | 0.54 | **0.20** | — | — |
| (2,4) | 18.16 | 17.82 | 17.82 | 22.41 |
| (2,5) | 1.89 | 1.90 | **1.94** | — |
| (3,4) | **0.21** | — | — | — |
| (3,5) | 5.86 | 5.29 | 5.50 | 5.49 |
| (4,5) | 73.17 | 72.60 | 72.60 | 77.23 |
| $C_k(0.05)$ | 2.57 | 2.53 | 2.49 | 2.44 |

### E. Worked Example

The weight function chosen is $w_k = \cos(\pi k/2M)$, $k = -M, \ldots, M$ with $M = 64$. Numerical evaluation of $C_u$ and $D_u$ when $u(x) = \cos(\pi x)$ gives $C_u = 0.617$ and $D_u = 0.446$. We consider Model A of Section II.C with missing edges $\{(2,3), (2,5), (3,4)\}$. With $N = 1024$ for simulations of the VAR process, we ran Matsuda's algorithm with significance level $\alpha = 0.05$.

Let $(V, E_0)$ be the complete graph. The test statistics $Z_N(T_k, T_{k+1}^i)$ for the potential models and the critical levels $C_k(0.05)$ at which they are tested are given in Table I. The steps are interpreted as follows:

$k = 0$: Not all test statistics are above the critical level, so the process does not stop; $(V, E_1)$ is set to the graph with the edge $\{(3,4)\}$ missing as this had the lowest corresponding test statistic.

$k = 1$: Likewise $(V, E_2)$ is set to the graph with the edges $\{(2,3), (3,4)\}$ missing as $(2,3)$ had the lowest corresponding test statistic.

$k = 2$: Likewise $(V, E_3)$ is set to the graph with the edges $\{(2,3), (2,5), (3,4)\}$ missing as (2, 5) had the lowest corresponding test statistic.

$k = 3$: At this step all the statistics are above $C_3(0.05)$; we stop the process here and take $(V, E_3)$ as the estimated graph.

This procedure gave the true final graph for the model.

## IV. An Efficient Testing Procedure

### A. Multiple Hypothesis Testing

We now introduce a new and much more efficient approach for identifying the true graphical model for $\{\boldsymbol{X}_t\}$. While still based on the test statistic defined in (9), our method doesn't update at each iteration. Essentially, we carry out Matsuda's method only for $k = 0$, taking $(V, E_0)$ as the complete graph. If the value of the statistic $Z_N(\boldsymbol{T}_0, \boldsymbol{T}_1^i)$ corresponding to graph $(V, E_1^i)$ is below an appropriate critical level, it is deemed a correct graph and the missing edge $i$ should also be missing in the estimated graphical model. We construct our estimated model by removing insignificant edges via a MHT.

Our null hypotheses are of the form $H_i : (V, E_1^i)$ is correct. The alternative hypothesis in each case is the fully connected or complete graph. Each test is thus concerned with whether an edge exists between two vertices specified by the value of $i$.

*Proposition 2:* If the graph $(V, E_1^i)$ is correct for edges corresponding to $i = i_1, \ldots, i_s$ and incorrect for all others, then the graphical model $(V, E)$ for $\{\boldsymbol{X}_t\}$ is the graph with only edges $\{i_1, \ldots, i_s\}$ missing.

*Proof:* If graph $(V, E_1^i)$ is correct and $i$ corresponds to the edge $(j, k)$, then by definition $S^{jk}(f) = 0$ for $-1/2 \leq f < 1/2$ where $\boldsymbol{S}(f)$ is the spectral matrix of the true graphical model. This means that edge $(j, k)$ must also be missing in $(V, E)$ and this is the case for all $i = i_1, \ldots, i_s$. Conversely, if $(V, E_1^i)$ is incorrect, $S^{jk}(f) \neq 0$ and $(j, k)$ must necessarily be in $(V, E)$, hence the result. ■

We can list the $L = p(p-1)/2$ hypotheses in an obvious way:

$$\begin{aligned}
H_1 :&\quad (V, E_1^1) \text{ is correct; } (1,2) \notin E \\
&\quad \vdots \\
H_{p-1} :&\quad \left(V, E_1^{p-1}\right) \text{ is correct; } (1,p) \notin E \\
H_p :&\quad (V, E_1^p) \text{ is correct; } (2,3) \notin E \\
&\quad \vdots \\
H_L :&\quad (V, E_1^L) \text{ is correct; } (p-1,p) \notin E.
\end{aligned}$$

Multiple hypothesis testing may be addressed via the maximin stepdown procedure [14, Sec. 9.2]. With $Z_N^i \equiv Z_N(\boldsymbol{T}_0, \boldsymbol{T}_1^i)$ for $i = 1, \ldots, L$ and *ordered* test statistics $Z_N^{(1)} \leq \cdots \leq Z_N^{(L)}$ the corresponding hypotheses $H_{(1)}, \ldots, H_{(L)}$ can be tested using the maximin stepdown procedure:

- *Step 1*: if $Z_N^{(L)} < C_L$, accept $H_1, \ldots, H_L$.
- *Step 2*: if $Z_N^{(L)} \geq C_L$ but $Z_N^{(L-1)} < C_{L-1}$, reject $H_{(L)}$ and accept $H_{(1)}, \ldots, H_{(L-1)}$

$\vdots$

- *Step l*: if $Z_N^{(L)} \geq C_L, \ldots, Z_N^{(L-l+2)} \geq C_{(L-l+2)}$, but $Z_N^{(L-l+1)} < C_{(L-l+1)}$ reject $H_{(L)}, \ldots, H_{(L-l+2)}$ and accept $H_{(1)}, \ldots, H_{(L-l+1)}$.

$\vdots$

- *Step* $L+1$: if $Z_N^{(L)} \geq C_L, \ldots, Z_N^{(1)} \geq C_1$, reject $H_1, \ldots, H_L$.

*Remark 3:* For each of these tests $\hat{\boldsymbol{T}}_0(f) = \hat{\boldsymbol{S}}(f)$ and $\hat{\boldsymbol{T}}_1^{-1}(f)$ has only a single zero constraint so that finding it does not require the iterative scheme in [27]. Consequently, the test statistics may be assembled very easily and efficiently.

### B. Critical Levels

The choice of the critical values $C_1, \ldots, C_L$ is related to the idea of the family-wise error rate (FWER). If $Y$ is the number of true null hypotheses that are falsely rejected, then the FWER is defined as $P(Y \geq 1)$, i.e., the probability that at least one true null hypothesis will be falsely rejected. It is desired that FWER $\leq \alpha$ for all possible constellations of true and false hypotheses, the so-called strong error control ([14], (9.3)). This can be achieved using the (conservative) Holm approach [14, p. 363]: at each level the critical value can be evaluated using $C_i(\alpha) = F^{-1}(1 - \frac{\alpha}{i})$, where $F(\cdot)$ denotes the common distribution function of the test statistic under the null hypothesis, which from (10) is in fact $\Phi(\,\cdot\,)$, the standard Gaussian distribution function, in our case. So we choose our critical values according to the easily computed formula

$$C_i(\alpha) = \Phi^{-1}\left(1 - \frac{\alpha}{i}\right). \qquad (13)$$

TABLE II
ORDERED STATISTICS $Z_N^{(i)}$ AND CRITICAL LEVELS $C_i(0.05)$ FOR MHT

| $i$ | Missing Edge | $Z_N^{(i)}$ | $C_i(0.05)$ |
|---|---|---|---|
| 10 | (4,5) | 73.17 | 2.58 |
| 9 | (1,5) | 67.92 | 2.54 |
| 8 | (1,2) | 53.71 | 2.50 |
| 7 | (1,4) | 22.25 | 2.45 |
| 6 | (2,4) | 18.16 | 2.39 |
| 5 | (1,3) | 12.72 | 2.33 |
| 4 | (3,5) | 5.86 | 2.24 |
| 3 | (2,5) | 1.89 | 2.13 |
| 2 | (2,3) | 0.54 | 1.96 |
| 1 | (3,4) | 0.21 | 1.64 |

### C. Worked Examples

Using the same VAR$_5$(1) observations as in Section III.E, we list our $L = 10$ hypotheses:

$$\begin{aligned}
H_1 :&\quad \text{Edge does not exist between } (1,2) \\
H_2 :&\quad \text{Edge does not exist between } (1,3) \\
\vdots&\quad \qquad\qquad \vdots \\
H_{10} :&\quad \text{Edge does not exist between } (4,5)
\end{aligned}$$

Ordering the test statistics and including the critical levels $C_i(0.05)$ of (13) gives Table II. We can see that $Z_N^{(10)} \geq C_{10}(0.05), \ldots, Z_N^{(4)} \geq C_4(0.05)$ and $Z_N^{(3)} < C_3(0.05)$, so we reject $H_{(10)} \ldots H_{(4)}$ and accept $H_{(3)} \ldots H_{(1)}$. Note that this means our estimated graphical model is the graph with edges $\{(2,5), (2,3), (3,4)\}$ missing, the true graph for the model.

We also compared behaviors of Model B of Section II.C using $x = 0$, with Model C, the only parametric difference being that $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \neq \boldsymbol{I}_5$ for Model C. The former has missing edges $\{(2,3), (2,5)\}$ the latter has only (2, 3) missing. Constructing a table like Table II for each we find for Model B that edges (2,3) and (2,5) have associated statistics 1.49 and −0.31 and are classified as missing, all other hypotheses are rejected. For Model C edge (2,3) has associated statistics 1.07 and is classified as missing, all other hypotheses are rejected. So again the true graphs were found.

## V. EFFICIENCY CONTRAST

*Proposition 3:* The number of test statistics calculated in the Matsuda algorithm is $O(p^4)$ and in the MHT is $O(p^2)$.

*Proof:* For Matsuda's algorithm, assuming the final output is the true graphical model with $k$ missing edges,

$$\begin{aligned}
\frac{p(p-1)}{2} + \left[\frac{p(p-1)}{2} - 1\right] + \cdots + \left[\frac{p(p-1)}{2} - k\right] \\
= (k+1)\frac{p(p-1)}{2} - \frac{k(k-1)}{2} \qquad (14)
\end{aligned}$$

test statistics are calculated, where $k \in \{0, \ldots, p(p-1)/2\}$. Setting the ratio of non-edges to total possible edges to $a$, we can write $k = a\frac{p(p-1)}{2}$ for $0 \le a \le 1$. Then substituting into (14), the total number of test statistics needing to be calculated, $n$ say, satisfies

$$n = p^4\left[\frac{a}{4} - \frac{a^2}{8}\right] + o(p^4),$$

where $o(p^4)$ denotes terms of smaller order than $p^4$. For sparsity take $1/2 < a < 1$, then asymptotically in $p$,

$$\frac{3p^4}{32} < n < \frac{p^4}{8},$$

i.e., $O(p^4)$. For the MHT, regardless of the number of missing edges in the model, we always calculate $n = p(p-1)/2$ statistics, so asymptotically, $n \approx p^2/2$, i.e., $O(p^2)$. ■

Clearly the sample size, $N$, and length of weight sequence, $2M+1$, will affect the time it takes to calculate each test statistic. Also, if there is only one missing edge in our model, as is the case in the MHT, we do not need to iterate in order to find the matrix satisfying the constraints in (5). If there is more than one missing edge, as in all steps of the Matsuda algorithm excluding the first, iteration is required as set out in [27]. As the number of iterations must increase as more edges are removed from the model for a good estimate, we will denote this number at each stage as $l_k$. ($l_k = 1$ in the MHT as we only have to iterate once). It can be shown by considering the steps in the construction process that computation time for each statistic is $\sim 2NM + Np^2 l_k$.

Combining this with the number of test statistics needed to be calculated above, Matsuda's algorithm has a time $T_1 \sim 2NMp^4 + Np^6 l_k$ and for the MHT,

$$T_2 \sim 2NMp^2 + Np^4. \tag{15}$$

So the calculation times $T$ for the tests would be expected to be

$$T = \begin{cases} O(p^6) & \text{for Matsuda's algorithm;} \\ O(p^4) & \text{for the MHT.} \end{cases} \tag{16}$$

## VI. Practical Comparison for Small Dimensions

For small values of $p$ we are able to make direct practical comparisons of the two algorithms as Matsuda's can still be calculated in a reasonable time period.

### A. Timings

Fig. 2 compares calculation times $T$ in seconds, for the tests for $N = 1024, M = 32$. Fig. 2(a) plots $T_1^{1/6}$ versus $p$ for Matsuda's algorithm, while Fig. 2(b) plots $T_2^{1/4}$ versus $p$ for the MHT. In both plots these times increase linearly with $p$ as expected.

Fig. 2(c) shows the ratio $T_1/T_2$, illustrating the rapid increase in computation time for Matsuda's algorithm with $p$, compared to the MHT approach. These results were derived by randomly generating a $\mathrm{VAR}_p(1)$ model matrix $\mathbf{\Phi}_1$ (see Appendix-A) for each $p$ value considered, and then recording the completion time of each algorithm—Matsuda's or MHT—for that model.

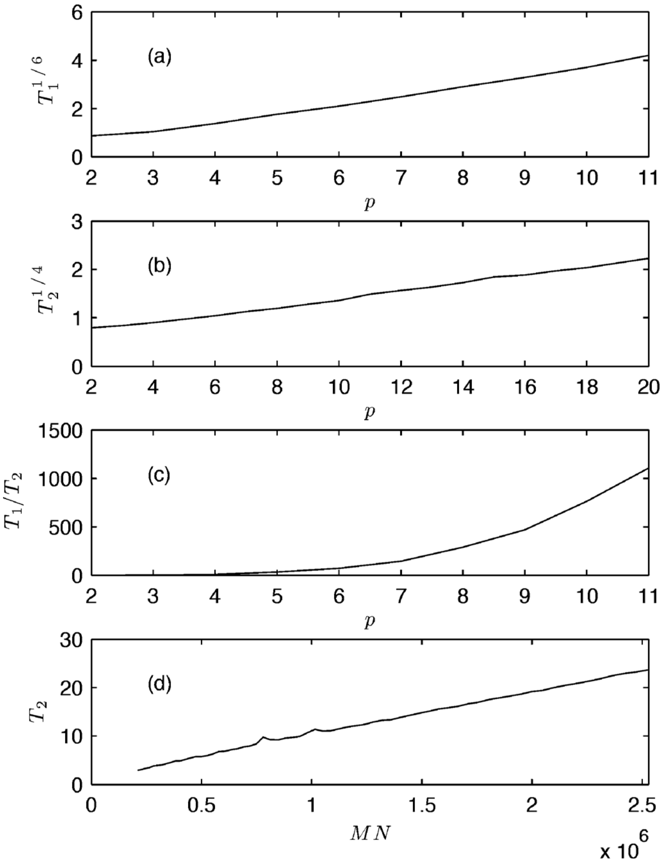

Fig. 2. Calculation timings in seconds: (a) $T_1$ for Matsuda's algorithm, to the one-sixth power, versus $p$, (b) $T_2$ for the MHT, to the one-quarter power, versus $p$, (c) the ratio of computation times $T_1/T_2$ versus $p$, and (d) $T_2$ for the MHT versus $MN$. Here $N = 1024, M = 32$.

Fig. 2(d) shows, for $p = 5$ fixed and the MHT, a plot of $T_2$ versus $MN$, where $M = N/32$ and $N$ increases from 200 to 9 000. From (15)

$$T_2 \sim 2NMp^2 + \frac{NMp^4}{M} \Rightarrow \frac{\mathrm{d}T_2}{\mathrm{d}(NM)} \sim 2p^2 + \frac{32p^4}{N},$$

which for large $N$ means that $T_2$ should have a constant gradient with $MN$, as seen in Fig. 2(d). These results were derived by randomly generating a single $\mathrm{VAR}_5(1)$ model matrix $\mathbf{\Phi}_1$, (Appendix-A), and then recording the completion time for the MHT algorithm for that model using the different $M, N$ combinations specified.

### B. Power

We will compare the results of the MHT approach against Matsuda's algorithm using two different models. To do this we utilise the concepts of (i) FWER, defined in Section IV.B, and (ii) effective power, the probability of rejecting all false hypotheses [21].

The first model is the $\mathrm{VAR}_5(1)$ model of (4) and we consider the cases $x = 0$ (missing edges $\{(2,3),(2,5)\}$), and 0.1, (single missing edge $\{(2,5)\}$), as used in [16].

We considered combinations $(N, M)$ of (512, 16), (1024, 32), (2048, 64). Results are based on 600 replications for each $(N, M)$ pair.

TABLE III
AVERAGE AND STANDARD ERROR OF VALUES OF THE MODEL B ($x = 0$) TEST STATISTIC $Z_N^i$ FOR EACH EDGE TEST WITH $N = 2048, M = 64$

| Edge | Average | Standard Error |
|---|---|---|
| (1,2) | 26.93 | 4.57 |
| (1,3) | 37.94 | 5.25 |
| (1,4) | 12.55 | 3.10 |
| (1,5) | 41.39 | 5.63 |
| **(2,3)** | **0.25** | **1.08** |
| (2,4) | 33.21 | 5.03 |
| **(2,5)** | **0.34** | **1.05** |
| **(3,4)** | **1.00** | **1.21** |
| (3,5) | 13.40 | 3.39 |
| (4,5) | 15.39 | 3.68 |

For $x = 0$ to compare the algorithms, we only consider the edges (2,3),(2,5),(3,4). This is due to the fact that these produce the three borderline statistics and while others may sometime fall outside the critical region—i.e., we reject them as edges—this is infrequent enough that simply for comparison purposes it is worth saving time by ignoring these. This approach is supported by the results in Table III which used the values $N = 2048$ and $M = 64$. (In the computations the test statistics for other edges were essentially taken to be infinity.)

The results displayed in Fig. 3 were constructed as follows. For the multiple hypothesis test, $\alpha$ was varied between 0 and 0.5 in steps of 0.00125, and used as in (13). The MHT was carried out for each of the 600 replications followed by the two steps:

1. the FWER was recorded as the proportion of the replications for which at least one true null hypothesis was falsely rejected;
2. the effective power of the test was recorded as the proportion of replications for which (3, 4) was not included as a missing edge. This is essentially the power of the sub-test on the hypotheses claiming edges (2,3), (2,5), (3,4) to be missing, since of these the only hypothesis that is false is the (3, 4) one; see Table IV.

For Matsuda's algorithm a parameter $\beta$ was created and varied between 0 and 0.5 in steps of 0.00125, and then $\alpha$ formed from $\alpha = \beta^5$; this $\alpha$ is the quantity used in (12). This approach allowed us to concentrate more $\alpha$ values near zero, resulting in a more even grid for the resultant FWER. Matsuda's algorithm was carried out for each of the 600 replications and the FWER and effective power recorded.

Figs. 3(a), (c) and (e) show the relationship between the FWER and effective power for the MHT (solid line) and Matsuda's algorithm (dashed line). As can be seen, there is no significant difference in the power of the test for the two methods.

For the case $x = 0.1$ we see from Table IV that the hypotheses stating (2, 3) and (3, 4) to be missing edges are both false. So the same basic procedure is carried out as for $x = 0$ but now the effective power is computed as the probability of rejecting both the hypotheses involving (2, 3) and (3, 4). The results are shown in Figs. 3(b), (d) and (f) from which it is seen that again the MHT does at least as well as Matsuda's algorithm.

Turning to model A of Section II.C, with $\mathbf{\Phi}_1$ given in (3) and missing edges $\{(2,3),(2,5),(3,4)\}$, we can see in Table V that the only other 'boundary edge' is (3, 5).

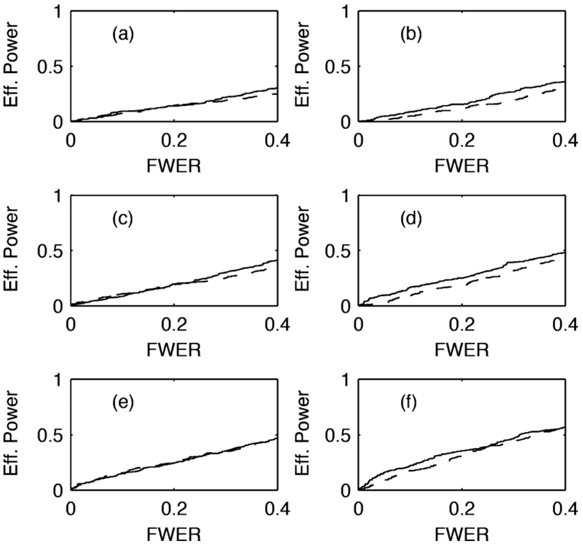

Fig. 3. FWER versus effective power for the MHT (solid lines) and Matsuda's algorithm (dashed line) for Model B, (4), with (a) $N = 512, M = 16, x = 0$ (b) $N = 512, M = 16, x = 0.1$, (c) $N = 1024, M = 32, x = 0$ and (d) $N = 1024, M = 32, x = 0.1$, (e) $N = 2048, M = 64, x = 0$ and (f) $N = 2048, M = 64, x = 0.1$.

TABLE IV
STATE OF THE MISSING EDGE HYPOTHESES FOR MODEL B WHEN $x = 0$ AND $x = 0.1$

| $x$ | missing edge hypothesis | | |
|---|---|---|---|
| | (2,3) | (2,5) | (3,4) |
| 0 | True | True | False |
| 0.1 | False | True | False |

TABLE V
AVERAGE AND STANDARD ERROR OF VALUES OF THE MODEL 2 TEST STATISTIC $Z_N^i$ FOR EACH EDGE TEST WITH $N = 2048$ AND $M = 64$

| Edge | Average | Standard Error |
|---|---|---|
| (1,2) | 50.00 | 5.93 |
| (1,3) | 15.74 | 3.52 |
| (1,4) | 22.02 | 4.34 |
| (1,5) | 64.12 | 6.79 |
| **(2,3)** | **0.29** | **1.06** |
| (2,4) | 15.66 | 3.38 |
| **(2,5)** | **0.27** | **1.06** |
| **(3,4)** | **0.32** | **1.05** |
| **(3,5)** | **3.86** | **1.95** |
| (4,5) | 66.09 | 6.61 |

TABLE VI
AVERAGE TYPE I AND II PERCENTAGE ERRORS

| | $p = 10:29$ | $p = 30$ | $p = 30:50$ |
|---|---|---|---|
| Type I | 2.2 | 3.0 | 4.1 |
| Type II | 1.3 | 2.4 | 2.9 |

Again we considered combinations $(N, M)$ of (512, 16), (1024, 32), (2048, 64) and used 600 replications for each $(N, M)$ pair. The results were calculated using the same method as above, the only difference being the effective power is now the power of the sub-test on hypotheses claiming the edges (2,3),(2,5),(3,4),(3,5) to be missing. Of these, the false hypothesis is that stating (3, 5) to be a missing edge.

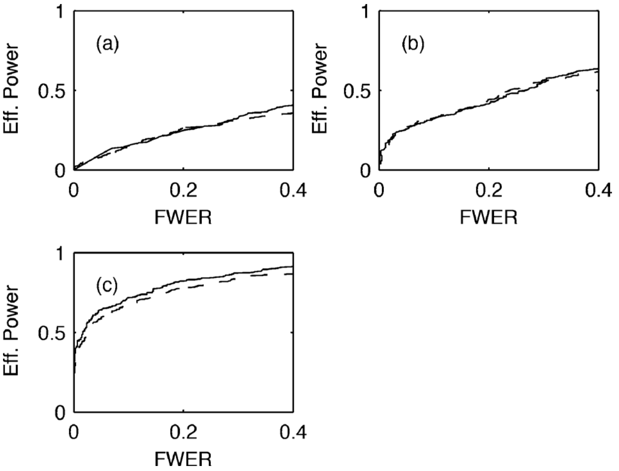

Fig. 4. FWER versus effective power for the MHT (solid lines) and Matsuda's algorithm (dashed line) for Model A, (3), and (a) $N = 512, M = 16$ (b) $N = 1024, M = 32$ and (c) $N = 2048, M = 64$.

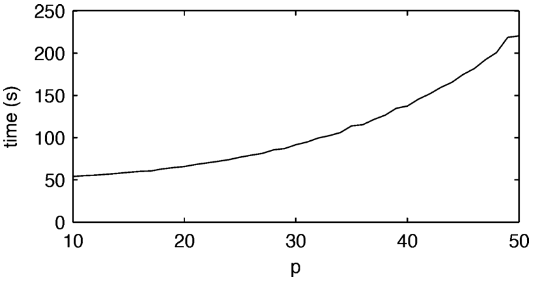

Fig. 5. Calculation timings in seconds for the MHT algorithm as $p$ varies from 10 to 50. Here $N = 2048$ and $M = 128$.

Fig. 4 compares the FWER and effective power for the MHT and Matsuda's algorithm. Again, there is no significant difference in the power of the test for the two methods.

## VII. MHT Algorithm for Higher Dimensions

We have shown that the MHT approach performs well for a relatively small number of dimensions $p$. We now look at higher dimensions.

### A. Timings

It might be thought that the inefficiency of Matsuda's algorithm is not of concern for such moderately large $p$, given modern computing power. However Fig. 5 gives timings (see Section VI.A) for the MHT algorithm in seconds for $p$ from 10 to 50 (using a 3 GHz processor). Here $N = 2048$ and $M = 128$. For $p = 50$ the time taken was about 220s; if this is scaled up (crudely) for Matsuda's algorithm by $p^2 = 2500$ we arrive at a time of over 6 days.

### B. Accuracy

Table VI reports the average type I and type II percentage errors encountered in the model estimation when $\alpha = 0.05$. Here averaging is (i) over the 20 estimated models for $p = 10 : 29$ (first column), (ii) over 100 repeat simulations for the single case $p = 30$ (second column), and (iii) over the 21 estimated models for $p = 30 : 50$ (third column). The type I percentage error is here the ratio 100(number of edges accepted when missing)/(number missing) and the type II percentage error is the ratio 100(number of edges deleted when present in the true graph)/(number present).

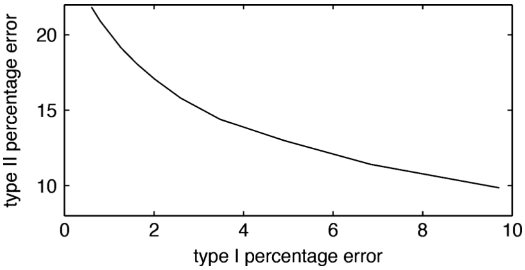

Fig. 6. Type I and II percentage errors for $p = 150$ as $\alpha$ is varied. Here $N = 2048, M = 512$.

Fig. 6 gives the type I and II percentage errors when $p$ is fixed at the large value $p = 150$ and $\alpha$ is varied. These results were derived using a $\mathrm{VAR}_p(1)$ model matrix $\mathbf{\Phi}_1$ (see Appendix A) giving rise to a true graphical model with 36% of connections present. The results seem quite satisfactory and behave in the reciprocal way expected.

### C. Parallelizability

The full algorithm consists of three stages:

1. Compute the weighted periodogram and store it in memory (along with its inverse).
2. Calculate all the necessary test statistics.
3. Select the graphical model by way of the multiple hypothesis test.

Once the test statistics have been computed the time taken to select the graphical model, step 3, is negligible.

Parallelization can be used to greatly speed up step 2, the calculation of the test statistics. In contrast to Matsuda's implementation there is no dependency between the calculation of each of the test statistics. On a multicore CPU a test statistic can be assigned to each core, and upon completion the next statistic needing calculation is assigned. Fig. 7 illustrates that the overall timing is close to linear in the reciprocal of the number of cores used.

The projected intercept of the line in Fig. 7, approximately 33 s, indicates the overheads from the parallelization and calculation of the weighted periodogram, and provides a limit on how fast the algorithm can run. Step 1, contributes to this overhead, and requires care as it can vary enormously in magnitude depending on implementation. For example, with $p = 150$, $N = 2048$ and $M = 256$ our CPU optimized implementation for step 1 took 3.4 s while a non-naïve, but nonetheless unoptimized implementation, took over 420 s.

The key message is that our algorithm is perfect for parallelization and consequent huge speed gains.

## VIII. Application to EEG Data

We now apply the MHT method to electroencephalogram (EEG) data, (resting conditions with eyes closed), for 33 males, 19 diagnosed with negative-syndrome schizophrenia, and 24 controls. This rare heritage clinical dataset from unmedicated patients was discussed in detail in [19]. Interest is in detecting

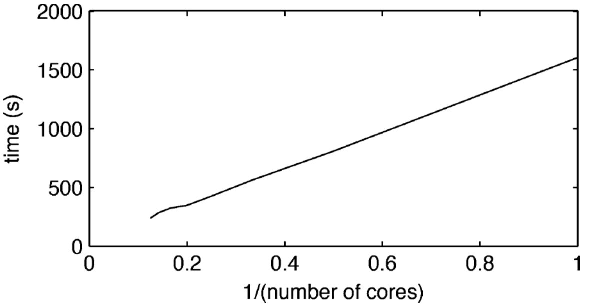

Fig. 7. Calculation timings in seconds for the MHT algorithm for $p = 60, N = 2048, M = 512$ against the reciprocal number of cores, as the number of cores varies from 1 (right of plot) to 8 (leftmost).

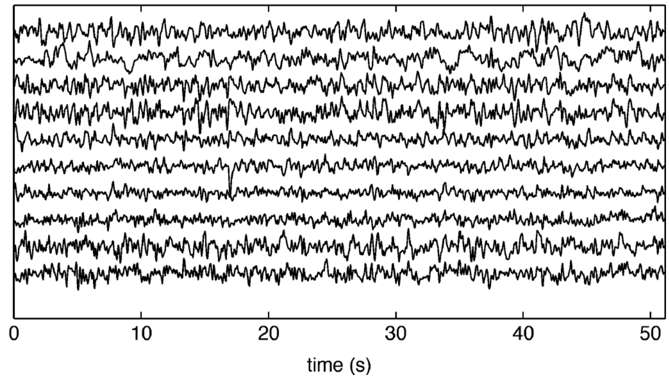

Fig. 8. Ten channel EEG time series for one of the negative-syndrome patients.

any differences in patterns of brain connectivity between the groups.

For each individual EEG was recorded on the scalp at 10 sites so that $\{\boldsymbol{X}_t\}$ is a $p = 10$ vector-valued process. There are $2^{p(p-1)/2} = 2^{45}$ possible graph structures, and $p(p-1)/2 = 45$ possible connections between the series (edges to the graph). Each possible connection was assigned a connection index from 1 to 45 as given in [19].

For illustration purposes, the ten channel time series for one of the negative-syndrome patients is shown in Fig. 8. For each of the negative-syndrome patients the MHT algorithm was used to determine whether an index-$i$ connection existed, and the percentage of the group of patients exhibiting this connection was recorded. The same was done for the control group. Fig. 9 gives the resulting percentages for each connection and both groups. For 3/4 of the connections the percentage is lower for the controls, suggesting patients exhibit a tendency towards higher connectivity, a result consistent with [19] where completely different methodology was used.

## IX. Concluding Discussion

Matsuda's approach to identification of a graphical model involves an appealing Kullback-Leibler statistic but, while improving on exhaustive search approaches, his implementation using a backward stepwise selection is extremely heavy computationally. This paper introduced a multiple hypothesis test implementation using Matsuda's statistic. The number of statistics needing to be calculated is reduced by $O(p^2)$ and the computational burden for evaluating the test statistics themselves is notably reduced as iterative fitting algorithms are no longer required.

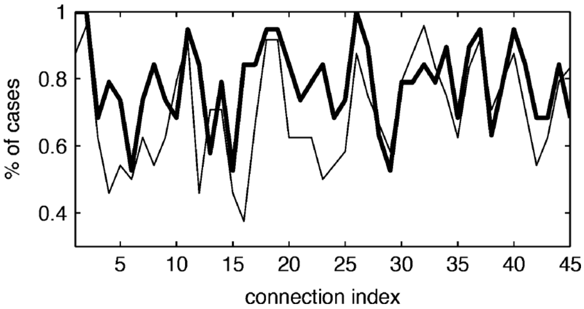

Fig. 9. Percentage of negative-syndrome patients (heavy line) and controls (thin line) exhibiting a specified connection. $(N = 1024, M = 20, \alpha = 0.01)$.

The MHT approach allows us to derive a more relevant control on the error rate in contrast to the stepwise procedure where the error rate used in each test step doesn't have a clear link to the total error of the procedure. The type I error rate we are controlling is the probability of failing to delete an edge when it is missing in the true graphical model. It may be more intuitive to define the error as deleting an edge that is contained in the true graph. In order to do this we would have to accurately know the distribution of the test statistic under this alternative, but unfortunately we don't know this.

The conservative nature of the Holm approach can in theory be somewhat offset by using an adaptive approach, (explained in detail by Guo [12]), particularly for large $p$. The result is a more powerful test than the standard Holm procedure and although the FWER will be higher, Guo showed it still controls the FWER asymptotically. We implemented this methodology but for our examples and the values of $p$ utilized, differences were very small; however, this approach is undoubtedly worthy of further investigation.

It is possible using our method to conduct an efficient stepwise approach by running the MHT and keeping all edges that clearly exist (i.e., have a very large test statistic), thus defining a new $\boldsymbol{T}_0$ to that used previously. Much of the work is thus completed. Then the MHT can be re-run to test models differing from $\boldsymbol{T}_0$ by one edge, but such additional steps require the iterative scheme [27].

Finally, we have shown that the algorithm scales very well—is highly parallelizable—with appropriate computing resources. Future work would involve rendering the algorithm for efficient calculation on high performance computing hardware such as GPUs.

## Appendix

### A. Random Model Construction

For our simulations random $\text{VAR}_p(1)$ models were constructed by randomly formulating $p \times p$ matrices $\boldsymbol{\Phi}_1$ with the number of zero entries specified as follows.

For a given $p$ value a $p \times p$ matrix $\boldsymbol{\Phi}_1$ was constructed with null entries. All diagonal elements and non-diagonal elements in position $(i, j)$ for which $(i + j)_{\text{mod}k} = 1$ were populated by random values sampled from the $\mathcal{N}(0, 1)$

distribution. The matrix was then subject to spectral decomposition and any eigenvalues with modulus greater than unity were replaced by their reciprocals and $\mathbf{\Phi}_1$ reconstructed using the modified eigenvalues. For such a $\mathbf{\Phi}_1$ we know $\det\{\boldsymbol{I}_p - \mathbf{\Phi}_1 z\} \neq 0$ for all $|z| \leq 1$, [15, pp. 15 & 653] and so a stationary process results. The choice of $k$ controls the sparsity; our default choice $k = 5$ makes approximately 64% of the $\mathbf{\Phi}_1$ matrix entries zero for $p = 10 : 50$.

## Acknowledgment

R. J. Wolstenholme thanks the EPSRC (U.K.) for financial support. Helpful comments and observations by three reviewers were much appreciated.

## References

[1] F. R. Bach and M. I. Jordan, “Learning graphical models for stationary time series,” *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2189–2199, 2004.

[2] M. T. Bahadori and Y. Liu, “An examination of practical Granger causality inference,” in *Proc. SIAM Int. Conf. Data Min.*, Austin, TX, USA, May 2–4, 2013, pp. 467–475.

[3] D. R. Brillinger, “Remarks concerning graphical models for time series and point processes,” *Revista de Econometria (Brazilian Rev. Econometr.)*, vol. 16, pp. 1–23, 1996.

[4] R. Dahlhaus, “Graphical interaction models for multivariate time series,” *Metrika*, vol. 51, pp. 157–172, 2000.

[5] R. A. Davis, P. Zang, and T. Zheng, “Sparse vector autoregressive modeling,” ArXiv preprint arXiv:1207.0520 [Online]. Available: http://arxiv.org/abs/1207.0520

[6] D. M. Edwards, *Introduction to Graphical Modelling*, 2nd ed. New York, NY, USA: Springer, 2000.

[7] M. Eichler, “Fitting graphical interaction models to multivariate time series,” in *Proc. 22nd Conf. Uncertainty in Artif. Intell.*, Arlington, VA, USA, 2006, pp. 147–154, AUAI Press.

[8] M. Fiecas and H. Ombao, “The generalized shrinkage estimator for the analysis of functional connectivity of brain signals,” *Ann. Appl. Statist.*, vol. 5, pp. 1102–1125, 2011.

[9] M. Fiecas, H. Ombao, C. Linkletter, W. Thompson, and J. Sanes, “Functional connectivity: Shrinkage estimation and randomization test,” *NeuroImage*, vol. 49, pp. 3005–3014, 2010.

[10] R. Fried and V. Didelez, “Decomposability and selection of graphical models for time series,” *Biometrika*, vol. 90, pp. 251–267, 2003.

[11] U. Gather, M. Imhoff, and R. Fried, “Graphical models for multivariate time series from intensive care monitoring,” *Statist. Med.*, vol. 21, pp. 2685–2701, 2002.

[12] W. Guo, “A note on adaptive Bonferroni and Holm procedures under dependence,” *Biometrika*, vol. 96, pp. 1012–1018, 2009.

[13] D. Kazakos and P. Papantoni-Kazakos, “Spectral distance measures between Gaussian processes,” *IEEE Trans. Autom. Control*, vol. 25, no. 5, pp. 950–959, 1980.

[14] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, 3rd ed. New York, NY, USA: Springer, 2005.

[15] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Berlin, Germany: Springer, 2006.

[16] Y. Matsuda, “A test statistic for graphical modelling of multivariate time series,” *Biometrika*, vol. 93, pp. 399–409, 2006.

[17] Y. Matsuda, Y. Yajima, and H. Tong, “Selecting models with different spectral density matrix structures by the cross-validated log likelihood criterion,” *Bernoulli*, vol. 12, pp. 221–249, 2006.

[18] T. Medkour, A. T. Walden, and A. P. Burgess, “Graphical modelling for brain connectivity via partial coherence,” *J. Neurosci. Methods*, vol. 180, pp. 374–383, 2009.

[19] T. Medkour, A. T. Walden, A. P. Burgess, and V. B. Strelets, “Brain connectivity in positive and negative syndrome schizophrenia,” *Neurosci.*, vol. 169, pp. 1779–1788, 2010.

[20] M. B. Priestley, *Spectral Analysis and Time Series*. London, UK: Academic Press, 1981.

[21] J. P. Shaffer, “Multiple hypothesis testing,” *Annu. Rev. Psychol.*, vol. 46, pp. 561–584, 1995.

[22] S. Song and P. J. Bickel, “Large vector autoregressions,” ArXiv preprint arXiv:1106.3915 [Online]. Available: http://arxiv.org/abs/1106.3915

[23] J. Songsiri, J. Dahl, and L. Vandenberghe, “Graphical models of autoregressive processes,” in *Convex Optimization in Signal Processing and Communications*, D. P. Palomar and Y. C. Eldar, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[24] J. Songsiri, J. Dahl, and L. Vandenberghe, “Topology selection in graphical models of autoregressive processes,” *J. Mach. Learn. Res.*, vol. 11, pp. 2671–2705, 2010.

[25] T. P. Speed and H. Kiiveri, “Gaussian Markov distributions over finite graphs,” *Ann. Statist.*, vol. 14, pp. 138–150, 1986.

[26] J. Timmer *et al.*, “Cross-spectral analysis of tremor time series,” *Int. J. Bifurcation Chaos*, vol. 10, pp. 2595–2610, 2000.

[27] N. Wermuth and E. Scheidt, “Fitting a covariance selection model to a matrix,” *Appl. Statist.*, vol. 26, pp. 88–92, 1977.

**R. J. Wolstenholme** received the M.Sci. degree in mathematics from Imperial College London, U.K., in 2012 and is currently working towards a Ph.D. in statistics. His research interests include graphical modelling of multichannel time series.

**A. T. Walden** (A’86–M’07–SM’11) received the B.Sc. degree in mathematics from the University of Wales, Bangor, U.K., in 1977, and the M.Sc. and Ph.D. degrees in statistics from the University of Southampton, Southampton, U.K., in 1979 and 1982, respectively. He was a Research Scientist at BP, London, U.K., from 1981 to 1990, and then joined the Department of Mathematics at Imperial College London, London, U.K., where he is currently a Professor of statistics.