16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

IDENTIFICATION OF SPARSE MULTIVARIATE AUTOREGRESSIVE MODELS

Florin Popescu

Fraunhofer Institute FIRST, Intelligent Data Analysis (IDA) Laboratory
Kekuléstrasse 7, 12489, Berlin, Germany
phone: + (49) 30 6392 1884, fax: + (49) 30 6392 1805, email: florin.popescu@first.fraunhofer.de
web: http://www.first.fraunhofer.de/brain2robot

ABSTRACT

A heuristic search method is presented by which a multivariate auto-regressive (MVAR) process is identified such that its model order, sparse structure and noise covariance is accurately recovered. A novel minimum description length (MDL) formulation of time-series linear regression is derived and applied to the problem of identifying (and coding) sparse AR matrix structures such that sparsification is largely achieved in a single initial step and improved iteratively. The method was tested against synthetic data generated by known sparse MVAR processes, compared with commonly used model selection criteria (AIC, BIC) used for identification, suggesting that it is significantly more accurate and does not overfit.

1. INTRODUCTION

The modelling of time series data has two main practical purposes: prediction and analysis. Whereas for the former, issues related to stability are important, for the latter they usually are not directly related to the tasks at hand, which can be classification of a particular recording among possible target types, data compression or spectral estimation. One commonly used tool for such purposes is linear auto-regressive (AR) modelling, otherwise called multivariate auto-regression (MVAR). Computing the AR coefficients for stationary Gaussian processes is straightforward, once the model structure is known (a list of which outputs depend on which inputs at which lags), and can be accomplished algorithmically by multivariate linear regression, among other approaches. Determining the model order, though, remains a challenging problem, despite much research devoted it and the availability of several practical algorithms for parameter estimation and model selection. This work addresses a problem which includes model order estimation, but goes beyond it in terms of complexity, namely towards that of finding the sparsity structure of an AR process from observations of data, and doing so in the presence of mixed sources, i.e. linear combinations of outputs from a standard MVAR process with uncorrelated random inputs. The challenge is finding a balance between model complexity and fitting error which allows for more reliable analysis. The requirement of sparsity comes from the practical numerical problem of the quadratic growth of the number of parameters w.r.t. the number of output variables, which can overwhelm AR estimation procedures. The minimal requirement of a model structure estimation technique is that it replicates the structure of known data generating processes with some acceptable degree of accuracy, exclusively through the analysis of generated data. In other words, the identified model should be very similar to the generating model.

The relative complexity of the problem of identifying the correct sparsity structure of an MVAR generating model can be understood intuitively. Whatever the model selection criterion we use, there is some kind of quantitative measure (an objective function) which we must minimize, which depends on the observed data and whose independent variable is the identified model structure. In the case of generating model order K we may imagine a stepwise search of models of order 1,2… K, K+1 which will hopefully exhibit a distinct minimum at order K. However, for a D dimensional system there are $2^{\wedge}(D^2(K+1))$ combinations of non-zero AR parameters that we need to search over in a similar fashion, which is prohibitively expensive. In the absence of a convex objective function, some type of type of realizable heuristic search procedure must be implemented which demonstrates, empirically, an encouraging measure of success. Whereas previous approaches, with applications in biomedical imaging signal analysis [1, 2] have concentrated on Bayesian approaches to this problem, the current work focuses on minimum description length .

2. MULTIVARIATE AUTO-REGRESSION

2.1 Equivalent forms of AR models

The most general (and common) form of writing the governing equation of an AR process, or an AR model is:

$$x_n = \sum_{i=1}^{K} A_i x_{n-i} + b + Be_n \qquad (1)$$

where x is a D-dimensional vector and correspondingly A_i and B are $D \times D$ matrices while b is a D-vector called a *bias*. This is a form referred to herein as a *pre-mix* AR form. If we assume e to be a vector of i.i.d. noise inputs of unit variance we should make further assumptions on the type of matrix B can be (if $B \neq I$ we call the system *mixed*):

$$E_n(e_{n,p}^T B^T B e_{n,q}) = \left(B^T B\right)_{pq} \simeq C_{pq} = C_{qp}$$

$$E_n(e_{n,p}^T B^T B e_{m,q\neq p}) = 0$$

where $B^TB = C$ is a $D \times D$ *covariance* matrix. The matrix B thus defined can be decomposed into a unitary orthogonal *rotation* matrix R and a diagonal *scaling* matrix S (B=RS).

$$x_n = \sum_{i=1}^{K} A_i x_{n-i} + b + RSe_n$$

Premultiplying by R^{-1} (by orthogonality $R^{-1} = R^T$)

$$R^T x_n = \sum_{i=1}^{K} R^T A_i x_{n-i} + R^T b + Se_n$$

Which can be re-written as:

$$\begin{aligned} y_n &= \sum_{i=1}^{K} R^T A_i R y_{n-i} + R^T b + Se_n \\ y_n &= \sum_{i=1}^{K} A_{R,i} y_{n-i} + b_R + Se_n \\ x_n &= Ry_n \end{aligned} \qquad (2)$$

This is what we will refer to as the *post-mix* form and it is the type of AR structure that we are interested in this paper, corresponding to physical processes in which we assume that independent sources interact dynamically (A_i) and are mixed spatially (*R*) into a set of observations.

2.2 Estimation of AR parameters

Although equations (1) and (2) represent the same process, and thus pre- and post- mix versions have the same model order, rotation changes the sparsity structure. For reasons which will be later apparent, a somewhat unusual procedure for estimating AR parameters is presented, despite the existence of well-established estimation techniques [3], which are not trivial, especially if *B* in (1) is not diagonal. Starting with the somewhat modified post-mix form:

$$x_n = A_{U,0} x_n + \sum_{i=1}^{K} A_{U,i} x_{n-i} + b_U + S_U e_{U,n} \qquad (3)$$

where the null-lag matrix $A_{U,0}$ is a strictly upper diagonal:

$$A_{U,0,p,q} = 0 \quad if \quad q \le p$$

This has a distinct computational advantage over (1): it can be re-arranged into a set of *D* standard linear regression problems given *N* observations (so-called NRML approach):

$$\begin{aligned} x_{K..N,d} &= W_d(x_{1..N-K,1..D}) \begin{bmatrix} a_{U,d} \\ b_{U,d} \end{bmatrix} + \varepsilon_{U,K..N,d} \\ s_d(\boldsymbol{\theta}, \mathbf{x}) &\triangleq \|\boldsymbol{\varepsilon}_d\| = \sqrt{\boldsymbol{\varepsilon}_d^2 \boldsymbol{\varepsilon}_d} \quad \text{where} \\ \mathbf{x}_d &= \mathbf{W}_d \boldsymbol{\theta}_d + \boldsymbol{\varepsilon}_d = \mathbf{W}_d \boldsymbol{\theta}_d + s_d(\boldsymbol{\theta}, \mathbf{x}) \mathbf{e}_d \end{aligned} \qquad (4)$$

where *W* is constructed from 1's, 0's and entries of *x* and the vector $\mathbf{a}_{U,}$ concatenates all *d*-th rows of the matrices A_U and s_d is defined as the 2 norm of the residual in each dimension.

Due to the strictly upper diagonal structure of $A_{U,0}$ the residuals $e_{U,K..N,d}$ are uncorrelated across dimension *d* [4]. Actually, eqn. (3) is actually yet another form (the *unmixed* form) of (2). To see this:

$$x_n = A_{U,0} x_n + \sum_{i=1}^{K} A_{U,i} x_{n-i} + b_U + S_U e_{U,n}$$

$$S_U^{-1} x_n = S_U^{-1} A_{U,0} x_n + S_U^{-1} \sum_{i=1}^{K} A_{U,i} x_{n-i} + S_U^{-1} b_U + e_{U,n}$$

$$\boxed{S_U^{-1}(I - A_{U,0})} x_n = S_U^{-1} \sum_{i=1}^{K} A_{U,i} x_{n-i} + S_U^{-1} b_U + e_{U,n}$$

Performing the singular value decomposition on the boxed LHS matrix (keeping in mind *U* and *V* are orthogonal)

$$\boxed{U_U \Sigma_U V_U^T} x_n = S_U^{-1} \sum_{i=1}^{K} A_{U,i} x_{n-i} + S_U^{-1} b_U + e_{U,n}$$

$$V_U^T x_U = \Sigma_U^{-1} U_U^T S_U^{-1} \sum_{i=1}^{K} A_{U,i} x_{n-i} + \Sigma_U^{-1} U_U^T S_U^{-1} b_U + \Sigma_U^{-1} \left(U_U^T e_{U,n} \right)$$

Performing the coordinate transformation $y_n = \boxed{V^T} x_n$

$$y_n = \sum_{i=1}^{K} \boxed{\Sigma_U^{-1} U_U^T S_U^{-1} A_{U,i} V_U} y_{n-i} + \boxed{\Sigma_U^{-1} U_U^T S_U^{-1}} b_U + \boxed{\Sigma_U^{-1}} e_n$$

We have the same form as (2), if we take care to re-arrange *V* such that it is not only orthogonal but a rotation matrix also. Note also that an orthogonal transformation of a unit variance, uncorrelated noise vector is also unit variance and uncorrelated $U_U^T e_{U,n} = e_n \rightarrow e_n^T e_n = e_U^T U_U U_U^T e_{U,n} = e_U^T e_{U,n}$

Thus a MVAR estimation procedure is established involving *D* independent linear regressions – for *mixed* AR systems. It is the nature of linear regression itself that provides the opportunity for principled sparsification.

3. MDL FOR REGRESSION

The Minimum Description Length principle has been proposed by Rissanen [5] and applied to MVAR model order estimation, and usually is expressed in terms of the overall number of variables used and the number of training points (not structure). A more precise bound on coding length was used, specifying codes, code lengths and approximations (expressed in bits) for each of the estimated parameters θ and for the presumably random vector of residuals ε [6], based on previous work by the author.

For the parameters, a prefix-free code based on fractional-exponent binary representation (called α-code) was used, whose length is denoted by the function $\bar{\alpha}$

$$\begin{aligned} &\theta_i - \Delta_i \le \alpha^{-1}\left(\alpha\left(\theta_i, \Delta_i\right)\right) \le \theta_i + \Delta_i \\ &\bar{\alpha}(\theta_i, \Delta_i) \simeq c_1 \log\left(\left| c_2 \left(\log^2(\log|\theta_i| + c_3) \right) - 1 \right| + \right. \\ &\left. \log\left(erf\left(c_4 (1 - \Delta_i) |\theta_i|^{-1} + 1 \right) |\theta_i| + \Delta_i \right) - \log \Delta_i \right) + 1 \end{aligned}$$

The approximation contains 4 numerically fit O(1) constants. As α is prefix-free, the code for a vector of rationals $\alpha(\boldsymbol{\theta})$ is simply the concatenation of the codes of its elements $\alpha(\theta_i)$.

For the residuals, it was assumed that the recorded values are quantized with resolution Δ_X which is known either through the properties of the recording equipment, the problem representation, or other post-hoc analysis. The coding problem, i.e. a scheme which allows the counting of errors of any size, becomes one of counting the maximal expected number of hypercubes of size Δ_X that intersect an *N* dimensional sphere. If $s_{d,\mathbf{m}}$ is the 2-norm of the residual

vector ε_d to be coded, $s_{d,\mathbf{m}} \gg \Delta_{\mathbf{x}}$ and N is large, the following approximation holds:

$$\bar{h}_{\Delta_{\mathbf{x}}}(s_d{}^2(\varepsilon_d), N) \cong \log\left(\frac{\pi^{N/2}}{\Gamma(N/2+1)}\right) + \frac{N}{2}\log\left(\frac{s_d{}^2}{\Delta_{\mathbf{x}}^2}\right)$$

The *h* code is a spiral-spherical universal counting scheme and is compact for random vectors which are generated by spherical or elliptical probability distributions (such as the Gaussian). Given that our residuals are actually composed of *D* uncorrelated thus independent vectors (as assumed), the total coding length for residuals is a sum over channels.

$$\bar{h}(\boldsymbol{\varepsilon}) = \sum_{d=1}^{D} \bar{h}_{\Delta_{\mathbf{x}}}(s_d{}^2(\varepsilon_d), N)$$

The complete (loss-less) code length approximation of the data, is to be minimized is:

$$\lambda_{\mathbf{m}}(\boldsymbol{\theta}, \boldsymbol{\Delta}, \mathbf{x}) = \bar{\alpha}(\boldsymbol{\theta}, \boldsymbol{\Delta}) + \sum_{d=1}^{D} \bar{h}_{\Delta_{\mathbf{x}}}\left(s_{d,\mathbf{m}}^2\left(\alpha^{-1}\left(\alpha(\boldsymbol{\theta}, \boldsymbol{\Delta})\right), \mathbf{x}\right), N\right)$$

Note that this equation requires the function $s_d(\cdot)$, which relates the size or the approximation error of some regression to the parameters and their storage precision), for some model **m** which specifies how the parameters encode **x**. Thus principle can be applied to any type of regression fit.

4. MDL FOR LINEAR REGRESSION

The α code mentioned above has a rather nice property in that the numerical value of the de-coding $\alpha^{-1}(\alpha(\theta_i, \Delta_i)) = \theta_i + \delta_i$ is uniformly distributed on the interval $(\theta_i\text{-}\Delta_i,\ \theta_i\text{+}\Delta_i)$ with $p(\delta_i)=(2\ \Delta_i)^{-1}$, over all possible pairs of 1. This allows us to take the expected value of the increase in the norm of the residual. Calculating the growth of residual norm w.r.t. parameter accuracy $\boldsymbol{\delta}$ is straightforward:

$$s_d^2\left(\alpha^{-1}\left(\alpha(\boldsymbol{\theta}_d, \boldsymbol{\Delta}_d)\right), \mathbf{x}\right) =$$
$$\left(\mathbf{x}_d - \mathbf{W}_d \alpha^{-1}\left(\alpha(\boldsymbol{\theta}_d, \boldsymbol{\Delta}_d)\right)\right)^T \left(\mathbf{x}_d - \mathbf{W}_d \alpha^{-1}\left(\alpha(\boldsymbol{\theta}_d, \boldsymbol{\Delta}_d)\right)\right)$$
$$\left(\mathbf{x}_d - \mathbf{W}_d\left(\boldsymbol{\theta}_d + \boldsymbol{\delta}_d\right)\right)^T \left(\mathbf{x}_d - \mathbf{W}_d\left(\boldsymbol{\theta}_d + \boldsymbol{\delta}_d\right)\right)$$

In the case of (4), if we choose $\boldsymbol{\theta}_d = \left(\mathbf{W}_d^T \mathbf{W}_d\right)^{-1} \mathbf{W}_d^T \mathbf{x}_d$ it minimizes 2-norm of the residual $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$, so the gradient of this value is w.r.t $\boldsymbol{\theta}_d$ zero, meaning any perturbation from $\boldsymbol{\theta}_d$ results in strictly quadratic growth:

$$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + \boldsymbol{\delta}_d^T \mathbf{W}_d^T \mathbf{W}_d \boldsymbol{\delta}_d \triangleq \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} + \boldsymbol{\delta}_d^T \mathbf{Q}_d \boldsymbol{\delta}_d$$

Taking the expected value over possible perturbations (i.e. the outputs of the coder-decoder):

$$E(\boldsymbol{\delta}_{d,i}^T (\mathbf{Q}_d)_{i,i} \boldsymbol{\delta}_{d,i}) = \int_{-\Delta_i}^{\Delta_i} p(\delta_{d,i})(\mathbf{Q}_d)_{i,i} \delta_{d,i}^2 d\delta_{d,i}$$
$$= \left(2\Delta_{d,i}\right)^{-1} \frac{2}{3}(\mathbf{Q}_d)_{i,i} \Delta_{d,i}^3 = \frac{1}{3}(\mathbf{Q}_d)_{i,i} \Delta_{d,i}^2$$
$$E\left(\boldsymbol{\delta}_{d,i}^T (\mathbf{Q}_d)_{i,j} \boldsymbol{\delta}_{d,j}\right)_{i \neq j} = c \int_{-\Delta_j}^{\Delta_j} \int_{-\Delta_i}^{\Delta_i} (\mathbf{Q}_d)_{i,j} \delta_{d,i} \delta_{d,j} d\delta_{d,i} d\delta_{d,j} = 0$$
$$\therefore E(\boldsymbol{\delta}_d^T \mathbf{Q}_d \boldsymbol{\delta}_d) = E(\sum_{i,j} \boldsymbol{\delta}_{d,i}^T (\mathbf{Q}_d)_{i,j} \boldsymbol{\delta}_{d,j}) = \frac{1}{3}\sum_i (\mathbf{Q}_d)_{i,i} \left(\Delta_i^2\right)$$

Finally, a differentiable expression for the MDL optimization function for linear regression can be written.

$$E_\delta \lambda_{\mathbf{m}}(\boldsymbol{\theta}, \boldsymbol{\Delta}, \mathbf{x}) = \bar{\alpha}(\boldsymbol{\theta}, \boldsymbol{\Delta}) + \sum_d E_\delta \bar{h}_{\Delta_{\mathbf{x}}}\left(s_{d,\mathbf{m}_{AR}}^2 + \boldsymbol{\delta}_d^T \mathbf{Q}_{d,\mathbf{m}} \boldsymbol{\delta}_d, N\right) \cong$$
$$\sum_i \bar{\alpha}(\theta_i, \Delta_i) + \sum_d \bar{h}_{\Delta_{\mathbf{x}}}\left(s_{d,\mathbf{m}}^2(\boldsymbol{\theta}, x) + \frac{1}{3}\sum_i (\mathbf{Q}_{d,\mathbf{m}})_{i,i} \Delta_i^2, N\right) \quad (5)$$

The approximation $E_\varepsilon(\bar{h}(x+\varepsilon)) \cong \bar{h}(x + E_\varepsilon(\varepsilon))$ is valid if $\bar{h}$ is locally linear – being O(log(*x*)) that assumption is not unreasonable. The 'optimal' AR model thus defined is:

$$\{\mathbf{m}_{AR}^*, \Delta^*\} = \arg\min_{\mathbf{m}_{AR}, \boldsymbol{\Delta}} E_\delta \lambda(\boldsymbol{\theta}, \boldsymbol{\Delta}, \mathbf{x}, \mathbf{m}_{AR})$$

5. OPTIMIZATION PROCEDURE

The objective function describes the expected description length of the time series **x** given a time series model **m** at precision Δ_X. In this work, a search over post-mix AR models is undertaken, so it is worth specifying that a model $\mathbf{m}_{AR}$ is a set of variables which tell us how to reconstruct the data given an error vector.

$$\mathbf{m}_{AR,p} = \{\boldsymbol{\omega}, \mathbf{s}, \mathbf{a}, \boldsymbol{\theta}\}_p$$

where, relative to (1) the angle vector $\boldsymbol{\omega}$ specifies the rotation matrix by the polar transformation R=exp(*T*(**ω**)) where *T*(**ω**) rearranges the angle vector into the required anti-symmetric matrix and **s** is the diagonal of *S*. The positive integer vector **a** is an index set into an enumeration of all elements of $A_{0,1,\ldots\infty}$ (implicitly defining *K*). The *active* parameters indexed by **a** are those which are included in the the model: the shorter **a** is, the sparser the model.

A simplification to the strict MDL principle we made was to assume that coding length of **ω** and **s** vary minimally or meaninglessly amongst candidate AR models (i.e. we can add a constant factor which depends only on *D* to (5)). The objective function $E_\delta\lambda_m$ is not convex w.r.t. $\boldsymbol{\Delta}$ and the optimization procedure for any given AR model is not trivial. Even if there existed a procedure which would give us the optimal $\boldsymbol{\Delta}$ for a given **m**, an exhaustive search over model structure space is prohibitively expensive. Given a certain model, though, that any parameters for which after the optimization step, $|\theta_i| < \Delta_i$ means that those parameters can be set to 0 - in the AR context this means they can be eliminated from the active set a new model is recomputed. Such a step is called a *cull*. In a *rotation*, if any elements of $A_{U,0,q}$ are non-zero at a given iteration *q*, a new *R* is computed such that $A_{U,0,q+1}$ is zero as described in (2.2). Finally, some of the less significant parameters can be eliminated – e.g. those with low $\bar{\alpha}(\theta_i, \Delta_i)$ - this is called a *trim*.

The basic algorithm used was the following:

1. R_1=I. Fit $\mathbf{m}_{\mathrm{AR1}}$.
2. For a set of N_{TRIM} least significant parameters trim all combinations of them and pick $\mathbf{m}_{\mathrm{ARq+1}}$ the active parameter set with lowest MDL until MDL grows.
3. If rotation is not necessary, terminate MDL search. If a rotation is necessary, compute and apply R_{q+1}, while re-activating all parameters of lower lag than the current maximum active lag. Cull thereafter until active no parameter is zero. Go to 2.

After either step 1. or 2. culling was performed recursively until no active parameters' coded values equalled 0.

6. NUMERICAL RESULTS

For validation of the procedure, 500 data sets were analysed, produced by random-valued but stable AR models whose order varied randomly from 2 to 5, whose elements were $\boldsymbol{N}(0,2/D)$, whose sparsity (bias included) varied uniformly from empty to full,. $\Delta_{\mathbf{X}}$ was set at 0.01, the biases were $\boldsymbol{N}(0,1)$ and the gains were random with $1\text{-}\boldsymbol{U}(0,1)$, rotation angles random with $\boldsymbol{U}(0°,30°)$ where $\boldsymbol{N}$ and $\boldsymbol{U}$ stand for the normal and uniform distributions.

	N=200	N=2000	N=20000
$\bar{h}(\mathbf{x})(ND)^{-1}$ (bits)	**8.40**±0.41	**8.69**±0.44	**8.70**±0.45
$\bar{h}(\boldsymbol{\varepsilon}_{\mathrm{SIM}})(ND)^{-1}$	**7.90**±0.13	**8.12**±0.13	**8.12**±0.13
$\bar{h}(\boldsymbol{\varepsilon}_{\mathrm{KNOWN}})(ND)^{-1}$	**8.51**±0.17	**8.16**±0.14	**8.14**±0.14
$\bar{h}(\boldsymbol{\varepsilon}_{\mathrm{MDLU}})(ND)^{-1}$	**8.32**±0.16	**8.16**±0.14	**8.13**±0.13
$\bar{h}(\boldsymbol{\varepsilon}_{\mathrm{MDLM}})(ND)^{-1}$	**8.44**±0.27	**8.30**±0.23	**8.13**±0.14
$\Delta\phi_{\mathrm{SIM,MDLU}}$ (°)	**16.3**±5.8	**16.4**±5.6	**16.2**±5.7
$\Delta\phi_{\mathrm{SIM,MDLM}}$	**14.5**±6.6	**10.8**±6.9	**6.89**±6.5
$\lvert\theta_{\mathrm{MDLU}}\rvert-\lvert\theta_{\mathrm{SIM}}\rvert$	**0.23**±2.6	**7.31**±5.2	**11.5**±7.7
$\lvert\theta_{\mathrm{MDLM}}\rvert-\lvert\theta_{\mathrm{SIM}}\rvert$	**-2.21**±3.5	**-0.87**±3.2	**2.25**±2.6
$\lvert\theta_{\mathrm{SBC}}\rvert-\lvert\theta_{\mathrm{SIM}}\rvert$	**25.8**±12	**26.6**±12	**26.7**±12
$\lvert\theta_{\mathrm{FPE}}\rvert-\lvert\theta_{\mathrm{SIM}}\rvert$	**27.2**±12	**27.2**±12	**27.1**±12
$K_{\mathrm{MDLU}}=K_{\mathrm{SIM}}$ (%)	13<**75**<12	4.2<**90**<5.4	2.4<**92**<5.2
$K_{\mathrm{MDLM}}=K_{\mathrm{SIM}}$	24<**68**<7.8	12<**84**<4	2.6<**92**<5.2
$K_{\mathrm{SBC}}=K_{\mathrm{SIM}}$	6.8<**42**<51	1<**48**<51	0.2<**49**<51
$K_{\mathrm{FPE}}=K_{\mathrm{SIM}}$	1.4<**44**<54	0<**45**<55	0<**47**<53

Table 1. **Simulation results**. The subscripts refer to the following: simulated system (SIM), MDL optimization with pre-knowledge of model structure (KNOWN), MDL optimization without rotations, but with upper diagonal entries in $A_{U,0}$ non-zero(MDLU), Schwartz criterion (SBC), Akaike (FPE), and finally MDL optimization with rotations (MDLM) – its corresponding rows are highlighted. X±y refers to mean and mean absolute deviation, whereas x<Y<z mean percentage less than, equal to, and greater than.

Data size vectors of 200, 2000 and 20,000 points with $\boldsymbol{N}(0,1)$ noise were constructed. The entire procedure was programmed in MATLAB (Mathworks, USA). The optimization of eqn. (5) was performed by the Levenberg-Marquardt algorithm. Of interest was relative performance to known model selection criteria: Schwarz's Bayesian Information Criterion (SBC) and Akaike's Final Prediction Error (FPE) criterion (also known as BIC and AIC [7]). The maximum model order considered in all approaches was 10. The routines of the package ARFIT [8] were used to fit pre-mixed AR models with SBC and FPE.

The quantities of interest were first of all $\bar{h}(\mathbf{x})$ vs. $\bar{h}(\boldsymbol{\varepsilon})$, i.e. the coding cost of the signal assuming it is random vs. the coding cost of the residuals, expressed in bits/data point (entropy rate), where the number of data points for residuals is taken to be ND. Of further interest were the errors in estimating the rotation matrix R_{SIM}, defined as:

$$\Delta\phi(R_{S1M},R) \triangleq \tan^{-1}\left(0.5\min\left\|\left(R_{SIM}{}^{T}R\right)_{d,..}-(\pm1,0,0)\right\|\right)$$

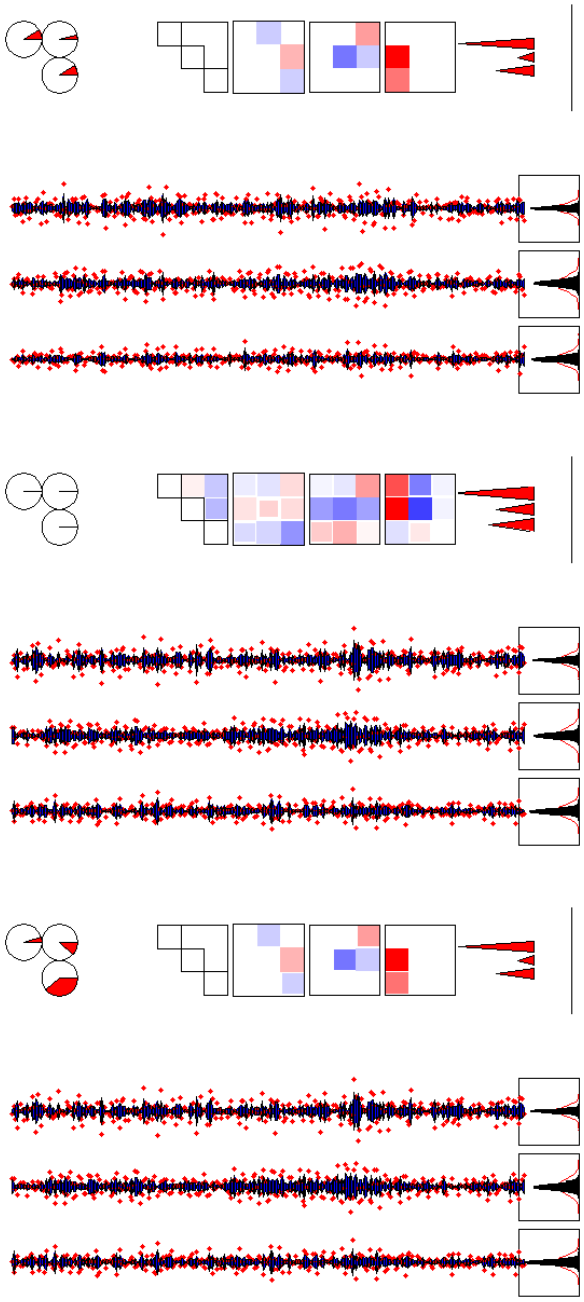

Figure 1. Sample simulated system and its identification. Panels, top to bottom: the simulated (SIM), the MDLU and MDLM systems. Each panel, top row, left to right: rotation matrix representation (angles), Hinton diagrams of matrices $A_{0..K}$ - size of square proportional to $\alpha(\theta_i)$, a triangular bar chart of channel gains (dB). Bottom 3 rows: sample of signal (dots) and residuals (solid line), and a histogram of signal (line) and of residuals (solid).

This value measures effective rotation (the angle by which a full set of orthogonal vectors differ under the two respective rotations). Regarding sparsity, we recorded the number of parameters identified $|\theta|$ as well as the model order K which is identified. A summary of results is presented in Table 1. Mean $\bar{\alpha}$ was .019 bits/pt and estimation accuracy was 12.9, 11.7 and 4.4% with increasing N. A graphic demonstration of a simulated system and its identified structure is shown in Figure 1. Note that for this particular system although the MDL or h lengths are very close (SIM 8.1502, MDLU 8.158, MDLM 8.154) the MLDM system identified, besides being correct, is far sparser than MDLU. Rotation angles are not the same but the SIM and MDLM rotation *matrices* are nearly orthogonal ($\Delta\phi$=0.38°) – MDLM systems may be equivalent under 90° or 180° rotations or channel swapping.

7. DISCUSSION

The numerical results showed that the proposed method (MDLM) identifies the correct model order with high accuracy and deviates on average by 2 or 3 parameters for the models simulated (with no discernible trend towards either over- or under- fitting), with higher accuracy for longer input data sets and with entropy rates almost undistinguishable from those simulated. Without rotation (MDLU) slightly worse performance for information rates and number of parameters is obtained while the rotation error is significantly higher (16 vs. 7°). In contrast, the classical model order criteria overfit significantly in both order and number parameters – the latter being expected as they are not usually meant to identify sparse structures.

Like SBC and FPE, MDL [9] has traditionally been expressed in terms of numbers of parameters in previous work [7]. The proposed parameter-by-parameter approach is somewhat in line with the work described in [10] but with quite different accounting of coding lengths and with introduction of previously unaccounted for signal de-mixing (rotation) which is important from a compression perspective. If the rotation matrix of the post-mix form is accurately determined, so is the noise covariance of the equivalent pre-mix form. The compression rates achieved are not high: the simulations ensured this deliberately, since we are generally interested in analysing signals which are highly stochastic. Note that the parameter coding lengths were much smaller than those for residuals yet MDLM fit well.

A further novelty was the minimization of what cannot be described as a bound [7], but rather the expected coding length of the data set given uncertainties in the output value of the coding-decoding process, which has the following intuitive justification: any coding scheme for a parameter limited to a certain interval *should* code for a rational number which can appear with uniform likelihood within the interval – otherwise it would be inefficient. The more parameters a model has, the closer the total code length for parameters is to its predicted average. More importantly the proposed approach does not (as in previous MDL formulations) perform mini-max on upper bound estimates of coding length, which would be somewhat conservative and may lead to underfit although both formulations would exhibit strong consistency [11], i.e. that they 'work' given enough data. However, a full discussion of this issue is beyond the scope of this work - many other convolved derivations such as those involving universal coding schemes were omitted for brevity and clarity herein. The objective function used means, simply: the most likely minimum description length given an optimal continuous-valued parameter coding scheme over bounded uncertainty intervals. Numerically, the predicted description length for a set of parameters coded at a certain precision was very close to that predicted by eqn. (5).

Of singular importance is that the algorithm identified the structure of 'hidden' sparse AR systems well. Naturally, the AR model is only an abstracted schema which may mis-represent the physical processes generating observed data. Further work is necessary on compressing and classifying real-world signals, and extending the proposed approach to non-stationary, non-Gaussian and nonlinear systems.

ACKNOWLEDGMENTS

Supported by the EC grant MEXT-CT-2004-014194.

REFERENCES

[1] L. Harrison, W. D. Penny, and K. J. Friston, "Multivariate Autoregressive Modelling of fMRI time series," *Neuroimage*, vol. 19, pp. 1477--1491, 2003.

[2] P. A. Valdés-Sosa, J. M. Sanchez-Bornot, A. Lage-Castellanos, M. Vega-Hernández, J. Bosch-Bayard, L. Melie-García, and E. Canales-Rodríguez, "Estimating brain functional connectivity with sparse multivariate autoregression," *Philos Trans R Soc Lond B Biol Sci*, vol. 360, pp. 969-981, 2005.

[3] A. Schlögl, "Comparison of Multivariate Autoregressive Estimators," *Signal processing*, pp. 2426-9, 2006.

[4] M. Pourahmadi, *Foundations of time series analysis and prediction theory*. New York: Wiley-Interscience., 2001.

[5] J. Rissanen, "Modeling by shortest data description.," *Automatica*, vol. 14, pp. 465–471, 1978.

[6] F. Popescu, "Universal codes: their construction, approximation and relevance in minimum description length (MDL) formulations.," Fraunhofer FIRST Technical Reports 5/2008, 2008.

[7] A. D. Lanterman, "Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model order estimation," *International Statistical Review*, vol. 69, pp. 185-212, 2001.

[8] A. Neumaier and T. Schneider, "Estimation of parameters and eigenmodes of multivariate autoregressive models," *ACM Trans Math Soft* vol. 27, pp. 27–57, 2001.

[9] M. Hansen and B. Yu, "Minimum description length model selection criteria for generalized linear models," *Science and Statistics: Festschrift for Terry Speed. IMS Lecture Notes -- Monograph Series*, vol. 40, 2002.

[10] K. Judd and A. Mees, "On selecting models for nonlinear time series," *Physica D*, vol. 82 pp. 426-444, 1995.

[11] G. Qian and C. Field, "Law of iterated logarithm and consistent model selection criterion in logistic regression.," *Stat Probability Lett,* vol. 56, 2002.