Vector Autoregressive Order Selection in Practice

Piet M. T. Broersen

***Abstract*—Vector time series analysis takes the same model order and model type for the different signals involved. Selection criteria have been developed to select the best order to simultaneously predict the different components of the vector. The prediction of single channels might require a different order or type for the best accuracy of each separate signal. That can become a problem in multichannel analysis if the individual signals have completely different model orders. Therefore, univariate and multichannel spectra are not similar. Furthermore, the selected order may vary in practice with the number of signals that are included in a vector. A turbulence example shows the results of order selection and the consequences in estimating the coherency between the same two components from vector signals with dimensions two and five.**

***Index Terms*—Autoregressive model, coherence estimation, magnitude-squared coherence (MSC), order selection, time series analysis.**

I. Introduction

MULTICHANNEL random data can be analyzed with periodograms [1], vector autoregressive (VAR) models [2], [3], and a more general class of time series models [4]. Classical or periodogram analysis suffers from the subjective choice of the bandwidth. Many parameter estimation methods for the general class of time series models are not completely reliable [4]. The subspace approach for general vector time series modeling can be attractive, but no practical experience is available with order selection [4]. Therefore, multivariate VAR models will mostly be used. Several methods were developed for VAR estimation [2]. In comparative investigations, a preference for the Nuttall–Strand method [3] was found because of the stationarity of the estimated models and the favorable numerical properties [5]. Recent results with the Nuttall–Strand method [5] support previous experiences [2].

The number of parameters strongly increases if more signals are simultaneously analyzed. Order selection is a special problem for multichannel problems. The best vector order generally cannot be derived with order selection applied to the univariate signals in the vector [2]. The theoretical analysis of VAR models is often given as an asymptotical approximation [6]. Then, all estimation methods and many order selection criteria seem to be equivalent, while practical experience shows important differences in finite samples. It is known from univariate signal processing that finite-sample effects become of influence if the number of estimated parameters is greater than one-tenth of the sample size [7]. Finite-sample effects were also reported for VAR models [8]. Moreover, a compromise between overfit and underfit was established, which leads to a combined information criterion (CIC) [9]. The principles to determine overfit and underfit have been first derived for univariate models [10]. Some ambiguities in determining the costs of underfit will be discussed in this paper.

Much research in multivariate dynamic models is published in econometric applications. VAR models are very common tools in empirical work in economy [11]. Multivariate analysis typically starts from a univariate analysis. This is followed by considering small subsystems [11]. This approach seems to be more successful than eliminating variables from a general large model with many variables included. A bottom-up approach to multivariate time series analysis starts from analyzing individual variables or small groups of variables. Traditionally, larger econometric models are constructed from verified smaller models [11].

Software is available for VAR modeling, which includes order selection [12]. Generally, the dimension of the vector is determined by the number of measured signals or by the purpose of the investigation. Only the maximum VAR order has to be chosen by the data analyst; the best model order is automatically selected. Moreover, information is available about the fit of all estimated model orders. This gives the possibility to enlarge the highest autoregressive (AR) candidate order if one of the highest of all candidate orders would be occasionally selected. The model fit, expressed in the prediction error or in CIC, should be regularly increasing above the best order, which gives extra information about the quality of the selected model order [8]. If the region of regular increase is not yet reached, it is advisable to try higher model orders as candidates.

This paper is an extended version of a contribution to the 2007 IEEE Instrumentation and Measurement Technology Conference [13], where an example with turbulence data obtained with direct numerical simulations was treated [14]. The Navier–Stokes equations were solved on a shared memory supercomputer SGI Origin 3800, which required a total of 15 500 CPU hours. A study of the flow around a cylinder gives information about heat and mass transfer to a clothed human limb in extreme outdoor conditions. One velocity component was previously analyzed with time series analysis [7]. This showed that the selected time series estimates had interesting properties. The results of a univariate analysis will be compared with a multichannel analysis with automatic order selection, with two or more signals simultaneously in the vector. The question is whether the different turbulent velocity components have coherence. Special attention will be given to the influence

Manuscript received July 2, 2007; revised September 8, 2008. First published April 24, 2009; current version published July 17, 2009. The Associate Editor coordinating the review process for this paper was Dr. Robert Gao.

The author is with the Department of Multi-Scale Physics, Delft University of Technology, 2628 BW Delft, The Netherlands (e-mail: P.M.T.Broersen@tudelft.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIM.2009.2015631

of the vector dimension and the sample size on the selected model order. Furthermore, an example with real-life data from the Pacific Fisheries Environmental Laboratory [15] about the amount of fish in the ocean will be examined.

II. Vector AR Models

A discrete-time vector time series x_n is a vector from a vector space X as a function of the integer n. The dimension of the space X is m. The individual components of the vector x_n are denoted $[x_n]_i$. A VAR(p) vector process is a stationary stochastic signal that is generated by the following difference equation [1], [2]:

$$x_n + A_1 x_{n-1} + \cdots + A_p x_{n-p} = \varepsilon_n. \tag{1}$$

The A_k are the $m \times m$ VAR parameter matrices, which are fully characterized by their matrix elements $[A_k]_{ij}$. The number p of VAR parameter matrices A_k is the AR order. The generating signal ε_n is a multivariate white-noise signal with the diagonal correlation matrix P_ε. The mean value of all signals is supposed to be zero; otherwise, the mean is subtracted. The variance σ_ε^2 of ε_n is defined as the trace of P_ε [9]

$$\sigma_\varepsilon^2 = \mathrm{tr}(P_\varepsilon). \tag{2}$$

Generally, only second-order moments are considered, involving one or two individual components. It is assumed that the vector process is weakly stationary. The power spectral density matrix of x_n has m^2 autospectra and cross spectra and is denoted $H(\omega)$ [2]

$$H(\omega) = \begin{pmatrix} h_{11}(\omega) & \cdots & h_{1m}(\omega) \\ \vdots & \ddots & \vdots \\ h_{m1}(\omega) & \cdots & h_{mm}(\omega) \end{pmatrix}. \tag{3}$$

The power spectral density matrix $H(\omega)$, as well as the autocorrelations and cross correlations, can be calculated from the VAR parameter matrices A_k [2]. The complex coherency of two components $[x_n]_i$ and $[x_n]_j$ is defined as

$$w_{ij}(\omega) = \frac{h_{ij}(\omega)}{\{h_{ii}(\omega) h_{jj}(\omega)\}^{1/2}}. \tag{4}$$

It has amplitude and phase, but the amplitude is generally given as the magnitude-squared coherence (MSC)

$$\mathrm{MSC} = |w_{ij}(\omega)|^2 = \frac{|h_{ij}(\omega)|^2}{h_{ii}(\omega) h_{jj}(\omega)}. \tag{5}$$

The cross-correlation function $r_{ij}(q)$ between the two vector components $[x_n]_i$ and $[x_n]_j$ is defined for the lag q as

$$r_{ij}(q) = E\left\{[x_{n+q}]_i [x_n]_j\right\}. \tag{6}$$

Together, these correlations determine the correlation matrix $R(q)$, given by

$$R(q) = \begin{pmatrix} r_{11}(q) & \cdots & r_{1m}(q) \\ \vdots & \ddots & \vdots \\ r_{m1}(q) & \cdots & r_{mm}(q) \end{pmatrix}. \tag{7}$$

True parameters yield the true spectra and correlations. Estimated spectra and correlation functions can be calculated with estimated parameter matrices [2].

An order selection criterion is required for the selection of the best VAR order from candidate orders between 0 and L, where the maximum L is defined by the data analyst. L should have no influence on the selected order if chosen high enough. An asymptotically based order selection criterion is based on the fit RES(k) of the estimated VAR(k) model to the data, together with a penalty factor α for each additionally estimated parameter. The residual fit is given by

$$\mathrm{RES}(k) = \det \hat{P}_k \tag{8}$$

where $\hat{P}_k$ is the correlation matrix for the lag $q = 0$ of the estimated residual vectors $\hat{\varepsilon}_n^k$ that have been used in the minimization procedure to calculate the estimates of the parameters for order k. These residuals for order k are given with the observed signal and the estimated parameter matrices $\hat{A}_i$ as

$$\hat{\varepsilon}_n^k = x_n + \hat{A}_1 x_{n-1} + \cdots + \hat{A}_k x_{n-k}. \tag{9}$$

The generalized information criterion GIC(k, α) is an asymptotical criterion for order selection defined as [9]

$$\mathrm{GIC}(k, \alpha) = N \ln \mathrm{RES}(k) + \alpha m^2 k \tag{10}$$

where α is a constant penalty factor for each estimated parameter. The value of α is determined as a compromise between the probabilities and the costs of overfit and underfit [9]. The cost of overfit is determined for a VAR(p) process by the accuracy of models of orders $p+1$ and higher with too many estimated parameters that have zero as their true values. The cost of underfit is given by the deterministic error of the VAR($p-1$) model of one order too low that belongs to a critical true parameter matrix A_p. The critical value of A_p would give [9]

$$\mathrm{E}\left\{\mathrm{GIC}(p, \alpha)\right\} = \mathrm{E}\left\{\mathrm{GIC}(p-1, \alpha)\right\}. \tag{11}$$

In m-dimensional VAR modeling, one-order underfit involves m^2 parameters. Defining the cost of underfit requires an assumption about a critical value for the sum of squares of all m^2 parameters at the same time. That critical sum for the true parameters is given by $(\alpha - 1)m^2/N$ [9]. This can be the square of a single parameter element if all other elements of the parameter matrix are assumed to be zero but can also be the sum of m^2 equal squared elements. Order selection criteria should look for a compromise between the selection of too low and too high orders. A good choice in univariate order selection is $\alpha = 3$ [7], and a good choice in multichannel selection is $\alpha = 2 + 1/m^2$ [9].

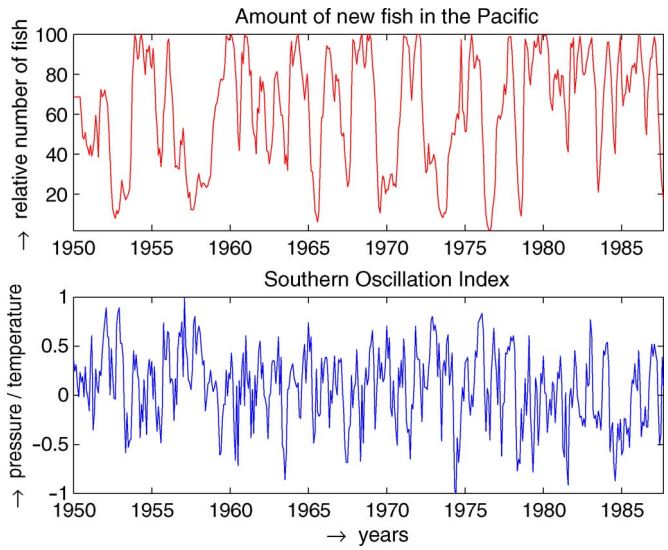


Fig. 1. Monthly observations of estimated new fish and SOI.

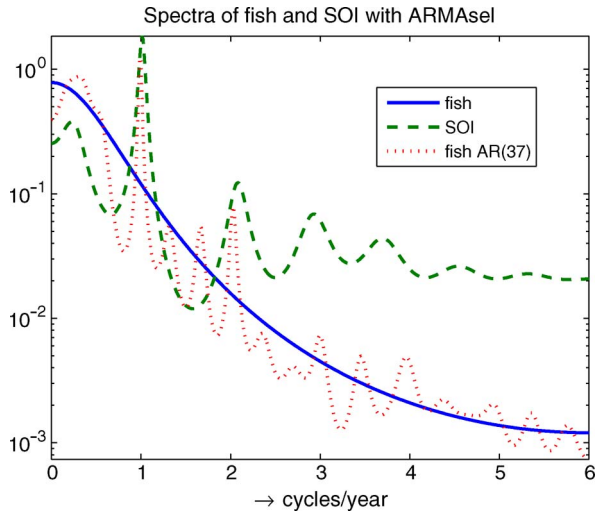


Fig. 2. Spectral densities of fish and SOI for the individually selected AR models and for the AR(37) fish model.

The finite-sample criterion CIC(k) for VAR processes is defined as [9]

$$\mathrm{CIC}(k) = N \ln \mathrm{RES}(k) + Nm \left(\prod_{i=1}^{k} \frac{1+mv_i}{1-mv_i} - 1 \right) + k \quad (12)$$

where N is the number of observations of the vector process, i.e., the length of x_n. The finite-sample variance coefficients v_i for the Nuttall–Strand estimates have to be substituted in (11). They make the penalty dependent on the model order i and are given by [8]

$$v_i = 1/(N - im + 1). \quad (13)$$

The order k with the smallest value for CIC(k) or GIC(k, α) is selected. The order selection criterion CIC(k) is adapted to finite-sample estimation and takes care of the compromise between overfit and underfit [9].

III. Order Selection and Vector Dimension

Order selection has been studied in much greater detail for univariate than for multichannel random data. In this section, a smooth transition from univariate data to more vector dimensions is made by constructing independent univariate AR data records with the same spectrum. A number of univariate signals can be combined into one vector process. For an example process, fish data that are known from the literature have been used [15], [16]. The same data were previously used [13] for an extensive univariate analysis. The data consist of 453 monthly values of the amount of new fish in the Pacific and of an environmental series called the southern oscillation index (SOI). The SOI is a measure for the variation in the air pressure that is related to sea surface temperatures in the central Pacific. The Pacific Fisheries Environmental Laboratory collected the data to study the relation between seasonality in fish and temperature [16]. Fig. 1 displays the data. It is immediately seen that the SOI has a rather strong yearly periodic component. The fish data are less pronounced. They are certainly not normally distributed, with a great number of observations close to the maximum value of 100. Sometimes, a periodicity of one year seems to be present, and at other times, this periodicity is certainly not observable.

Both signals have been individually analyzed with the ARMAsel program for automatic time series analysis [7]. For the fish data, the AR(2) model was selected with the finite-sample criterion, and for SOI, the AR(15) model was used. If the asymptotical Akaike information criterion (AIC) [1] has been used, the fish data would give order AR(37). The high-order AR spectrum gives more details for the fish data. This strong influence of the order selection criterion for the fish data is an indication that a periodicity is on the boundary of statistical significance for the given amount of data. If less data were available, AR(2) would be selected, and AR(37) would almost always be selected for larger data sets. The observed sample size is in a transient area. Fig. 2 gives the estimated spectra. Obviously, the spectral peak at 1 cycle/year is not present in the AR(2) spectrum of the fish data but is strong in the other spectra. Those two spectra also share peaks at 2 and 3 cycles/year. However, the other local peaks in the spectra seem not to be related. It turned out that all estimated AR models of the fish data with orders between 2 and 37 had about the same spectral accuracy [13]. The magnitude of the estimated parameters is on the boundary of statistical significance for about 500 observations.

New univariate AR(37) data with a normal distribution can be generated with the parameters of the AR(37) fish data model. They were combined into a vector process for a simulation study [13]. They have the special characteristic that the AR prediction error is about the same for the model orders between 2 and 37, for N is about 500. Those newly generated simulation data are known to be AR(37). Moreover, for N is about 500, many models are on the boundary of statistical significance. Univariate order selection from less than 500 data will yield an order between 2 and 37. Smaller sample sizes generally give lower orders in order selection. If more than 2000 new

AR(37) observations are generated, order 37 will be selected in almost all realizations. Therefore, lower orders will never be selected. Only occasionally will overfit models with a few extra parameters be selected.

For a test of vector order selection, ten independent AR(37) signals have been generated with that given spectrum, each of which is N observations long. Those signals were combined into vector processes of dimensions m from one to ten. All true parameter matrices A_k have the AR(37) model on the diagonal elements, whereas all off-diagonal elements of A_k are zero. It is clear that the true spectral matrix (3) will have a diagonal with the given AR(37) spectrum as its expectation and with zeros for all off-diagonal cross spectra. Moreover, as each of the signals in the vector has the true order 37, the selected order for a vector of those signals should also be 37 if the number of observations N is greater than about 2000.

For different values of N and for all dimensions between two and ten, the VAR model order has been selected with GIC(k, α) of (10) and CIC(k) of (12). Values between 1.25 and 3 were used for α. For N greater than 5000 and for $1.25 < \alpha < 3$, the selected VAR order was always 37 for all values of the penalty α and for all dimensions between two and ten. The AR parameters are very significant, and neither overfit nor underfit was found in any simulation run. Order selection was not a problem at all for those large sample sizes.

In univariate order selection, the given values of α would often select an overfit order higher than 37 for $N = 5000$. In particular, $\alpha = 1.25$ and $\alpha = 1.5$ would give a high probability of overfit [10]. This demonstrates that the probability of *overfit* is smaller for multichannel vector data than for univariate data.

In simulations with different values of N (but less than 1000) and for all vector dimensions, the performance of GIC(k, α) for various values of α has been characterized as follows [13]. Low values of α give overfit, and mostly, the highest of all candidate orders is selected. High values of α give underfit, and mostly, AR(2) is selected. Generally, the order 37 is selected for only a specific range of values of α. Unfortunately, that range of α depends on the sample size N, the dimension m, and the significance of the generating AR parameters. No single value of α in GIC(k, α) gives better results than CIC(k). In other words, for various N, m, and generating processes, CIC(k) outperforms GIC(k, α) for every fixed value of α. Only for very large sample sizes may the performance be the same. This was also the result of a finite-sample analysis [8].

In univariate signal processing, the probability of one-order overfit is determined by a χ^2 distribution with 1 DOF [7], [10]. The costs of one-order overfit are mainly given by the probability that the 1-D χ^2 distribution exceeds the penalty factor α. The choice of the best value of α is based on an evaluation of the cost of overfit in comparison with the cost of one-order underfit if the last parameter is on the boundary of statistical significance [10].

In multichannel data, overfit is found if the selected order is greater than the true VAR order. In this example, where all univariate orders are the same and there is no coherence between the individual signals in the vector, the VAR order is the same as the univariate order. If the true VAR order is p, all m^2 elements of the parameter matrix A_{p+1} would have the same expectation zero. The probability of one-order overfit in those m-dimensional vector signals with the selection criterion GIC(k, α) of (10) is determined by the probability that a χ^2 distribution with m^2 DOF will exceed αm^2. This is much smaller than the probability that $\chi^2 > \alpha$ that applies for the univariate case $m = 1$. A refined analysis of multichannel overfit has been given in [9]. For greater values of m, the cost of overfit strongly decreases because the probability of overfit becomes smaller.

Now, suppose that only one single vector component has the true order p and it has a nonzero parameter $[A_p]_{ii}$ at that order. All other true parameters of A_p are zero. Vector order p would be selected with GIC(p, α) if

$$N \ln \mathrm{RES}(p-1) - N \ln \mathrm{RES}(p) > \alpha m^2. \tag{14}$$

Therefore, this single univariate signal requires a multichannel residual reduction that is m^2 times greater than what a univariate signal would need to select the order p in this example. That would not be a problem in (14) if N is high enough, because underfit then becomes unlikely. However, the selection can prefer underfitting orders for much higher values of N in vector order selection than in univariate order selection. Moreover, underfitting becomes more likely if the vector dimension m increases. The properties of underfit in univariate signals and in vector signals are only comparable if all m^2 elements of A_p simultaneously have the same critical value.

The number of parameters equals $m^2 L$ for the vector with dimension m if the maximum candidate order is L. No numerical problems would have been encountered if the number of estimated parameters is about the same as the total number of observations. To avoid problems with order selection, however, it is advisable to limit the highest candidate order to a VAR order L that requires a number of estimated parameters that is less than half the total number of mN observations [13], i.e., $L < N/(2m)$. Furthermore, the asymptotical performance of order selection criteria is reached for a smaller value of N if the dimension m is low. The boundary value of statistical significance for individual parameters increases with the vector dimension in this experiment, where the vector is built with independent AR(37) signals.

IV. Turbulence Data

A. Univariate Signal Processing

Solutions of the Navier–Stokes equations of a turbulent subcritical flow around a circular cylinder surrounded by a porous layer were the subject of a numerical study [14]. The porous layer at a small distance models the clothing, and the cylinder models the arm of a fireman. The flow in the space between the porous layer and the cylinder determines the heat transfer. The pressure and the velocity V in three perpendicular directions x, y, and z around a porous cylinder can be computed in a direct numerical simulation study [14]. Fig. 3 gives five signals: the pressure, the three orthogonal velocity components, and an extra velocity at a different location, denoted V_{x2}. The time has been normalized with T_{St}, the dimensionless

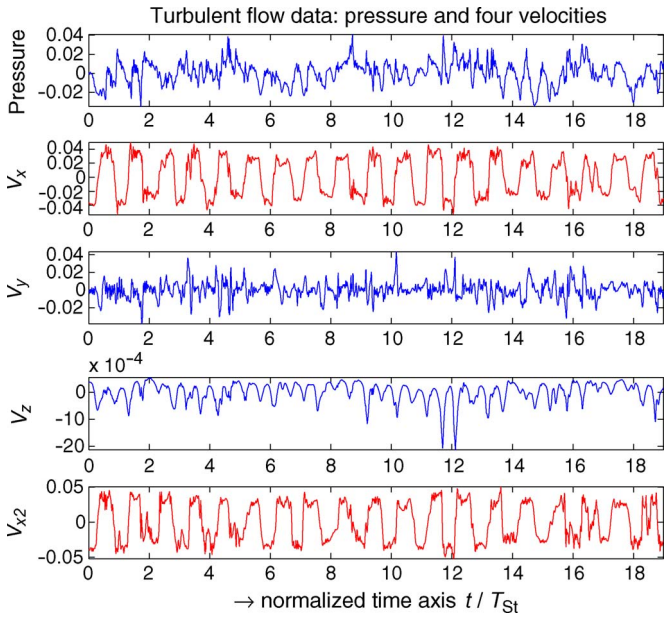


Fig. 3. Five turbulence signals obtained with direct numerical simulations.

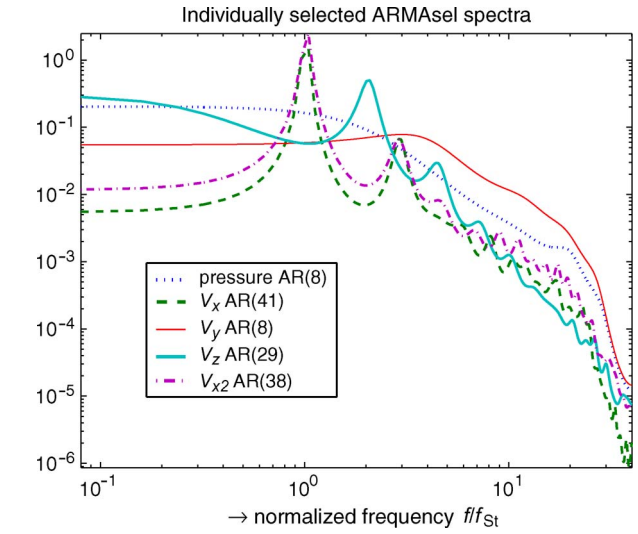


Fig. 4. Normalized univariate power spectral densities of five turbulence signals, each having 1524 observations. Orders 8, 41, 8, 29, and 38 have been selected, respectively, for the univariate signals.

Strouhal time, which, in turn, is given by $1/f_{\mathrm{St}}$. The dimensionless Strouhal number f_{St} determines the vortex shedding frequency [14].

Fig. 4 gives the five univariate spectra, estimated from 1524 observations. The AR orders have been selected by the automatic ARMAsel program [7] for the univariate signals. The two velocities V_x and V_{x2} have the highest selected orders, and their spectra show similar peaks at the normalized frequency of one. V_z also has peaks but at different frequencies, with a strong peak at a frequency of two; the other two signals have no clear periodicity. If the AR order is taken as eight or ten for all five signals, the estimated spectra would all have the character and shape of the pressure spectrum or of the V_y spectrum in Fig. 4, i.e., without peaks.

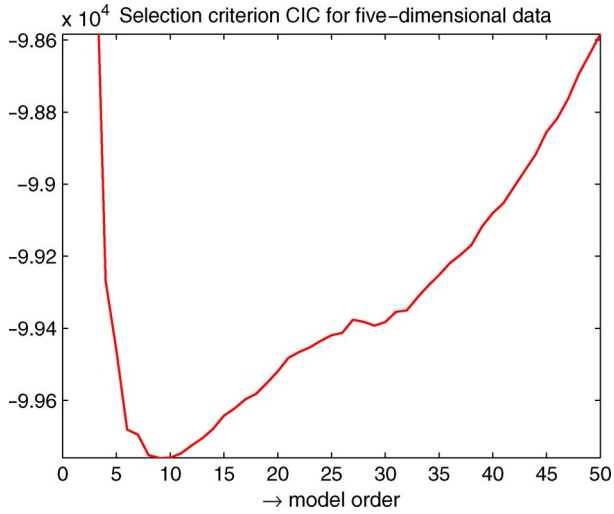


Fig. 5. Values of the selection criterion CIC for all candidate orders of 5-D turbulence signals, as a function of the model order. Order 9 is selected.

B. Multichannel Processing

Additional information in multichannel processing is provided by the cross spectra, cross correlations, and the MSC of (5). The literature mostly deals with second-order moments for spectra [2], but it gives definitions like (3) and (7) for VAR processes with more dimensions. However, the higher order moments for dimensions greater than two are generally not taken into account. For normally distributed processes, those higher order moments do not give additional information [1]. This paper studies the influence of the number of channels on order selection and on the estimated coherency. Therefore, all five turbulence signals will be simultaneously treated as a vector; furthermore, another vector process with only the two V_x signals will be studied.

VAR order 9 was selected with CIC(k) for the vector with five signals. The vector accuracies for other orders are shown in Fig. 5. The spectra of the individual vector components have no peaks for VAR order 9. It turns out that all individual autospectra $h_{ii}(\omega)$ computed with (3) have a character like V_y in Fig. 4, i.e., without peaks. The order has also been selected with the asymptotical criterion GIC(k, α) of (10). This is the famous criterion AIC [2] for $\alpha = 2$. AIC selected order 10 for the given data. By taking $\alpha = 1.75$, the criterion selected order 32 for the data, with spectral peaks for V_x and V_z and without peaks for the pressure signal and V_y. For $\alpha = 1.5$, order 41 was selected. However, this is merely an illustration of the fact that in almost all circumstances, a value for α can be found for which the selected order is higher or lower than that with CIC(k). It is not advisable to use those values of α smaller than two that are prone to select overfitting models in many circumstances.

Fig. 6 shows the MSC of all vector components with V_x. It is very small for the combination of V_x with all other signals, except with V_{x2}, which is the other velocity in the x-direction. Taking higher model orders for the vector signal gives many more small details in the estimated coherencies. However, it has no significant influence on the level of the first three coherency estimates in Fig. 6; those coherencies are always small.

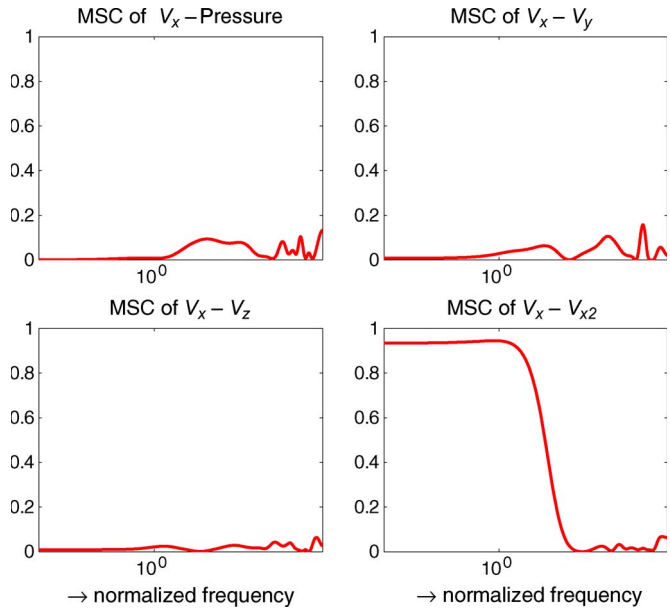


Fig. 6. MSC of the second signal in Fig. 3 with the four other signals, computed as the result of one single evaluation with dimension five. The selected order with CIC(k) was nine.

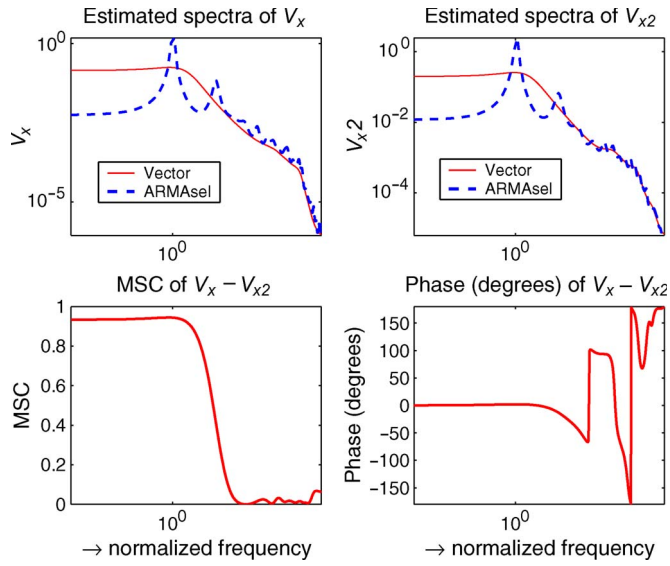


Fig. 7. Selected univariate and multichannel spectra of the second and fifth signal in Fig. 3. The dimension of the VAR vector was five. The order selected with CIC(k) was nine. The lower plots give the MSC and its phase.

The upper plots in Fig. 7 give the estimated spectra of the two velocities V_x and V_{x2} for two situations. The first is estimated from the VAR model with (3) for the selected order 9. The second is estimated from the univariate model with the ARMAsel algorithm, as already presented in Fig. 4. The vector estimates look very much like univariate estimates that could be computed for AR order 10. Therefore, no peaks are present, and the univariate order selection preferred much higher AR orders with peaks. The MSC and the phase of the coherency in the lower part of Fig. 7 show a high coherency for the whole frequency range below the normalized frequency of two. It is obvious that the VAR order, selected for the 5-D vector, is too low because the estimated vector spectra miss many spectral details.

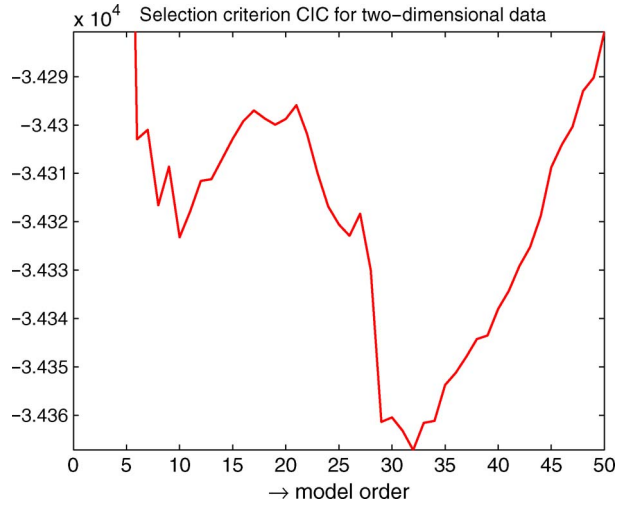


Fig. 8. Values of the selection criterion CIC for all candidate orders of 2-D turbulence signals, as a function of the model order; only the signals V_x and V_{x2} have been included in the vector. Order 32 is selected.

C. 2-D and 5-D Vector Processing

To investigate the influence of the vector dimension, the two interesting signals V_x and V_{x2} with a significant coherency are taken together as one 2-D vector. Fig. 8 gives the CIC as a function of the VAR order. Order 32 has now been selected. It is clear that a local minimum is present around order 9 or 10, which is the order selected for the 5-D vector. Likewise, Fig. 5 shows a slight ripple at orders around 30. It has been verified that the selected orders are, in both cases, independent of the highest candidate order, as long as that was higher than 9 or 32, respectively. Even with 200 as the highest candidate order, the algorithm selected the same orders and did not show any computational problem.

CIC(k) selected order 9 for the signals in the complete 5-D vector and order 32 for the 2-D problem. The vector and the univariate ARMAsel spectral estimates are different in the upper plots in Fig. 7 and are close in Fig. 9. The MSC for $V_x - V_{x2}$ is high and almost a constant in Fig. 7 for the normalized frequency range below two. However, the 2-D evaluation in Fig. 9 shows maxima at the spectral peaks, whereas the 5-D MSC estimate was flat. A comparison of Figs. 7 and 9 demonstrates the problems of automatic order selection in multichannel problems. This behavior could be expected after the simulations with different dimensions in Section III. For higher vector dimensions, the selected orders can become lower if the true process parameters are slightly greater than the boundary of univariate statistical significance.

Finally, the multichannel coherencies for the 5-D vector for orders 9 and 32 have been compared to the same coherencies for the 2-D vector. Fig. 10 gives the computed coherencies. Both coherencies are quite close for orders 9 and 32, whether they are computed from 2-D or 5-D vectors. The conclusion may be that the poor estimate of the coherency details for the 5-D signal vectors is not a computational problem for high dimensions but an order selection problem.

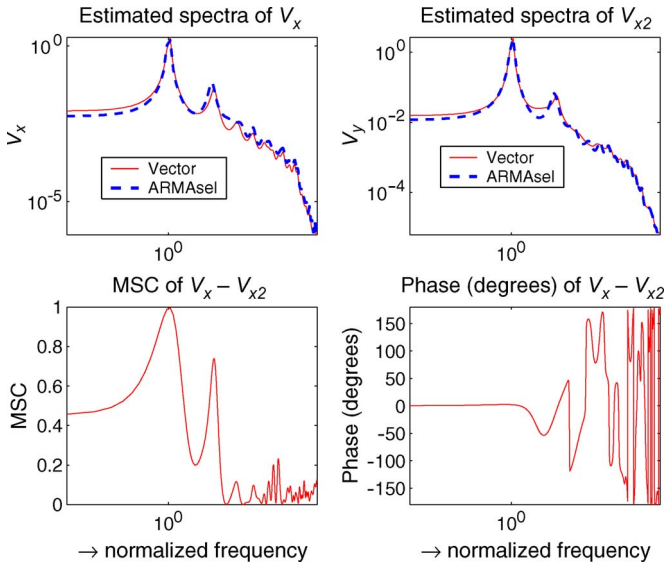


Fig. 9. Selected univariate and multichannel spectra and the MSC of the second and fifth signal in Fig. 3. The order selected with CIC(k) was 32 with only those two signals.

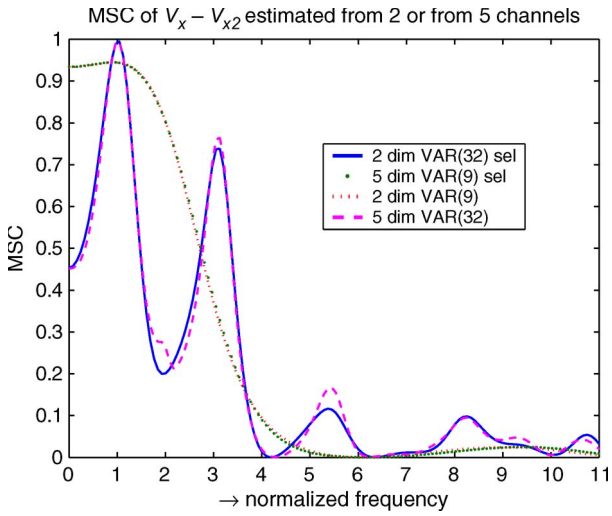


Fig. 10. MSC of the second and fifth signal in Fig. 3. The dimension of the VAR vector was two with only those two signals or five with all signals included. The order selected with CIC(k) was nine for dimension five and 32 for dimension two. The MSC for dimension five and order 32 has been computed with a fixed-order estimated VAR(32) model, without order selection.

A close look at the VAR(32) coherencies around the normalized frequency of two reveals a difference in Fig. 10. The univariate spectra in the x-direction have a local minimum at the frequency of two in Fig. 4. However, the spectrum in the z-direction has a strong maximum there, in the same figure. The investigation of the higher fixed VAR order 32 for the coherency with two and five dimensions gives a strong indication that the wiggle at the normalized frequency of two is caused by feed-across from V_z at that frequency for the given data. The wiggle becomes a local maximum in the coherency for VAR models of orders higher than about 35. It has been shown before that a sharp peak in one component of a vector signal can cause a spurious peak in the autospectrum of the other vector components [2].

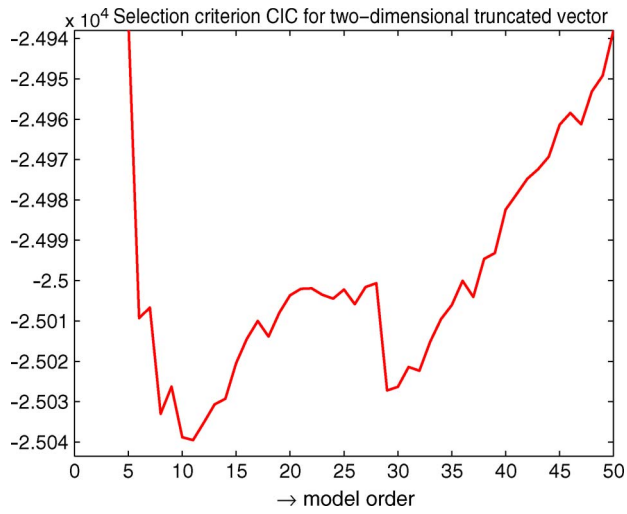


Fig. 11. Values of the selection criterion CIC for all candidate orders of the truncated 2-D turbulence signals V_x and V_{x2}, as a function of the model order. Only 1100 observations have been used. Order 11 is selected.

V. 2-D Processing of Truncated Turbulence Data

Univariate order selection is generally not applicable to vector processes. This can be demonstrated with a simple VAR(1) process [2] that has the following parameter matrix:

$$A_1 = \begin{pmatrix} -0.85 & 0.75 \\ -0.65 & -0.55 \end{pmatrix}. \tag{15}$$

This VAR(1) process of dimension two generates two univariate AR moving average ARMA(2, 1) processes [2], with the same pair of complex conjugated poles at 0.122 Hz. The order selected for the vector process is almost always the true VAR(1) model for all sample sizes greater than $N = 50$. The best univariate models are two ARMA(2, 1) processes. That model is mostly selected with the order and type selection of the automatic ARMAsel algorithm [7] for time series analysis. An ARMA(2, 1) process is theoretically equivalent with an AR(∞) model with many small parameters. In practice, the best univariate orders for estimated AR models of that VAR(1) process will increase with the number of observations. If the univariate candidates for order selection are restricted to AR models, the selected order depends on the sample size. The selected AR orders varied from two to nine for N from 50 to 100 000, with increasing orders for a larger data size. This demonstrates that univariate selected AR orders can be rather different from VAR orders. Generally, the best VAR order cannot be derived with univariate order selection.

In the previous section, the VAR order selected with CIC was nine for 5-D data and 32 for two dimensions, each with 1524 observations. The selected order depends not only on the dimension but also on the sample size. Now, suppose that only 1100 observations are available for the signals in Fig. 3, just because the experiment was stopped earlier. It may be expected that the true process characteristics are the same for 1100 and 1524 observations. Fig. 11 gives the CIC for those truncated data, where VAR order 11 is selected. A comparison of Figs. 8 and 11 shows that the global and local CIC minima have been interchanged. It follows from (10) that the residual reduction in

RES(k) should become greater to select the same order from a smaller sample size.

The univariate AR orders selected for 1100 observations are 39 and 37 for V_x and V_{x2}, respectively. The spectra strongly resemble those in Fig. 4 for $N = 1524$. However, the spectra computed from the selected VAR(11) model look like the vector spectra in Fig. 7. The local minimum at order 29 in Fig. 11 and the strong difference between the character of selected univariate and vector spectra are a good argument to try a model order of about 30 to estimate the coherency of the truncated signals of length 1100. If the spectral character of the same signal with a selected univariate model and a vector model is similar, the vector order selection for dimension two can be trusted. If the character is completely different, a vector order can be found for which that difference is made smaller or almost eliminated.

It has been shown in Section III that higher vector dimensions can give problems in order selection that do not appear for smaller dimensions with the same sample size. Section V showed that this is the case for 1524 turbulence data. There, the selection problem could be solved by analyzing a vector of dimension two. It is recommended to restrict vector order selection to a maximum dimension of two. However, this section demonstrates that even for two dimensions in the vector, order selection may have undesirable results. Undesirable means that the spectral characters computed from a component of a vector signal and its univariate equivalent are very different. Asymptotically, for large sample sizes, this is not possible, and for finite samples, it is undesirable. The solution is to start with the univariate analysis. Afterward, the VAR model order is selected. If the univariate spectra and the vector autospectra are similar, there is no reason to deviate from the selected model orders. If the characters of the spectra are not similar, look at what happens at higher VAR orders. It might be interesting to compute the coherency for a VAR order for which the character of the components of the vector signal resembles the univariate spectra.

The final question is whether this practical solution can lead to erroneous conclusions. That can be studied by truncating the turbulence data to still smaller sample sizes. The statistical significance of spectral details diminishes for smaller sample sizes. Univariate order selection has been severely tested to demonstrate that the selected model includes all details that are statistically significant and no more. Therefore, details that are selected in a univariate analysis should also be present in a vector analysis with the same signal. Only due to feed-across from other signals in the vector can components have more details than their selected univariate AR models.

All details of the spectrum of V_x lose their significance in univariate data if not enough samples are available. This can be simulated with the practical turbulence data in Section IV by using only truncated signals, pretending that they are all the data available. Taking $N = 100$ would select five, five, and three for the 2-D VAR order and the univariate AR orders, respectively. For $N = 200$, the selected orders are six, seven, and four. For $N = 500$, the selected orders are six, six, and eight. The estimated MSC for all those sample sizes looks like the VAR(9) results in Fig. 10. For N greater than 700, the univariate orders become 39 and 37, and the results become similar to what is found for $N = 1100$, with a selected VAR order of about 10. For $N > 1150$, the selected VAR order suddenly jumps to around 30, because the second minimum in CIC becomes the global minimum, like in Fig. 8. The truncated data example shows that details become statistically significant earlier in univariate data, before VAR order selection includes them.

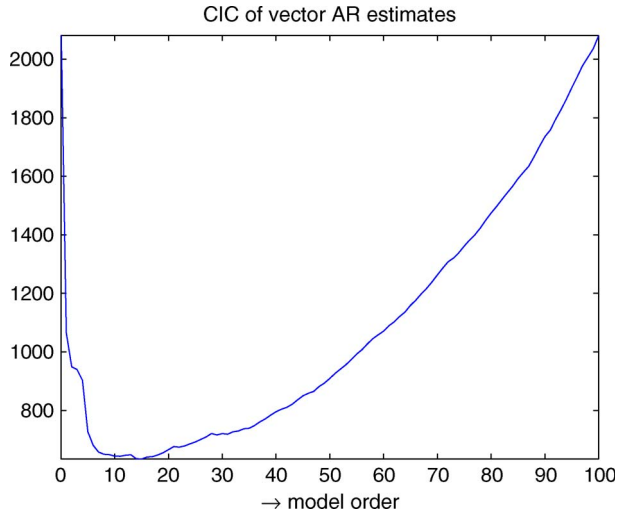


Fig. 12. Values of the selection criterion CIC(k) for all candidate orders k of the fish and SOI data in Fig. 1. Order 15 is selected.

VI. Vector Models for Fish and SOI Data

The univariate spectra of fish and SOI data are shown in Fig. 2. The selected AR order for the fish was two, but the prediction accuracy of all models of orders 2 to 37 was quite close. The parameters of those model orders were on the boundary of statistical significance. The selected order for the SOI data was 15. The periodicity with a strong spectral peak is visible in the selected SOI spectrum but is not significant for the fish data. It is interesting to see what will happen if those two signals are combined into a vector signal.

Fig. 12 shows that the multichannel order selection criterion CIC(k) has its minimum for order 15. The algorithm in [12] could estimate 400 parameter values (AR order 100 with four parameters in each parameter matrix A_k) from $N = 453$ observations without numerical problems. The upward-curved behavior of CIC above order 15 is due to the finite-sample character. Higher order parameters are estimated from fewer residuals and are less accurate than the asymptotical accuracy measures. Therefore, the inaccuracy grows faster at higher orders.

Fig. 13 shows the results of the multichannel spectral analysis for the selected vector order 15. Both the multichannel vector spectra (3) and the automatically selected univariate spectra in Fig. 2 are given in the upper figures. The difference for the spectra of the SOI data is negligible; the difference for multichannel and univariate order selection for fish is very remarkable. Vector order selection gives the yearly peak in the fish data, which is also the peak with the maximum coherency. Furthermore, local maxima are found in the coherency.

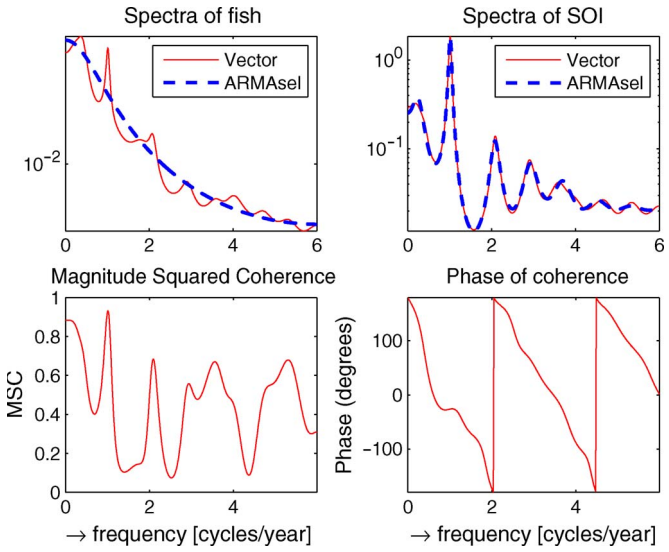


Fig. 13. Estimated spectra of fish and SOI and the MSC with the phase of the coherency. Multichannel order selection selected order 15 for those data.

The peaks for 2 and 3 cycles/year are expected, in contrast with those for 3.5 and 5.4. The coherency is high for the whole lower frequency range, which is a strong indication that El Niño affects the fish population. The selected order 15 gives a very regular and reliable appearance of both the magnitude and the phase of the coherency. Taking higher VAR orders would give an irregular shape for the magnitude and the phase of the coherency.

VII. Conclusion

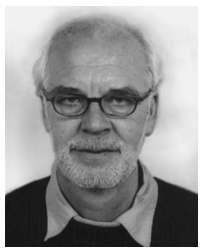
Order selection can have different results for univariate and vector signal processing. The study of univariate spectra and autocorrelations avoids the feed-across mixing of sharp peaks between signals, which may appear in vector processing. Therefore, it is advisable to always estimate the univariate spectra for their selected univariate orders. However, only vector processing can compute the coherency between signals, and the estimation of the coherency requires the combined VAR estimate. The order of the VAR model can be selected with the finite-sample criterion CIC.

Simultaneously taking more than two signals in one vector hardly has any advantage for normally distributed signals, but it can have undesirable effects in order selection. The results of order selection would only be independent of the dimension in an asymptotical approximation. In practice, larger data sets improve the performance of order selection for higher vector dimensions. However, the feed-across mixing from more sources is still possible. These effects disappear or diminish if the dimension of the vector signal is taken as two. The coherency between two signals is best estimated by limiting the vector dimension to two signals.

If the univariate spectra and the vector autospectra of the vector components are similar, there is no reason to deviate from the selected VAR model order. Order selection will give reliable results for the coherency. If the characters of these spectra are not similar, it is advisable to evaluate some models with a higher VAR order and to analyze the resulting spectra and coherencies. The probability of overfit is small for vector data, and it is not useful to evaluate lower order VAR models.

References

[1] M. B. Priestley, *Spectral Analysis and Time Series*. London, U.K.: Academic, 1981.
[2] S. L. Marple, *Digital Spectral Analysis With Applications*. Englewood Cliffs, NJ: Prentice–Hall, 1987.
[3] O. N. Strand, "Multichannel complex maximum entropy (autoregressive) spectral analysis," *IEEE Trans. Autom. Control*, vol. AC-22, no. 4, pp. 634–640, Aug. 1977.
[4] J. Mari, P. Stoica, and T. McKelvey, "Vector ARMA estimation, a reliable subspace approach," *IEEE Trans. Signal Process.*, vol. 48, no. 7, pp. 2092–2104, Jul. 2000.
[5] A. Schlögl, "A comparison of multivariate autoregressive estimators," *Signal Process.*, vol. 86, no. 9, pp. 2426–2429, Sep. 2006.
[6] H. Lütkepohl, *Introduction to Multiple Time Series Analysis*. Berlin, Germany: Springer-Verlag, 1991.
[7] P. M. T. Broersen, *Automatic Autocorrelation and Spectral Analysis*. London, U.K.: Springer-Verlag, 2006.
[8] S. de Waele and P. M. T. Broersen, "Finite sample effects in vector autoregressive modeling," *IEEE Trans. Instrum. Meas.*, vol. 51, no. 5, pp. 917–922, Oct. 2002.
[9] S. de Waele and P. M. T. Broersen, "Order selection for vector autoregressive models," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 427–433, Feb. 2003.
[10] P. M. T. Broersen and H. E. Wensink, "On the penalty factor for autoregressive order selection in finite samples," *IEEE Trans. Signal Process.*, vol. 44, no. 3, pp. 748–752, Mar. 1996.
[11] H. Lütkepohl, "General-to-specific or specific-to-general modelling? An opinion on current econometric terminology," *J. Econom.*, vol. 136, no. 1, pp. 319–324, Jan. 2007.
[12] S. de Waele, *Automatic Spectral Analysis Matlab Toolbox*, 2003. [Online]. Available: www.mathworks.com/matlabcentral/fileexchange
[13] P. M. T. Broersen, "Multichannel autoregressive order selection in practice," in *Proc. IEEE/IMTC Conf.*, Warsaw, Poland, 2007, pp. 134–139.
[14] M. P. Sobera, C. R. Kleijn, and H. E. A. van den Akker, "Subcritical flow past a circular cylinder surrounded by a porous layer," *Phys. Fluids*, vol. 18, no. 3, pp. 038 106-1–038 106-4, Mar. 2006.
[15] S. Bograd, F. Schwing, R. Mendelssohn, and P. Green-Jessen, "On the changing seasonality over the North Pacific," *Geophys. Res. Lett.*, vol. 29, no. 9, pp. 47.1–47.4, 2002.
[16] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications*. New York: Springer-Verlag, 2000.



Piet M. T. Broersen was born in Zijdewind, The Netherlands, in 1944. He received the M.Sc. degree in applied physics and the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 1968 and 1976, respectively.

He is currently with the Department of Multi-Scale Physics, Delft University of Technology. He developed statistical measures to let measured data speak for themselves in time series models, as a practical solution for the spectral and autocorrelation analysis of stochastic data. His main research interest is in the automatic and unambiguous identification of the character of stationary random data.