Proceedings of the 42nd IEEE
Conference on Decision and Control
Maui, Hawaii USA, December 2003

Estimation of confidence regions for the parameters of ARMA models - guaranteed non-asymptotic results

M.C. Campi
Department of Electrical Engineering and Automation,
University of Brescia, Via Branze 38, 25123 Brescia, Italy.
Email: campi@ing.unibs.it

Erik Weyer
CSSIP, Department of Electrical and Electronic Engineering,
The University of Melbourne, Parkville, VIC 3010, Australia.
Email: e.weyer@ee.mu.oz.au

***Abstract*— In this paper we consider the problem of estimating confidence regions for the parameters of ARMA models. Based on subsampling techniques and building on earlier exact finite sample results due to Hartigan, we compute the exact probability that the true parameters belong to certain regions in the parameter space. By intersecting these regions, a confidence region containing the true parameters with guaranteed probability is then obtained. All results hold true for a finite number of data points and no asymptotic theory is used. The usefulness of the approach is illustrated in a simulation example.**

I. Introduction

It is widely recognised that a model is of limited use if no certification of its quality is delivered together with the model itself. In principle, a model can be used as if it were the true system provided that it is so accurate that the system-model discrepancy is negligible. However, this is seldom the case, and the model accuracy should also be taken into account when the model is used in practice. For instance, in prediction, the prediction error is formed by two components: the stochastic fluctuation due to noise and the systematic error due to model inaccuracy.

To be credited with usefulness, a model uncertainty evaluation technique should meet two requirements:

i) It must hold under general conditions;
ii) It should provide tight evaluations of the system uncertainty.

In connection with **i)**, we note that in real situations the presence of restrictive assumptions is awkward. For example, assuming a specific distribution of the noise (e.g. that it is bounded or that it is Gaussian) generates problems at two different levels: first, the theory looses in applicability; second, verifying the assumption may be difficult in a given application. Point **ii)** is important because loose uncertainty evaluations generate conservativeness in the belief that the model is less reliable than it actually is. For example, a robust controller looses in performance as the level of uncertainty increases.

Despite the recognised need for model uncertainty evaluation methods, there is a basic lack of methodologies able to provide *guaranteed* results. This is mainly due to the inherent theoretical difficulties encountered in the development of such methodologies.

One point that needs to be kept in mind is that, in system identification (e.g. Ljung (1999)), one always uses a *finite* number of data points. And, in fact, uncertainty in the model is due to such a finiteness. Likewise, for the evaluation of model quality one will only have a finite amount of data available. Thus, a sound uncertainty evaluation method must provide results valid when the number of data is finite, and, possibly, small.

Quite often, uncertainty evaluations are based on the asymptotic theory of system identification. It is common experience of theorists and practitioners that this theory - though applied heuristically with a finite number of data points - in many situations delivers sensible results. On the other hand, the correctness of the results is not guaranteed, and contributions (Bittanti et al (2002)) have appeared that show that the asymptotic theory may as well fail to be reliable in certain situations. Moreover, when the available data is scarce, using asymptotic results makes no sense. Thus, there is a need for developing techniques that provide results guaranteed for finite data samples.

In this paper we study ARMA models and develop a methodology for the evaluation of their accuracy which is rigorously valid for any size of the data sample. The theory in this paper calls for the assumptions that the model class is rich enough to contain the true system. However, this assumption must be put under the correct light: it is important to note that this assumption regards the model class used for model quality assessment, not the model class used for the actual identification. In fact, the method developed in this paper does not deliver a nominal model; instead, it allows us to determine an uncertainty region in the parameter space. Thus, one can use one model class (possibly of restricted complexity) for identification, and then assess the reliability of the obtained model by considering the full-order model class. As these models are of different orders the reliability assessment can be suitably performed in the frequency domain.

The mathematical approach of this paper is inspired by the work of Hartigan (Hartigan (1969,1970)) in the statistical literature. In Hartigan (1969), Hartigan considered the problem of estimating a constant from noisy measurements and introduced the idea that sample estimates based on a certain group theoretical property exhibit special distribution

0-7803-7924-1/03/$17.00 ©2003 IEEE

characteristics, valid for a finite number of measurements. Though this idea has generated moderate resonance in the statistical literature and has not been explored at all by the identification community, it contains the seed for important achievements in finite sample-based system identification. The present paper departs from the original work of Hartigan in that we consider more general random sequences (and this allows us to deal with *dynamical* systems). Yet, the main underlying idea is still within Hartigan's framework. Thus, this paper can also be seen as a contribution in the direction of fertilizing the area of system identification with ideas imported from a certain area of the statistical literature.

Our earlier finite sample results (e.g. Weyer and Campi (2002)) were data independent, in the sense that they were uniform with respect to the considered class of data generating systems, and they could essentially be evaluated without any data. Because of the uniformity, it was realised that the results could be quite conservative for the particular system at hand. The approach presented here is data based and uses data generated by the actual system at hand, and hence avoid the problems due to uniformity. Finite sample results using a data based approach has also been developed in Campi et al (2002), and of course many popular techniques such as bootstrap are data based. However, few rigorous finite sample results exists for bootstrap methods.

The paper is organised as follows. In the next section we give a simple preview example illustrating the main idea in the approach. In section III we consider ARMA models and give the algorithm for construction of the confidence region and the theoretical results giving the probability that the true parameters belong to the constructed region. A simulation example demonstrating the usefulness of the method is given in section III-D before conclusions are given.

II. A PREVIEW EXAMPLE

In this section, a preview example is given that illustrates the type of results developed. Consider the system

$$y_t + a_0 y_{t-1} = w_t, \tag{1}$$

where $a_0 = 0.2$ and $\{w_t\}$ is an independent sequence of uniformly distributed random variables between -1 and 1. 1026 data points were generated according to (1). Our goal is to form a confidence regions for a_0 from the available data set.

Rewrite the system as a model with generic parameter a:

$$y_t + a y_{t-1} = w_t.$$

The predictor and prediction error associated with the model are

$$\hat{y}_t(a) = -a y_{t-1}, \quad \epsilon_t(a) = y_t - \hat{y}_t(a) = y_t + a y_{t-1}.$$

Next we compute the prediction errors $\epsilon_t(a)$ for $t = 1, \ldots, 1025$ and calculate

$$f_{t-1}(a) = \epsilon_{t-1}(a)\epsilon_t(a), \quad t = 2, \ldots, 1025.$$

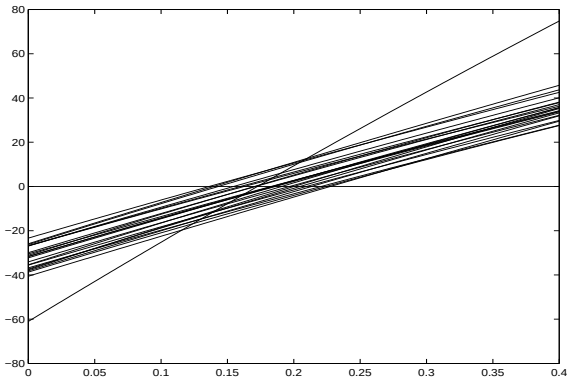

Fig. 1.

Using the $f_{t-1}(a)$'s, we want to form empirical estimates of the correlation $E[\epsilon_{t-1}(a)\epsilon_t(a)]$. Such estimates however, need be constructed very carefully. First, we generate a set G of subsets of $I = \{1, \ldots, 1024\}$ which is a group with respect to the symmetric difference, i.e. $(I_i \cup I_j) - (I_i \cap I_j) \in G$, if $I_i, I_j \in G$. The generated group has 2048 elements and apart from the set I itself and the empty set, each set in G has 512 elements. The sets in G are denoted $I_1, \ldots, I_{2048}$. The incident matrix for a group is a matrix whose (i, j) element is 1 if $j \in I_i$ and zero otherwise. An incident matrix $\bar{R}$ for the 2047 nonempty sets are generated as follows. Let $R(1) = [1]$, and recursively compute

$$R(2n) = \begin{bmatrix} R(n) & R(n) \\ R(n) & J - R(n) \\ 0 & e^T \end{bmatrix}$$

where J and e are, respectively, a matrix and a vector of all ones. Then, $\bar{R} = R(1024)$.

The estimates of the correlation $E[\epsilon_{t-1}(a)\epsilon_t(a)]$ (in fact a re-scaled version as no normalization is present) are then given by

$$g_i(a) = \sum_{k \in I_i} f_k(a), \quad i = 1, \ldots, 2048 \tag{2}$$

($g_i(a) = 0$ if $I_i = \emptyset$). A few $g_i(a)$ functions are plotted in Figure 1. The line with the steepest slope is obtained for the set I itself. This is natural since I contains twice as many elements as any other set of the group, and (2) is not normalised with the number of terms in the summation.

Now, the idea is that for the true a_0, $\epsilon_t(a_0) = w_t$ is white noise and it is very unlikely that all the $g_i(a_0)$ functions but a few will be less than zero or greater than zero. Grounded on this idea, we discard the rightmost and leftmost regions where only 51 functions out of the calculated 2048 are less than zero or greater than zero. The resulting interval [0.12, 0.24], is the confidence region for a_0. It is a rigorous fact (stated in Theorem 3.1) that this confidence region has probability $1 - 51 * 2/2048 = 0.9502 > 95\%$ to contain the true parameter

value a_0. Notice that there is no need to normalise the sums (2) since we are only interested in whether $g_i(a)$ is greater or smaller than 0.

A verification of the theoretical confidence result was performed by running the same simulation 1000 times. The empirical frequency of a_0 being in the confidence interval was 0.956, in good agreement with the theoretical result.

III. Confidence regions for ARMA models

A. Data generating system

The ARMA system that generates the data is given by

$$A^\circ(z^{-1})y_t = C^\circ(z^{-1})w_t,$$

where

$$A^\circ(z^{-1}) = 1 + a_1^\circ z^{-1} + \cdots + a_n^\circ z^{-n} \quad (3)$$

$$C^\circ(z^{-1}) = 1 + c_1^\circ z^{-1} + \cdots + c_p^\circ z^{-p} \quad (4)$$

are stable polynomials with no common factors and $\{w_t\}$ is a zero-mean white wide-sense stationary sequence of random variables with spectral density $\Phi_w(\omega) = \lambda_w^2 > 0$. Notice that no a-priori knowledge of the noise level is assumed. In particular λ_w^2 does not need to be known.

B. Model structure

The model class is $A(z^{-1},\theta)y_t = C(z^{-1},\theta)w_t$, $\theta \in \Theta$, where $A(z^{-1},\theta)$ and $C(z^{-1},\theta)$ are the same as in (3) and (4) except that a_i° and c_i° are substituted by a_i and c_i, $\theta = [a_1 \ \cdots \ a_n \ c_1 \ \cdots \ c_p]^T$ and $C(z^{-1},\theta)$ is stable for any $\theta \in \Theta$.

C. Construction of the uncertainty region

We commence by introducing a procedure for the determination of a certain set Θ_r. Later on in this section, this procedure will be integrated in an algorithm which constructs the parameter uncertainty region $\hat{\Theta}$ by intersecting Θ_r sets.

Procedure for the construction of Θ_r

1) Compute $\epsilon_t(\theta) = y_t - \hat{y}_t(\theta) = A(z^{-1},\theta)/C(z^{-1},\theta)y_t$, where t ranges over a finite interval, say, $[1, H]$;
2) Select an $r \geq 1$. For $t = 1 + r, \ldots, N + r = H$, compute $f_{t-r}(\theta) = \epsilon_{t-r}(\theta)\epsilon_t(\theta)$;
3) Let $I = \{1, \ldots, N\}$ and consider a collection G of subsets $I_i \subseteq I$, $i = 1, \ldots, M$, forming a group under the symmetric difference operation (i.e. $(I_i \cup I_j) - (I_i \cap I_j) \in G$, if $I_i, I_j \in G$). Compute $g_i(\theta) = \sum_{k \in I_i} f_k(\theta), \quad i = 1, \ldots, M$;
4) Select an integer q in the interval $[1, (M+1)/2)$ and find the region Θ_r such that at least q of the $g_i(\theta)$ functions are bigger than zero and at least q are smaller than zero.

Remark 3.1: In the procedure, the group G can be freely selected. Thus, if $I = \{1, 2, 3, 4\}$, a suitable group is $G = \{\{1, 2\}, \{3, 4\}, \emptyset, \{1, 2, 3, 4\}\}$; another one is $G = \{\{1\}, \{2, 3, 4\}, \emptyset, \{1, 2, 3, 4\}\}$; yet another one is $G =$ all subsets of I. While the theory presented holds for any choice, the quality of the result in the uncertainty region assessment is affected by the choice made. Moreover, the feasible choices are limited by computational considerations. For example, the set of all subsets cannot be normally chosen as it is a truly large set.

The intuitive idea behind this algorithm is that, for $\theta = \theta^\circ$, the functions $g_i(\theta)$ assume positive or negative value at random ($\epsilon(t, \theta_0)$ is white noise), so that it is unlikely that almost all of them are positive or that almost all of them are negative. Since point 4 in the construction of Θ_r discards regions where all $g_i(\theta)$'s but a small fraction (q should be taken to be small compared to M, see Theorem 3.1 below) are of the same sign, we expect that $\theta^\circ \in \Theta_r$ with high probability. This is put on solid mathematical grounds in the next theorem.

THEOREM 3.1: Assume that variables w_t admit a density (so that $Pr\{w_t = c\} = 0$, for any real c) and that they are symmetrically distributed around zero. Then, the set Θ_r constructed above is such that: $Pr\{\theta^\circ \in \Theta_r\} = 1 - 2q/M = 1 - \delta_r$.

Note that by choosing different values of q, $1 - 2q/M$ can take on different values for different choices of r, hence we have used the notation δ_r. The proof is given in the appendix.

Remark 3.2: The only reason for requiring that the variables w_t admit a density is to avoid that the functions $g_i(\theta)$ defined in point 3 can take on the same value with nonzero probability. Though this condition can be dropped, we have preferred to maintain it to avoid unduly complications.

When the $\{w_t\}$ process is independent and identically but not symmetrically distributed, we can obtain symmetrically distributed data by considering the difference between two subsequent data points.

The noise assumption is mild enough to accomodate a number of situations. In particular, one can describe possible outliers by allowing the noise to take on large values with small probability.

Theorem 3.1 quantifies the probability that θ° belongs to the region Θ_r. It holds for any finite N and introduces no conservativeness at all, since such a probability is *exactly* equal to $1 - \delta_r$. Theorem 3.1 deals only with one side of the medal in the study of uncertainty evaluation techniques. A good evaluation method must have two properties: the provided region must have guaranteed probability (and this is what Theorem 3.1 delivers); the region must be restricted, and, in particular, it should concentrate around θ° as the number of data points increases. We next provide a result that shows how the second property can be fulfilled (again, the proof is given in the appendix).

THEOREM 3.2: Let $\epsilon_t(\theta) = \frac{A(z^{-1},\theta)}{C(z^{-1},\theta)}y_t$ be the prediction error associated with the considered model class. Then, $\theta = \theta^\circ = [a_1^\circ \ \cdots \ a_n^\circ \ c_1^\circ \ \cdots \ c_p^\circ]^T$ is the unique solution

to the set of equations:

$$E[\epsilon_{t-r}(\theta)\epsilon_t(\theta)] = 0, \quad r = 1, \ldots, n+p. \tag{5}$$

where E is the expectation operator.

Theorem 3.2 claims that if we simultaneously impose $n+p$ correlation conditions, then the only solution is the true θ°. Guided by this idea, we consider $n+p$ *sample* correlation conditions and, correspondingly, apply the "Procedure for the construction of Θ_r" for $r = 1, \ldots, n+p$. As $N \to \infty$, the functions $g_i(\theta) \to E[\epsilon_{t-r}(\theta)\epsilon_t(\theta)]$, provided that the number of elements in each set I_i also tends to infinity. (It is easy to construct groups with this property. Construction of good groups has been considered in Gordon (1974).) This means that each region Θ_r gets smaller and the intersection of them gives an uncertainty region shrinking around the true parameter θ°. This leads to the following algorithm.

Algorithm for the construction of $\hat{\Theta}$

1) For $r = 1, \ldots, n+p$, construct Θ_r as above.
2) Let $\hat{\Theta} = \cap_{r=1}^{n+p} \Theta_r$.

We conclude this section with a fact which is immediate from Theorem 3.1. and stated for the sake of completeness.

THEOREM 3.3: Under the assumptions of Theorem 3.1, the set constructed in the "Algorithm for the construction of $\hat{\Theta}$" is such that: $Pr\{\theta^\circ \in \hat{\Theta}\} \geq 1 - \sum_{r=1}^{n+p} \delta_r$ where δ_r is defined in Theorem 3.1.

The inequality in the Theorem is due to that the sets $\{\theta^\circ \notin \Theta_r\}$ may be overlapping for different r's, see simulation example in section III-D.

D. Simulation example

Consider the ARMA-system

$$y_t + a_0 y_{t-1} = w_t + c_0 w_{t-1}, \tag{6}$$

where $a_0 = -0.5, c_0 = 0.2$ and $\{w_t\}$ is an independent sequence of zero mean normally distributed random variables with variance 1. 1026 data points were generated according to (6). As a model class we used $y_t + a y_{t-1} = w_t + c w_{t-1}$ with associate predictor and prediction error given by

$$\begin{aligned} \hat{y}_t(a,c) &= -c\hat{y}_{t-1}(a,c) + (c-a)y_{t-1}, \\ \epsilon_t(a,c) &= y_t - \hat{y}_t(a,c) = y_t + a y_{t-1} - c\epsilon_{t-1}(a,c). \end{aligned}$$

In order to form a confidence region for (a_0, c_0) we calculated

$$\begin{aligned} f_{t-1,1}(a,c) &= \epsilon_{t-1}(a,c)\epsilon_t(a,c), \quad t = 2, \ldots, 1025 \\ f_{t-2,2}(a,c) &= \epsilon_{t-2}(a,c)\epsilon_t(a,c), \quad t = 3, \ldots, 1026 \end{aligned}$$

and then computed

$$\begin{aligned} g_{i,1}(a,c) &= \sum_{k \in I_i} f_{k,1}(a,c), \quad i = 1, \ldots, 2048 \\ g_{i,2}(a,c) &= \sum_{k \in I_i} f_{k,2}(a,c), \quad i = 1, \ldots, 2048 \end{aligned}$$

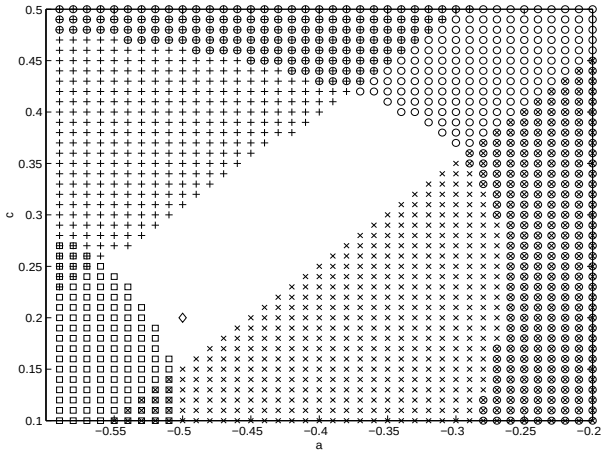

Fig. 2. Confi dence region for (a_0, c_0)

using the same group as in the preview example. Next we discarded those values of a and c for which zero was among the 25 largest and smallest values of $g_{i,1}(a,c)$ and $g_{i,2}(a,c)$. Then according to Theorem 3.3 (a_0, c_0) belongs to the constructed region with probability at least $1 - \frac{2 \cdot 2 \cdot 25}{2048} = 0.9512$.

The obtained confidence region is the blank area in Figure 2. The area marked with x is where 0 is among the 25 smallest values of $g_{i,1}$, the area marked with + is where 0 is among the 25 largest values of $g_{i,1}$. Likewise for $g_{i,2}$ with the squares representing when 0 belongs to the 25 largest elements and the circles the 25 smallest. The true value (a_0, c_0) is marked with a diamond in the middle of the blank region. As we can see, each step in the construction of the confidence region excludes a particular region.

IV. CONCLUSIONS

In this paper we have derived an algorithm for construction of confidence regions for ARMA models. The algorithm is based on computing empirical correlation functions using subsamples and discarding regions in the parameter space where only a small fraction of the empirical functions are greater/smaller than zero. Building on finite sample results from Hartigan (1969) we derived bounds, valid for a finite number of data points, on the probability that the true model parameters belong to the constructed region. The approach can be extended to ARMAX systems, and the approach bears promise for further development of rigorous finite sample results useful in practical applications.

V. ACKNOWLEDGMENTS

This work is partly supported by MIUR under the project "New methods for Identification and Adaptive Control for Industrial Systems".

VI. REFERENCES

[1] Åström K.J. and T. Söderström (1974). "Uniqueness of the maximum likelihood estimates of the parameters of an ARMA model". *IEEE Trans. on Automatic Control*, Vol. 29, no.6, pp. 769-773

[2] Bittanti S., M.C. Campi and S. Garatti (2002). "New results on the asymptotic theory of system identification for the assessment of the quality of estimated models". In *Proc. 41st Conf. on Decision and Control*, Las Vegas, USA, Dec. 2002.

[3] Campi, M.C, S.K. Ooi, and E. Weyer (2002). "Non-asymptotic quality assessment of generalised FIR models" *Proceedings of IEEE Conference on Decision and Control*, pp. 3416-3421, Las Vegas, Nevada, USA, December 2002.

[4] Gordon L. (1974). "Completely Separating Groups in Subsampling", *Annals of Statistics* Vol. 2, pp. 572-578.

[5] Hartigan J. A. (1969). "Using Subsample Values as Typical Values", *Journal of American Statistical Association*, Vol. 64, pp. 1303-1317.

[6] Hartigan J. A. (1970). "Exact Confidence Intervals in Regression Problems with Independent Symmetric Errors", *Annals of Mathematical Statistics*, Vol. 41, pp. 1992-1998.

[7] Ljung, L. (1999). *System Identification - Theory for the User.* 2nd Ed. Prentice Hall.

[8] Weyer E., and M.C. Campi (2002). "Non-asymptotic confidence ellipsoids for the least squares estimate." *Automatica*. Vol 38., no. 9, pp. 1539-1547.

APPENDIX

A. Proof of Theorem 3.1

The proof is divided into a few steps in the form of propositions.

Proposition 1.1: Let $\{w_t\}$ be a sequence of independent random variables with symmetric distribution around zero. Let $I = \{1, \ldots, N\}$, and let G be a collection of subsets $I_i \subseteq I$, $i = 1, \ldots, M$, forming a group under the symmetric difference operation (i.e. $(I_i \cup I_j) - (I_i \cap I_j) \in G$, if $I_i, I_j \in G$). Pick any $\bar{I} \in G$ and an integer r. Then, the set of variables

$$\left\{ \sum_{k \in I_i} w_k w_{k+r}, \quad i = 1, \ldots, M \right\} \tag{7}$$

has the same M-dimensional joint distribution as the set of variables

$$\left\{ \sum_{k \in I_i} w_k w_{k+r} - \sum_{k \in \bar{I}} w_k w_{k+r}, \quad i = 1, \ldots, M \right\}, \tag{8}$$

provided that the order of the variables is suitably rearranged.

Before providing the proof, we give a simple example illustrating the idea. Suppose that $I = \{1, 2, 3, 4\}$, $r = 1$, and $G = \{\{1, 2\}, \{3, 4\}, \emptyset, \{1, 2, 3, 4\}\}$. Take $\bar{I} = \{1, 2\}$. Proposition 1.1 says that (by convention, $\sum_{k \in \emptyset} w_k w_{k+r} = 0$):

$$\{w_1 w_2 + w_2 w_3,\ w_3 w_4 + w_4 w_5,\ 0,\ w_1 w_2 + w_2 w_3 + w_3 w_4 + w_4 w_5\} \tag{9}$$

has the same distribution as

$$\{0,\ w_3 w_4 + w_4 w_5 - w_1 w_2 - w_2 w_3,\ -w_1 w_2 - w_2 w_3,\ w_3 w_4 + w_4 w_5\}.$$

Proof: Consider the variables

$$w_1 w_{1+r}\ w_2 w_{2+r}\ w_3 w_{3+r}\ \cdots\ w_N w_{N+r} \tag{10}$$

and re-organize them in the following chains

$$\begin{array}{llll} w_1 w_{1+r} & w_{1+r} w_{1+2r} & w_{1+2r} w_{1+3r} & \cdots \quad (chain\ 1) \\ w_2 w_{2+r} & w_{2+r} w_{2+2r} & w_{2+2r} w_{2+3r} & \cdots \quad (chain\ 2) \\ \vdots & & & \\ w_r w_{2r} & w_{2r} w_{3r} & w_{3r} w_{4r} & \cdots \quad (chain\ r) \end{array}$$

We consider the various chains in turn. Consider first chain no. 1 and scan its elements from left to right. When an element belonging to the set $\{w_k w_{k+r}, k \in \bar{I}\}$ - say $w_{\bar{k}} w_{\bar{k}+r}$ - is encountered, introduce the new variable $\tilde{w}_{\bar{k}+r} = -w_{\bar{k}+r}$, and rewrite the element as $-w_{\bar{k}} \tilde{w}_{\bar{k}+r}$. The next element is then rewritten as $\tilde{w}_{\bar{k}+r} \tilde{w}_{\bar{k}+2r}$ with $\tilde{w}_{\bar{k}+2r} = -w_{\bar{k}+2r}$. So is the next one: $\tilde{w}_{\bar{k}+2r} \tilde{w}_{\bar{k}+3r}$, and we proceed this way until another element in $\{w_k w_{k+r}, k \in \bar{I}\}$ - say $w_{\bar{\bar{k}}} w_{\bar{\bar{k}}+r}$ - is encountered. This element is rewritten as $-\tilde{w}_{\bar{\bar{k}}} w_{\bar{\bar{k}}+r}$, where $\tilde{w}_{\bar{\bar{k}}} = -w_{\bar{\bar{k}}}$, interrupting the sequence of sign change. We proceed scanning the first chain and we start changing the sign again when we encounter the next element in $\{w_k w_{k+r}, k \in \bar{I}\}$. The procedure terminates when all elements in the first chain have been considered. Then, we start over again with chain no. 2, and then chain no. 3 and so on. When all chains have been scanned, we reorder all variables in a sequence, similarly to (10). The resulting sequence is in fact the same as (10), except that some variables have been rewritten with a '˜' and, correspondingly, some signs have been changed.

Next, we rewrite all elements in (8) with the new notation (i.e., substituting w_t with $-\tilde{w}_t$, if w_t has been substituted by $-\tilde{w}_t$). It can be seen by inspection that the rewritten variables in (8) take on the same form as the variables in (7) (though in a rearranged order) except that some variables appear with the '˜ '. The theorem conclusion is then drawn by observing that the '˜' is immaterial as far as the distributions is concerned since the w_t's are symmetrically distributed around zero, so that w_t and $\tilde{w}_t = -w_t$ have the same distribution.

The next proposition proves that the variables in the set (7) exhibit a precise ordering property.

Proposition 1.2: Let $\{w_t\}$ be a sequence of independent random variables with symmetric distribution around

zero and such that $Pr\{w_t = c\} = 0$, for any t and any real c. Let $I = \{1, \dots, N\}$, and let G be a collection of subsets $I_i \subseteq I$, $i = 1, \dots, M$, forming a group under the symmetric difference operation, and pick an integer r.
Then, the set of variables in (7) has the following property: each variable in the set has the same probability $1/M$ to be in the j-th position (i.e. there are exactly $j-1$ other variables in the set (7) smaller than the variable under consideration) and this holds for any choice of j between 1 and M.

Thus, if we consider the situation described before the proof of Proposition 1.1, it means that the variables in (9) have the same probability of being in a generic j-th position. In other words, if we were asked to bet on one of the variables to be e.g. smaller than all others, our probability of success would not be affected by the choice we make.

Proof: Pick a variable in the set (7), say $\sum_{k\in\bar{I}} w_k w_{k+r}$, $\bar{I} \in G$. This variable is in the j-th position if the inequality

$$\sum_{k\in\bar{I}} w_k w_{k+r} > \sum_{k\in I_i} w_k w_{k+r}$$

is satisfied for exactly $j-1$ choices of $I_i \in G$. But, this is equivalent to say that

$$\sum_{k\in I_i} w_k w_{k+r} - \sum_{k\in\bar{I}} w_k w_{k+r} < 0$$

holds for $j-1$ selections of I_i. Now, using Proposition 1.1 we have:

$$\begin{aligned} &Pr\left\{\sum_{k\in I_i} w_k w_{k+r} - \sum_{k\in\bar{I}} w_k w_{k+r} < 0 \right. \\ &\qquad\qquad \left. \text{for } j-1 \text{ selections of } I_i\right\} \\ &= Pr\left\{\sum_{k\in I_i} w_k w_{k+r} < 0 \quad \text{for } j-1 \text{ selections of } I_i\right\} \end{aligned}$$

showing that the probability of the event on the left hand side does not depend on the chosen $\bar{I}$. So, any $\bar{I}$ has the same probability that $\sum_{k\in\bar{I}} w_k w_{k+r}$ is in the j-th position and, there being M the possible choices of $\bar{I}$, the probability is $1/M$.

We now come to the proof of Theorem 3.1. Consider the event

$$\begin{aligned} A = &\left\{\sum_{k\in I_i} w_k w_{k+r} < 0 \right. \\ &\qquad \left. \text{for at most } q-1 \text{ selections of } I_i\right\} \cup \\ &\left\{\sum_{k\in I_i} w_k w_{k+r} > 0 \quad \text{for at most } q-1 \text{ selections of } I_i\right\} \\ &= \{0 \text{ is in the 1-st or 2-nd or } \cdots \text{ or } q\text{-th position}\} \\ &\quad \cup \{0 \text{ is in the } M\text{-th or } (M-1)\text{-th or } \cdots \\ &\qquad\qquad \text{or } (M-q+1)\text{-th position}\} \end{aligned}$$

In view of Proposition 1.2 (note that 0 is one variable in set (7)),

$$Pr(A) = 2q/M. \tag{11}$$

Note that $w_t = \epsilon_t(\theta^\circ)$, so that $\sum_{k\in I_i} w_k w_{k+r} = g_i(\theta^\circ)$ (recall the definition of $g_i(\theta)$ in the "Procedure for the construction of Θ_r"). Suppose that the probabilistic outcome s has been selected in A. Then, either $g_i(\theta^\circ) > 0$ for at most $q-1$ selection of I_i or it is < 0 for at most $q-1$ selection of I_i, so that $\theta^\circ \notin \Theta_r$ (recall the construction of Θ_r). Vice versa, if $s \notin A$, then $g_i(\theta^\circ) > 0$ for at least q selection of I_i and it is < 0 again for at least q selection of I_i, yielding $\theta^\circ \in \Theta_r$. Using (11), the conclusion is drawn that $Pr\{\theta^\circ \in \Theta_r\} = 1 - \frac{2q}{M}$ and the proof is completed.

B. Proof of Theorem 3.2

In the proof, we use the following lemma, taken from Åström and Söderström (1974).

Lemma 1.1: Consider the function

$$f(z) = \frac{g(z)}{\Pi_{i=1}^{\ell}(z-u_i)^{t_i}}$$

where g is analytic inside and on the unit circle, the numbers u_i are distinct and $t_i \geq 1$. Assume that

$$\oint f(z) z^{k-1} dz = 0, \quad k = 1, \dots, q,$$

where the integration path is the unit circle and $q = \sum_{i=1}^{\ell} t_i$. Then, f is analytic inside and on the unit circle.

We now turn to the proof of Theorem 3.2. Condition (5) can be re-written as

$$\begin{aligned} 0 &= \int_{-\pi}^{\pi} \Phi_\epsilon(\omega) e^{i\omega r} d\omega \\ &= \int_{-\pi}^{\pi} \left|\frac{A(e^{-i\omega},\theta)}{C(e^{-i\omega},\theta)} \frac{C^\circ(e^{-i\omega})}{A^\circ(e^{-i\omega})}\right|^2 \lambda_w^2 e^{i\omega r} d\omega \\ &= \oint \frac{z^n A(z^{-1},\theta) z^p C^\circ(z^{-1})}{z^p C(z^{-1},\theta) z^n A^\circ(z^{-1})} \cdot \frac{A(z,\theta)C^\circ(z)}{C(z,\theta)A^\circ(z)} \frac{\lambda_w^2}{i} z^{r-1} dz \\ &= \oint \frac{g(z)}{\Pi_1^{\ell}(z-u_i)^{t_i}} z^{r-1} dz = 0, \quad r = 1, \dots, n+p, \end{aligned}$$

where $g(z) = \frac{z^n A(z^{-1},\theta) z^p C^\circ(z^{-1}) A(z,\theta) C^\circ(z)}{C(z,\theta) A^\circ(z) i} \frac{\lambda_w^2}{i}$ is analytic inside and on the unit circle, the numbers u_i are the distinct zeros of $z^p C(z^{-1},\theta) z^n A^\circ(z^{-1})$ and t_i is their multiplicity. Then, by applying Lemma 1.1 with $q = n+p$, we conclude that $\frac{g(z)}{\Pi_{i=1}^{\ell}(z-u_i)^{t_i}}$ is analytic inside and on the unit circle. In turn, this implies that the zeros of $z^p C(z^{-1},\theta) z^n A^\circ(z^{-1})$ - which are all inside the unit circle - are canceled by those of $z^n A(z^{-1},\theta) z^p C^\circ(z^{-1})$. Since $z^n A^\circ(z^{-1})$ and $z^p C^\circ(z^{-1})$ have no common factors, this gives $C(z^{-1},\theta) = C^\circ(z^{-1})$ and $A(z^{-1},\theta) = A^\circ(z^{-1})$, concluding the proof.