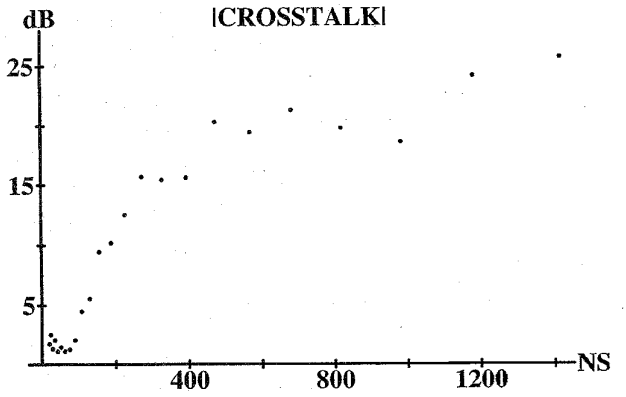

Fig. 1. Absolute value of the crosstalk with respect to the number of samples (NS) used to estimate the cross cumulants. Each point is the average of 10 experiments.

to estimate the cross-cumulants. Each point in Fig. 1 corresponds to the average over 10 experiments, in which the mixing matrix is randomly chosen: The matrix entries m_{ij} $(i \neq j)$ are random numbers in the range $[-1, +1]$. With 500 samples, a residual crosstalk of about -20 dB is obtained. In the case of nonstationary signals, cross-cumulant estimation must be done on few samples and has a larger variance. Consequently, it can lead to more inaccurate estimation of the mixing matrix. We still obtained an interesting performance: a residual crosstalk of about -15 to -20 dB, with various signals (colored noise, speech) and statistics estimated over 500 samples.

In this correspondence, we proved that the mixing matrix can be estimated using fourth-ordercross-cumulants, for two mixtures of two non-Gaussian sources. Solutions are obtained by rooting a fourth-order polynomial equation. Using second-order cross-cumulants allows us to simplify the method; the solution is then obtained by rooting two second-order polynomial equations and gives the result if one source is Gaussian. The methods are then quite simple, but its roots are very sensitive to the accuracy of the estimated cumulants. In fact, this direct solution is less accurate than indirect methods, especially adaptive algorithms. Moreover, we restricted the study to the separation of two sources, and theoretical solutions for three sources or more seems not easily tractable. However, in the case of two mixtures of two sources, it may give a good starting point with a small computation cost for any adaptive algorithm.

REFERENCES

[1] J.-F. Cardoso, "Blind identification of independent signals," in *Proc. Workshop Higher Order Spectral Anal.*, Vail, CO, June 1989.

[2] P. Comon, "Separation of sources using higher order cumulants," in *Proc. SPIE Advanced Algorithms and Architectures for Signal Processing IV*, San Diego, vol. 1152, Aug. 8–10, 1989.

[3] ——, "Separation of stochastic processes," in *Proc. Workshop Higher Order Spectral Anal.*, Vail, CO, June 1989.

[4] ——, "Independent component analysis: A new concept?" *Signal Processing*, vol. 36, no. 3, pp 287–314, 1994.

[5] C. Jutten and J. Hérault, "Blind separation of sources, Part I: An adaptive algorithm based on a neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–10, 1991.

[6] J.-L. Lacoume and P. Ruiz, "Sources identification: A solution based on cumulants," in *Proc. IEEE Acoust., Speech, Signal Processing Workshop V*, Minneapolis, MN, Aug. 1988.

[7] E. Moreau and O. Macchi, "New self-adaptive algorithms for source separation based on contrast functions," *Proc. IEEE Signal Processing Workshop Higher Order Statistics*, South Lake Tahoe, CA, USA, June 1993, pp. 215–219.

[8] D. T. Pham and Ph. Garat, "Séparation aveugle de sources temporellement corréllées," in *Proc. XIV Colloque GRETSI*, Juan-Les-Pins, France, Sept. 1993, pp. 317–320.

[9] D. T. Pham, Ph. Garat, and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach," in *Signal Processing IV, Theories and Applications*, J. Vandewalle, R. Boite, M. Moonen, and A. Oosterlink, Eds. Brussels: Elsevier, Aug. 1992.

[10] R. L. L. Tong and Y. H. V. C. Soon, "Indeterminacy and identifiability of blind identification," *IEEE Trans. Circuits Syst.*, vol. 38, pp. 499–509, 1991.

On the Penalty Factor for Autoregressive Order Selection in Finite Samples

P. M. T. Broersen and H. E. Wensink

***Abstract*— The order selection criterion that selects models with the smallest squared error of prediction is the best. The finite sample theory describes equivalents for asymptotic order selection criteria that are better in the finite sample practice. This correction for finite sample statistics is the most important. Afterwards, a preference in order selection criteria can be obtained by computing an optimal value for the penalty factor based on a subjective balance of the risks of overfitting and underfitting.**

I. INTRODUCTION

Recently, a finite sample theory for the selection of an autoregressive model order has been presented [1]. It provides a comprehensive description of the peculiarities of estimation and order selection in finite and small samples of autoregressive time series. A sample is called finite if the maximum model order that is interesting is greater than about $0.1N$, where N is the sample size.

The asymptotical theory describes theoretical statistical expectations as multiples of $1/N$. Its validity and accuracy is poor in finite data series when the model order p is not very small with respect to N. The inclusion of higher order terms in the Taylor expansion fails to give the necessary improvements. Another approach, the finite sample theory, is based upon the actual observed behavior of the estimates. For four autoregressive (AR) estimation methods, a correction is provided in replacing $1/N$ by the finite sample variance coefficients, which are empirical formulas for a method-dependent description of the actual degrees of freedom in parameter estimation.

The order selection problems have been discussed in a former paper [1]. Finite sample equivalents have been given for the FPE criterion [2] and for selection criteria with the logarithm of the residual variance. They are asymptotically equivalent to the existing criteria, but their performance turns out to be better in finite samples. The remaining problem is the optimal value of the penalty factor α for selecting the best predicting model for a variety of processes, which is the subject of this paper. Which model order is best for a given process depends on *four* indicators: sample size N, estimation method, true model order, and true values of the process parameters.

Manuscript received February 24, 1994; revised September 18, 1995. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Monique Fargues.

The authors are with the Department of Applied Physics, Delft University of Technology, Delft, The Netherlands.

Publisher Item Identifier S 1053-587X(96)02410-3.

Hence, selection of the true order is not always desirable and we need a measure for model quality.

In this work, we introduce the selection error as a quality criterion for selected models. In contrast to the mathematical derivation of consistent criteria with fixed parameter values and increasing sample size, we define a critical parameter value depending on the sample size N and on α. In this way, we find theoretically a preference for the penalty α, which is still based on a subjective choice for the balance between overfitting and underfitting. We compare results of asymptotical derivations with finite sample simulations to verify that they have practical relevance. Finally, we give an asymptotical result for the expected decrease in FPE due to overfit, showing that expectations of criteria after selection are no longer equal to expectations for fixed orders without selection.

II. AR Estimation

An autoregressive process of order K is given by parameters $1, a_1, \cdots, a_K$ and zero mean i.i.d. innovations ϵ_n with variance σ_ϵ^2. The true order K may be infinite. Only stationary processes are considered where the roots of the characteristic polynomial lie within the unit circle. A model of order p has parameters $\hat{a}_i(p)$, together forming the $(p+1) \times 1$ vector $\hat{\boldsymbol{a}}(p)$, with $\hat{a}_0(p) = 1$. A model order has to be selected from a number of candidate orders $p = 0, 1, \cdots, L$. The residual variance $S^2(p)$ is defined as the mean square fit of the model to the data from which the parameters have been estimated, and decreases always for higher model orders. The predictive capacity of the model is expressed by the prediction error PE(p). Let the matrix $\boldsymbol{R}$ describe the exact covariance structure of the true Kth-order AR process. PE(p) has been defined with the $(p+1) \times (p+1)$ submatrix $\boldsymbol{R}(p)$ of the true $\boldsymbol{R}$ as follows:

$$\mathrm{PE}(p) = \hat{\boldsymbol{a}}^{\boldsymbol{T}}(\boldsymbol{p})\, \boldsymbol{R}(\boldsymbol{p})\, \hat{\boldsymbol{a}}(\boldsymbol{p}) \tag{1}$$

and has as asymptotical expectation $\sigma_\epsilon^2(1+p/N)$ for $p \geq K$ [1], [2]. PE(p) is the natural measure for the fit of a model to a given process where $\boldsymbol{R}$ is known, for instance in simulations.

Elements of information theory have been used to derive the order selection criterion AIC(p) [3]. Consistent and minimally consistent criteria are found by changing the penalty factor 2 of AIC(p) into $\ln(N)$ [4] and $2 \ln \ln(N)$ [5], respectively. These criteria can be described together as a generalized information criterion, GIC(p, α) with α as the penalty factor [1], as follows:

$$\mathrm{GIC}(p, \alpha) = \ln [S^2(p)] + \alpha \frac{p}{N}. \tag{2}$$

When a sample is finite, also the estimation method and the parameter order i influence the estimates. Therefore, the finite sample variance coefficients $v(i, \cdot)$ have been developed based on the actual degrees of freedom that play a role in a given estimation method. Four AR estimation methods have been treated: the Yule–Walker method (YW), the method of Burg (Burg), the least squares method that minimizes forward and backward residuals (LSFB), and the least squares method that minimizes forward residuals only (LSF). The $v(i, \cdot)$ are given by [1] and [6], as follows:

$$\begin{aligned} v(i, \mathrm{YW}) &= \frac{N-i}{N(N+2)} \\ v(i, \mathrm{Burg}) &= \frac{1}{N+1-i} \\ v(i, \mathrm{LSFB}) &= \frac{1}{N+1.5-1.5i} \\ v(i, \mathrm{LSF}) &= \frac{1}{N+2-2i} \qquad i \geq 1. \end{aligned} \tag{3}$$

Apart from this, $v(0, \cdot)$ is defined to be $1/N$ if the estimated mean of the time series is subtracted; otherwise, $v(0, \cdot)$ is defined to be zero.

TABLE I
ASYMPTOTIC VALUE OF THE SELECTION RISK SR$(L-K, \alpha)$ FOR OVERFFIT, FOR $K = 0$ AND $L = 100$

α	SR	α	SR	α	SR
.0	100.000	2.6	1.288	6	.132
.5	98.964	2.8	1.014	7	.080
1.0	56.788	2.915	.915	8	.049
1.5	7.524	3.0	.851	9	.031
2.0	2.568	3.2	.723	10	.019
2.2	1.936	4	.411	12	.007
2.4	1.520	5	.226	20	.000

As finite sample equivalent of GIC(p, α), the finite sample criterion [1] has been introduced, as follows:

$$\mathrm{FIC}(p, \alpha) = \ln [S^2(p)] + \alpha \sum_{i=0}^{p} v(i, \cdot). \tag{4}$$

In almost all simulations, the average performance of FIC(p, α) is better than that of GIC(p, α) for each estimation method, sample size, and value of α [1].

III. Efficiency of Order Selection

The PE(p) of (1) depends strongly on the sample size. We introduce the selection error $SE(p)$ as a quality criterion for a selected model of order p, corrected for the sample size and input power, as

$$\mathrm{SE}(p) = N[\mathrm{PE}(p)/\sigma_\epsilon^2 - 1]. \tag{5}$$

For fixed order p, $E[\mathrm{SE}(p)]$ equals asymptotically p for $p \geq K$, because the expectation of the scaled PE(p) is then given by $1+p/N$ [1], [2]. We use the index p to denote a fixed order. In simulations with selection, however, the selected order $M(\alpha)$ differs from run to run. Criteria FIC(p, α), each with a different value for the penalty α, select their own model order $M(\alpha)$ in each run. The averages SE$\{M(\alpha)\}$ over the simulation runs involve different model orders $M(\alpha)$ because it is highly improbable that a criterion selects the same model order in every run.

The possibilities of selecting a wrong model order can be separated into overfitting and underfitting. For a good order selection criterion the risks of overfitting and underfitting should together be small. However, popular order selection criteria minimize the risks of either underfitting or overfitting. In consistent criteria, overfitting is made small at the cost of underfitting. In contrast, asymptotic efficient criteria with $\alpha = 2$ minimize the risk of underfitting [7] at the cost of overfitting. When orders are actually selected, the expectation of the selection error increases with a selection component in comparison to expectations in fixed order situations. Shibata [7], [8] determined the asymptotic distributions, Prob$(M = p)$ for orders $M \geq K$, selected with the FPE(p) or AIC(p) [7]. These depend for one order overfit on the probability that the square of a normally distributed zero-mean parameter estimate exceeds the significance level of α in the χ^2 distribution [7], and similar combinations for higher overfit orders. The extra stochastic contribution to $E[\mathrm{SE}(M)]$ for $M > K$ is caused by selection. The selection error can in the case of overfitting be written as $E[\mathrm{SE}(M)] = E[\mathrm{SE}(K)] + E[\mathrm{SR}(L-K, \alpha)]$. In this way, we define SR$(L-K, \alpha)$ as the selection risk; $L-K$ is the number of overfit order candidates. The asymptotical expectation of this risk

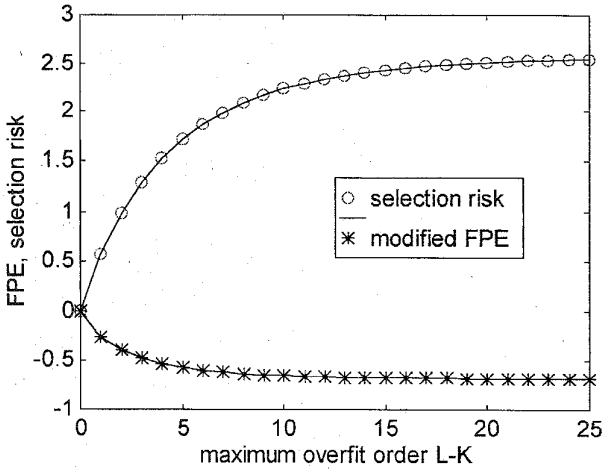

Fig. 1. Selection risk and modified FPE for $\alpha = 2$.

is found with $E[\mathrm{SE}(K)] = K$ and a result of Shibata [8] as

$$E[\mathrm{SR}(L-K,\alpha)] = \sum_{m=1}^{L-K} \mathrm{Prob}(\chi^2_{m+2} > \alpha, m), \qquad M \geq K. \tag{6}$$

Table I gives values for $E[\mathrm{SR}(L-K,\alpha)]$ for $L-K=100$. It shows that, when α is less than two, the overfitting risk is enormous, whereas when α takes values greater than seven the risk becomes practically zero. Hence, in consistent order selection criteria, where α is made a function of N, e.g. $\ln(N)$ or $2\ln\ln(N)$, the selection risk due to overfitting becomes zero when $N \to \infty$.

An important side effect of order selection is that expectations for selected models of order M are no longer equal to the presampling expectations for fixed model orders p. We will illustrate this for FPE(p), which will asymptotically select the same order as AIC(p). Without selection, FPE(p) is an unbiased estimate for the prediction error for model orders $p \geq K$, with expectation $\sigma_\epsilon^2(1+p/N)$ [2]. However, the FPE(M) of a *selected* model with $M > K$ has a lower value than FPE(K) in the case of overfit; otherwise, no overfitted model would be selected. We derive an asymptotical result for a scaled FPE(M) in the Appendix as

$$\begin{aligned} &E[N\{\mathrm{FPE}(M)/\sigma_\epsilon^2 - 1\}] \\ &\quad = K - E[\mathrm{SR}(L-K,2)] \\ &\qquad + 2\sum_{i=0}^{L-K} i\,\mathrm{Prob}(M = K+i) \qquad K \leq M \leq L. \end{aligned} \tag{7}$$

Remember that the result above would be M without selection. Both the theoretically expected increase (6) of the selection risk $\mathrm{SR}(L-K,2)$ and the decrease with respect to K in FPE(M) of (7) are presented in Fig. 1, where the possibility of underfitting is neglected. The asymptotical value for $\mathrm{SR}(L-K,2)$ for $L \to \infty$ is about 2.57, which is already achieved for 20 overfit orders. Also, the modified FPE(M) of (7) has reached a constant value then, being -0.69. This result shows that a function giving an unbiased estimate for prediction error prior to selection is biased after selection.

We have investigated the applicability of (6) and (7) to finite samples in many simulations with white noise, which is an ideal example for studying the effects of overfitting without interference with underfitting problems. Averages over 10 000 simulation runs are given for Burg estimates $N = 100, K = 0, L = 15$. The average SE(M) of the selected models was the same for selection with GIC($p,2$) and FPE(p), the value was 3.32; the average scaled FPE(p) was -0.80. The asymptotical theoretical values are 2.43 and -0.67, respectively, for $L-K=15$. In selection with FIC($p,2$), the average SE(M) was 2.96 in the same simulations, showing that the finite sample criterion gives a better result with lower SE(M). For greater sample sizes, simulation results of GIC($p,2$) and FIC($p,2$) are closer and are also nearer to the asymptotical theoretical value.

We suppose now that a choice has to be made between two orders, $K-1$ and K. We define a critical value $\tilde{a}_K$ as the value of a_K that equalizes $E[\mathrm{FIC}(K-1,\alpha)]$ and $E[\mathrm{FIC}(K,\alpha)]$, so it depends on the value of α. It can easily be derived with the finite sample theory [1], [9] and with (4) that

$$\begin{aligned} E[\mathrm{FIC}(K,\alpha)] &= E[\mathrm{FIC}(K-1,\alpha)] \\ &\quad + \ln(1-\tilde{a}_K^2)(1-v(K,\cdot)) + \alpha v(K),\cdot). \end{aligned} \tag{8}$$

The critical value for the last parameter follows from this equation as

$$\tilde{a}_K^2 = 1 - \frac{\exp\{-\alpha v(K,\cdot)\}}{1-v(k),\cdot)}. \tag{9}$$

For a process with this critical value, FIC(p,α) will select the orders $K-1$ and K with about the same probability. For more specific conclusions, knowledge of the finite sample distribution function of FIC(p,α) would be necessary, and that is not available. The appearance of the finite sample variance coefficient $v(K,\cdot)$ in (9) illustrates the influence of the four indicators for the best model order: the true order K; the sample size N; the estimation method; and the value of the last parameter. An asymptotical expression for (9) becomes $\tilde{a}_K = \sqrt{\{(\alpha-1)/N\}}$ by substituting $1/N$ for $v(K,\cdot)$ and neglecting higher order terms.

If the models with orders K and $K-1$ have the same expectation of the selection criterion in (8), it is interesting to determine the difference in the quality of those models as given by the selection error. The asymptotical expectation of the prediction error for order K is related to $E[\mathrm{PE}(K-1)]$ as

$$\begin{aligned} E[\mathrm{PE}(K)] &= \sigma_\epsilon^2(1+K/N) \approx E[\mathrm{PE}(K-1)] \\ &\quad \cdot [1-a_K^2][1+1/N] \end{aligned} \tag{10}$$

where the approximation in the second step is allowed because terms with N^{-2} and higher are neglected. Substituting the asymptotical value $\tilde{a}_K$, it follows that $E[\mathrm{PE}(K-1)] - E[\mathrm{PE}(K)]$ is $\sigma_\epsilon^2(\alpha-2)/N$ for that critical value. We define the difference $\mathrm{SE}(K-1) - \mathrm{SE}(K)$ if the last parameter of the process has the critical value for that penalty factor, as the selection risk for underfitting $\mathrm{SR}(-1,\alpha)$. This yields, with (5), the following expression for the expectation:

$$E[\mathrm{SR}(-1,\alpha)] = \alpha - 2. \tag{11}$$

Values of α less than two give a negative underfitting risk; $\alpha = 2$ gives no risk. $E[\mathrm{SR}(-1,\ln N)]$ becomes $\ln(N) - 2$ for order $K-1$ in the consistent criterion, so $\alpha = \ln N$ gives infinite risk for $N \to \infty$. For values larger than the critical $\tilde{a}_K$, order K will more often be selected, whereas for smaller values this will be $K-1$. The asymptotical derivation can be considered as an approximation for finite sample sizes where the more complicated expression (9) for the critical parameter value should be used and no neglect of higher order terms is allowed.

IV. A Possible Choice for α

The penalty factor α influences the ratio between the two components of the selection risk—the risk of overfitting and the risk of underfitting. Obvious artifacts in best values for α are $\alpha = \infty$ for white noise, because the zeroth order model is best, and $\alpha = 0$ for

TABLE II
OPTIMAL COEFFICIENTS α FOR FIC(M,α) FOR $K=1, L=N/3$, BASED ON SIMULATIONS

N	YW	Burg	LSFB	LSF
20	2.9	3.1	3.3	3.3
30	2.9	3.0	3.2	3.3
50	2.9	3.0	3.2	3.2
100	2.9	3.0	3.2	3.1
250	2.9	2.9	3.1	3.1
1000	2.9	2.9	3.0	3.0

$L = K$ because overfitting is impossible then and underfitting will not take place if the penalty is zero. In general, the greatest α without getting underfit will be the best choice for a given process, because that will give the smallest overfit risk. Table I shows that the selection risk due to overfitting is smaller when α increases. However, (11) shows that the risk due to underfitting increases with α.

An optimal value of α for *unknown* processes will depend on the balance between underfit and overfit. Underfit risk will only appear in practice if the true parameters have values of about the critical magnitude. In contrast, the risk of overfit will always apply if the maximum order L is not too restricted. We will try to minimize the maximum of selection error in practice by balancing both heterogeneous risks. Consider the case where the Kth parameter of the AR(K) process has the critical value $\tilde{a}_K$. *We define as the optimal α for unknown processes that specific value for which the expected selection risk due to overfitting is equal to the expected risk in underfitting one order with the critical parameter.*

In asymptotical theory this value is found as that value for which $E[\text{SR}(L-K,\alpha)]$ in Table I equals $E[\text{SR}(-1,\alpha)]$ of (11), which yields $\alpha = 2.915$. For the finite sample FIC(M,α) the results are roughly the same. Table II presents the simulation results of an evaluation for an AR(1) process. The value of α is determined with $v(1,\cdot)$ in (9) by making the underfitting risk $E[\text{SE}(1)-\text{SE}(0)]$ equal to the overfitting risk as found in simulations. The optimal α values in the finite sample FIC(M,α) are given for different values of N and different estimation methods. The best α is shown to depend slightly on the estimation method and on the sample size, but is always close to 2.915, the asymptotical value that was found for GIC(M,α). We therefore conclude that a penalty factor α equal to three is a good compromise in finite samples.

The balance between both risks can be established by means of Table I and $E[\text{SR}(-1,\alpha)] = \alpha - 2$. The ratio of the underfitting risk $\alpha - 2$ and the overfitting risk is zero for $\alpha = 2, 1$ for $\alpha = 2.915$, 4.9 for $\alpha = 4$, and 421 for $\alpha = 10$. Hence, only a small range of α values gives a ratio of approximately one. Other sensible definitions of the underfit risk would not yield completely different values for α. In consistent methods α is a function of the sample size N, resulting in a dependence of the balance on this sample size. There might be a good reason to take α greater than two, but our analysis gives no reason at all to make it dependent on N. An argument to take very high values for α is that the constant selection risk due to overfitting is made small, whereas the loss due to underfitting will only occur in practice if the true parameters have near-critical value. An opposite argument to be careful with higher values of α is that the loss of underfitting becomes much greater than considered if not only the last parameter, but also a number of previous true parameters are near their critical values. In asymptotical theory, the maximum selection risk due to underfitting can become $K(\alpha-2)$ if all true parameters have critical values. This illustrates why $\alpha = 2$ is efficient: For α greater than two, the maximum possible increase in the selection risk due to underfitting goes to infinity for $K \to \infty$.

When everything about the data-generating process is known *a priori*, the largest value of α should be chosen such that the last process parameter would be incorporated in the model and no more; thus, order K would be selected. However, in these circumstances no selection is required because everything had to be known in advance. In the case with an unknown value for the last parameter, FIC$(p,3)$ is a good compromise. The reason that, in finite samples, consistent criteria seem to work well is found in the fact that the penalty is close to three for a range of observation lengths. Simulations with a variety of true processes and four estimation method always demonstrated a good, and usually the best, performance of FIC$(p,3)$ in comparison with FIC(p,α) and GIC(p,α) for many different values of α.

V. CONCLUSIONS

The possibilities of selecting a wrong order can be separated into overfitting and underfitting. The risk of overfitting is determined by statistics. When an order is selected that is too low, the risk of underfitting depends on the true process by the deterministic values of the last parameters. An order selection criterion performs well in practice if the penalty factor is such that a balance is created between both risks. A good predicting model is then found with FIC$(p,3)$, with penalty $\alpha = 3$. It balances the expected statistical increase of prediction accuracy due to overfitting with the deterministic risk of underfitting one order.

APPENDIX

The selection error SE(M) is related to PE(M) by (5). The selection risk SR$(L-K,\alpha)$ describes the increase of SE(M) in the case of overfit. Hence, the prediction error, corrected for overfit above order K and properly scaled, is given by

$$E[\text{PE}(M)] = \sigma_\epsilon^2\{1 + K/N + E[\text{SR}(L-K,\alpha)]/N\}. \quad \text{(A1)}$$

For white noise, the expected residual variance that belongs to models with this value of PE(M) follows easily because $E[\text{PE}(M)+S^2(M)]$ remains constant $2\sigma_\epsilon^2$ in the asymptotic theory. Residuals for models of order $M \geq K$ are white, so this applies also to overfitted models asymptotically, yielding

$$E[S^2(M)] = \sigma_\epsilon^2\{1 - K/N - E[\text{SR}(L-K,\alpha)]/N\}. \quad \text{(A2)}$$

As FPE(p) is given by $(N+p)/(N-p)*S^2(p)$, which approximately equals $S^2(p)\{1+2p/N\}$, it follows that

$$E\{\text{FPE}(M)] = E[S^2(M)]\left\{1 + 2\sum_{i=K}^{L}\frac{i\,\text{Prob}(M=i)}{N}\right\} \quad \text{(A3)}$$

assuming only orders K and higher are selected. Neglecting terms of N^{-2} and higher, (7) is found.

REFERENCES

[1] P. M. T. Broersen and H. E. Wensink, "On finite sample theory for autoregressive model order selection," *IEEE Trans. Signal Processing*, vol. 41, no. 1, pp. 194–204, 1993.

[2] H. Akaike, "Statistical predictor identification," *Annals Inst. Statist. Math.*, vol. 22, pp. 203–217, 1970.

[3] ——, "A new look at statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.

[4] G. Schwarz, "Estimating the dimension of a model," *Annals Statistics*, pp. 461–464, 1978.

[5] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J.R. Stat. Soc. Ser. B*, vol. B-41, pp. 190–195, 1979.

[6] P. M. T. Broersen, "Selecting the order of autoregressive models from small samples," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 874–879, 1985.
[7] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, vol. 63, pp. 117–126, 1976.
[8] ——— "A theoretical view of the use of AIC," in *Time Series Analysis, Theory and Praxis 4*, O. D. Anderson, Ed. New York: : Elsevier North-Holland Sci., 1983.
[9] P. M. T. Broersen and H. E. Wensink, "A framework for autoregressive theory," in *Proc. EUSIPCO-92, Brussels, Belgium*, pp. 775–778.

The Optimal Focusing Subspace for Coherent Signal Subspace Processing

S. Valaee and P. Kabal

Abstract—**In this correspondence, we introduce a technique to determine an optimal focusing frequency for the direction-of-arrival estimation of wideband signals using the coherent signal-subspace processing method. We minimize the subspace fitting error to select an optimal focusing frequency. Direct optimization of this criterion can be computationally complex—the complexity increases with the number of frequency samples. An alternative technique is introduced that performs nearly as well as the optimal method. This suboptimal technique is based on minimizing a tight bound to the error. The computational complexity of the suboptimal method is independent of the number of frequency samples. The simulation results show that the proposed method reduces both the bias of estimation and the resolution threshold signal-to-noise ratio (SNR).**

I. Introduction

Array processing techniques can be used to locate wideband signals. A wideband signal has a bandwidth comparable to the center frequency. Several methods for the processing of wideband signals using an array of sensors have been proposed in the literature [11], [1], [6]. The first step in some of these techniques is to obtain samples of the signal in the frequency domain. These samples are found by applying a discrete Fourier transformation to the time samples of the signal or by using a filter bank. The samples of the spectrum can be uniformly or nonuniformly spaced.

Many array processing techniques use the concept of the *signal subspace*. The signal subspace is the span of the location vectors of the array for fixed directions-of-arrival (DOA's). Since each location vector is a function of frequency, the signal subspace depends on the frequency of the observation. For wideband signals, the signal subspaces at different frequencies do not overlap, and as a result, the observation vectors at the frequency bins cannot be directly added to each other. Wang and Kaveh [11] propose *focusing* of the observation vectors. Focusing involves transforming the signal subspaces at different frequency bins into a predefined subspace (called the *focusing subspace*). They choose an arbitrary frequency, say, the center frequency of the spectrum of the signals, and transform the subspace at each frequency bin into the subspace created by the span of the location vectors at the focusing frequency. Then, they use a high-resolution algorithm such as MUSIC [8] to estimate the DOA's of the sources. They show that focusing reduces the resolution threshold signal-to-noise ratio (SNR), which is defined as the SNR for a prescribed probability of resolution. They also show that if the integral of the signal covariance matrix taken over the frequency spectrum is full rank, the method can be applied to coherent signal localization. Hung and Kaveh [5] use a unitary variant of the CSM algorithm to avoid the focusing loss. They use the center frequency for focusing.

Swingler and Krolik [9] prove that for a single-source scenario, it is possible to have an unbiased estimate of the DOA's if the centroid of the source spectrum is selected as the focusing frequency. In [10], we showed that for multiple sources, the CSM algorithm cannot provide unbiased estimates of the DOA's. In this work, we propose a method to select the focusing subspace. The method is based on minimizing a subspace fitting error. The subspace fitting error for each frequency bin is defined as the distance between the focusing matrix and the transformed location matrix. Later, we minimize a tight bound to the error. The simulation results show that using the method proposed here reduces the resolution threshold SNR and the bias of the DOA's estimates.

Manuscript received December 7, 1993; revised August 9, 1995. This work was supported in part by a grant from Natural Sciences and Engineering Research Council of Canada (NSERC). The associate editor coordinating the review of this paper and approving it for publication was Prof. Fu Li.

S. Valaee is with INRS-Telécommunications, Université du Québec, Verdun, Canada H3E 1H6.
P. Kabal is with the Department of Electrical Engineering, McGill University, Montreal, Canada H3A 2A7.

Publisher Item Identifier S 1053-587X(96)02588-1.

II. The CSM Algorithm

Consider an array of p sensors exposed to $q < p$ far-field wideband sources. The output of the sensors in the frequency domain is represented by

$$\mathbf{x}(f) = \mathbf{A}(f, \boldsymbol{\theta})\mathbf{s}(f) + \mathbf{n}(f) \tag{1}$$

where $\mathbf{s}(f)$ and $\mathbf{n}(f)$ are the Fourier transforms of the signal and noise vectors, and $\mathbf{A}(f, \boldsymbol{\theta}) = [\mathbf{a}(f, \theta_1) \cdots \mathbf{a}(f, \theta_q)]$ is the full-rank $p \times q$ matrix of location vectors. It is assumed that the signal and noise samples are independent identically-distributed sequences of complex Gaussian random vectors with unknown covariance matrices $\mathbf{S}(f)$ and $\sigma^2\mathbf{I}$, respectively. With these assumptions, the covariance matrix of the observation vector at the frequency f_j is given by

$$\mathbf{R}(f_j) = \mathbf{A}(f_j, \boldsymbol{\theta})\mathbf{S}(f_j)\mathbf{A}^H(f_j, \boldsymbol{\theta}) + \sigma^2\mathbf{I} \tag{2}$$

where the superscript H represents the Hermitian transposition. For simplicity of notation, we suppress the frequency variable and represent $\mathbf{R}(f_j)$ by $\mathbf{R}_j$, $\mathbf{A}(f_j, \boldsymbol{\theta})$ by $\mathbf{A}_j$, and so on.

The CSM algorithm [11] is based on forming new observation vectors $\mathbf{y}_j$ such that

$$\mathbf{y}_j = \mathbf{T}_j\mathbf{x}_j \tag{3}$$

where the $\mathbf{T}_j$'s are called the *focusing matrices*. In the unitary variant of the CSM algorithm [5], the $\mathbf{T}_j$, $j = 1, \ldots, J$ are selected from

$$\min_{\mathbf{T}_j} \|\mathbf{A}_0 - \mathbf{T}_j\mathbf{A}_j\|, \quad \text{subject to } \mathbf{T}_j^H\mathbf{T}_j = \mathbf{I} \tag{4}$$

where $\|\cdot\|$ is the Frobenius matrix norm [4]. The solution to this minimization is given by [4], [5]

$$\mathbf{T}_j = \mathbf{V}_j\mathbf{W}_j^H \tag{5}$$

where $\mathbf{V}_j$ and $\mathbf{W}_j$ are the left and the right singular vectors of $\mathbf{A}_0\mathbf{A}_j^H$. In (5), $\mathbf{A}_j$ and $\mathbf{A}_0$ are assumed to be known. In practice,