Rejoinder

E. J. Hannan

A good deal of the discussion of the paper relates to the use of criteria such as AIC or BIC. The latter derives, via Gaussian assumptions, from (4.2). Of course the Gaussian assumptions are not necessary, as Jorma Rissanen points out. However it is not easy to prescribe a probability law for a stochastic process. That could be done by taking the innovations as independent but that also is a fiction. In addition, the statistical analysis often has a special purpose. The result of these two difficulties is that methods that are rather *ad hoc*, but still general and effective, will always be important. Fourier methods are an example of this, to some extent, as also is the technique, involving the use of the third order spectrum, in David Brillinger's comments. Raj Bhansali also deals with a special problem, namely s step ahead prediction, $s > 1$, which could not easily be treated via (4.2). Of course one cannot deal with very special techniques in a general survey, even if one had the wit to think of them.

Ritei Shibata and I are in agreement, I think, about the above and the relation of the purpose of the analysis to the method used. I cannot quite see his objection to Rissanen's encoding argument which, in a sense, treats data and parameters in the same way, since Ritei Shibata favors a Bayesian argument that does much the same. Rissanen's argument does not require finiteness of the true order. Indeed the notion of true order is rejected. Rissanen would use a prior for the autoregressive order, for example allotting $c2^{-\log^* h}$ to order h, $\log^* h = \log h + \log\log h + \cdots$, up to the last positive term. Of course this series converges but very slowly and it is not asserted that the truth lies in the model set. The results Ritei Shibata quotes about order of consistency relate to autoregressions. It is not clear to me that the boot cannot be on the other foot for ARMA model fitting. After all if there was a true finite order (or something very near to that) some overfitting that AIC might induce could result in false, nearly matching, poles and zeros. These could be troublesome if, say, pole placement was the end purpose of the statistical analysis.

It must be agreed that the structure theory in Section 2 of the paper has been little used in statistical practice. One reason for this may be a lack of familiarity with the theory and this the paper, partly, sought to redress. Another reason would be a lack of ready access to algorithms and programs. Once the dimension, n, of the output is increased the "curse of dimensionality" has its effects, even for an AR. Of course determining Kronecker indices allows the dimension of the parameter space to vary over a fine grid of integer values. For example for $n = 3$ all dimensions occur except 1, 2, 3, 7. Of course there is an arbitrary quality about Kronecker indices. One way to exorcise the curse of dimensionality is to use special knowledge about the elements of A, B, C in (2.3) so that only for $S = I$ will the change of basis, $x(t) \to Sx(t)$, in the state space leave A, B, C in the special form. Such prior knowledge may often be available as David Brillinger points out. However, there will be cases when prior knowledge is too vague for this. A related phenomenon to the use of prior constraints is that for $n = 1$ the systems are listed with $p \equiv q$. One may feel that $q \ll p$ will do. A way to handle this is to find an estimate, $\hat{d}$, of $d = \max(p, q)$ and then to examine, using AIC or BIC, pairs $(\hat{d}, q)$ $q < \hat{d}$ (or $(p, \hat{d})$, $p < \hat{d}$, for that matter). If T is large, when $\hat{d}$ will also be relatively big, this will be simpler than looking at (p, q), $p \le \hat{d}$, $q \le \hat{d}$. David Brillinger wants a heavier penalty on large q. One should perhaps be careful not to allow the investigator too much leeway to indulge his prejudices. (Referring to a related phenomenon, there is some evidence that careless use of rules for rejecting outliers has led to errors.) The same kind of objection can be raised to a proliferation of criteria, of which Rainer Dahlhaus introduces another in his (1). This criterion has appeal, as does also $m_h(T)$ in Section 5 of the paper. However, in relation to a special purpose both might do badly. (See the discussion below in $m_h(T)$ in Section 5.) AIC, BIC have the virtue that they (or their generalizations such as (4.2)) have a sound general principle behind them.

The idea of data reduction as a central statistical aim is an old one and underlies Rissanen's theory. It is not to be accepted uncritically but it also should not be rejected out of hand because it differs from received statistical theory. The consistency results in Section 5 are only of suggestive value, as are all theorems since reality is so complex. Such results are useful also in the development of further theory, albeit only of suggestive value. The same kind of theory, in any case, leads to (5.10) which is not constrained in its applications to a case of a true ARMA model.

I do not agree with Rainer Dahlhaus' statement about $\Phi(j)$ and $\Phi_h(j)$. Two situations can be contrasted. One is the fitting of an ARMA model of fixed order to nonARMA data. Then, θ being the parameter vector, one may show that $\hat{\theta} \to \Theta_0$, in the sense that any subsequence has a sub-subsequence converging to

a point of Θ_0. Here Θ_0 is essentially the set of points at which the supremum is attained of the limit, as $T \to \infty$, of the likelihood. The contrasting case is that in Section 5 where h may increase indefinitely as $T \to \infty$ (and quite fast since the result Rainer Dahlhaus refers to is uniform in $h \leq c(T/\log T)^{1/2}$). In any case if $\Phi_h(j)$ replaces $\Phi(j)$ the same result holds, without the c that occurs on the right side.

There are a number of other suggestions with which one can hardly disagree, such as prewhitening, as suggested by David Brillinger. It seems a good general principle that any trick that helps in spectral estimation will help also with rational transfer function estimation, but that idea seems to have taken some time to dawn upon us, or perhaps only on me. Rainer Dahlhaus' emphasis on tapering is in the same vein. Of course zeros near the unit circle are somewhat more troublesome computationally than poles but such poles can cause large biases.

I agree with Victor Solo that the canonical correlation technique in Section 6 is likely to be very inefficient. Its virtue is that it gives first estimates of the Kronecker indices and the system parameters. The italicized statement below (6.2) which Victor Solo queried arises from the following. In the ARMAX case we have two transfer functions, $k(z)$, $l(z)$, the latter being the transfer function from observed inputs, $z(t)$, to outputs. (See (1.1′).) The corresponding Hankel matrix has blocks $[K(j), L(j)]$, where $L(j)$ is the coefficient matrix of z^j in $l(z)$. All of the structure theory in Section 2 goes through, much as before. However, usually we may not say $E\{z(s)z(t)'\} = 0$, $s \neq t$. Thus one cannot uncritically generalize the canonical correlation theory in Section 6 so as to find Kronecker indices. This can be seen from the analogue of (2.2) for which the residual vector will now contain $z(t+1)$, $z(t+2)$, $\cdots$, and will not be orthogonal to the generalization of y_t, which will contain $z(t)$, $z(t-1)$ etc. Of course this does not mean that the use of canonical correlation theory is impossible.

I am interested in Victor Solo's comments about ARMA models in relation to a sinusoid plus noise, $\varepsilon(t)$. The idea is old of course and relates to Prony's method. If $p = q = 2$, for a simple sinusoid at frequency ω_0 then

$$y(t) - 2\cos\omega_0 \cdot y(t-1) + y(t-2) = \varepsilon(t) - 2\cos\omega_0 \cdot \varepsilon(t-1) + \varepsilon(t-2).$$

Fitting this on the basis of Gaussianity one is, to the first order, minimizing

$$\int_{-\pi}^{\pi} I(\omega) \left| \frac{1 + \alpha(1)e^{i\omega} + \alpha(2)e^{i2\omega}}{1 + \beta(1)e^{i\omega} + \beta(2)e^{i2\omega}} \right|^2 d\omega,$$

$$I(\omega) = \frac{1}{T}\left|\sum_1^T y(t)e^{it\omega}\right|^2.$$

As $T \to \infty$ we shall have $(\hat\alpha(1), \hat\alpha(2)) - (\hat\beta(1), \hat\beta(2)) \to 0$ but also each converging to values producing a zero, in the factor multiplying $I(\omega)$ in the integrand, at ω_0. thus $\hat\alpha(1), \hat\beta(1) \to -2\cos\omega_0$, $\hat\alpha(2), \hat\beta(2) \to 1$. However, $\hat\alpha(1), \hat\alpha(2)$ will converge faster so that a zero develops. The interesting observation is that the $\hat\alpha(j)$ are in error by $O(T^{-3/2})$ for then ω_0 may be estimated to that accuracy. It is well known that the $\hat\omega$ that locates the maximum of $I(\omega)$ has this accuracy but the use of the ARMA model lends itself to real time calculation and hence to the estimation of the changing frequency of a frequency-modulated signal. An accurate asymptotic analysis would be valuable.

Finally I come to a number of comments associated with Section 3. I agree fully with what Jorma Rissanen says. Moreover it really is not easy to imagine cases where the high order dynamic system does not contain elements depending on statistical estimates. The determination of some kind of Hankel norm approximation seems usually to depend on recovering a function analytic in the unit disc from its phase function on the circle. I believe that this relation is rather unstable so that errors in the original high order system could propagate wildly. The approach via Hankel norm approximation does not seem to me to fit in with the data analytic approach in the remainder of the paper or, at least, I can't see how to fit it in. For this reason I cannot see how to incorporate the approach via (1) in Rainer Dahlhaus' comments with the Hankel norm approach. There could only be the most tenuous relation between relative errors in singular values and transfer functions.

I incorporated the theory in Section 3 because it is about rational transfer function approximation and because it is deep. The classical Wiener-Kolmogoroff prediction theory is not fully relevant to practice because it uses the infinite past. However it brings an understanding that is central to the statistical theory of stationary time series. The same may turn out to be true of the kind of theory in Section 3. Of course that theory has associations with Padé approximations and may have begun from that. However I know too little about this to make any useful comment.