Coherent Forecast with Nonlinear Econometric Models

Calzolari, Giorgio and Panattoni, Lorenzo
IBM Scientific Center, Pisa, Italy.

12. June 1988

Coherent Forecast with Nonlinear Econometric Models

by **Giorgio CALZOLARI** *and* **Lorenzo PANATTONI**

Centro Scientifico IBM
via S.Maria 67

56100 Pisa
☎ *050-47383*

The drawbacks of forecasts obtained with the usual deterministic solution methods in nonlinear systems of stochastic equations have been widely investigated in the literature. Most of the proposed therapies are based on some estimation of the conditional mean of the endogenous variables in the forecast period. This however provides a solution to the problem which does not respect the internal coherency of the model, and in particular does not satisfy nonlinear identities. At the same time, for analogy with univariate skewed distributions, the conditional mean may be expected to lie on the *wrong* side of the deterministic solution, meaning that it moves towards values of the variables where the probability density is lower, rather than towards the most probable values.

In a previous study, Brillet, Calzolari and Panattoni (1986) proposed to estimate the *mode* of the joint distribution of the endogenous variables as an alternative optimal predictor. The method proposed in that paper maximized the joint density of a subset of the endogenous variables, corresponding to stochastic equations only (analogously to FIML estimation where, at least conceptually, identities are first substituted into stochastic equations, and then the likelihood is defined and maximized). Experimental results were given in that paper for a large scale macroeconomic model of the French economy developed at INSEE. After suggestions and comments received from F.J. Henk Don, Steven G. Hall, Frederic P. Sterbenz and Kenneth F. Wallis, we tackle here the problem from a more general point of view. Mirella Damiani supplied the data and the model experimented with at the end of the paper. Giorgio Letta suggested us the mathematical framework that underlines the new method. We are greatly indebted to all of them, but retain full responsibility for any errors. A full paper is in preparation, and will be presented at the *1988 European Meeting of the Econometric Society* (Bologna).

The more general approach here proposed is designed to maintain the identities. The model with identities is viewed as a mapping between the space of the random errors and an hypersurface in the higher dimensional space of the endogenous variables; maximization is performed on such a hypersurface. Experimental results on the two alternative *mode predictors* are provided for a macro model of the Italian economy.

Let the simultaneous equation model be represented as

$$f(y_t, x_t, a) = \begin{bmatrix} u_t \\ 0 \end{bmatrix} \qquad t = 1,2,\ldots,T \tag{1}$$

where y_t is the $M \times 1$ vector of endogenous variables at time t, x_t is the vector of predetermined variables at time t, a is the vector of all unknown structural coefficients in the model, and f is the $M \times 1$ vector of structural form operators. The model is supposed to contain $m \leq M$ stochastic equations and $M - m$ identities. The $m \times 1$ vector of random error terms at time t,

$$u_t = \begin{bmatrix} u_{1,t} \\ u_{2,t} \\ \vdots \\ u_{m,t} \end{bmatrix} \tag{2}$$

is assumed to be independently and identically distributed as $N(0, \Sigma)$ with the $m \times m$ covariance matrix Σ completely unknown, apart from being symmetric and positive definite. The vector u_t is followed by a vector of $M - m$ zeroes in the structural form equations (1). We also decompose the vector y_t in two subvectors of length m and $M - m$, respectively

(3) $$y_t = \begin{bmatrix} y_t^{(1)} \\ y_t^{(2)} \end{bmatrix} = \begin{bmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{m,t} \\ y_{m+1,t} \\ \vdots \\ y_{M,t} \end{bmatrix}.$$

There is a certain freedom in ordering the endogenous variables inside the subvectors $y_t^{(1)}$ and $y_t^{(2)}$. A choice that is rather obvious, but not necessarily unique, is to put into $y_{1,t}$ the variable which is *explained* by the structural equation whose error term is $u_{1,t}$, put into $y_{2,t}$ the variable which is *explained* by the structural equation whose error term is $u_{2,t}$, and so on, till $y_{m,t}$ and $u_{m,t}$. We have filled in this way the subvector $y_t^{(1)}$. The subvector $y_t^{(2)}$ is filled in some way from the remaining $M - m$ endogenous variables.

It is usually assumed that a simultaneous equations system like (1) uniquely defines the values of the elements of y_t once values for the coefficients, the predetermined variables, and the disturbance terms are given. This means that the structural form equations (1) implicitly define a system of reduced form equations

(4) $$y_t = g(x_t, a, u_t)$$

Since all predictors are *conditional* on model's parameters and predetermined variables, we shall often indicate the reduced form simply as

(5) $$y_t = g(u_t)$$

The way the problem is tackled here reminds the *full information maximum likelihood* estimation of simultaneous equations systems (e.g. Amemiya, 1983). Since the model (1) includes $M - m$ identities, we first substitute these identities into the m stochastic equations. The notations could be made more accurate by dividing the vector f in two subvectors of functions, and considering more carefully the mappings implied by them. A more precise notation, however, is not strictly necessary in what follows, and is not introduced for the sake of simplicity. The resulting model may be written as

(6) $$p(y_t^{(1)}, x_t, a) = u_t \qquad t = 1,2,\ldots,T$$

where the $m \times 1$ vector of functions p is obtained from the vector of functions f (eq.1) after substitutions.

We assume that, given the coefficients and the predetermined variables, the functions vector p such that $u_t = p(y_t^{(1)})$ is a continuous and differentiable *one to one* mapping from a subset of R^m *onto* the whole R^m (u_t is multivariate normal, and therefore it spans the whole R^m space, while $y_t^{(1)}$ may be restricted to a subspace: for example, some of its elements may not assume negative values) and that the inverse function is also continuous and differentiable.

The joint density of the elements of $y_t^{(1)}$ can thus be obtained, as usual, from the density of u_t and the Jacobian determinant. Apart from an additive constant, the log-density of the m - dimensional random vector $y_t^{(1)}$ is

(7) $$L_t = -\frac{1}{2}\log|\Sigma| + \log\left|\left/\frac{\partial p_t}{\partial y_t^{(1)\prime}}\right/\right| - \frac{1}{2}p_t'\Sigma^{-1}p_t$$

where the vertical bar indicates *absolute value*, while the slash indicates *determinant*.

To obtain the *mode predictor* at time h, we must first compute the subvector $y_h^{(1)}$ that maximizes (7), given the predetermined variables x_h, the coefficients a and the covariance matrix Σ (for a and Σ, of course, we shall use the available estimates $\hat{a}$ and $\hat{\Sigma}$). Given $y_h^{(1)}$, we must then solve the subsystem of identities, obtaining $y_h^{(2)}$, and thus the complete $M \times 1$ predictor for all the endogenous variables.

It is well known from the theory of maximum likelihood estimation that the boring operation of substituting the identities can be bypassed (see, for example, Rothenberg and Leenders, 1964, pp.71-72). In fact, the same values (and therefore the same maximum) of L_t can be obtained if we partition the $M \times M$ Jacobian corresponding to all equations into 4 blocks

(8) $$J = \frac{\partial f_h}{\partial y_h'} = \begin{bmatrix} J_{1,1} & J_{1,2} \\ J_{2,1} & J_{2,2} \end{bmatrix}$$

(the 1,1 block corresponds to stochastic equations) and instead of $|\partial p_h/\partial y_h^{(1)\prime}|$ we use in (7) the ratio between the determinant of the $M \times M$ Jacobian matrix of the complete system, and the determinant of the $(M - m) \times (M - m)$ submatrix corresponding to identities

(9) $$L_t = -\frac{1}{2}\log|\Sigma| + \log\left|\frac{|J|}{|J_{2,2}|}\right| - \frac{1}{2}p_t'\Sigma^{-1}p_t$$

To simplify notations, in what follows the *time subscripts* t and h will sometimes be suppressed, when this does not create confusion.

For an easier comparison with the approach of next section, we modify the expression of the density given in equation (7). First, rather than viewing L as a function of $y^{(1)}$, we view it as a function of u. To do this, we simply substitute u to p in the last term, and consider the Jacobian matrix of first derivatives of the elements of $y^{(1)}$ with respect to the elements of u (which is simply the inverse of the matrix $\partial p/\partial y^{(1)\prime}$ used above)

$$(10) \qquad -\frac{1}{2}\log|\Sigma| - \log\left|\left|\frac{\partial y^{(1)}}{\partial u'}\right|\right| - \frac{1}{2}u'\Sigma^{-1}u$$

Then we simply substitute to the absolute value of the Jacobian determinant the square root of the determinant of the product between the transpose of the Jacobian matrix and the matrix itself

$$(11) \qquad -\frac{1}{2}\log|\Sigma| - \log\left|\frac{\partial y^{(1)\prime}}{\partial u}\,\frac{\partial y^{(1)}}{\partial u'}\right|^{\frac{1}{2}} - \frac{1}{2}u'\Sigma^{-1}u$$

Maximizing (11) in the forecast period, h

$$(12) \qquad \underset{u\in R^m}{Max}\left[-\frac{1}{2}\log|\Sigma| - \log\left|\frac{\partial y^{(1)\prime}}{\partial u}\,\frac{\partial y^{(1)}}{\partial u'}\right|^{\frac{1}{2}} - \frac{1}{2}u'\Sigma^{-1}u\right]$$

provides a value for the $m \times 1$ vector of random errors u_h. Inserting this vector into the model (1) and solving the model at time h (that is, at least conceptually, the same as inserting it into the reduced form equations 4 or 5), we obviously get the same $M \times 1$ predictor y_h discussed above.

The values of the endogenous variables computed in this way provide only a partial answer to our problem. In fact, the value computed for the first subvector ($y_h^{(1)}$) is the mode of the joint distribution of the elements of this subvector, but the second subvector ($y_h^{(2)}$) is simply calculated to preserve the internal coherency of the model (the whole vector y_h is, in fact, a solution of the system). Suppose now that we interpret the *mode predictor* as the *most likely joint value* of the endogenous variables in the forecast period. Then only $y_h^{(1)}$ can be interpreted in this way, but the whole vector y_h is *not* the *most likely joint value* of *all* the endogenous variables.

In order to maximize the joint density function of *all* the endogenous variables of the model, we must first consider the type of mapping implied by the model, then some concepts on the measure of a regular (hyper)surface, and finally introduce a suitable definition of the probability density, with respect to this measure. To clarify the problem, let us consider the simplest cases.

Suppose that our model consists of two equations, the former is stochastic while the latter is an identity. Given coefficients and predetermined variables, solving the model we get the value of the *two* endogenous variables for *one* value of the error term. If the solution is unique and the error term is normally distributed, this gives a mapping from R^1 *into* R^2. The reverse is not true; we can, in fact, enter $y_{1,t}$ and $y_{2,t}$ into the first equation and compute $u_{1,t}$ as a *residual* (and this will certainly be unique), but we cannot take two arbitrary values for the endogenous variables, since they are constrained by the second equation (a point that is well known to the model builders: coherent historical data *must* satisfy all the identities). The values of $y_{1,t}$ and $y_{2,t}$ must lie on a curve in the two-dimensional plane. Therefore our model can be viewed as a mapping from R^1 *onto* a one-dimensional subset of R^2, that is a curve in the plane.

Suppose that our model consists of three equations, two of which are stochastic. If for any value of the random errors $u_{1,t}$ and $u_{2,t}$ (in R^2) the model provides a unique solution for $y_{1,t}$ $y_{2,t}$ and $y_{3,t}$, the values of the y's will neither span the full R^3 space, nor any three-dimensional subset of R^3, but will lie on a surface. The model, therefore, can be regarded as a mapping from R^2 *onto* a two-dimensional surface in R^3.

In the general case of M equations, m of which are stochastic, the model (in the forecast period, given predetermined variables and parameters) can be viewed as a mapping from R^m *onto* an m-dimensional hypersurface in the R^M space (being $m \leq M$). We call Y this hypersurface. Under the two assumptions

1) for any u in R^m the solution y is unique;

2) for any y in Y we get a unique vector u;

then we are dealing with a *one to one* mapping between R^m and the m-dimensional hypersurface Y in R^M. As in equation (5), we indicate this mapping as $g: R^m \rightarrow Y$ (the reduced form) and its inverse as $g^{-1}: Y \rightarrow R^m$ (practically the first $m \times 1$ subvector of the structural form functions vector f).

The first assumption is somewhat restrictive, but widely adopted and accepted in the treatment of nonlinear systems. The second, on the contrary, is not restrictive at all, given the way in which model builders write the structural form equations of a model like (1). In fact, if a vector of endogenous variables belongs to Y, it certainly satisfies the constraint of the $M - m$ identities. In such a case, if we plug this vector into the m stochastic equations, we get a unique vector of *residuals*.

Having defined the nature of this *one to one* mapping, we now assume, as in the previous section, continuity and differentiability (in both directions), and finally assume that the $M \times m$ Jacobian matrix $G(u) = \partial y/\partial u' = \partial g(u)/\partial u'$ has full rank ($= m$).

In a $M \times m$ rectangular matrix, a substitute for the notion of determinant of a square matrix is the concept of *modulus* (it will be indicated with vertical bars). It is defined as the square root of the sum of all the $\binom{M}{m}$ squared determinants of order m obtained by cancelling $M - m$ rows. A simple way of computing the modulus, even when M and m are large, is obtained applying a well known theorem of linear algebra on the determinant of the product of two rectangular matrices; it is the square root of the determinant of the product between the $m \times M$ transpose and the matrix itself. For our Jacobian, $G(u)$, it is

$$|G(u)| = \left|\frac{\partial y}{\partial u'}\right| = \left|\frac{\partial g(u)}{\partial u'}\right| = \left/\left[\frac{\partial y'}{\partial u}\ \frac{\partial y}{\partial u'}\right]\right/^{\frac{1}{2}} \tag{13}$$

Having assumed full rank ($= m$) for the Jacobian, the modulus is $\neq 0$. The m - dimensional hypersurface Y defined by the model is, under all these assumptions, a *regular hypersurface.*

We need now to define a measure on the regular m - dimensional hypersurface Y. As well known, the M -dimensional Lebesgue measure of any m - dimensional hypersurface in $\boldsymbol{R}^M$ ($m < M$) is zero (e.g. Sikorski, 1969, p.299). This is consistent with intuition: the two-dimensional measure (area) of a curve in the plane is zero; the three-dimensional measure (volume) of a curve or of a surface in $\boldsymbol{R}^3$ is also zero, etc. But we can define a special one-dimensional measure on curves, consistent with the intuitive notion of length; or a special two-dimensional measure on surfaces, consistent with the intuitive notion of area. Existence and uniqueness of such a measure, for a regular m -dimensional hypersurface Y in $\boldsymbol{R}^M$ ($m \leq M$) is stated, for example, in Sikorski (1969, p.327, theorem 1.1), and is given by

$$d\mu_Y = |G(u)|\, du = \left|\frac{\partial y}{\partial u'}\right| du = \left|\frac{\partial g(u)}{\partial u'}\right| du \tag{14}$$

This theorem can be applied to derive a probability density with respect to the hypersurface measure μ_Y. If $\varphi(u)$ is the probability density for u in $\boldsymbol{R}^m$, the measure of the probability that y belongs to an element of the hypersurface Y, resulting (through $y = g(u)$) from the element du of $\boldsymbol{R}^m$, is given by

$$d\pi_Y = \varphi(u)\, du \tag{15}$$

and can easily be developed as follows

$$d\pi_Y = \varphi(u)\, du = \frac{\varphi(u)}{|G(u)|}\,|G(u)|\, du \tag{16}$$

$$= \frac{\varphi\left\{g^{-1}[y]\right\}}{\left|G\left\{g^{-1}[y]\right\}\right|}\, d\mu_Y = \psi_Y(y)\, d\mu_Y$$

where

$$\psi_Y(y) = \frac{\varphi\left\{g^{-1}[y]\right\}}{\left|G\left\{g^{-1}[y]\right\}\right|} \tag{17}$$

is taken as the definition of the probability density of the vector y on the regular m -dimensional hypersurface Y. It is rather obvious that $\psi_Y(y)$ respects the usual conditions for density functions, being always ≥ 0 and being equal 1 its integral, over the whole Y, with respect to the m -dimensional hypersurface measure μ_Y .

We now indicate with L_Y the log-density

$$L_Y = \log\left[\psi_Y(y)\right] \tag{18}$$

that will obviously be $L_{Y,t}$ for the values of y at time t, or $L_{Y,h}$ in the forecast period, h . Being φ the density of the m -variate normal $N(0, \Sigma)$, we have, apart from a constant term

$$L_Y = -\frac{1}{2}\log|\Sigma| - \log\left|G\left\{g^{-1}[y]\right\}\right| - \frac{1}{2}\left[g^{-1}(y)\right]'\Sigma^{-1}\left[g^{-1}(y)\right] \tag{19}$$

L_Y in equation (19) is *conceptually* defined as a function of the $M \times 1$ vector y, whose domain is the m - dimensional hypersurface Y in $\mathbb{R}^M$. The $M \times 1$ vector y, that maximizes L_Y in Y (that is, y is *constrained* on Y), is the *mode predictor*

$$\underset{y \in Y}{Max} \left\{ -\frac{1}{2} \log |\Sigma| - \log \left| G\left\{g^{-1}[y]\right\} \right| - \frac{1}{2} \left[g^{-1}(y)\right]' \Sigma^{-1} \left[g^{-1}(y)\right] \right\} \tag{20}$$

If we interpret it as the *most likely joint value* of the endogenous variables in the forecast period, such an interpretation applies to *all* the endogenous variables, unlike the result derived in the previous section. Numerical values of L_Y can easily be calculated by replacing $g^{-1}(y)$ with u

$$-\frac{1}{2} \log |\Sigma| - \log |G(u)| - \frac{1}{2} u' \Sigma^{-1} u \tag{21}$$

Equation (21) provides a function of the $m \times 1$ vector u, that must be maximized for u *without constraints* in $\mathbb{R}^m$

$$\underset{u \in R^m}{Max} \left[-\frac{1}{2} \log |\Sigma| - \log |G(u)| - \frac{1}{2} u' \Sigma^{-1} u \right] \tag{22}$$

in the forecast period, h. Inserting the resulting vector u_h into the model (1), then solving the model in the forecast period, we obtain the $M \times 1$ *mode predictor* y_h.

From a technical point of view, the process is quite similar to what we did in for equation (12). The only difference is that in equation (12) we have the modulus of the $m \times m$ square matrix $\partial y^{(1)}/\partial u'$, while here we have the modulus of the $M \times m$ rectangular matrix $\partial y/\partial u'$.

The two functions to be maximized for the two methods described above are given in equations (11) and (21), respectively. Maximization is to be done with respect to the vector u_h, while predetermined variables are given, as well as coefficients and covariance matrix of the random error process (set at their estimated values $\hat{a}$ and $\hat{\Sigma}$). The first term in both equations is a constant. We must therefore compute the vector u_h which maximizes the sum of the second and third term of equations (11) and (21). The third term in both equations is particularly simple, being a positive definite quadratic form, whose Hessian is the matrix Σ^{-1}.

We have used the well known updating formula due to Broyden, Fletcher, Goldfarb and Shanno (BFGS, see for example Dennis and More', 1977). The algorithm is based on an iterative updating of an initial ($m \times m$) positive definite matrix. Since the computational efficiency is greatly improved if the initial matrix approximates the Hessian, it was rather obvious in both cases to use the available estimate of Σ^{-1}.

The algorithm also requires the evaluation at each step of the gradient of the function. This revealed to be a rather serious problem and at this stage it has been solved by the numerical computation of first derivatives. This approach however has two main drawbacks. First of all it requires a long computation time, but nevertheless it came out to be computationally more efficient than rival maximizing algorithms which require only the computation of the function value. To quantify this aspect we can mention that in the case of the Italian model here considered (see below for description) the maximization of the quantities defined by equations (11) and (21), with a tolerance of ten significant digits on the function's value, in both cases required five iterations. The computation globally took about 14 minutes of CPU time on an IBM 4341 computer in both cases. However for the first method we can maximize the expression of the density given in equation (9), rather than (11); in this case the computation is much faster (only 50 seconds). This is mainly due to the fact that the elements of the Jacobian matrices in equations (11) and (21) are obtained from differences between two solutions of the simultaneous equation system, while the corresponding elements in equation (9) are simply computed from the differences between the residuals of the single equations: obviously the latter computation is much faster than the former. Unfortunately, we are unable to provide a similar simplification for the other method.

A second main drawback lies in a possible lack of accuracy in the computation of the derivatives. In order to assess the robustness of the results versus the way in which derivatives are computed, several formulas (i.e. two, three and five points formulas with different sizes of the increment) were experimented with. The results proved to be very robust versus both the choice of the formula and the choice of the numerical increment, provided that a centered formula is used.

For some of the main endogenous variables of the model (listed in table 1), table 2 presents in *per thousand* form the differences between the deterministic predictor and the two *mode predictors* previously defined (the two columns are labelled *M-mode* and *m-mode*, respectively).

$$\frac{y_{i,h}^{(det)} - y_{i,h}^{(mode)}}{y_{i,h}^{(det)}} \cdot 1000 \tag{23}$$

Comparisons with the *mean predictors* (not displayed here; results are in preparation) suggest that *mean* and *mode* are not necessarily on opposite sides with respect to the deterministic solution. Therefore, abandoning the deterministic forecast in favor of the (more efficient) *mean predictor*, we are not necessarily moving towards a region of less likely values.

Between the two *mode predictors*, the global approach *(M-mode)* produces values that are considerably more distant from the deterministic solution. The distance is also considerably larger than for the *mean predictors*.

Table 1

List of the main endogenous variables

PREMIN$I	=	Deflator for domestic private consumption (growth rate).
PREXP$I	=	Export price index (growth rate).
P2$I	=	Production per hour in the industrial sector (growth rate).
S$I	=	Wage per hour in the industrial sector (growth rate).
INVIDK$I	=	Gross fixed investment in the industrial sector (growth rate).
PR7	=	Profit rate.
PREING$I	=	Wholesale price index (growth rate).
VAINDK$I	=	Value added in the industrial sector at constant prices (growth rate).
DIST$I	=	Unemployment (growth rate).
OCCD$I	=	Total dependent employment (growth rate).
OCSRD$I	=	Employment in the sector of services (growth rate).
EXFOBK$I	=	Exports at constant prices (growth rate).
IMFOBK$I	=	Imports at constant prices (growth rate).
PILK$I	=	Gross domestic product at constant prices (growth rate).
COFAMK_I	=	Domestic private consumptions.

Table 2

One-period forecast at 1986
Per thousand deviations from deterministic forecasts

	M - mode	*m - mode*
PREMIN$I	-.3975	-.0445
PREXP$I	9.592	.6288
P2$I	-33.57	-4.008
S$I	1.271	-.0214
INVIDK$I	8.838	-1.402
PR7	-.8109	-.1591
PREING$I	2.931	.0883
VAINDK$I	-10.79	-2.864
DIST$I	54.41	2.169
OCCD$I	26.64	-.3355
OCSRD$I	1.788	-.0155
EXFOBK$I	1.308	.3365
IMFOBK$I	-15.72	-2.809
PILK$I	5.811	-1.721
COFAMK_I	11.47	-2.439

Amemiya,T. (1983): "Non-Linear Regression Models", in *Handbook of Econometrics*, ed. by Z.Griliches and M.D.Intriligator. Amsterdam: North-Holland, Vol.I, 333-389.

Brillet,J.L., G.Calzolari, and L.Panattoni (1986): "Coherent Optimal Prediction with Large Nonlinear Systems: An Example Based on a French Model". Paris: INSEE, Service des Programmes, discussion paper presented at the *European Meeting of the Econometric Society*, Budapest, September 1-5.

Damiani,M. (1987): "Uno Schema Macroeconomico di Simulazione per l'Economia Italiana". Universita' di Perugia, Facolta' di Scienze Politiche, mimeo.

Dennis,J.E., and J.J. More' (1977): "Quasi-Newton Methods, Motivation and Theory", *SIAM Review* 19, 46-89.

Rothenberg,T.J., and C.T.Leenders (1964): "Efficient Estimation of Simultaneous Equation Systems", *Econometrica* 32, 57-76.

Sikorski,R. (1969): *Advanced Calculus. Functions of Several Variables*. Warszawa: PWN - Polish Scientific Publishers.

Sylos Labini,P. (1987): "The Theory of Unemployment, too, is Historically Conditioned", *Banca Nazionale del Lavoro Quarterly Review* 163, 379-435.