4 Multivariate forecasting methods

Usually, multivariate forecasting methods rely on *models* in the statistical sense of the word, though there have been some attempts at generalizing extrapolation methods to the multivariate case. This does not necessarily imply that these methods rely on *models* in the economic sense of the word. Rather, one may classify multivariate methods with regard to whether they are *atheoretical*, such as time-series models, or *structural* or theory-based. Words can be confusing, however, as the so-called 'structural time-series methods' suggested by HARVEY are actually atheoretical and not theory-based. Truly structural forecasting models with an economic background will be left for the next unit.

4.1 Is multivariate better than univariate?

Multivariate methods are very important in economics and much less so in other applications of forecasting. In standard textbooks on time-series analysis, multivariate extensions are given a marginal position only. Empirical examples outside economics are rare. Exceptions are data sets with a predator-prey background, such as the notorious data on the population of the Canadian lynx and the snowshoe hare. In contrast, the multivariate view is central in economics, where single variables are traditionally viewed in the context of relationships to other variables. Contrary to other disciplines, economists may even reject the idea of univariate time-series modeling on grounds of the theoretical interdependence, which appears to be an exaggerated position.

In forecasting, and even in economics, multivariate models are not necessarily better than univariate ones. While multivariate models are convenient in modeling interesting interdependencies and achieve a better (not worse) fit within a given sample, it is often found that univariate methods outperform multivariate methods *out of sample*. Among others, one may name as possible reasons:

1. Multivariate models have more parameters than univariate ones. Every additional parameter is an *unknown* quantity and has to be estimated. This estimation brings in an additional source of error due to sampling variation.

2. The number of potential candidates for multivariate models exceeds its univariate counterpart. Model selection is therefore more complex and lengthier and more susceptible to errors, which then affect prediction.

3. It is difficult to generalize nonlinear procedures to the multivariate case. Generally, multivariate models must have a simpler structure than univariate ones, to overcome the additional complexity that is imposed by being multivariate. For example, while a researcher may use a nonlinear model for univariate data, she may refrain from using the multivariate counterpart or such a generalization may not have been developed. Then, multivariate models will miss the nonlinearities that are handled properly by the univariate models.

4. Outliers can have a more serious effect on multivariate than one univariate forecasts. Moreover, it is easier to spot and control outliers in the univariate context.

An additional complication is *conditional forecasting*, which means that a variable Y is predicted, while values for a different variable X are assumed over the prediction interval. If X is a policy variable, it may make sense to regard it as fixed and not to forecast it. Even then, true-life X may react to future values of Y, while that reaction is ignored in the conditional forecast. The arguments of 'feedback' by CHATFIELD and the so-called Lucas critique are closely related. Important contributions on the feedback problem, with some relation to forecasting, are due to GRANGER (Granger causality) and to ENGLE&HENDRY&RICHARD (weak and strong exogeneity). If X is merely another variable that is also predicted over the prediction interval, though maybe using a simple univariate extrapolation method, these forecasts may be wrong and there will be another source of error.

If the forecasting model is designed such that forecasts of one or more variables of type X are not generated, while one or more variables of type Y are modeled as being dependent on X, the system is called *open-loop*. If all variables are modeled as dependent on each other and on lags, the system is called *closed-loop*. Even closed-loop systems may allow for deterministic terms, such as constants, trend, or other variables that can be assumed as known without error at any future time point.

4.2 Static and dynamic forecasts

Even more important than for univariate models is a clear distinction of *in-sample model fitting* and *out-of-sample forecasting*. If $1, \ldots, T$ serves as the time range for a 'training set', a model is selected on the basis of these observations, the model parameters are then estimated from the very same set, and predictions are then generated for $2, \ldots, T$ according to the generation mechanism of the model, the 'prediction errors' will be mere *residuals*. If variables are predicted over the time range $T+1, \ldots, T+h$ without using any further information on the 'test set', the differences between observations (if available) and predicted values will be true out-of-sample prediction errors. The exercise yields a sequence of a one-step, a two-step, ..., and a h–step prediction, which are sometimes called a *dynamic forecast*. There are many intermediate cases.

In *static forecasts*, true observations are used for all lagged values of variables. This situation does *not* correspond to a forecaster who only uses information on the training set. Prediction errors from static forecasts are difficult to classify. One would expect that they tend to increase from $T+1$ to $T+h$ if the forecasting model misses an important part of the DGP, which results in the impression of a time-changing DGP. Otherwise, the errors can be regarded as one-step out-of-sample prediction errors. If observations are used for explanatory current variables in regression-type prediction models, the resulting error comes closer to a residual. The expression *ex-post prediction* is used for both variants. Sometimes it is also used for mere in-sample fitting and calculating residuals.

Other intermediate cases are obtained if, for example, a model is selected on the basis of the total sample $1, \ldots, T+h$, while parameters are estimated from $1, \ldots, T$ only. The resulting forecast is not really out-of-sample, and its prediction error tends to under-estimate the true errors. Such exercises are often presented in research work and are sometimes labeled incorrectly as out-of-sample prediction.

4.3 Cross correlations

An important exploratory tool for modeling multivariate time series is the *cross correlation function* (CCF). The CCF generalizes the ACF to the multivariate case. Thus, its main purpose is to find linear dynamic relationships in time series data that have been generated from stationary processes.

In analogy to the univariate case, a multivariate process X_t is called (covariance) stationary if

1. $\mathrm{E}\, X_t = \mu$, i.e. the mean is a time-constant n–vector;
2. $\mathrm{E}\,(X_t - \mu)(X_t - \mu)' = \Sigma$, i.e. the variance is a time-constant positive definite $n \times n$–matrix Σ;
3. $\mathrm{E}\,(X_t - \mu)(X_{t+k} - \mu)' = \Gamma(k)$, i.e. the covariances over time depend on the lag only, with non-symmetric $n \times n$–matrices $\Gamma(k)$.

The matrix-valued function of $k \in \mathbb{Z} : k \mapsto P(k)$, which assigns correlation matrices instead of covariance matrices, is then called the CCF. One may note that $p_{ij}(k) = p_{ji}(-k)$, such that also the CCF is informative for non-negative k only or, alternatively, for positive and negative k and $i \leq j$. For estimates of the CCF in computer programs, the latter version is usually preferred. Definitions vary across programs and authors, such that $P(k)$ may be equivalent to somebody else's $P(-k)$. This is important insofar as a predominance of, for example, negative values $p_{ij}(-k)$ over the positive values $p_{ij}(k)$ with regard to their size and significance is often taken as indicating a causal direction from the j–th variable to the i–th variable etc. This is a crude check, anyway, as Granger causality can be assessed reliably only from the CCF of pre-filtered values or from multivariate autoregressive or, preferably, moving-average representations.

Identification of vector autoregressive (VAR) or other multivariate models from the empirical CCF is difficult. Also the issue of significance bounds for the CCF values is tricky. Automatically generated bounds from computer packages only serve as rough guidelines. The main basic result is Bartlett's formula, which is given for the case of a bivariate process ($n = 2$) here:

$$\begin{aligned}\lim_{T\to\infty} T \operatorname{cov}\left(\hat{p}_{12}(h), \hat{p}_{12}(k)\right) = \sum_{j=-\infty}^{\infty} \Big[& p_{11}(j)\, p_{22}(j+k-h) + p_{12}(j+k)\, p_{21}(j-h) \\ & - p_{12}(h)\left\{p_{11}(j)\, p_{21}(j+k) + p_{22}(j)\, p_{21}(j-k)\right\} \\ & - p_{12}(k)\left\{p_{11}(j)\, p_{21}(j+h) + p_{22}(j)\, p_{21}(j-k)\right\} \\ & + p_{12}(h)\, p_{12}(k)\left\{\frac{1}{2}p_{11}^2(j) + p_{12}^2(j) + \frac{1}{2}p_{22}^2(j)\right\}\Big]\end{aligned}$$

This extremely complicated formula simplifies for low-order vector MA processes and, of course, for white noise. For example, if one of the two processes is

white noise and the true $p_{12}(h) = 0$ for all but finitely many h, then the asymptotic variance of $\hat{p}_{12}(h)$ will be 1 for these h. As for the ACF and correlograms, a very slow decay of the CCF indicates non-stationarity of the integrated or unit-root type. Even if integrated variables are indicated, taking first differences is not necessarily recommended. The problem of taking differences in a multivariate setting has led to the extended literature on *cointegration*.

Extensions of the usual cross-correlogram analysis are the cross-correlogram analysis of prewhitened variables and phase plots. If variables are individually prewhitened, the diagonal elements of the CCF matrix will be trivial, while the off-diagonal elements may be more representative of the true dynamic interaction among variables. For discrete-time variables, a *phase plot* refers to a scatter plot of a variable y_t and a lag, such as y_{t-1} or y_{t-j}. If y_t has been generated by a first-order autoregression $y_t = \phi y_{t-1} + \varepsilon_t$, points should show a straight line with the slope corresponding to ϕ. Such phase plots are particularly valuable if a nonlinear time-series relationship is suspected. With multiple time series, also phase plots of y_{it} versus $y_{j,t-k}$ can be considered, although the number of possible diagrams becomes large and the variety of diagrams can become confusing.

4.4 Single-equation models

Two types of so-called single-equation models can be considered for multivariate forecasting: regression models and transfer-function models. Both types are 'open-loop' models and model a dynamic relationship of an 'endogenous' variable that depends on one or several 'explanatory' variables. The methods are helpful in forecasting only if future values of the explanatory variables are known, if forecasting of explanatory variables is particularly easy, or at least if there is no dynamic feedback from the endogenous to the explanatory variables. In the latter case, the condition is similar to the requirement that there is *no Granger causality* running from the endogenous to the explanatory variables and that there is *Granger causality* from the explanatory to the endogenous variables. The concepts may not be entirely equivalent, as Granger causality often only refers to linear relationships and additional conditions may be imposed by the parameterization of interest. These issues have been investigated in the literature on *exogeneity* (ENGLE&HENDRY&RICHARD).

The linear *regression model* can be assumed as known. In a time-series setting, errors are often correlated over time, such that ordinary least squares

(OLS) may not be appropriate and some kind of generalized least squares (GLS) procedure should be considered. These GLS models assume two equations, with a time-series model for the errors from the 'means' equation, for example an ARMA model:

$$\begin{aligned} y_t &= \beta' x_t + u_t \\ \phi(B)\, u_t &= \theta(B)\, \varepsilon_t. \end{aligned}$$

The specification $\theta(B) = 1$ and $\phi(B) = 1 - \phi B$ yields the classical GLS model that is used in the Cochrane-Orcutt or Prais-Winsten estimation procedures. When these models are applied for forecasting, one does not only need predicted values for x but also for u. Because u_t is unobserved, it is usually replaced by in-sample residuals from the means equation and by out-of-sample predictions based on these residuals. The difference between true errors u and estimated residuals $\hat{u}$ brings in another source of uncertainty, adding to the sampling variation in the coefficient estimates $\hat{\beta}$, $\hat{\phi}$, and $\hat{\theta}$. In many cases, it is more advisable to replace the means equation by a dynamic version that contains lagged values of y_t and of some x_t as further explanatory variables, and whose errors are approximately white noise (*dynamic regression*). An efficient search for the optimal dynamic specification of such equations can be lengthy and may resemble the techniques used in transfer-function modeling. Thus, the two methods are not too different in spirit. It can be shown that static GLS models are a very restrictive version of dynamic regression models. Some authors recommend removing trends and seasonals from all variables or possible differencing before regression modeling, which however may be problematic and may entail a deterioration of predictive accuracy.

Transfer-function models are based on dynamic relationships among mean- and possibly trend-corrected variables y_t (output) and x_t (input) of the form

$$\delta(B)\, y_t = \omega(B)\, x_t + \theta(B)\, \varepsilon_t. \tag{1}$$

Three different lag polynomials must be determined, which can be difficult. A common suggestion is to start from an ARMA or ARIMA model for the input variable x_t and to apply the identified 'filter' to both input and output. It is then much easier to determine the remaining polynomial. Of special interest is the case that $\omega_j = 0$ for $j < d$. This indicates that x affects y with a *delay* and that, for any h–step forecast for y_t with $h \leq d$, forecasts for the input x_t are not required. In this case, x is also called a 'leading

indicator' for y. In analogy to the GLS regression model, efficient forecasting needs predictions for the unobserved error series, which are constructed by extrapolating in-sample residuals.

4.5 Vector autoregressions and VARMA models

These models usually treat all n variables in a vector variable y as 'endogenous'. For stationary multivariate processes, a convenient parameterization is the vector ARMA (VARMA) model

$$\Phi(B)\, y_t = \Theta(B)\, \varepsilon_t \tag{2}$$

with the multivariate white noise series ε_t, defined by $\mathrm{E}\left(\varepsilon_t \varepsilon'_{t-k}\right) = 0$ if $k \neq 0$. The model is stable if all roots of $\det(\Phi(z)) = 0$ are larger than one in absolute value. The model is invertible, i.e. y can be expressed as a convergent infinite-order vector autoregression, if all roots of $\det(\Theta(z)) = 0$ are also larger than one. Uniqueness requires the roots of $\det(\Theta(z)) = 0$ to be larger or equal to one. Unfortunately, this condition is necessary but not sufficient for a unique representation. This is one of the reasons why vector ARMA models are rarely used in practice.

For $\Theta(z) \equiv 1$, one has the vector AR or VAR model

$$\Phi(B)\, y_t = \varepsilon_t,$$

which is quite popular in economics. The simple VAR(1) model for the bivariate case $n = 2$ can be written as

$$y_t = \Phi y_{t-1} + \varepsilon_t,$$

with the 2×2–matrix

$$\Phi = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}.$$

Of special interest are cases such as $\phi_{12} = 0$. Then, the first variable y_1 does not depend on past values of the second variable, although it may be correlated with it and it may depend on its own past via ϕ_{11}. There is no Granger causality from y_2 to y_1. The model reduces to a transfer-function model. y_1 can be forecasted via the univariate model $\hat{y}_{1,t-1}(1) = \hat{\phi}_{11} y_{1,t-1}$ and generally as

$$\hat{y}_{1,t-1}(h) = \hat{\phi}_{11}^{h} y_{1,t-1}.$$

Building on this forecast, a forecast for y_2 can be constructed from the second equation. Similar remarks apply to the reverse case $\phi_{21} = 0$.

For the general VAR(1) case, h–step forecasts are obtained from

$$\hat{y}_{t-1}(h) = \hat{\Phi}^h y_{t-1},$$

where $\Phi^2 = \Phi\Phi$, $\Phi^3 = \Phi^2\Phi$ etc. Similarly, forecasts from VAR(p) models can be constructed in strict analogy to the univariate case. The forecasts require coefficient estimates and a lag order p, which are determined from the 'training' sample ending in $t-1$. While cross correlograms may be helpful in finding a good lag order p, many researchers choose it by minimizing multivariate information criteria. Estimation can then be conducted by OLS, which can be shown to be efficient in the absence of further coefficient restrictions.

VARX and VARMAX models are extensions of the VAR and VARMA framework, which allow for exogenous ('X') variables whose dynamics are not specified or whose dynamics at least does not depend on the modeled 'endogenous' variables y. For forecasting, the X variables require an extrapolation technique or assumptions on their future behavior.

4.6 Cointegration

Although VAR modeling traditionally assumes stationarity of all series, it is not generally recommended to difference non-stationary components individually, as such a step may destroy important dynamic information. The critical issue is *cointegration*. In short, when there is cointegration, cointegrated models should be used for forecasting. When there is no cointegration, series with an integrated appearance should be differenced.

Cointegration is said to exist in a vector time series y_t when some of the components are individually first-order integrated (I(1)) while the remaining components may be stationary, and there is a linear combination of components that can be expressed by a vector β such that $\beta' y$ is stationary (I(0)). Cointegration is non-trivial when β has non-zero entries at the integrated components. Apart from this *CI(1,1) cointegration*, which reduces integration order from one to zero, higher-order cointegration can be defined but its practical applicability is reduced and examples for applications are rare. Note that there are slightly different definitions of CI(1,1) cointegration in the literature but that the above definition is the most convenient for multivariate VAR modeling.

The issues at stake can be best motivated by considering a VAR system that possibly contains some integrated and some stationary components

$$\Phi(B)\, y_t = \varepsilon_t.$$

It is easily shown that such systems can always be transformed into an *error-correction form*

$$\Delta y_t = \Phi y_{t-1} + \Gamma_1 \Delta y_{t-1} + \ldots + \Gamma_{p-1} \Delta y_{t-p+1} + \varepsilon_t, \tag{3}$$

where the matrices Φ and $\Gamma_1, \ldots, \Gamma_{p-1}$ are one-one functions of the autoregressive coefficient matrices $\Phi_1, \ldots, \Phi_p$. Note that the lag order of the polynomial $\Gamma(z) = I - \Gamma_1 z - \ldots - \Gamma_{p-1} z^{p-1}$ is $p-1$, while the lag order of $\Phi(z)$ is p. The crucial matrix is Φ. If the rank of Φ is zero, there is no cointegration and the system is best handled in differences. If the rank of Φ is n, the system is stable and should be handled in the original form without differencing. If the rank is between 1 and $p-1$, there is cointegration. In this case, the rank of Φ yields the number of linearly independent cointegrating vectors in the system. Empirically, the rank of Φ is determined by a sequence of tests with the null hypothesis 'rank $\Phi \leq k$' and the alternative 'rank $\Phi > k$'. All details are found in JOHANSEN's monograph. This rank determination procedure has been integrated into many econometric software packages, such as RATS and EViews.

Once the rank has been determined, estimation of (3) is conducted by *reduced-rank regression*, a technique that was developed by T.W. ANDERSON. Usually, this estimation procedure has been implemented in econometric software together with the rank-determination sequence. Then, estimates $\hat{\Phi}$ and $\hat{\Gamma}_1, \ldots, \hat{\Gamma}_{p-1}$ can be substituted for the true matrices and h–step forecasts are retrieved easily.

Ignoring cointegration issues and naïvely differencing non-stationary components implicitly assumes the restriction $\Phi = 0$. If the true $\Phi \neq 0$, this does not only imply *misspecification* due to an incorrect restriction but this misspecification has particular consequences. The forecasts will not reflect the property that $\beta' y$ is stationary, in contrast with the sample. In economics, the usual interpretation is that 'equilibrium conditions are ignored'. For example, consumption or investment may amount to a relatively constant share of national income in the sample, while that share may rise or fall in the forecast interval. Forecasts may violate theory-based plausibility conditions. Using simulation and algebraic techniques, the consequences of cointegration

for forecasting were analyzed by CLEMENTS&HENDRY and in two influential studies by ENGLE&YOO and by CHRISTOFFERSEN&DIEBOLD. While ENGLE&YOO demonstrated that cointegration may only be important for prediction at longer horizons, such as $h > 10$, CHRISTOFFERSEN&DIEBOLD showed that such results depend on which variable is targeted in prediction. Accurate prediction of the original components may be a different task from predicting stationary variables such as Δy and $\beta' y$. In some specifications, cointegration may even be important for one-step prediction, while, in other specifications, even long-horizon predictions are best conducted on the basis of the original 'level' VAR $\Phi(B) y_t = \varepsilon_t$, while the differences VAR $\Gamma(B) \Delta y_t = \varepsilon_t$ will typically suffer from serious misspecification. As a bottom line, it is recommended to use cointegration techniques and the implied prediction models. When no cointegration is found, one should use VARs in differences.

4.7 Multivariate state-space models

State-space models, such as HARVEY's 'structural' time-series models, can be generalized to the multivariate case. CHATFIELD draws particular attention to the SUTSE (*seemingly unrelated time-series equations*) model with common trends

$$\begin{aligned} y_t &= \Theta \mu_t + \mu^* + n_t, \\ \mu_t &= \mu_{t-1} + w_t. \end{aligned}$$

In this model, the matrix Θ is assumed as rectangular $n \times r$ with $r < n$, such that the second equation updates an r–dimensional integrated process. Some restrictions on μ^* and Θ are needed to identify the model. It can be shown that $n-r$ cointegrating relationships are implied by this model. The random-walk process μ_t is viewed as the 'common trend' of the system, which yields one out of several similar though conflicting definitions of 'common trends'. In strict analogy to the remarks on univariate state-space models, also these models are not really generalizations of VAR or vector ARIMA models and they do not really allow for time-varying structure. They are equivalent to certain vector ARIMA models and merely use a different parameterization.

References

[1] Anderson, T.W. (1951) 'Estimating linear restrictions on regression coefficients for multivariate normal distributions'. *Journal of the American Statistical Association* **85**, 813–823.

[2] Box, G.E.P., and G.M. Jenkins (1970) *Time Series Analysis, Forecasting, and Control*. Holden-Day.

[3] Brockwell, P.J., and R.A. Davis (1991) *Time Series: Theory and Methods*. 2nd edition, Springer-Verlag.

[4] Chatfield, C. (2001) *Time-series Forecasting*. Chapman & Hall.

[5] Christoffersen, P.F., and Diebold, F.X. (1998) 'Cointegration and long-horizon forecasting', *Journal of Business & Economics Statistics* **16**, 450–458.

[6] Clements, M., and Hendry, D.F. (1998) *Forecasting economic time series*. Cambridge University Press.

[7] Dickey, D.A., and Fuller, W.A. (1979) 'Distribution of the estimators for autoregressive time series with a unit root' *Journal of the American Statistical Association* **74**, 427–431.

[8] Engle, R.F. (1992) 'Autoregressive conditional heteroskedasticity with estimates of variance of United Kingdom inflation' *Econometrica* **50**, 987–1007.

[9] Engle, R.F., Hendry, D.F., and Richard, J.-F. (1983) 'Exogeneity'. *Econometrica* **51**, 277–304.

[10] Engle, R.F., and Yoo, B.S. (1987) 'Forecasting and Testing in Cointegrated Systems', *Journal of Econometrics* **35**, 143–159.

[11] Franses, P.H. (1996) *Periodicity and Stochastic Trends in Economic Time Series*. Oxford University Press.

[12] Franses, P.H., and VanDijk, D. (2000) *Non-linear time series models in empirical finance*. Cambridge University Press.

[13] GARDNER, E.S. Jr. (1985) 'Exponential Smoothing: The State of the Art' *Journal of Forecasting* **4**, 1–28.

[14] GARDNER, E.S.JR., AND MCKENZIE, E. (1985) 'Forecasting trends in time series' *Management Science* **31**, 1237–1246.

[15] GRANGER, C.W.J. (1969) 'Investigating Causal Relations by Econometric Models and Cross-Spectral Methods' *Econometrica* **37**, 424–438.

[16] HARVEY, A.C. (1989) *Forecasting, Structural Time Series, and the Kalman Filter*. Cambridge University Press.

[17] HYLLEBERG, S., ENGLE, R.F., GRANGER, C.W.J., AND YOO, B.S. (1990) 'Seasonal integration and cointegration' *Journal of Econometrics* **44**, 215–238.

[18] JOHANSEN, S. (1995) *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.

[19] TAYLOR, S. (1986) *Modelling Financial Time Series*. John Wiley & Sons.

[20] TONG, H. (1990) *Non-linear Time Series: A Dynamical Systems Approach*. Oxford University Press.