Document de travail

DIRECT MULTI-STEP ESTIMATION AND FORECASTING

N° 2005-10
Juillet 2005

Guillaume CHEVILLON
Analysis and Forecasting Department, OFCE
Economics Department, University of Oxford

Corresponding e-mail: guillaume.chevillon@sciences-po.fr; téléphone: +33 1 44 18 54 36. I am grateful to David Hendry, Bent Nielsen and Mike Clements for helpful comments and suggestions.

Observatoire Français des Conjonctures Économiques
69, Quai d'Orsay 75340 Paris Cedex 07
Tel : 01 44 18 54 00 Fax : 01 45 56 06 15
E-mail: ofce@ofce.sciences-po.fr Web: http://www.ofce.sciences-po.fr

Abstract

This paper surveys the literature on multi-step forecasting when the model or the estimation method focuses *directly* on the link between the forecast origin and the horizon of interest. Among diverse contributions, we show how the current consensual concepts have emerged. We present an exhaustive review of the existing results, including a conclusive review of the circumstances favourable to *direct* multi-step forecasting, namely different forms of non-stationarity. We also provide a unifying framework which allows us to analyse the sources of forecast errors and hence of accuracy improvements from *direct* over *iterated* multi-step forecasting.

Keywords: Multi-step Forecasting, Direct estimation, Varying Horizon, Structural breaks, Non-stationarity.

Résumé

Cet article constitue un exposé des développements concernant la prévision à horizon variable lorsque les modèles ou méthodes d'estimation visent directement le lien entre l'origine de la prévision et l'horizon considéré. Au sein des diverses contributions, nous montrons comment les concepts qui font actuellement consensus ont progressivement émergé. Nous présentons de manière exhaustive les résultats existants, en particulier concernant les circonstances favorables à l'usage de la méthode d'estimation directe, à savoir différentes formes de non-stationnarité. Nous fournissons également un cadre unificateur qui permet d'analyser les différentes sources d'erreur de prévision, et ainsi d'amélioration de la précision de la prévision grâce à la méthode directe (par opposition à l'itération de prévisions à horizon unitaire).

Mots-Clef: Prévision à horizon variable, Estimation directe, Multi-étapes, Chocs structurels, Non-stationnarité.

1 Introduction

Economic forecasting is a task distinct from that of modelling because it has been shown (see inter alia Clements and Hendry, 1999, Allen and Fildes, 2001 and Fildes and Stekler, 2002) that causal models do not necessarily forecast better that non-causal alternatives. Rather, causal models often suffer forecast failure and many adjustment techniques have been developed such as intercept corrections (see e.g. Clements and Hendry, 1998a). Other routes include the use of non-congruent models or 'naive' formulations—such as constant growth or random walk hypotheses—which often enhance accuracy owing to their robustness to instability (structural breaks, regime change, economic policy shifts, technological discoveries...) which generate misspecification in the economic models.

When a modeler wishes to produce forecasts at several horizons, an intuitively appealing idea, 'direct multi-step estimation' (DMS), consists in matching model design with the criterion used for its evaluation. Hence, DMS directly minimizes the desired multi-step function of the in-sample errors and offers a potential way to avoid some of the aforementioned difficulties. By contrast, the standard procedure uses one-step estimation—via minimizing the squares of the in-sample one-step ahead residuals—from which multi-step forecasts are obtained by 'iterated multi-step' (denoted here by IMS). One intuition behind DMS is that a model which is misspecified for the data generating process (DGP) need not be a satisfactory forecasting device. However, misspecification is insufficient: predictors like constant growth are misspecified but robust. Here, the desired robustness is to misspecification of the model dynamics or vis-à-vis unnoticed parameter change. Among model misspecifications which might sustain DMS, unnoticed unit roots stand out; neglected serial correlation of the disturbances also provide a rationale at short horizons. In stationary processes, DMS could enhance forecast accuracy, but gains fade rapidly as the horizon increase.

The idea of multi-step estimation has a long history and its developments have followed many paths. Two main DMS approaches have been studied: first, for the parametric technique, the same model parameters are estimated via minimizing distinct horizon-dependent criteria; the techniques used in this case are most often nonlinear, and the model may or not be assumed misspecified. By contrast, non-parametric DMS focuses on the parameters of a different—misspecified beyond the first step—model at each horizon.

The purpose of this article is to review the main contributions to the multi-step forecasting literature and to show how they arose in order to provide a unifying treatment of the many distinct existing results. We first need to explain what we mean by *direct* multi-step forecasting as its definition has emerged only progressively and we think preferable not to define it at this stage but after reviewing the various contributions: this concept has historically served as an umbrella for different approaches and only one has proved useful in forecasting. Let us only clarify for now that the traditional estimation method consists in estimating, for a vector variable $\mathbf{x}_t$, the equation relating it to its past and, potentially, to additional variables. If we denote by $\mathcal{F}_t$, the sigma-field representing the information available at time t, the traditional method seeks to model and estimate how $\mathbf{x}_t$ is generated given $\mathcal{F}_{t-1}$, or $\mathbf{x}_t|\mathcal{F}_{t-1}$, so as to produce an equation such that:

$$\mathbf{x}_t = \widehat{\mathbf{f}}\left(\mathcal{F}_{t-1}\right), \text{ for } t \leq T. \tag{1}$$

From a date T, when $\mathcal{F}_T$ is available, it is therefore possible, using the estimated (1), to generate a forecast for $T+1$, namely

$$\widehat{\mathbf{x}}_{T+1|T} = \widehat{\mathbf{f}}\left(\mathcal{F}_T\right),$$

assuming that the intertemporal link between $\mathbf{x}_t$ and $\mathcal{F}_{t-1}$ will remain valid in the future. When $\mathcal{F}_T$ is generated only by $\{\mathbf{x}_t\}_{t\leq T}$, the same assumption about $\mathbf{x}_{T+h}$ and $\mathcal{F}_{T+h-1}$, for $h > 1$, leads to replacing $\mathcal{F}_{T+1}$ by $\widehat{\mathcal{F}}_{T+1}$ which we regard as pseudo information relating to $\left\{..., \mathbf{x}_T, \widehat{\mathbf{x}}_{T+1|T}\right\}$ where $\widehat{\mathbf{x}}_{T+1|T}$ is assumed to be authentic information (in reality $\widehat{\mathcal{F}}_{T+1} = \mathcal{F}_T$), so that we produce:

$$\widehat{\mathbf{x}}_{T+2|T} = \widehat{\mathbf{f}}\left(\widehat{\mathcal{F}}_{T+1}\right),$$

and so on, for higher forecast horizons. We define the resulting forecasts as iterated multi-step or IMS.

By contrast, an alternative method consists in directly estimating the relationship of interest at the hth horizon, namely $\mathbf{x}_t|\mathcal{F}_{t-h}$, so that a DMS forecast is generated by

$$\widetilde{\mathbf{x}}_{T+h|T} = \widetilde{\mathbf{k}}_h\left(\mathcal{F}_T\right).$$

Care must be paid to the terms used: the one-step (1S) parameter estimates (which coincide for both IMS and DMS at $h = 1$) are those obtained for $\widehat{\mathbf{f}}\left(\cdot\right)$, they imply some IMS counterparts by combination and powering-up. By contrast the DMS parameters are *directly* estimated. Thus, the main distinction between the two methods is that IMS forecasting necessitates only one estimation procedure but the estimates are modified for each horizon, whereas DMS needs re-estimation for each h, but then such estimates are directly usable. We note, and will see below, that some authors define DMS as estimation of the one-step parameters using a non-linear criterion based on $\mathbf{x}_t|\mathcal{F}_{t-h}$; this, seemingly, uncouples estimation and forecasting and does not correspond to our choice of definition, although both are related, which will lead us to consider this case too. We provide below the seven key steps of the progressive research which explain what the state of knowledge now is, each literature strain provides the opportunity for a discussion.

We organise our analysis as follows: section 2 explains the first instances when it was suggested to resort to some dynamic, rather than one-step, estimation. The next section makes explicit the results regarding the inefficiency from using a multi-step procedure to estimate the one-step ahead parameters of a well-specified model (which we call parametric DMS) and we show the need for model misspecification in §4. We turn to non-parametric estimation in section 6. The main theoretical results regarding forecasting are presented in section 5, robustness towards breaks is analysed in §7 and section 8 concludes the review of literature. After reviewing all the progress made in the literature and the many aspects covered, we finally present our analysis of the general framework appropriate for the analysis of direct multi-step estimation and forecasting in section 9 and show that it explains how to interpret the results found in existing literature.

2 Early suggestions: estimate the dynamic 'solution path'.

The first instance when some dynamic estimation was suggested is found in Cox (1961) who compares the mean-square forecast errors from an Exponentially Weighted Moving Average and an AR(1) model with an intercept when the true data generating process is either AR or ARMA with an intercept. He shows that, if the mean of the process to be forecast is allowed to shift, the parameters of the prediction model should depend on the forecasting horizon so that robustness can be achieved. He suggests combining the EWMA and the AR forecasting techniques with weights which vary with the horizon.

At the turn of the 1970s, several authors start focusing on the difficulties in estimating dynamic models. Their concern is that of predetermined variables and their interest lies in the design of estimation techniques which take full advantage of the dynamic structure of the series.

Klein (1971) suggests a multi-step estimator which minimizes the 'solution path' as mentioned in Haavelmo (1940). His idea is that in general if the data generating process follows an AR(1) (which can readily be extended to include more lags or exogenous variables):

$$y_t = \alpha y_{t-1} + \epsilon_t, \quad \text{for } t = 1, ..., T, \text{ and } |\alpha| < 1,$$

and it is wished to obtain forecasts of $y_{T+h} = \alpha^h y_T + \sum_{i=0}^{h-1} \alpha^i \epsilon_{T+h-i}$, for $h = 1, ..., H$, then least-squares estimation of the model leads to minimizing the criterion function:

$$\sum_{h=1}^{H} \sum_{t=1}^{T-h} \left(\sum_{i=0}^{h-1} \alpha^i \epsilon_{t+h-i} \right)^2 = \sum_{h=1}^{H} \sum_{t=1}^{T-h} \left(y_{t+h} - \alpha^h y_t \right)^2,$$

with respect to the coefficient α. In a simulation experiment, the author lets several parameters vary and his findings are that *(i)* multi-step methods seem to perform better in smaller samples (here 50 vs. 400), *(ii)* adding a trendless exogenous variable seems to help DMS, but a trending variable does not, and *(iii)* the initial observation does not affect the previous results. In applying this dynamic estimation method to the Wharton model, he finds that he can reduce the mean average prediction error (MAPE) from 6.29% to 5.33% in 2-step ahead out-of-sample forecasting, when comparing it to an IV estimation with principal components.

Hartley (1972) studies the properties of the dynamic least squares estimator (DLS for him) suggested by Klein (1971) in the univariate AR(1) case. He shows that the new estimator is more robust to residual autocorrelation than OLS.

Assuming that the process can be written, for $t = 1, ..., T$, as

$$y_t = \alpha y_{t-1} + \epsilon_t, \tag{2}$$

$$\epsilon_t = \rho \epsilon_{t-1} + u_t, \tag{3}$$

where y_0 is fixed, $\epsilon_0 = 0$, $\{u_t\}$ is an independently and identically distributed (i.i.d.) process whose elements have zero mean, variance σ^2 and finite third and fourth moments, $|\alpha| < 1$ and $|\rho| < 1$, Hartley shows that if the dynamic solution path

$$y_t = \alpha^t y_0 + \sum_{i=1}^{t} \alpha^{t-i} \epsilon_i,$$

is estimated by generalised least squares (GLS), then it is the same as OLS when $\rho = 0$. Denoting the OLS and DLS estimators of α by respectively $\widehat{\alpha}$ and $\widetilde{\alpha}$, then

$$\widehat{\alpha} - \alpha \underset{T\to\infty}{\to} \frac{\rho\left(1-\alpha^2\right)}{1+\alpha\rho},$$

but $\widetilde{\alpha}$ does not converge unless it is assumed that $y_0 = O_p\left(T^k\right)$, for $k > 0$, under which circumstance, if $\rho = 0$:

$$\widetilde{\alpha} - \alpha \underset{T\to\infty}{\to} \mathsf{N}\left[0, \left(1-\alpha^2\right)\frac{1+3\alpha^2+\alpha^4}{\left(1+\alpha^2\right)^2}\frac{\sigma^2}{y_0^2}\right],$$

so that the asymptotic variance of the DLS estimator is of order $1/T^{2k}$. The author shows that when $\rho \neq 0$ and $y_0 = O_p\left(T^k\right)$, there exists a function $f(\cdot,\cdot)$ such that

$$\lim_{T\to\infty} \mathsf{Var}\left[\widetilde{\alpha} - \alpha\right] = f(\alpha,\rho)\frac{\sigma^2}{y_0^2}.$$

Thus the DLS estimator is consistent, whereas OLS is not. Hartley notes that the assumption about the initial observation is satisfied even with very low k. Yet the variance cannot be made arbitrarily small since σ should then increase with y_0. He also notices that if the errors follow an MA(1) rather than an AR(1), then the DLS estimator is the Maximum Likelihood Estimator.

Johnston, Klein, and Shinjo (1974) notice that dynamic models which incorporate lagged values of the endogenous variable may lead to a contradiction between the assumptions made for estimation and for forecasting. Indeed, it is common practice since the work by Mann and Wald to consider that the lags of the endogenous variable can be asymptotically treated as 'exogenous', or predetermined. However, when formulating a forecast at several periods in the future, the intermediate lags—between the forecast origin and the period of the forecast—can no longer be seen to be predetermined and this aspect ought to be taken into consideration. They build their work on the previous results by Haavelmo (1944) who shows that for the case of a stationary AR(1) process with no drift:

$$y_t = \alpha y_{t-1} + e_t, \tag{4}$$

the optimal—in the sense of minimizing a quadratic loss function in e_{T+1} and e_{T+2}—prediction formulae for $T+1$ and $T+2$ from an end-of-sample forecast origin y_T are given by:

$$\begin{aligned} y_{T+1} &= \alpha y_T, \\ y_{T+2} &= \alpha^2 y_T. \end{aligned}$$

The significant aspect is that the maximum likelihood estimate of α is that of α^2 too.

Johnston, Klein, and Shinjo compare several forecasting methods for the AR(1) with different parameter values and apply their techniques to the Wharton model. Their idea is to compute estimators and resulting forecasts which incorporate the dynamic structure of the data generating process. The hypothesis is that 'systems using up to pth order generated lag values as instruments or regressors will perform best in p period point prediction'. In general, the DLS estimator for a model such as

$$\mathbf{A}(L)\,\mathbf{y}_t = \boldsymbol{\epsilon}_t,$$

where $\mathbf{A}(L)$ is a matrix polynomial and L the lag operator, is that which minimizes the criterion:

$$\mathrm{tr}\sum\nolimits_{t=1}^{T}\left(\mathbf{A}(L)^{-1}\boldsymbol{\epsilon}_t\right)\left(\mathbf{A}(L)^{-1}\boldsymbol{\epsilon}_t\right)'.$$

In the univariate AR(1) case from (4), this is:

$$\widetilde{\alpha} = \underset{\alpha}{\mathrm{argmin}}\sum\nolimits_{t=1}^{T}\left(y_t - y_0\alpha^t\right)^2. \tag{5}$$

The procedure used by the authors for actual minimization is a grid search. The results of Monte Carlo simulations with fixed or stochastic initial values and various stationary values of α show that the variance of the DLS estimator is higher than that of OLS when the initial value is stochastic, but lower for a fixed initial value. In terms of mean-square forecast error (MSFE), their results are that for small samples (either 20 or 50 observations and the forecast horizons, respectively, of 5 or 10 periods) DLS outperforms OLS when the initial value is fixed, but when the latter is stochastic, the forecast loss is lower for DLS only for very small samples.

The authors then use Two-Stage Least Squares estimators for the Wharton model. They match the values of the lag used for the endogenous variable as an instrument and the horizon at which it is desired to forecast. Unfortunately, their results for out-of-sample prediction are somewhat inconclusive. Some gains are obtained at short horizons and seem to improve with the lead in the forecast for Personal Income and Total Consumption but not for the G.N.P. deflator, Investment in Non-farm inventories and Unemployment rate. The G.N.P. deflator is the only variable for which the within-sample residual standard error and the post-sample root-MSFE are of the same magnitude. The latter is much larger than the former as regards the other variables.

Discussion

The concept of solution path is the first attempt to provide an estimation method which embodies the dynamics of the process. Unfortunately, its dependence on the initial observation makes it impractical since it can be strongly contaminated by measurement errors and it asymptotically relies on the artificial assumption that the initial observation increases with the sample size. Thus this methodology is of little use for both stationary and integrated processes. Yet it paves the way to multi-step estimation where it is not the same initial observation which is used, but the same lag; thus leading, instead of (5), to

$$\widetilde{\alpha}_h = \underset{\alpha_h}{\operatorname{argmin}} \sum\nolimits_{t=h}^{T} \left(y_t - y_{t-h}\alpha^h\right)^2 .$$

And, now, there is no longer any need for an exploding initial value. This augurs all the better for the use of DMS since the first simulation results by Johnston, Klein, and Shinjo seem to confirm that such methods fare well in small samples, where the initial value need not be of magnitude different from the rest of the observations.

3 Inefficiency of DMS estimation in a well-specified model.

The first authors who analyse multi-step estimation techniques compare their asymptotic properties to those of other well established methods when the model is well-specified for the stationary DGP.

Johnston (1974) analyses the forecasting properties of the multi-step estimator (Dynamic Estimator for him) suggested by Johnston, Klein, and Shinjo (1974) and compares them to those of a one-step ahead estimator. His framework is that of a well specified dynamic vector model with exogenous variables and mean zero errors:

$$\mathbf{y}_t = \mathbf{z}_t \boldsymbol{\theta} + \boldsymbol{\epsilon}_t, \tag{6}$$

where $\mathbf{z}_t = (\mathbf{y}_t, \mathbf{y}_{t-1}, \mathbf{x}_t)$ and $\boldsymbol{\theta}' = (\mathbf{A}_0', \mathbf{A}_1', \mathbf{B}')$, with zero diagonal entries for $\mathbf{A}_0$. For an estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$**,** $\widehat{\mathbf{y}}_{T+h,h}$ is the forecast of $\mathbf{y}_{T+h}$ conditional on $\{\mathbf{y}_t\}_{t\leq T}$, $\{\mathbf{x}_t\}_{t\leq T+h}$ and $\widehat{\boldsymbol{\theta}}$, as obtained by some least-squares technique. The user's loss function is assumed to be quadratic and given by:

$$L\left(T, \underline{h}, \overline{h}\right) = \sum\nolimits_{h=\underline{h}}^{\overline{h}} w_h \left(\mathbf{y}_{T+h} - \widehat{\mathbf{y}}_{T+h,h}\right) \mathbf{Q} \left(\mathbf{y}_{T+h} - \widehat{\mathbf{y}}_{T+h,h}\right)', \tag{7}$$

where $\mathbf{Q}$ is a positive definite matrix weighting the importance of forecast errors across equations and the set of weights, $\{w_h\}$, expresses the relative importance of the forecasting horizons ($\forall h$ $w_h \geq 0$). The direct multi-step estimator is then defined as:

$$\widehat{\boldsymbol{\theta}}_{DMS} = \underset{\widehat{\boldsymbol{\theta}}}{\operatorname{argmin}} \left\{ \sum_{t=1}^{T-\overline{h}} L\left(t, \underline{h}, \overline{h}\right) \right\}. \tag{8}$$

It corresponds to the DLS, whose distribution was derived by Hartley (1972) in the univariate case. And when the model is well-specified—i.e. $\rho = 0$ in (3)—this estimator is asymptotically less efficient than the one-step OLS, thus being consistent with the claim in Haavelmo (1944) that, when the error loss function is quadratic, the rankings of the estimators in prediction and estimation efficiency match one another. The author's Ph.D. thesis showed that, asymptotically, the 'optimal' estimator is invariant to the choice of—quadratic—prediction error loss function. Johnston sustains that, in practice, multi-step estimation can be justified if it is more efficient than an alternative computationally equivalent estimator. Yet, as the paper proves, the asymptotically most efficient—in terms of minimum variance—estimator from (8) is given by $\overline{h} = \underline{h} = 1$ (the author only considers the case where $\overline{h} = \underline{h}$, $\mathbf{Q} = \mathbf{I}$, $w_h = 1$ $\forall h$, and where $\widehat{\boldsymbol{\theta}}_{DMS}$ is obtained by iterated minimization of (8), given an initial estimate $\widehat{\boldsymbol{\theta}}_0$ which provides $\widehat{\mathbf{z}}_t\left(\widehat{\boldsymbol{\theta}}_0\right)$, until convergence). The main result is that:

$$\Sigma_{\overline{h}} - \Sigma_{\overline{h-1}} \geq 0,$$

where $\Sigma_{\overline{h}}$ is the asymptotic variance of the multi-step estimator (scaled by $\sqrt{T}$) for $\overline{h} = \underline{h}$. Thus the one-step OLS estimator has *minimum asymptotic variance* and is hence efficient. The author mentions some unpublished simulation results which confirm this finding even in small samples. He notes, however, that small sample biases should be taken into account since they will make the estimator variance and MSFE to differ.

Kabaila (1981) is interested in comparing the asymptotic efficiency of the IMS and DMS estimators in general non-linear processes. His assumptions are that the process under consideration $\{y_t\}$ is strictly stationary and generated by:

$$y_t = f\left(y_{t-1}, y_{t-2}, ...; \theta_0\right) + \epsilon_t,$$

where the process $\{\epsilon_t\}$ is i.i.d. and its elements have zero expectation and variance σ^2, $\theta_0 \in \Theta \subset R^p$, and y_t is a measurable function of $\{\epsilon_t\}$. The dots in $f(\cdot)$ indicate that the initial values can be of any form and number. Obviously, $f(y_{t-1},...;\theta_0) = E[y_t|y_{t-1},...;\theta_0]$. Define, similarly,

$$g_k(y_{t-k}, y_{t-k-1},...;\theta_0) = E[y_t|y_{t-k},...;\theta_0], \quad \text{for } k > 1.$$

The function $h_{k,t}(\cdot)$ is defined as that obtained by backward substitution of the $f(\cdot)$ and ϵ_{t-j} for $j = 0,...,k-1$, such that:

$$g_k(y_{t-k}, y_{t-k-1},...;\theta_0) = E[h_{k,t}(\theta_0)|y_{t-k},...].$$

Kabaila makes an additional assumption, namely that for all $\theta \in \Theta$:

$$h_{k,t}(\theta) = \sum_{i=1}^{k-1} \epsilon_{t-i} U_{t,i}(\theta) + V_{t-k}(\theta),$$

where $U_{t,i}(\theta)$ $(i > k-1)$ is a function of θ, the ϵ_j and y_m for $j \leq t-i-1$ and $m \leq t-k$; $U_{t,k-1}(\theta)$ and $V_{t-k}(\theta)$ are functions of $y_{t-k}, y_{t-k-1}, ...$

Let $\widehat{\theta}_T$ and $\widetilde{\theta}_{k,T}$ denote minimizers—with respect to θ—of some approximations to the in-sample (for a sample of size T) sum of the squared residuals, respectively $y_t - f(y_{t-1}, y_{t-2}...;\theta_0)$ and $y_t - g_k(y_{t-k}, y_{t-k-1},...;\theta_0)$. By approximation, it is meant that the initial values $y_{-1},...$ may not be known and this is reflected in the objective function.

Provided that the asymptotic variances of the estimators (which exist) are nonsingular, Kabaila proves that the 1S estimator is efficient, as an estimator of θ.

Discussion

These authors are interested in comparing some parameter estimators which account for some of the dynamics of the process. This is one of the two strains of multi-step estimation and, unfortunately, brings no benefits. Here the *same* parameter is to be estimated by either one-step or h-step methods. It is simply the objective functions that differ, in so far as the h-step criterion is a non-linear composition of the one-step. Indeed in both cases, the h-step fitted values—$\widehat{\mathbf{y}}_{T+h,h}$ or $h_{k,t}(\theta)$—are computed using the same model as that for 1S. Under these assumptions, the authors show that one-step estimation is asymptotically more efficient than this type of DMS. The two contributions are thus essential, since they show that for DMS to provide any gains, one of the four following assumptions must be made: (*i*) the model is misspecified, (*ii*) different models are used for 1S and DMS, (*iii*) it is the implied (powered-up) multi-step parameters which are of interest, not the one-step estimated by a multi-step criterion or (*iv*) the gains are to be found in small samples. These assumptions are studied by other authors as we see below and will lead to the preferred approach to direct multi-step estimation which no longer aims to estimate the 1S model via multi-step techniques.

4 Parametric DMS estimation under misspecification.

The first contributions to the parametric approach to multi-step forecasting suggest some forms of model misspecification which could provide a sufficient rationale for the use of DMS. By parametric, it is meant that the one-step ahead parameters are the object of interest but that they are estimated by minimizing functions of the multi-step errors.

Stoica and Nehorai (1989) extend the concept of direct multi-step estimation which was suggested by Findley to ARMA models:

$$A(L)\,y_t = C(L)\,\epsilon_t,$$

where $A(L) = \sum_{i=0}^{p} a_i L^i$ and $C(L) = \sum_{i=0}^{p} c_i L^i$. The forecasts $\widehat{y}_{T+h,h}$ are computed as the conditional expectation of y_{T+h} given y_T for the model with parameter $\theta = (a_0, ..., a_p, c_0, ..., c_p)$. Define $B_h(L) = \sum_{i=0}^{h-1} b_i L^i$ and $D_h(L) = \sum_{i=0}^{p} d_i L^i$, such that:

$$C(L) = A(L)\,B_h(L) + L^h D_h(L),$$

so that

$$y_t = \left(B_h(L) + \frac{D_h(L)}{A(L)} L^h \right) \epsilon_t.$$

The h–step ahead forecast error is thus given by

$$e_{T+h,h} = y_{T+h} - \widehat{y}_{T+h,h} = B_h(L)\,\epsilon_{T+h}.$$

The authors define the multi-step parameter estimator as that which minimizes a function of the in-sample squared multi-step forecast errors,

$$\widehat{\theta}_h = \underset{\widetilde{\boldsymbol{\theta}}_h \in \Theta}{\operatorname{argmin}}\, \mathsf{F}\left(V_{1,T}, ..., V_{h,T}\right),$$

where $V_{k,T} = T^{-1} \sum_{t=1}^{T-k} e_{t+k,k}^2$, for $k = 1, ..., h$. They provide various algorithms to obtain the non-linear estimates.

Under the assumption that the DGP follows an ARMA(p, p) , Stoica and Nehorai present several results, namely that (i) the one-step estimator $\widehat{\theta}_1$ for $\mathsf{F}(u) = u$, is consistent and asymptotically efficient among the class of estimators whose covariance matrices depend only on the second-order properties of the data; and (ii) that the only stationary point of $V_{1,\infty}$ is θ^*, the true parameter value. By contrast, they note that the multi-step criterion may have several minima. The consequence is that for there being any gain from using multi-step estimation, the main assumptions have to be modified. Thus, if it is assumed that the true DGP is not known, it is still possible under weak conditions to show that $\widehat{\theta}_h$ converges to some value which leads asymptotically to the 'best'—in the sense of minimizing $\mathsf{F}(V_{1,\infty}, ..., V_{h,\infty})$—forecasts. The use of DMS can therefore be justified in practice.

The authors conduct a Monte Carlo experiment in which they analyse the forecasts obtained for four models:

$$
\begin{aligned}
&ARMA\,(3,3): && \left\{ \begin{array}{l} y_t - 0.95 y_{t-1} + 0.81 y_{t-2} - 0.7695 y_{t-3} \\ = \epsilon_t - 0.97 \epsilon_{t-1} - 0.775 \epsilon_{t-2} + 0.6732 \epsilon_{t-3}; \end{array} \right. \\
&BLAR\,(1): && y_t = 0.4 y_{t-1} + \epsilon_t + 0.8 y_{t-1} \epsilon_{t-1}; \\
&TMA\,(3): && y_t = \left\{ \begin{array}{ll} \epsilon_t + 0.15 \epsilon_{t-1}, & \text{if } \epsilon_t < 0, \\ \epsilon_t - 0.97 \epsilon_{t-1} + 0.81 \epsilon_{t-2} - 0.7857 \epsilon_{t-3}, & \text{if } \epsilon_t \geq 0; \end{array} \right. \\
&ARMA\,(2,2): && y_t - 0.98 y_{t-2} = \epsilon_t - 0.87 \epsilon_{t-1} - 0.775 \epsilon_{t-2}.
\end{aligned}
$$

They estimate the models over samples of size 200 and forecast over the horizons $h = 1, .., 4$. The forecasting models are either an AR(4) or an AR(8) , except for the ARMA$(2, 2)$ model for which they either try an AR(1) or an AR(6). The multi-step estimators are computed for the four horizons at once. Their results are that the first three model provide no rationale for the use of multi-step estimation, other than the fact that the forecast accuracy is essentially the same for IMS and DMS. By contrast the fourth model forecast by an AR(1) does indeed provide a gain for DMS. It must be noted that the gain is for horizons 2 and 4. The slope estimates are 0.26 for IMS and 0.30 for DMS. The authors conclude that under-parameterization seems to benefit DMS.

Discussion

Although Stoica and Nehorai do not make explicit the difference between the model parameters and their powered-up multi-step counterparts, they show the importance of the hypothesis of model misspecification as a justification for the use of DMS. Here, it is the same model which is used at all forecast horizons, but the estimation method matches the desired outcome. In their simulations, the authors find that an ARMA$(2, 2)$ estimated by an AR(1) model can lead to more accurate forecasts when using DMS. Their conclusion relating to under-parameterization mirrors that of Bhansali (1999); it is surprising that the very specific form of DGP they use should not strike them: it exhibits a root close to unity. It is thus possible that non-stationarity may appear as a feature benefitting DMS.

5 Efficiency in matching criteria for estimation and forecast evaluation.

Analyzing the ARIMA time series reported in Madrikakis (1982), Weiss and Andersen (1984) compare the forecasting properties of various estimation methods when the forecast accuracy criterion varies. In particular, they compare the one-step and multi-step ahead forecasts. They find that when a one-step ahead forecast accuracy loss function is used, it is preferable to use one-step ahead estimation (and then OLS, Least-Absolute Deviation seem similar for either MSFE or Mean Absolute Error (MAE) criteria). Similarly, when the forecast accuracy is measured by the absolute percentage trace of a matrix of the forecast errors at several horizons, the best amongst the four estimation methods which they use (the multi-step trace, one-step ahead OLS, one-step MAE and one-step Mean Absolute Percentage Error) is the multi-step trace. They, thus, find some significant improvement from matching estimation and forecasting horizons.

Weiss (1991) builds upon the earlier work on multi-step estimation for forecasting and derives conditions under which this technique is asymptotically 'optimal', in a sense that he defines. He builds on the work by Johnston, where, in model (6), he allows for more lags of the endogenous variable. He also defines the error terms as a function of the parameter $\boldsymbol{\theta}$:

$$\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}_t(\boldsymbol{\theta}) = \mathbf{y}_t - \mathbf{z}_t\boldsymbol{\theta},$$

where $\mathbf{x}_t$, in $\mathbf{z}_t = (\mathbf{y}_t, \mathbf{y}_{t-1}, ..., \mathbf{y}_{t-p}, \mathbf{x}_t)$, is a vector of variables strongly exogenous with respect to $\boldsymbol{\theta}$ (see Engle, Hendry, and Richard, 1983). The model is not assumed to coincide with the DGP, and any of the following may be present: irrelevant regressors, omitted variables, serial correlation, misspecified functional form, etc. The author works under fairly mild assumptions allowing for a uniform Law of Large Numbers (LLN) and a Central Limit Theorem (CLT). The forecasts are given, as in Johnston (1974), as the conditional expectation computed by assuming that the model is well specified and, similarly, the forecast evaluation criterion is $\mathsf{E}\left[L\left(T,1,\overline{h}\right)\right]$, where the expectation is taken with respect to the true process, $\mathbf{Q} = \mathbf{I}$, and $L(\cdot,\cdot,\cdot)$ is defined in (7). $\widehat{\boldsymbol{\theta}}_{DMS}$ is defined as in (8), where the parameter space, $\boldsymbol{\Theta}$, is assumed compact. The inclusion of lags of $\mathbf{y}_t$ in $\mathbf{z}_t$ implies that $\widehat{\boldsymbol{\theta}}_{DMS}$ is not the simple weighted least-squares estimator. Weiss assumes that a uniform LLN will hold for $G_{T,\overline{h}}(\boldsymbol{\theta}) = \sum_{t=1}^{T-\overline{h}} L\left(t,1,\overline{h}\right)$ and that its limit coincides with that of the forecast evaluation criterion, denoted $\overline{G}_{T,\overline{h}}(\boldsymbol{\theta}) = \mathsf{E}\left[L\left(T,1,\overline{h}\right)\right]$. The author then proves that, given a minimizer of the continuous function $\overline{G}_{T,\overline{h}}(\boldsymbol{\theta})$ on $\boldsymbol{\Theta}$, which exists on a compact set and is denoted by $\widetilde{\boldsymbol{\theta}}$,

$$G_{T,\overline{h}}\left(\widehat{\boldsymbol{\theta}}_{DMS}\right) - \overline{G}_{T,\overline{h}}\left(\widetilde{\boldsymbol{\theta}}\right) \underset{T\to\infty}{\overset{a.s.}{\to}} 0.$$

If the sequence of $\left\{\widetilde{\boldsymbol{\theta}}\right\}_{T=1}^{\infty}$ is identifiably unique,[1] then $\widehat{\boldsymbol{\theta}}_{DMS}$ is strongly consistent for $\widetilde{\boldsymbol{\theta}}$, i.e.

$$\widehat{\boldsymbol{\theta}}_{DMS} - \widetilde{\boldsymbol{\theta}} \underset{T\to\infty}{\overset{a.s.}{\to}} 0.$$

and there exists a scaling matrix $K_{\overline{h}}$ such that:

$$T^{1/2}K_{\overline{h}}\left(\widehat{\boldsymbol{\theta}}_{DMS} - \widetilde{\boldsymbol{\theta}}\right) \underset{T\to\infty}{\overset{\mathsf{L}}{\to}} \mathsf{N}(\mathbf{0},\mathbf{I}).$$

Thus the multi-step estimator is asymptotically optimal, in the sense that it minimizes the desired criterion function. In small samples, two opposite effects are present: the variance of the multi-step estimator should be larger than that of the one–step ahead, but the bias should be smaller. A

[1] i.e. if and only if $\forall\eta > 0$, $\liminf_{T\to\infty}\left\{\min_{\boldsymbol{\theta}\in\mathbf{N}_T^C(\eta)}\left[\overline{G}_{T,\overline{h}}(\boldsymbol{\theta}) - \overline{G}_{T,\overline{h}}\left(\widetilde{\boldsymbol{\theta}}\right)\right]\right\} > 0$, where $\mathbf{N}_T(\eta)$ is a neighbourhood of $\widetilde{\boldsymbol{\theta}}$ of radius η such that its complement $\mathbf{N}_T^C(\eta)$ is a compact set of $\boldsymbol{\Theta}$.

Monte Carlo simulation thus attempts to exemplify the results for a sample of 100 observations with random initial observations. The data generating process is univariate autoregressive with distributed lags (ADL):

$$y_t = \alpha_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \gamma_0 z_t + \gamma_1 z_{t-1} + \epsilon_t,$$

where z_t follows a stationary AR(1): $z_t = \psi z_{t-1} + u_t$, with $|\psi| < 1$. The errors—$\{\epsilon_t\}$ and $\{u_t\}$—are i.i.d. standard normal and independent. The intercept α_0 is set to zero and estimated in all cases. The cases studied are either a well-specified ADL$(2,1)$ model or one with some degree of misspecification: omitted regressors y_{t-2}, z_{t-1}, or $\{z_t, z_{t-1}\}$; or wrong functional form (estimation of the log of the data for which the intercept is non-zero). The only cases that provide a rationale for using the multi-step estimator are those when either $\beta_1 = 1$ and β_2 is close to zero (and not estimated) and when z_{t-1} is omitted as a regressor. Thus it seems that DMS performs better when the DGP is better modelled as a random walk than as a stationary AR(2) and when some residual autocorrelation is omitted.

Weiss (1996) analyses the forecasting properties of models which are estimated using the cost function used also for the appraisal of the forecast. The main idea is that when this criterion is quadratic, then the optimal forecast is the expectation, conditional on the information set. But this result is only valid as long as this loss function is also used for the evaluation of the forecast. Granger (1969) had considered predicting several steps into the future and recommended some techniques. For instance, letting $C_F(\cdot)$ denote the forecast evaluation cost function, if it is desired to predict y_{t+h} given $\{y_i\}_{i \leq t}$, the forecaster could minimize the in-sample bias term from a linear predictor:

$$\min_{a,b_i} \sum_t C_F \left(y_{t+h} - \sum_{j=0}^{m} b_j y_{t-j} - a \right).$$

Alternatively, if minimization is difficult to carry out, it would be sensible to first estimate $(b_1, ..., b_m)$, by OLS, forming $\widetilde{y}_{t+h} = y_{t+h} - \sum_{j=0}^{m} \widehat{b}_j y_{t-j}$, and then to minimize $\sum_t C_F(\widetilde{y}_{t+h} - a)$ with respect to a. Such methods were proposed because Granger thought that it should be asymptotically sensible to use the same criteria for both estimation and evaluation. In his article, Weiss focuses on one-step ahead forecasts and derives the optimal predictors; yet the Monte Carlo that he provides do not show a substantial improvement.

Discussion

One of the important contributions of Weiss (1991) is that his definition of optimality is not that the estimator should have the lowest possible asymptotic variance but that it achieves the lowest in-sample (multi-step) MSFE. This shift of focus is crucial to the appraisal of DMS methods and it seems natural to evaluate a method by assessing how well it achieves the aim for which it is designed. His simulations point to the fact that for DMS to perform better than IMS, the series must be non-stationary, either of stochastic—via unit roots—or of deterministic—since location shifts imply residual autocorrelation—form.

6 Design of non-parametric DMS models.

This first strain of direct estimation focuses on fitting different models for forecasting at different horizons. Research along these lines attempts to establish reasonable 'good' criteria for choosing the order p of the 'best' AR(p), to use for forecasting. The 'non-parametric' terminology is explained in Bhansali (1999).

Findley (1983) provides a theoretical rationale for adapting the forecasting models to the forecasting horizon and suggests one type of technique which he applies to some standard time series from Box and Jenkins (1976). The author starts by considering the case when an AR(1) model is used for prediction of a variable h steps ahead. Denoting by $\{\rho_k\}$ the autocorrelation sequence of the process $\{y_t\}$, the parameter ψ_h which minimizes the MSFE

$$\mathsf{E}\left[\left(y_{T+h} - \widehat{y}_{T+h,h}\right)^2\right], \tag{9}$$

where $\widehat{y}_{T+h,h} = \psi_h y_T$, is simply the autocorrelation $\psi_h = \rho_h$, in the stationary case. If $\{y_t\}$ does indeed follow an AR(1), $y_t = \phi y_{t-1} + \epsilon_t$, where $\phi < 1$, then, naturally, we need to choose $\psi_h = \phi^h$ and $\phi = \rho_1$. If $\{y_t\}$ follows any other process but we still fit an AR(1) as above, in order to minimize (9), we must set

$$\begin{aligned} \phi &= \left(\rho_h\right)^{1/h}, &&\text{if } h \text{ is odd or } \rho_h > 0, \\ \phi &= 0, &&\text{if } h \text{ is even and } \rho_h < 0. \end{aligned}$$

Thus the 'optimal' model depends on the desired lead in forecasting. Notice that if h is even and $\rho_h < 0$, it is preferable not to fit an AR(1) model but rather to use $\psi_h = \rho_h$. It, therefore, seems that the formula most appropriate to multi-step forecasting cannot always be derived from an ARMA model. The results would asymptotically be the same if the estimators were computed to maximize the forecast log-likelihood. Findley remarks that when the forecast accuracy criterion combines several horizons, the degree of complexity is much higher. In order to improve forecast accuracy, it may seem desirable to use several lags of the variable. Findley suggests an h-step Akaike Information Criterion (AIC_h) in order to select the order of the AR(p) to be fitted (for p smaller than some $p_{\max}$). The order p is thus given by:

$$p = \underset{1\leq p\leq p_{\max}}{\operatorname{argmin}} \left\{\mathrm{AIC}_h\left(p\right)\right\},$$

where:

$$\begin{aligned}\mathrm{AIC}_h\left(p\right) &= T_0 \log\left[2\pi.SSQ\left(\widehat{\phi}_1,...,\widehat{\phi}_p\right)/T_0\right] + T_0 + 2\left(p+1\right),\\ T_0 &= T - p_{\max} - h + 1,\end{aligned}$$

and $\left(\widehat{\phi}_1,...,\widehat{\phi}_p\right)$ is computed as the set of coefficients which minimizes the in-sample sum of the squared h-step ahead residuals:

$$SSQ\left(\phi_1,...,\phi_p\right) = \sum\nolimits_{t=p_{\max}}^{T-h}\left(y_{t+h} - \sum\nolimits_{k=1}^{p}\phi_k y_{t-k+1}\right)^2.$$

The author applies his results to two standard time series: series C and E from Box and Jenkins (1976), where autoregressive models are fitted using the AIC_h criterion. The results exhibit an average gain for the proposed multi-step methods in terms of MSFE of about 4% for series C at horizons 5 and 10, and respectively 2.6% and 10.6% for series E at horizons 5 and 10.

Liu (1996) suggests to modify the standard fitting criteria for the order of an autoregressive process to allow for the inclusion of multi-step forecast errors. He proposes to partition the data set into non-overlapping vectors of length h, where h is the maximum desired forecast horizon. Estimating the resulting VAR by weighted least-squares is shown by the author to be leading asymptotically to the same estimates as those of a univariate model, but at a loss of efficiency. In a Monte Carlo simulation for samples of size 80 and 240, Liu compared the ratios of 2- and 4-step ahead root MSFEs. The results showed little improvement by using the multi-step methods, whether the data were generated by either a zero-mean stationary AR(1) or an ARI$(1,1)$. The author applies his methods to forecasting the quarterly U.S. (174 observations) and monthly Taiwan (192 obs.) unemployment rates, the log of quarterly real U.S. G.N.P. (179 obs.) and the monthly

U.S. consumer price index for food (241 obs). Several overlapping samples for each data set were used where the estimation was conducted for fixed sample sizes of respectively 100, 120, 100, 160 observations. Three main results emerge: first, when the one-step method is preferred, the loss from using a DMS method is low, except when the multi-step order is determined by a modified Fisher Information Criterion for which the loss can be up to 12.5%; second the multivariate procedure is always preferred for the trending variables in levels (the U.S. G.N.P. and food index), but not necessarily in differences; and third the DMS method is preferred for the monthly Taiwan unemployment rate but not for the quarterly U.S. For the latter result, the author suggests as an explanation that monthly data exhibit more time dependence.

Bhansali (1999) surveys the developments in multi-step criteria for the design of forecasting models. He first distinguishes two different approaches: a parametric and a non-parametric. In the former, the modeler attempts to establish what the true data generating process is, and estimates its k parameters via some multi-step technique (for instance by maximum likelihood); by contrast a non-parametric procedure approximates the unknown DGP by some process whose number of parameters is allowed to diverge, say $k(T)$, where T is the sample size and $k(T) = o(T)$. Assume that a process $\{y_t\}$ is approximated or modelled as a linear function of k lags, so that at an end-of-sample forecast origin T, it is wished to predict y_{T+h}, where $h \geq 1$. Write the resulting forecast $\widetilde{y}_{T+h,h}$ as

$$\widetilde{y}_{T+h,h} = \sum_{i=0}^{k} \widetilde{\alpha}_{h,i} y_{T-i},$$

where the $\widetilde{\alpha}_{h,i}$ are estimated by regressing y_{t+h} on $(y_t, ..., y_{t-k})$ from a hypothesized model (the forecast generating process), for fixed $h \geq 1$:

$$y_{T+h} = \sum_{i=0}^{k} \alpha_{h,i} y_{T-i}.$$

For notational simplicity, the dependence of the $\alpha_{h,i}$ on k is omitted. Define the unconditional mean square forecast error:

$$\widetilde{V}_{h,k}^{DMS} = \mathsf{E}\left[\left(y_{T+h} - \widetilde{y}_{T+h,h}\right)^2\right].$$

Similarly, letting $y_{T+1,1} = \sum_{i=0}^{k} \alpha_{1,i} y_{T-i}$ and noting that

$$y_{T+2} = \alpha_{1,0} y_{T+1,1} + \alpha_{1,1} y_T + \alpha_{1,2} y_{T-1} + ... = \sum_{i=0}^{k-1} \left(\alpha_{1,0}\alpha_{1,i} + \alpha_{1,i+1}\right) y_{T-i} + \alpha_{1,0}\alpha_{1,k} y_{T-k},$$

it is possible, by iterated substitution, to find a set of non-linear function $\beta_{h,i}$ of the set of $\{\alpha_{1,i}\}$ such that:

$$y_{T+h} = \sum_{i=0}^{k} \beta_{h,i} y_{T-i}.$$

Denote by $\widehat{\beta}_{h,i}$ the function of the estimated $\widetilde{\alpha}_{1,i}$ so that $\widehat{y}_{T+h,h} = \sum_{i=0}^{k} \widehat{\beta}_{h,i} y_{T-i}$. And then, the IMS MSFE is defined as:

$$\widehat{V}_{h,k}^{IMS} = \mathsf{E}\left[\left(y_{T+h} - \widehat{y}_{T+h,h} \right)^2 \right].$$

Note that $\{y_t\}$ can expressed as an autoregressive process of order p according to Wiener–Kolmogorov's theorem, where p can be infinite. Then, if $k \geq p$, the theoretical (for known parameters) MSFEs coincide for DMS ($V_{h,k}^{DMS}$) and IMS ($V_{h,k}^{IMS}$); but if $k < p$, the latter is larger than the former, which in turn is larger than the 'true' MSFE from the correct—potentially infinitely parameterized—model (see Bhansali, 1996). Define γ_i as the ith autocorrelation of $\{y_t\}$ for $i \geq 1$ and γ_0 its variance (in stationary processes), then using an AR(1) as a forecasting model:

$$V_{2,1}^{IMS} / V_{2,1}^{DMS} = 1 + \frac{\left[\gamma_2 - (\gamma_1)^2 \right]^2}{1 - (\gamma_2)^2} \geq 1,$$

where the equality arises when the model is well specified. Similarly it can be shown that if the process follows an MA(1) with parameter θ, $V_{2,1}^{IMS} / V_{2,1}^{DMS} = 1 + \left(\frac{\theta}{1+\theta^2} \right)^4 > 1$.

Bhansali recalls the main asymptotic findings. For a well specified model, the 1S estimation procedure is asymptotically equivalent to maximum likelihood and in the case of Gaussian processes, achieves the Cramér–Rao bound; yet this is not the case when $k \neq p$. By contrast, DMS is asymptotically inefficient for a well-specified model. However, if $k(T) \to \infty$, the distributions of the DMS and IMS estimators coincide for $T \to \infty$, under some regularity conditions (see Bhansali, 1993 when $k(T) = o\left(T^{1/3}\right)$). In analyzing the ARMA$(1,1)$ model:

$$y_t - \rho y_{t-1} = \epsilon_t - \theta \epsilon_{t-1},$$

Bhansali notes that the two-step ahead forecast is given by:

$$y_{t+2} = -\rho(\theta - \rho) \sum_{i=1}^{\infty} \theta^i y_{t-i} = \tau (1 - \theta L)^{-1} y_t, \tag{10}$$

so that he recommends to minimize the in-sample sum of squared forecast errors:

$$\sum \left(y_{t+2} - \tau (1 - \theta L)^{-1} y_t \right)^2,$$

for (τ, θ) rather than for the original parameters (ρ, θ) since it is the multi-step parameters which present an interest. Another justification is given by Stoica and Soderstrom (1984) who show that the parameter estimates $\left(\widehat{\tau}, \widehat{\theta}\right)$ are unique whereas $\left(\widehat{\rho}, \widehat{\theta}\right)$ may not be so. We therefore call the model (10), with parameters (τ, θ), the forecast generating process or FGP.

When the process to forecast or the model used is non-stationary—like the structural time series in Harvey (1993)—Haywood and Tunnicliffe-Wilson (1997) extend the work by Tiao and Xu (1993) to direct multi-step estimation of spectral densities. Bhansali reviews the different criteria which can be used for deciding on the lag length k to be used for forecasting and notes that some asymptotic efficiency can be shown for the MSFE obtained by DMS when k is treated as a random variable function of a modified AIC. Finally, the author concludes that there exists a rationale for DMS when the model is under-parameterized for the DGP or when the latter is complex or belongs to a class admitting an infinite number of parameters. He remarks also that even if a model is fitted to the data and seems to pass the traditional diagnostic tests, there might be a DMS forecast generating process which, because it explicitly allows for moving average errors, improves and robustifies the forecasting performances.

Schorfheide (2003) extends and confirms these results by presenting the case of local misspecification, whereby the disturbances exhibit serial correlation that asymptotically vanish.

Bhansali (2002) applies DMS estimation to a Monte Carlo experiment of seasonally adjusted autoregressive AR(9) time series estimated over a sample of 99 observations. The FGP was selected using the criterion in Shibata (1980), but this often led to selecting a model of order 0. The author concludes that his simulation does not seem to advocate the use of direct estimation and assumes that removing the seasonality may have damped the serial dependence of the process, or that the sample used is too small, or finally that this result may simply depend on the specific series used in the simulation.

More recently Ing (2003) has shown, when estimating a stationary AR(p) process via a misspecified AR(k) model, and when, contrary to the assumption in Bhansali (1996) of independence between the estimation sample and the realizations to forecast, that if $k \geq p$ then IMS is asymptotically more efficient than DMS (in terms of MSFE) and for both methods a lower k is more efficient as long as it is not lower than p. By contrast, Ing showed that when $k < p$, for given h

and k :

$$\lim_{T\to\infty}\left(\mathsf{MSFE}_{IMS}-\mathsf{MSFE}_{DMS}\right)>0.$$

Non parametric direct multi-step estimation was also the focus of Clements and Hendry (1996) and Chevillon and Hendry (2005). But these authors analyse estimation for forecasting rather than the design of the DMS FGP. They shed light on the dynamic properties leading direct estimation to improve accuracy and hence we present their contributions in section 8 where we review the features advocating the use of DMS.

Discussion

There is an extended literature on non-parametric DMS where the authors focus particularly on designing information criteria and on estimating long-memory time series. Results concur to show that these methods need reasonably large samples and strong time dependence, hence a recent focus of researchers on forecasting fractionally integrated time series. Ing (2003) provides an analytical justification for these asymptotic results. The multivariate framework in Liu (1996) has the disadvantage that it partitions the set of observations in non-overlapping subsets and thus loses a lot of information. It therefore cannot be used when only one forecast horizon matters, and it is not sure that estimating all horizons at once yield any better results than forecasting each separately, thus taking full advantage of the DMS framework. The definition of non-parametric DMS by Bhansali is constrained to model design or choice. He thus omits work on the *estimation* approach to DMS where the focus is not, as in the parametric approach, on the 1S parameters, but on the multi-step parameters that matter for forecasting.

7 Robustness of multi-step forecasts from ARMA models.

Tiao and Xu (1993) develop an extensive analysis of the properties of the DMS forecasts generated by an exponential smoothing formula—the FGP—which is estimated when the (true) DGP follows an ARIMA(p, $d=0$ or 1, q). They, thus, extend the results by Cox (1961) and show that multi-step estimation may be preferable. They motivate their study by comparing IMS and DMS forecasting properties for series A, from Box and Jenkins (1976): they fit an ARIMA$(0,1,1)$ model where the moving average parameter, θ, is estimated by minimizing the in-sample multi-step squared residuals implied by the exponential smoothing formula. The estimates hence depend on the forecast horizon. The authors use the mean corrected series and let the sample size vary from 101 to 157 observations. They report the ratios of the average (over the resulting 57 outcomes) squared forecast error for the IMS over those from the DMS estimation technique. The forecast horizon varies from 2 to 40 and the ratio first decreases with the lead (thus benefitting IMS) until horizon $h = 7$, and then it establishes itself between about 1.3 and 1.6. It must be noted, though, that the estimate $\widehat{\theta}_h$ increases with h and, from observation $h = 15$ onwards, it is unity, thus implying that the forecast is simply the sample average.

The authors extend the framework in Cox (1961) to a process $\{y_t\}$ which follows an ARIMA(p, d, q),

$$\phi(L)(1-L)^d y_t = \xi(L)\epsilon_t, \tag{11}$$

where $\phi(L)$ and $\xi(L)$ are polynomials—whose roots are stationary—of orders, respectively, p and q, and d is either 0 or 1. The aim is to analyse the robustness of the h-step ahead forecasts when these are obtained by the exponential smoothing formula:

$$\widehat{y}_{T+h,h} = (1-\theta_h)\sum_{t=0}^{T-1}\theta_h^t y_{T-t}, \tag{12}$$

and the forecast error is given by:

$$\widehat{e}_{T+h,h} = y_{T+h} - \widehat{y}_{T+h,h}. \tag{13}$$

The asymptotic h–step ahead MSFE is, in $\mathbb{R}\cup\{-\infty,+\infty\}$,

$$\sigma^2(h,\theta) = \lim_{T\to\infty}\mathsf{E}\left[\widehat{e}^2_{T+h,h}\right].$$

The authors show that the MSFE can be decomposed into the sum of the variance of the h-step ahead forecast errors under the 'true' model plus the squared bias introduced by the misspecification. This allows them to derive the exact formula for $\sigma^2(h,\theta)$, which exists for $\theta \in (-1,1)$, when $d = 1$, and for $\theta \in (-1,1]$, for $d = 0$.

If $\{y_t\}$ follows an ARIMA$(1,0,1)$, then:

$$y_t - \phi y_{t-1} = \epsilon_t - \xi\epsilon_{t-1},$$

which is referred to as the (ϕ, ξ) model and the forecasting model is, then, denoted by $(1, \theta)$. Tiao and Xu, then, derive the minimum of $\sigma^2(h, \theta)$, for given h, and it is obtained for:

$$\theta_h^* = \begin{cases} (1 - \sqrt{c})(\phi + c)^{-1}, & \text{for } (\phi, \xi) \in \mathbf{S}, \\ 1, & \text{otherwise,} \end{cases}$$

where $c = (1 + \xi)(\phi - \xi)(1 - \phi\xi)\left[(1 + \phi)\phi^{h-1} - 1\right]$ and S is some region of $[-1, 1] \times [-1, 1]$ which they define. Let $r(h; \phi, \xi)$ be the ratio of $\sigma^2(h, \theta_h^*)$ over the MSFE implied by the true model; it is a measure of the efficiency loss. The authors show that $r(1; \phi, \xi) < 1.2$ over a wide region around $\phi = \xi$, or when $\phi > \xi > 0$, or when $\frac{2}{3} < \phi \leq 1$; and it is moderate over a large part of the parameter space, as is often the case in empirical work, and which is one of the reasons of the widespread use of the exponential smoothing formula. When (ϕ, ξ) vary, θ_h^* is unity when ϕ is negative, or when $\xi > \phi$. As regards the behaviour with respect to h, the supremum of $r(h; \phi, \xi)$, when $\phi > \xi > 0$, is increasing the horizon but bounded as $h \to \infty$ by $4/3$. When comparing the DMS and IMS forecasting performances, the authors mention that the ratio $\sigma^2(h, \theta_1^*)/\sigma^2(h, \theta_h^*)$ is increasing in h for $\phi > \xi > 0$ and it tends to 2 as the horizon goes to infinity.

Under the general ARIMA case, in (11), Tiao and Xu then prove the consistency of the estimate $\widehat{\theta}_h(T)$ of θ_h^* obtained by minimizing $(T-h)^{-1}\sum_{t=1}^{T-h}\widehat{e}_{t+h,h}^2$, the in-sample average of the squared forecast errors: they show that, under some regularity assumptions:

$$\widehat{\theta}_h(T) \underset{T \to \infty}{\to} \theta_h^*,$$

where θ_h^* is a—the, if unique—minimizer of $\sigma^2(h, \theta)$ over $(-1, 1]$. This result extends to forecasts generated from the FGP:

$$(1 - L)^{b_1}(1 - L^s)^{b_2} y_t = (1 - \theta_1 L)(1 - \theta_2 L^s) u_t,$$

where $\{u_t\}$ is assumed to be i.i.d. Gaussian white noise, $s \geq 1$, $b_1 = 0$ or 1, $b_2 = 0$ or 1, $b_1 + b_2 > 0$, and the data generating process of the series is

$$\phi(L)(1 - L)^{d_1}(1 - L^s)^{d_2} y_t = \xi(L)\epsilon_t.$$

The FGP includes here, inter alia, the ARIMA$(0, 2, 2)$ non-stationary smooth trend model ($s = 1, b_1 = b_2 = 1$), and the multiplicative non-stationary seasonal models ($s = 12, b_1 = b_2 = 1$) and ($s = 12, b_1 = 0, b_2 = 1$).

Focusing on the dependence relation between the parameter estimates for varying forecast horizons, the authors show that if the true DGP is $(1-L)\,y_t = (1-\theta_0 L)\,\epsilon_t$, where $\epsilon_t \sim \mathsf{IN}\left(0, \sigma_\epsilon^2\right)$, then

$$T^{1/2}\left(\frac{\left[\widehat{\theta}_1(T)-\theta_0\right]}{\left[1-\widehat{\theta}_1(T)\right]^{1/2}}, \frac{\left[\widehat{\theta}_2(T)-\widehat{\theta}_1(T)\right]}{1-\widehat{\theta}_1(T)}, ..., \frac{\left[\widehat{\theta}_h(T)-\widehat{\theta}_{h-1}(T)\right]}{1-\widehat{\theta}_1(T)}\right)' \underset{T\to\infty}{\overset{\mathsf{L}}{\to}} \mathsf{N}\left[\mathbf{0}_h, \mathbf{I}_h\right]. \qquad (14)$$

This means that when the FGP and DGP coincide, the loss of efficiency from using a multi-step estimation procedure is

$$\frac{\mathsf{Var}\left[T^{1/2}\left(\widehat{\theta}_h(T)-\theta_0\right)\right]}{\mathsf{Var}\left[T^{1/2}\left(\widehat{\theta}_1(T)-\theta_0\right)\right]} = 1 + (h-1)\left(1-\theta_0^2\right), \quad \text{for } h \geq 1.$$

The results from (14) imply that multi-step estimation can be used to generate diagnostic tests. The authors suggest two of them and compare them to the Box–Ljung and Dickey–Fuller statistics. Yet, although the results seem promising in small samples, they are not decisive.

The contribution of Tiao and Xu is, thus, to show that direct multi-step estimation can lead to more efficient forecasts when the model is misspecified. Yet, when the forecasting model and the DGP coincide, it is still asymptotically preferable to use IMS in large samples since DMS leads to an efficiency loss.

Tiao and Tsay (1994) provide some theoretical and empirical considerations for the use of multi-step ("adaptive") estimation for forecasting. Their focus is on long-memory processes which can be represented by ARFIMA models. They compare the resulting forecasts to those obtained via single-step or multi-step estimation of a stationary ARIMA model:

$$(1-\alpha L)\,y_t = (1-\rho L)\,\epsilon_t,$$

where $|\alpha| < 1, |\rho| < 1$ and ϵ_t is not modeled since it is known that the FGP is misspecified for the DGP. The resulting h-step ahead forecasts and forecast errors are given by:

$$\widehat{y}_{T+h,h} = \begin{cases} \alpha y_T - \rho\epsilon_T, & \text{for } h = 1, \\ \alpha^{h-1}\widehat{y}_{T+1,1}, & \text{for } h \geq 2, \end{cases} \quad \text{and } \widehat{e}_{T+h,h} = y_{T+h} - \alpha^{h-1}\left(\alpha y_T - \rho\epsilon_T\right),$$

which imply that the variances of the forecast errors are

$$\mathsf{Var}\left[\widehat{e}_{T+h,h}\right] = \begin{cases} \sigma_\epsilon^2, & \text{for } h = 1, \\ \begin{cases} \sigma_y^2\left(1-2\alpha^h\gamma_h+\alpha^{2h}\right) + \alpha^2(h-1)\rho^2\sigma_\epsilon^2 \\ +2\alpha^{h-1}\rho\,\mathsf{Cov}\left[y_{T+h}-\alpha^h y_T, \epsilon_t\right], \end{cases} & \text{for } h \geq 2, \end{cases} \qquad (15)$$

where σ_y^2 and σ_ϵ^2 are the variances of y_t and ϵ_t, respectively, and γ_h is the lag h autocorrelation of y_t. Thus, multi-step estimation would lead to minimizing the variance in (15), and optimal values would depend on the horizon. The authors compare the Monte Carlo MSFEs from one-step and multi-step estimation to the 'true' forecast error variances obtained by using the DGP:

$$(1-L)^d y_t = u_t \text{ and } u_t \sim \mathsf{IN}\left(0, \sigma_u^2\right),$$

for the two values $d = 0.25$ and 0.45 (i.e. close to the non-stationarity coefficient of 0.5). They do not mention the sample size used for estimation, but report the forecast statistics up to 200 hundred steps ahead. Their results show that the loss from using the misspecified DMS ARIMA is never more than 5% in terms of MSFE and, in fact, almost always less than 1% when $d = 0.25$. The gain from DMS versus IMS is not significant—yet positive—when $d = 0.25$, but it is so, and rapidly increasing with the horizon, for almost non-stationary processes: it is about 6% for $h = 4$, 13% at $h = 10$, 26% at $h = 20$, 57% at $h = 50$ and 70% for $h = 100$ or 200. In practice, though, the distribution of $\{y_t\}$ is not known and (15) cannot be computed; yet the modeler can still compute some estimates by minimizing the in-sample squares of the forecast errors $\widehat{e}_{t+h,h}$ for $t = 1, ..., T$.

Tiao and Tsay then apply their method to the prediction of the differences in the series of the U.S. monthly consumer price index for food from 01/1947 to 07/1978, which have been reported in previous studies to be well modelled by an ARFIMA$(0, 0.423, 0)$ process. Using samples of 80 observations, the authors estimate the models used for the Monte Carlo and compare the resulting empirical MSFEs, computed as the average of the squared out-of-sample forecast errors. Their results strongly favour multi-step estimation of an ARIMA$(1, 1)$ over the other two techniques at all horizons and especially at large ones ($h \geq 40$). Tiao and Tsay conclude by noting that one the advantages of DMS is its estimation simplicity and the fact that it can be extended to forecast linear aggregates of future observations.

In his comment on Tiao and Tsay (1994), Peña (1994) suggests another case when multi-step estimation leads to better forecasting properties. He assumes that the DGP is such that it presents an additive outlier (unknown to the modeler):

$$x_t = z_t + \omega.1_{\{t=T\}},$$

and that Δz_t follows an AR(1) process without intercept and defines:

$$y_t = \Delta x_t = \Delta z_t + \omega \left(1_{\{t=T\}} - 1_{\{t=T+1\}}\right).$$

Assume that the autoregressive coefficient of $\{y_t\}$, α, is estimated by minimizing the in-sample multi-step forecast errors. Denote the resulting estimator by $\widehat{\alpha}_h$, where

$$\widehat{\alpha}_h = \left(\frac{\sum y_{t+h} y_t}{\sum y_t^2}\right)^{1/h} = r_h^{1/h}.$$

Therefore, in the presence of the outlier

$$\begin{aligned}
\widehat{\alpha}_h &= \left(\frac{\omega\left(z_{T+h} + z_{T-h} - z_{T+h+1} - z_{T-h+1}\right) + \sum z_{t+h} z_t}{2\omega^2 + 2\omega\left(z_T - z_{T+1}\right) + \sum z_t^2}\right)^{1/h}, \quad \text{for } h > 1, \\
\widehat{\alpha}_1 &= \left(\frac{\omega\left(z_{T+1} + z_{T-1} - z_{T+2} - z_T\right) + \sum z_{t+1} z_t}{2\omega^2 + 2\omega\left(z_T - z_{T+1}\right) + \sum z_t^2}\right),
\end{aligned}$$

and $\widehat{\alpha}_1$ is, hence, more affected by the outlier than $\widehat{\alpha}_h$, $(h > 1)$. Multi-step estimation may thus provide more robust estimates of the true parameters. The author notes also that such an estimation method can be used for diagnostic purposes.

In a comment about the computation of forecast intervals, Tsay (1993) suggests the use of multi-step (adaptive, for him) forecasting. His rationale is that all statistical models are imperfect representations of the reality and that, when it comes to forecasting, local approximations are more relevant than global ones. The two main implications of these remarks are that the maximum likelihood principle does not apply and that since "different forecast horizons have different local characteristics," different models should be fitted for each forecast. The author then considers forecasting the U.S. quarterly unemployment rate as in Chatfield (1993). The estimates are computed by minimizing the in-sample sum of squares of the multi-step residuals obtained by assuming that the data generating process can be approximated by an AR(2) model. This method results in non-linear estimation. This follows the technique used in Tiao and Tsay (1994). Tsay provides the point forecasts up to 12-step ahead together with the 95% prediction interval from the in-sample empirical distribution of the multi-step residuals. He compares his results to those obtained by fitting an AR(1) and an ARIMA$(1, 1, 0)$ model. Estimation over a sample of 48 observations leads to the AR(2):

$$y_t = .4409 + 1.5963 y_{t-1} - .6689 y_{t-2} + \epsilon_t,$$

which implies that the series is nearly integrated with a root of 0.9274. The author finds that the multi-step point forecasts are more accurate than those obtained by the other models. The prediction interval does not necessarily increase with the horizon for the "adaptive" forecast and it has the same amplitude as that of the AR(2) but, contrary to the latter, always (except at 12 steps ahead) contains the true outcome. The ARIMA model does contain the realised value too, but the forecast interval is much larger than that of the previous two models. The author concludes that multi-step estimation does indeed yield a positive outcome.

In their article, Lin and Tsay (1996) study whether using cointegrating properties improve long-term forecasting. In an empirical analysis, they compare several forecasting techniques to an 'adaptive' procedure of multi-step forecasts, which they use as a benchmark since it does not postulate the existence of a true model. They use a non-stationary VAR(p) model for the vector of n variables:

$$\mathbf{x}_t = \boldsymbol{\tau} + \sum_{i=1}^{p} \boldsymbol{\Upsilon}_i \mathbf{x}_{t-i} + \boldsymbol{\epsilon}_t,$$

where the $\boldsymbol{\epsilon}_t \sim \mathsf{IN}(\mathbf{0}, \boldsymbol{\Sigma})$. It is assumed that, letting $\boldsymbol{\Upsilon}(L) = \sum_{i=1}^{p} \boldsymbol{\Upsilon}_i L^i$, the series $\{\boldsymbol{\Psi}_i\}_{\infty}^{0}$ is such that $[\boldsymbol{\Upsilon}(L)]^{-1} = \sum_{i=0}^{\infty} \boldsymbol{\Psi}_i L^i$. The h–step ahead forecast error is then given by

$$\mathbf{e}_{T+h,h} = \sum_{i=0}^{h-1} \boldsymbol{\Psi}_i \boldsymbol{\epsilon}_{T+h-i}.$$

The DMS multi-step parameter estimates are given by minimizing the in-sample sum of the squared $\mathbf{e}_{t,h}$. This implies a non-linear function of the elements of $\{\boldsymbol{\Upsilon}_i\}_{p}^{1}$. For simplicity, Lin and Tsay suggest to simply use the least squares projection of $\mathbf{x}_t$ onto the space spanned by $(\mathbf{x}_{t-h}, ..., \mathbf{x}_{t-h-p+1})$ and a constant, for $t = h+p, ..., T$. The computing time of this alternative estimator is much lower.

Lin and Tsay compare their DMS forecasts to those obtained by cointegrated VARs for seven financial and macro-economic data sets. The vector processes are of dimension varying from 3 to 5 and are estimated over samples of 230 to 440 observations. The criterion used for analysis is the square root of the average trace of the MSFE. Their results show that multi-step techniques provide a greater forecast accuracy (up to a 60% gain), but for long horizon (beyond 50) only two of the series still exhibit a gain from using DMS. The authors find it difficult to account for these results.

Discussion

The articles summarised here provide a vast array of justifications for, and successful examples of the use of DMS methods. They confirm that three main types of model misspecification benefit direct multi-step forecasting, namely misspecified unit-roots, neglected residual autocorrelation and omitted location shifts–although the latter two can be thought of as representations of the same phenomenon. They also suggest that the success or failure of DMS can be used a model specification test. Peña shows that DMS estimates are more robust to additive outliers than one-step, but the practical use of this feature for forecasting may not be so significant in practice if indeed a shift occurs. Finally, Lin and Tsay, like Bhansali (1999), contrast the two ways to proceed with DMS estimation and forecasting: via either (i) using the same FGP for both IMS and DMS, the DMS estimates being computed by minimizing the implied in-sample h–step residuals, which can be non-linear functions of the FGP parameters; or by (ii) using a different model at each horizon where it is the multi-step parameters—defined as the coefficients from a projection of y_t on the information set up to time $(t-h)$—which are estimated.

8 When does DMS work?

Clements and Hendry (1996) develop an extended analysis of multi-step estimation for stationary and integrated processes. Their focus is on VAR(1) models as in:

$$\mathbf{x}_t = \boldsymbol{\Upsilon}\mathbf{x}_{t-1} + \boldsymbol{\epsilon}_t, \tag{16}$$

where the n-vector process $\{\boldsymbol{\epsilon}_t\}$ satisfies $\mathsf{E}[\boldsymbol{\epsilon}_t] = \mathbf{0}.$ From an end-of-sample forecast origin T:

$$\mathbf{x}_{T+h} = \boldsymbol{\Upsilon}^h\mathbf{x}_T + \sum_{i=0}^{h-1}\boldsymbol{\Upsilon}^i\boldsymbol{\epsilon}_{T+h-i}, \tag{17}$$

and the IMS and DMS forecasts are given respectively by

$$\widehat{\mathbf{x}}_{T+h} = \widehat{\boldsymbol{\Upsilon}}^h\mathbf{x}_T, \quad \text{and} \quad \mathbf{x}_{T+h} = \widetilde{\boldsymbol{\Upsilon}}_h\mathbf{x}_T,$$

where $\widehat{\boldsymbol{\Upsilon}}$ and $\widetilde{\boldsymbol{\Upsilon}}_h$ are the estimators of $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Upsilon}^h$ obtained by minimizing, respectively, the 1–step and h–step ahead in-sample forecast errors. The authors note that the relative accuracy of DMS versus IMS is given by that of the powered estimate versus the estimated power. Direct estimation of

$\widetilde{\mathbf{\Upsilon}}_h$ has, therefore, some potential when $\widehat{\mathbf{\Upsilon}}$ is badly biased for $\mathbf{\Upsilon}$, or when $\mathsf{E}[\mathbf{x}_{T+1} \mid \mathbf{x}_T] = \mathbf{\Psi}\mathbf{x}_T$ but $\mathsf{E}[\mathbf{x}_{T+h} \mid \mathbf{x}_T] \neq \mathbf{\Psi}^h\mathbf{x}_T$. However, they remark also that in stationary processes, misspecification of the DGP is not sufficient to advocate the use of DMS, since $\widehat{\mathbf{\Upsilon}}$ is the OLS and $\widetilde{\mathbf{\Upsilon}}_h$ converges towards the unconditional expectation with $\mathbf{\Upsilon}^h$ tending to zero as h increases. Hence, increasing divergence between $\widehat{\mathbf{\Upsilon}}$ and $\widetilde{\mathbf{\Upsilon}}_h$ is unlikely. Moreover, DMS is inefficient in small samples, so that if $\boldsymbol{\epsilon}_t \sim \mathsf{IN}\left(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}_n\right)$, biases are unlikely to be enough for a gain to appear. Thus, Clements and Hendry note that if $\boldsymbol{\epsilon}_t$ follows a negative moving average, there may be some potential for DMS. They derive a taxonomy of forecast errors and show that the only terms in common for both methods are those of error accumulation, namely $\sum_{i=0}^{h-1} \mathbf{\Upsilon}^i\boldsymbol{\epsilon}_{T+h-i}$ in the framework above. Simulating the small sample estimation biases, they show, for several stationary values—0, 0.4 and 0.8—of the autoregressive coefficient in a univariate AR(1) process without intercept, that for sample sizes ranging from 10 to 100, the two step ahead DMS does not yield better estimates of the powered coefficient than the squared IMS.

This result is specific to finite samples as Chevillon and Hendry (2005) show: when estimating (16) with an additional drift by OLS and (17), with a drift also, by GMM, with a HAC covariance matrix estimator, DMS is asymptotically more efficient than IMS in the case of stationary processes with positive slope. Indeed, in the univariate case, denoting by $\widehat{e}_h$ and $\widetilde{e}_h$ the IMS and DMS forecast errors at horizon h using GMM estimation, and ρ the slope coefficient, these authors show that:

$$h\left(\mathsf{E}\left[\widehat{e}_h^2\right]/\mathsf{E}\left[\widetilde{e}_h^2\right] - 1\rho\right) \underset{h\to\infty}{\to} \frac{2\rho}{\left(1-\rho^2\right)\left(2-\rho\right)},$$

the latter being of the same sign as ρ, as long as $|\rho| < 1$. In the case of integrated processes, this result collapses and IMS always dominates DMS.

A Monte Carlo analysis by Clements and Hendry, of the forecasts from the 'nonseasonal Holt–Winters Model' illustrates the relative behaviours of the IMS and DMS techniques in their framework. The data is generated by the sum of unobserved components for the trend, intercept and irregular elements:

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \\ \mu_t &= \mu_{t-1} + \beta_t + \delta_{1t}, \\ \beta_t &= \beta_{t-1} + \delta_{2t}. \end{aligned}$$

The disturbances ε_t, δ_{1t} and δ_{2t} are assumed to be normally distributed and independent through time and from one another (at all lags and leads), with zero means and variances, respectively, σ_{ε}^2, $\sigma_{\delta_1}^2$ and $\sigma_{\delta_2}^2$. This model can be reduced to an ARIMA$(0,2,2)$—or restrictions thereof—with or without a drift, or a deterministic trend. It is also possible to allow for stationary autoregressive roots:

$$(1-\tau_1 L)(1-\tau_2 L)\, y_t = \delta_{2t} + (1-L)\,\delta_{1t}.$$

The authors use six AR forecasting models: AR(2) models in levels or differences, with or without an imposed unit-root, with or without an intercept The main results are that DMS and IMS are somewhat equivalent when the model either estimates the unit-root or neglects MA components. However, when these two effects are present, there is a gain for DMS (seemingly increasing with the horizon) unless the MA term is effectively cancelled by an AR root, or when the model is under-parameterized for the DGP. The forecasts from using the pseudo-true values of the parameters under the DGP considered allow to separate model misspecification and estimation uncertainty effects. In general, the misspecification effects are constant or even decrease with the horizon, and multi-step forecasts can be more accurate in the very near future if the forecast error function is better approximated. In terms of estimation, DMS is more accurate when one (or two) unit-root is present in the DGP but not imposed in the model, and in the presence of omitted MA errors (the conjunction of both seems necessary, as opposed to either alone). A significant gain is present also when an intercept is estimated in conjunction with the other two effects, especially when the FGP is an AR(1), for which IMS fares badly. These results help explain why Stoica and Nehorai (1989) found that a model close to an ARIMA$(0,1,2)$ approximated by an AR(1) leads to improved forecasting performance when using DMS but not when the FGP is an AR(6).

The authors then focus on the driftless ARIMA$(0,1,1)$ DGP, where the MA component is omitted in the forecasting models and the unit-root is estimated. They use four estimators for the hth power of the slope in $y_t = \rho y_{t-1} + \epsilon_t$ and $\epsilon_t = \zeta_t + \theta \zeta_{t-1}$, where $\zeta_t \sim \mathsf{IN}\left(0, \sigma_\zeta^2\right)$:

$$\begin{aligned}
\left(\widehat{\rho}_{1S}\right)^h &= \left(\frac{\sum y_t y_{t-1}}{\sum_t y_{t-1}^2}\right)^h \quad \text{and} \quad \widetilde{\rho}_{DMS_h} = \frac{\sum y_t y_{t-h}}{\sum_t y_{t-h}^2}, \\
\left(\widehat{\rho}_{IV}\right)^h &= \left(\frac{\sum y_t y_{t-2}}{\sum_t y_{t-1} y_{t-2}}\right)^h \quad \text{and} \quad \widetilde{\rho}_{IVDMS_h} = \frac{\sum y_t y_{t-h-1}}{\sum_t y_{t-h} y_{t-h-1}}.
\end{aligned}$$

They show that

$$\begin{aligned}
T\left(\left(\widehat{\rho}_{1S}\right)^{h}-1\right) &\Rightarrow \left(\int_0^1 W\left(r\right)^2 dr\right)^{-1} h\left[\frac{1}{2}\left(W\left(1\right)^2-1\right)+\frac{\theta}{\left(1+\theta\right)^2}\right],\\
T\left(\left(\widehat{\rho}_{IV}\right)^{h}-1\right) &\Rightarrow \left(\int_0^1 W\left(r\right)^2 dr\right)^{-1} \frac{h}{2}\left(W\left(1\right)^2-1\right),\\
T\left(\widetilde{\rho}_{DMS_h}-1\right) &\Rightarrow \left(\int_0^1 W\left(r\right)^2 dr\right)^{-1} h\left[\frac{1}{2}\left(W\left(1\right)^2-1\right)+\frac{\theta}{h\left(1+\theta\right)^2}\right],\\
T\left(\widetilde{\rho}_{IVDMS_h}-1\right) &\Rightarrow \left(\int_0^1 W\left(r\right)^2 dr\right)^{-1} \frac{h}{2}\left(W\left(1\right)^2-1\right),
\end{aligned}$$

and provide small sample approximations of the distributions. The leftward non-centrality of IMS therefore increases with h, whereas that of DMS does not. The instrumental estimators seem better. Simulations illustrate these results.

This framework is also analysed in Chevillon and Hendry (2005) who now allow for a drift in the random walk. This induces the presence of a deterministic trend which asymptotically dominates estimation, yielding the same asymptotic accuracy for both methods. In finite sample though, disparities appear: DMS is more accurate when the drift is 'small' compared to the variance of the disturbances and when the latter exhibit negative serial correlation. Introducing the concept of 'weak' trend whereby the drift coefficient vanishes to zero asymptotically at the rate of $O\left(T^{-1/2}\right)$, Chevillon (2005b) derives asymptotic distributions where he allows for both the stochastic and deterministic trends to have an impact on estimation. The model he uses is:

$$y_t=\left(\sum_{i=0}^{h-1}\rho^i\right)\tau_T+\rho^h y_{t-h}+\varepsilon_t,\quad for\ h\geq 1,$$

where $\tau_T=\dfrac{\psi}{\sqrt{T}}$, $\mathsf{Var}[\varepsilon_t]=\sigma_\varepsilon$ and $\sigma^2=\lim_{T\to\infty}T^{-1}\mathsf{E}\left[\sum_{t=1}^{T}\varepsilon_t\right]$. The resulting IMS,$\left(\widehat{\tau}_T^{\{h\}},\widehat{\rho}_T^{h}\right)$, and DMS,$\left(\widetilde{\tau}_{h,T},\widetilde{\rho}_{h,T}\right)$, estimators are such that

$$\begin{bmatrix}\sqrt{T}\left(\widetilde{\tau}_{h,T}-\tau_{h,T}\right)\\ T\left(\widetilde{\rho}_{h,T}-1\right)\end{bmatrix}-\begin{bmatrix}\sqrt{T}\left(\widehat{\tau}_T^{\{h\}}-\tau_{h,T}\right)\\ T\left(\widehat{\rho}_T^{h}-1\right)\end{bmatrix}\Rightarrow\frac{\left(h-1\right)\theta}{\int_0^1\left[K_{\psi,\phi}\left(r\right)\right]^2dr-\left(\int_0^1K_{\psi,\phi}\left(r\right)dr\right)^2}\begin{bmatrix}\int_0^1K_{\psi,\phi}\left(r\right)dr\\ -1\end{bmatrix},\tag{18}$$

where $K_{\psi,\phi}$ is a drifting Ornstein-Uhlenbeck process defined as

$$K_{\psi,\phi}\left(r\right)=\psi f_\phi\left(r\right)+\sigma\int_0^r e^{\phi\left(r-s\right)}dW\left(s\right),$$

with $W\left(r\right)$ a Wiener process on $[0,1]$ and:

$$f_\phi\left(\cdot\right):r\to\frac{e^{\phi r}-1}{\phi}\quad\text{if }\phi\neq 0,\quad\text{and }f_0\left(r\right)=r.\tag{19}$$

The difference between the two types of estimators is a function of $\|\psi, \sigma\|$. In turn , this translates in the forecast errors which the author shows to be complex functions of the forecast horizon and parameters. Analysis of the distributions and Monte Carlo simulation prove that the weak trend framework accurately represents the finite sample behaviours and that it is the ratio ψ/σ that defines what 'finite' means in this context.

Deterministic misspecification has also been shown to benefit direct multi-step estimation. As mentioned in Chevillon and Hendry (2005), occasional breaks in the level of a trending process can generate serial correlation of the residuals from a constant parameter model and lead to the cases studied by these authors. In an unpublished paper from his doctorate thesis, Chevillon also analyses the influence of recent unnoticed breaks. He shows that DMS is more efficient at estimating the dynamic properties relevant for forecasting and that the potential occurrence of deterministic shocks hence advocate using direct methods. This aspect is confirmed in an empirical forecasting exercise for the South African GDP over 1973-2000 where a multi-step method designed by Aron and Muellbauer (2002) and variants thereof beat all of 30 rival techniques (Chevillon 2005a).

Discussion

These authors confirm, with their Monte Carlo, and prove analytically what previous authors had found in specific cases, namely that estimated unit-roots, structural breaks and omitted negative residual autocorrelation are key to the success of DMS forecasting. As opposed to some other authors, they use as a DMS model the projection of the variable onto the space spanned by its lags at and beyond h: it is the same autoregressive dynamics which is estimated. Their simulations also shed light on earlier results. The influence of small drifts is shown and it is seen that in general DMS is to be preferred when the data are—stochastically or deterministically—non-stationary or when the available sample is too small for reliable inference.

9 Direct Multi-step estimation and forecasting

9.1 Design of forecast estimators

In this section, we provide a general definition for the two types of forecasts which we have studied so far, namely the iterated one-step ahead (IMS_h) and direct h-step (DMS_h). We borrow a framework for the design of forecast estimators from Ericsson and Marquez (1998) and extend it to allow for dynamic estimation. Here, the modeler is interested in n endogenous variables, $\mathbf{x}$, and assumes that they depend on their lagged values, up to some $p \geq 0$, on some weakly exogeneous—which they actually may or may not be—variables $\mathbf{z}$ and on some vector of c parameters $\boldsymbol{\varphi}$. The model specifies some error process $\{\boldsymbol{\epsilon}_t\}$ —the distribution thereof may depend on $\boldsymbol{\varphi}$ and exhibit any form of autocorrelation, heteroscedasticity or non-stationarity—and is assumed to be valid over a sample of size $T + H$, so that there exists a n-vector function $\mathbf{f}(\cdot)$, such that:

$$\mathbf{f}\left(\mathbf{x}_t, ..., \mathbf{x}_{t-p}, \mathbf{z}_t, \boldsymbol{\varphi}, \boldsymbol{\epsilon}_t\right) = \mathbf{0}, \quad \text{for } t = p, ..., T, ..., T + H. \tag{20}$$

The sample is split into two: estimation is conducted over the first T observations and this is used to forecast the remaining H. Equation (20) describes an open model and it is convenient to transform it in a reduced closed form, solving it for $\mathbf{x}_t$. We change the time subscript t to $T + i$, and assume—under mild conditions, amongst which linearity of $\mathbf{f}(\cdot)$ is most common—that there exists a suitable transform of $\mathbf{f}(\cdot)$, denoted by $\mathbf{g}(\cdot)$, such that positive values of i represent the dates for which we wish to obtain forecasts in:

$$\mathbf{x}_{T+i} = \mathbf{g}\left(\mathbf{x}_{T+i-1}, ..., \mathbf{x}_{T+i-p}, \mathbf{z}_{T+i}, \boldsymbol{\varphi}, \boldsymbol{\epsilon}_{T+i}\right), \quad \text{for } i = p - T, ..., -1, 0, 1, ..., H. \tag{21}$$

Notice that this framework—as delineated in (21)—is quite general, and it may be the case that specific models should be restrictions thereof. For instance, if $n = 1$, $\mathbf{g}(\cdot)$ reduces to a single equation; it may also be nonlinear and the model could be static—if $p = 0$—or exclude exogenous variables.

For forecasting at horizons $i > 1$, there is a need for assumptions about the vector $\mathbf{z}_t$: either it is assumed strongly exogenous and it is possible to obtain conditional forecasts (see Engle, Hendry, and Richard, 1983), or a model for its behaviour is used, and in fact $\mathbf{z}$ is incorporated in $\mathbf{x}$. The forecasts are defined by their horizon, i, the actual variable of interest—which can be a transform of

$\mathbf{x}_{T+i}$—and the specification of its distribution, as given here by (21).[2] What values of $\mathbf{x}_{T+i-1}$, ..., $\mathbf{x}_{T+i-p}$, $\mathbf{z}_{T+i}$, $\boldsymbol{\varphi}$ and $\boldsymbol{\epsilon}_{T+i}$ are used in forecasting affects the outcome. For instance, the one-step ahead forecast, from an end-of-sample forecast origin at T, is obtained when the actual values of $\mathbf{x}_T, ..., \mathbf{x}_{T-p}$ are used in $\mathbf{g}(\cdot)$. And then, for $i > 1$, by replacing $\mathbf{x}_{T+i-1}$ in the equation with its corresponding forecast, (21) leads to 'powered-up' one-step ahead forecasts, which will be denoted by IMS and (21), specifying the parameters $\boldsymbol{\varphi}$ and the distributions of the disturbances is thus the corresponding forecast generating process, or FGP.

Alternatively it is possible to directly estimate the data generating process h steps ahead, for a fixed $h > 1$, using a transformed representation of (20). We let $\mathbf{k}_h(\cdot)$ denote a suitable transform of $\mathbf{f}(\cdot)$—possibly including some composition—such that:

$$\mathbf{x}_{T+i} = \mathbf{k}_h(\mathbf{x}_{T+i-h}, ..., \mathbf{x}_{T+i-h-p+1}, \mathbf{w}_{T+i}, \boldsymbol{\phi}_h, \boldsymbol{\nu}_{h,T+i}), \tag{22}$$
$$\text{for } i = p-1+h-T, ..., -1, 0, 1, ..., H,$$

where $\boldsymbol{\phi}_h$, a c-vector of parameters, and $\boldsymbol{\nu}_{h,t}$, a n-vector of disturbances are re-parameterizations of $\boldsymbol{\varphi}$ and $\boldsymbol{\epsilon}_t$. The r-vector $\mathbf{w}_{t+i}$ is assumed to be a transform of $\{\mathbf{z}_t\}$ which achieves a property of strong exogeneity for the parameters of (22), namely $\boldsymbol{\phi}_h$. The forecasts $\{\widetilde{\mathbf{x}}_{T+i,h};\ i = 1, ..., H\}$ obtained using $\mathbf{k}_h(\cdot)$ are the multi-step forecasts of $\{\mathbf{x}_{T+i};\ i = 1, ..., H\}$, using dynamic—or direct—estimation, the h–step DMS forecasts, generated by the DMS FGP (22). The exogeneity status of $\{\mathbf{z}_{T+i}\}$ and $\mathbf{w}_{T+i}$ may be misspecified in practice; additional uncertainty is generated when forecasts are used instead of their true realised values, especially given that their own FGPs may not coincide with their DGPs.

If the modeler knew with certainty the data generating process and it coincided with her model (20), then both IMS and DMS FGPs would provide the same forecasts. In practice, unfortunately, (20), (21) and (22) would have to be estimated and depending on which methods are used for this purpose, the estimated parameters $\widehat{\boldsymbol{\varphi}}$ and $\widetilde{\boldsymbol{\phi}}_h$,[3] will lead to different forecasts. The interdependence between *estimation* and *forecasting* is therefore intrinsic to the concept of multi-step forecasting. This, in turn leads to a forecast error taxonomy.

[2] We assume here that the econometric modeller does not *intentionally* mis-specify her model. She therefore considers it to be the data generating process (DGP).

[3] We assume here that only these parameters are estimated, and that the functional forms are part of the models, so that we do not write $\widehat{\mathbf{g}}(\cdot)$ and $\widetilde{\mathbf{k}}(\cdot)$, as would happen in the 'non-parametric' models presented in Bhansali (2002).

9.2 A general forecast-error taxonomy

We now borrow from Clements and Hendry who suggest in Clements and Hendry (1998b) and Hendry (2000) a general forecast error taxonomy which helps us in assessing the advantages of multi-step estimation. We use the framework presented above but for ease of exposition modify it slightly. Notice that in (20), $\mathbf{f}_{\mathbf{x}}(\cdot)$, if it represents the true DGP, provides the—potentially time dependent—joint density of $\mathbf{x}_t$ at time t, conditional on $\mathbf{X}_{t-1}^{t-p} = (\mathbf{x}_{t-1}, ..., \mathbf{x}_{t-p})$, and $\mathbf{z}_t$. Assume, without loss of generality, that $\{\mathbf{z}_t\}$ contains only deterministic factors—such as intercepts, trends and indicators—and that all stochastic variables are included in $\{\mathbf{x}_t\}$. As previously, it is desired to forecast $\mathbf{x}_{T+h}$, or perhaps of function thereof (e.g. if $\mathbf{z}_t$ originally contained stochastic variables), over horizons $h = 1, ..., H$, from a forecast origin at T. Now, the dynamic model does not coincide with the data generating process and it specifies the distribution of $\mathbf{x}_t$ conditional on $\mathbf{X}_{t-1}^{t-r}$, with lag length r, deterministic components $\mathbf{d}_t$ and implicit stochastic specification defined by its parameters $\boldsymbol{\psi}_t$. This model is fitted over the sample $t = 0, ..., T$, so that parameter estimates are a function of the observations, represented by:

$$\widehat{\boldsymbol{\psi}}_T = \boldsymbol{\Psi}_T\left(\widetilde{\mathbf{X}}_T^0, \mathbf{D}_T^0\right), \tag{23}$$

where $\widetilde{\mathbf{X}}$ denotes the measured data and, as before $\mathbf{D}_t^0 = (\mathbf{d}_t, ..., \mathbf{d}_0)$. A sequence of forecasts $\{\widehat{\mathbf{x}}_{T+h|T}\}$ is produced as a result. The subscript on $\widehat{\boldsymbol{\psi}}$ in (23) denotes the influence of the sample size. Let $\boldsymbol{\psi}_T^e = \mathsf{E}_T\left[\widehat{\boldsymbol{\psi}}_T\right]$, where it exists. Because the underlying densities may be changing over time, all expectation operators must be time dated. Future values of the stochastic variables are unknown, but those of deterministic variables are known; there, therefore, exists a function $\mathbf{g}_h(\cdot)$ such that

$$\widehat{\mathbf{x}}_{T+h|T} = \mathbf{g}_h\left(\widetilde{\mathbf{X}}_T^{T-r+1}, \mathbf{D}_{T+h}^{T+1}, \widehat{\boldsymbol{\psi}}_T\right). \tag{24}$$

The corresponding h–step ahead expected forecast error is, thus, the expected value of $\mathbf{e}_{T+h|T} = \mathbf{x}_{T+h} - \widehat{\mathbf{x}}_{T+h|T}$, and is given by

$$\mathsf{E}_{T+h}\left[\mathbf{x}_{T+h} - \widehat{\mathbf{x}}_{T+h|T} \mid \mathbf{X}_T^0, \{\mathbf{Z}^*\}_{T+h}^0\right],$$

where the actual values of the deterministic factors over the forecast period (including any deterministic shifts) are denoted by $\{\mathbf{Z}^*\}_{T+h}^{T+1}$ and $\{\mathbf{Z}^*\}_{T+h}^0 = \left[\{\mathbf{Z}^*\}_{T+h}^{T+1}, \mathbf{Z}_T^0\right]$; and the expectation operator is dated $T+h$ to take account of the model specification of the deterministic components between $T+1$ and $T+h$. The expectation of $\widehat{\mathbf{x}}_{T+h|T}$, conditional on the information available at T and on the assumptions made about the interval $T+1, ..., T+h$, is the *model induced* conditional expectation. Define, from an origin T, the h–step disturbance:

$$\boldsymbol{\varepsilon}_{T+h|T} = \mathbf{x}_{T+h} - \mathsf{E}_{T+h}\left[\mathbf{x}_{T+h|T} \mid \mathbf{X}_T^0, \{\mathbf{Z}^*\}_{T+h}^0\right]. \tag{25}$$

By construction, $\mathsf{E}_{T+h}\left[\boldsymbol{\varepsilon}_{T+h|T} \mid \mathbf{X}_T^0, \{\mathbf{Z}^*\}_{T+h}^0\right] = 0$ and $\boldsymbol{\varepsilon}_{T+h|T}$ is therefore an innovation against all available information. However, even for correctly-observed sample data, it is not, in general, the case that

$$\mathsf{E}_{T+h}\left[\mathbf{e}_{T+h|T} \mid \mathbf{X}_T^0, \{\mathbf{Z}^*\}_{T+h}^0\right] = 0$$

as we, now, show.

Using (25), the forecast error $\mathbf{e}_{T+h|T} = \mathbf{x}_{T+h} - \widehat{\mathbf{x}}_{T+h|T}$ from the model based on (24), can be decomposed as

$$\begin{aligned}
\mathbf{e}_{T+h|T} = \quad & + \mathsf{E}_{T+h}\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T^0, \{\mathbf{Z}^*\}_{T+h}^0\right] - \mathsf{E}_{T+h}\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T^0, \mathbf{Z}_{T+h}^0\right] && (ia) \\
& + \mathsf{E}_{T+h}\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T^0, \mathbf{Z}_{T+h}^0\right] - \mathsf{E}_{T}\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T^0, \mathbf{Z}_{T+h}^0\right] && (ib) \\
& + \mathsf{E}_{T}\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T^0, \mathbf{Z}_{T+h}^0\right] - \mathsf{E}_{T}\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T^0, \mathbf{D}_{T+h}^0\right] && (iia) \\
& + \mathsf{E}_{T}\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T^1, \mathbf{D}_{T+h}^1\right] - \mathsf{E}_{T}\left[\widehat{\mathbf{x}}_{T+h|T} \mid \mathbf{X}_T^{T-r+1}, \mathbf{D}_{T+h}^{T+1}, \boldsymbol{\psi}_T^e\right] && (iib) \\
& + \mathsf{E}_{T}\left[\widehat{\mathbf{x}}_{T+h|T} \mid \mathbf{X}_T^{T-r+1}, \mathbf{D}_{T+h}^{T+1}, \boldsymbol{\psi}_T^e\right] - \mathsf{E}_{T}\left[\widehat{\mathbf{x}}_{T+h|T} \mid \widetilde{\mathbf{X}}_T^{T-r+1}, \mathbf{D}_{T+h}^{T+1}, \boldsymbol{\psi}_T^e\right] && (iii) \\
& + \mathsf{E}_{T}\left[\widehat{\mathbf{x}}_{T+h|T} \mid \widetilde{\mathbf{X}}_T^{T-r+1}, \mathbf{D}_{T+h}^{T+1}, \boldsymbol{\psi}_T^e\right] - \widehat{\mathbf{x}}_{T+h|T} && (iv) \\
& + \boldsymbol{\varepsilon}_{T+h|T} && (v)
\end{aligned}$$

The first two rows arise from structural change affecting deterministic (*ia*) and stochastic (*ib*) components respectively; the third and fourth , (*iia*) and (*iib*), from model misspecification decomposed by deterministic and stochastic elements; the fifth (*iii*) from forecast origin inaccuracy; (*iv*) represents estimation uncertainty; and the last row, (*v*), is the unpredictable stochastic component.

When $\{\mathbf{Z}^*\}_{T+h}^0 = \mathbf{Z}_{T+h}^0$ (i.e. in the absence of deterministic shifts), then (*ia*) is zero; and, in general, the converse holds, that (*ia*) being zero entails no deterministic shifts. When $\mathsf{E}_{T+h}[\cdot] = \mathsf{E}_T[\cdot]$ (so that there are no stochastic breaks), (*ib*) is zero; but (*ib*) can be zero despite stochastic breaks, provided these do not indirectly alter deterministic terms. When the deterministic terms in the model are correctly specified, so that $\mathbf{Z}^0_{T+h} = \mathbf{D}^0_{T+h}$ then (*iia*) is zero, and again the converse seems to hold. In the case of correct stochastic specification, so that $\boldsymbol{\psi}^e_T$ summarises the effects of $\mathbf{X}^1_T$, then (*iib*) is zero; but now the converse is not true: (*iib*) can be zero in seriously misspecified models. Next, when the data are accurate (especially at the forecast origin), so that $\mathbf{X} = \widetilde{\mathbf{X}}$, (*iii*) is zero but the converse is unclear. When estimated parameters have zero variances, so that $\widehat{\mathbf{x}}_{T+h|T} = \mathsf{E}_T\left[\widehat{\mathbf{x}}_{T+h|T} \mid \widetilde{\mathbf{X}}^{T-r+1}_T, \mathbf{D}^{T+1}_{T+h}, \boldsymbol{\psi}^e_T\right]$, then (*iv*) is zero and the converse holds *almost surely.* Finally (*v*) is zero if and only if the world is non-stochastic.

Thus, the taxonomy includes elements of the main sources of forecast error, partitioning these by whether or not the corresponding expectation is zero. For there to be a gain from DMS, it must be obtained through estimation uncertainty (*iv*), possibly interacting with misspecification of deterministic or stochastic elements, (*iia*) and (*iib*). This is why the literature has shown that direct mutli-step estimation is beneficial for forecasting essentially in two contexts: when the model is misspecified for the stochastic properties of the process (omitted unit-roots) or when deterministic properties alter and go unnoticed, as in the context of breaks, which may reinforce the previous type of misspecification via induced serial correlation of the residuals or long-memory.

10 Conclusion

This paper has presented a review of the existing work on direct multi-step estimation for forecasting at varying horizons. We have show that this strain of literature has produced a vast amount of theoretical and empirical evidence favouring the use of this technique. Unfortunately, the diversity of approaches had made it difficult to draw a definite conclusion about its when's and why's. Here, we have shown that from the early contributions, the analyses have evolved towards either using DMS criteria for the design of forecasting models, or proper DMS estimation. In the light of our review, although the gain from using IMS or DMS varies with the horizon and the stochastic properties of the data, it is clear that the latter technique can be asymptotically more efficient than the former even if the model is well-specified. This result is explained by the improvement in the variance of the multi-step estimator resulting from direct estimation. It thus appears that the misspecification of the error process in the case of DMS estimation is not so detrimental to the accuracy of the estimators. However, the limiting distributions reflect only partially the estimation properties of the methods. Indeed, in finite samples, the absence of bias can never be perfectly achieved and, hence, DMS can prove a successful technique for obtaining *actual* estimates and not only for reducing the multi-step variances—and indeed with respect to the latter IMS could prove more precise. There is little hope for success for DMS in finite samples when the data are stationary and the models are well-specified. By contrast, when the models may be misspecified, DMS provides accuracy gains, both asymptotically and in finite samples. As we discussed in a general framework which allowed for a study of the various causes of forecast error, the main features that advocate DMS use are stochastic or deterministic non-stationarity The literature showed that it could originate from breaks, unit-roots, or fractional integration.

We can broadly separate the future research agenda into two categories. On the one hand, the existing trend on analyses of models and circumstances will continue. The influence of breaks need be evaluated further, in particular using the link between occasional shocks and fractional cointegration. Co-breaking—linear combinations of variables which are insensitive to the breaks—would be valuable here. Non-linear estimation and breaks that occur after the forecast origin also need more study. On the other hand, a fruitful strain revolves around model design. Recent work on the link between in-sample regressor collinearity and out-of-sample forecast performance seems an interesting route to pursue. In particular, the progress made regarding forecasting using factor analysis—when more variables than observations are available—point towards studying DMS properties since IMS is not an option in this context.

References

Allen, P. G. and R. A. Fildes (2001). Econometric forecasting strategies and techniques. In J. S. Armstrong (Ed.), *Principles of Forecasting*, pp. 303–362. Boston: Kluwer Academic Publishers.

Aron, J. and J. Muellbauer (2002). Interest rate effects on output: evidence from a GDP forecasting model for South Africa. *IMF Staff Papers 49*, 185–213.

Bhansali, R. J. (1993). Order selection for linear time series models: a review. In T. Subba Rao (Ed.), *Developments in Time Series Analysis*, pp. 50–56. London: Chapman and Hall.

Bhansali, R. J. (1996). Asymptotically efficient autoregressive model selection for multistep prediction. *Annals of the Institute of Statistical Mathematics 48*, 577–602.

Bhansali, R. J. (1999). Parameter estimation and model selection for multistep prediction of time series: a review. In S. Gosh (Ed.), *Asymptotics, Nonparametrics and Time Series*, pp. 201–225. New York, NY: Marcel Dekker.

Bhansali, R. J. (2002). Multi-step forecasting. In M. P. Clements and D. F. Hendry (Eds.), *A Companion to Economic Forecasting*, pp. 206–221. Oxford: Blackwell Publishers.

Box, G. E. P. and G. M. Jenkins (1976). *Time Series Analysis, Forecasting an Control* (2nd ed.). San Francisco, CA: Holden–Day. First published, 1970.

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business and Economic Statistics 11*(2), 121–135.

Chevillon, G. (2005a). Multi-step forecasting in emerging economies: an investigation of the south african gdp. Oxford economics working papers, no 212.

Chevillon, G. (2005b). 'Weak' trend for estimation and forecasting at varying horizons in finite samples. Oxford economics working papers, no 210.

Chevillon, G. and D. F. Hendry (2005). Non-parametric direct multi-step estimation for forecasting economic processes. *International Journal of Forecasting 21*, 201–18.

Clements, M. P. and D. F. Hendry (1996). Multi-step estimation for forecasting. *Oxford Bulletin of Economics and Statistics 58*, 657–683.

Clements, M. P. and D. F. Hendry (1998a). Forecasting economic processes. *International Journal of Forecasting 14*, 111–131.

Clements, M. P. and D. F. Hendry (1998b). *Forecasting Economic Time Series.* Cambridge: Cambridge University Press.

Clements, M. P. and D. F. Hendry (1999). *Forecasting Non-Stationary Economic Time Series.* Cambridge, MA: The MIT Press.

Cox, D. R. (1961). Prediction by exponentially weighted moving averages and related methods. *Journal of the Royal Statistical Society B 23*, 414–422.

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

Ericsson, N. R. and J. Marquez (1998). A framework for economic forecasting. *Econometrics Journal 1*, C228–C226.

Fildes, R. A. and H. O. Stekler (2002). The state of macroeconomic forecasting. *Journal of Macroeconomics 24*, 435–468.

Findley, D. F. (1983). On the use of multiple models for multi-period forecasting. *Proceedings of Business and Economic Statistics, American Statistical Association*, 528–531.

Granger, C. W. J. (1969). Prediction with a generalized cost of error function. *Operations Research Quarterly 20*, 199–207.

Haavelmo, T. (1940). The inadequacy of testing dynamic theory by comparing theoretical solutions and observed cycles. *Econometrica 8*, 312–321.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica 12*(Supplement), 1–115.

Hartley, M. J. (1972). Optimal simulation path estimators. *International Economic Review 13*, 711–727.

Harvey, A. C. (1993). *Time Series Models* (2nd ed.). Hemel Hempstead: Harvester Wheatsheaf. First edition: 1981.

Haywood, J. and G. Tunnicliffe-Wilson (1997). Fitting time series model by minimizing multistep-ahead errors: a frequency domain approach. *Journal of the Royal Statistical Society B 59*, 237–254.

Hendry, D. F. (2000). A general forecast-error taxonomy. *Econometric Society World Congress 2000 Contributed Papers 0608, Econometric Society.*

Ing, C.-K. (2003). Multistep prediction in autoregressive processes. *Econometric Theory 19*, 254–279.

Johnston, H. N. (1974). A note on the estimation and prediction inefficiency of ‘dynamic’ estimators. *International Economic Review 15*, 251–255.

Johnston, H. N., L. Klein, and K. Shinjo (1974). Estimation and prediction in dynamic econometric models. In W. Sellekaerts (Ed.), *Essays in honor of Jan Tinbergen*. London: Macmillan.

Kabaila, P. V. (1981). Estimation based on one step ahead prediction versus estimation based on multi-step ahead prediction. *Stochastics 6*, 43–55.

Klein, L. R. (1971). *An essay on the theory of economic prediction*. Chicago, IL: Markham.

Lin, J. L. and R. S. Tsay (1996). Co-integration constraint and forecasting: An empirical examination. *Journal of Applied Econometrics 11*, 519–538.

Liu, S. I. (1996). Model selection for multiperiod forecasts. *Biometrika 83*(4), 861–873.

Madrikakis, S. e. (1982). The accuracy of time series methods: the results from a forecasting competition. *Journal of Forecasting 1*, 111–153.

Mann, H. B. and A. Wald (1943). On the statistical treatment of linear stochastic difference equations. *Econometrica 11*, 173–220.

Peña, D. (1994). Discussion: Second-generation time-series model, A comment. SeeTiao and Tsay (1994), pp. 133–140.

Schorfheide, F. (2003). VAR forecasting under local misspecification. *Journal of Econometrics forhcoming*.

Shibata (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics 8*, 147–164.

Stoica, P. and A. Nehorai (1989). On multi-step prediction errors methods for time series models. *Journal of Forecasting 13*, 109–131.

Stoica, P. and Soderstrom (1984). Uniqueness of estimated k-step prediction models of ARMA processes. *Systems and Control Letters 4*, 325–331.

Tiao, G. C. and R. S. Tsay (1994). Some advances in non-linear and adaptive modelling in time-series analysis. *Journal of Forecasting 13*, 109–131.

Tiao, G. C. and D. Xu (1993). Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. *Biometrika 80*, 623–641.

Tsay, R. S. (1993). Comment: Adpative forecasting. SeeChatfield (1993), pp. 140–142.

Weiss, A. A. (1991). Multi-step estimation and forecasting in dynamic models. *Journal of Econometrics 48*, 135–149.

Weiss, A. A. (1996). Estimating time series models using the relevant cost function. *Journal of Applied Econometrics 11*, 539–560.

Weiss, A. A. and A. P. Andersen (1984). Estimating time series models using the relevant forecast evaluation criterion. *Journal of the Royal Statistical Society A147*, 484–487.