Non-Parametric Direct Multi-step Estimation for Forecasting Economic Processes

Guillaume Chevillon[*]
Economics Department, Oxford University, and OFCE, Paris.

and

David F. Hendry[†]
Economics Department and Nuffield College, Oxford University.

May, 2004

Abstract

We evaluate the asymptotic and finite-sample properties of direct multi-step estimation (DMS) for forecasting at several horizons. For forecast accuracy gains from DMS in finite samples, mis-specification and non-stationarity of the DGP are necessary, but when a model is well-specified, iterating the one-step ahead forecasts may not be asymptotically preferable. If a model is mis-specified for a non-stationary DGP, in particular omitting either negative residual serial correlation or regime shifts, DMS can forecast more accurately. Monte Carlo simulations clarify the non-linear dependence of the estimation and forecast biases on the parameters of the DGP, and explain existing results.

Keywords: Adaptive estimation, multi-step estimation, dynamic forecasts, model mis-specification.
JEL *Classification*: C32, C51, C53.

[*] Corresponding e-mail: guillaume.chevillon@bnc.ox.ac.uk.
Corresponding address: Observatoire Français des Conjonctures Economiques, 69 Quai d'Orsay, 75340 Paris Cedex 07, France.

[†] We are grateful to John Muellbauer, Bent Nielsen, Neil Shephard, Valérie Chauvin, to seminar participants at Nuffield College, Oxford, and at the 2003 Congress of the Irish Economic Association and to two anonymous referees for helpful comments and suggestions.
Financial support from the ESRC to the second author under grant L138251009 is gratefully acknowledged.

1 Introduction

Modelling and forecasting are distinct tasks because congruent, causal models need not forecast better in practice than non-congruent or non-causal: see Clements and Hendry (1999), Allen and Fildes (2001) and Fildes and Stekler (2002). Rather, causal models often suffer forecast failure, as the widespread use of adjustments techniques, such as intercept corrections, reveals (see e.g. Clements and Hendry, 1998). Explanations for forecasting less well than 'naive' formulations (e.g., 'no change') mainly rest on model mis-specification interacting with irregularities in the economy (especially structural breaks and regime changes from legislative, technological or political shifts), and entail a need for robustness: the M3 competitions emphasize this aspect (see e.g., Fildes and Ord, 2002).

When forecasting at several horizons, 'direct multi-step estimation' (DMS) is intuitively appealing and offers a potential route to alleviating some of these difficulties: estimate a model which minimizes the desired multi-step function of the in-sample errors, thereby matching design and evaluation. One-step estimation is the standard procedure of minimizing the squares of the one-step ahead residuals, from which multi-step forecasts are obtained by 'iterated multi-step' (denoted here by IMS).

The idea of multi-step estimation has a long history: Klein (1971) and Johnston, Klein, and Shinjo (1974) suggested that DMS could be more efficient than IMS, following an idea originally applied by Cox (1961) to exponentially-weighted moving-average (EWMA) or integrated moving-average (IMA(1,1)) models. Johnston (1974) put a temporary end to that literature by concluding that, if the model is correctly specified and a quadratic loss function is used as an estimation efficiency criterion, the latter itself constitutes a "reliable indicator of prediction efficiency" so DMS is less efficient than IMS.[1] In discussing the Wharton model, however, he noted that DMS forecasts seemed more robust to the mis-specification of the error processes. Findley (1983), Weiss and Andersen (1984), and Weiss (1991) considered DMS criteria for ARIMA models, and the AR(1) process in particular. They showed that if the model was mis-specified and the loss function quadratic in the multi-step forecast errors, then the asymptotically optimal estimators (in the sense of minimizing the appropriate criteria) depended on the lead period in forecasting. They found, however, that the small-sample properties of DMS varied with the degree of mis-specification, and that the conclusion in Johnston (1974) about the relative efficiencies of IMS and DMS still held. Tsay (1993) and Lin and Tsay (1996) showed that some forms of empirical model mis-specification could justify the use of DMS (also called 'adaptive' estimation). These findings were recently corroborated by Kang (2003) and Marcellino, Stock, and Watson (2004) in extensive empirical studies where estimation was carried out on stationary variables (i.e. first differences in the case of integrated series) but forecasting

[1] This result was based on the findings by Haavelmo (1944), in the case of a scalar first-order autoregressive process, and Hartley (1972).

performance was evaluated for the level of the series. Other authors applied direct estimation to alternative approaches: Haywood and Tunnicliffe-Wilson (1997), for instance, found that it brought significant estimation improvements in the frequency domain for an IMA$(1, 1)$ process. When the number of lags of an AR(k) model is chosen by either a one-step or a multi-step criterion and the true DGP follows a stationary AR(p), Bhansali (1996, 1997) and Ing (2003) have shown that DMS can outperform IMS when the model is misspecified with $k < p$.

One intuition behind DMS is, therefore, that a model which is mis-specified for the data generating process (DGP) need not be a satisfactory forecasting device. However, mis-specification is insufficient: predictors like constant growth are mis-specified but robust. Here, the desired robustness is to mis-specification of the error process. Clements and Hendry (1996) investigated model mis-specifications which might sustain DMS, and found that for unit-root processes, neglected negative moving-average errors provided a rationale at short horizons. In stationary processes, DMS could enhance forecast accuracy, but gains fade rapidly as the horizon increase. Bhansali (1999) presents a detailed overview of recent developments in the DMS literature: he distinguishes between the 'parametric' and 'non-parametric' approaches. With the former class, 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 mis-specified. In contrast, and this is the method which has led to fewer analyses and which we study here, non-parametric DMS focuses on the parameters of a different model at each horizon.

The general properties of IMS and DMS can help gauge their implications for multi-period forecasting when DMS is the only available technique. For example, IMS is infeasible when there are more variables, n, than observations, T, so a factor analysis (or principal components) is carried out separately for each forecast horizon (see, e.g., Stock and Watson, 1999). Additionally, when some regressors have marginal distributions that are difficult to model and even harder to forecast—so are treated as 'exogeneous'—DMS avoids incorporating unreliable forecasts of such regressors into forecasts of variables of interest. We show below that, when the model is well-specified for a DGP which is either stationary or integrated of order 1—noted I(1)—DMS can improve upon IMS in terms of forecast accuracy. Moreover, some model mis-specification benefits DMS. The cases which we find favourable comprise estimated unit-roots with neglected negative residual autocorrelation, perhaps caused by occasional and discrete alterations in the growth rate of an integrated series (e.g., turning points in business cycles).

In this article, we build upon Clements and Hendry (1996) and extend their analysis. Section 2 defines DMS, and its benefits are then analysed in the next two sections, when the model is well specified and either stationary (section 3) or integrated (section 4) respectively. Then section 5 discusses the properties of DMS for forecasting from mis-specified models, focusing on neglected residual autocorrelation as in Clements and Hendry (1996), but now allowing for a deterministic term which we show alters the relative properties of IMS and DMS. A Monte Carlo simulation—in section 6—illustrates our results. Section 7 shows how our findings shed light on the existing literature and, finally, section 8 concludes. Appendices detailing the proofs are available from the authors on request.

2 Direct multi-step estimation

2.1 The data generating process

The properties of DMS depend on the type of DGP and potential mis-specifications considered. To clarify results, we use simple models which embody the salient characteristics found in empirical studies, so vector autoregressive models (VAR) comprise the class of DGPs for analysis, and more specifically a VAR(1), since any vector ARMA model can be re-written in this form. Consider the dynamic system for a vector of n variables $\mathbf{x}_t$:

$$\mathbf{x}_t = \boldsymbol{\tau} + \boldsymbol{\Gamma}\mathbf{x}_{t-1} + \boldsymbol{\epsilon}_t, \quad \text{for} \quad t \geq 1, \tag{1}$$

where the only requirement is that $\mathsf{E}[\boldsymbol{\epsilon}_t] = \mathbf{0}$. If $\boldsymbol{\tau} \neq 0$, the process has an intercept (or a drift): it can then be re-written in companion form as:

$$\begin{pmatrix} 1 \\ \mathbf{x}_t \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \boldsymbol{\tau} & \boldsymbol{\Gamma} \end{pmatrix} \begin{pmatrix} 1 \\ \mathbf{x}_{t-1} \end{pmatrix} + \begin{pmatrix} 0 \\ \boldsymbol{\epsilon}_t \end{pmatrix}, \quad \text{or} \quad \mathbf{X}_t = \boldsymbol{\Psi}\mathbf{X}_{t-1} + \boldsymbol{\mathcal{E}}_t. \tag{2}$$

Data processes are defined by their deterministic and stochastic properties, as known since Wold (1938). In economics, the former are restricted—after suitable transforms of the data—to the empirically relevant cases of a non-zero mean and/or a linear trend. In terms of stochastic properties, we examine here series that are either stationary or integrated of first order.

2.2 Estimation for forecasting

From an end-of-sample observation T (the forecast origin) the process $\{\mathbf{X}_t\}$ is determined h periods ahead by:

$$\mathbf{X}_{T+h} = \boldsymbol{\Psi}^h \mathbf{X}_T + \sum_{i=0}^{h-1} \boldsymbol{\Psi}^i \boldsymbol{\mathcal{E}}_{T+h-i}. \tag{3}$$

A forecast for the variable $\mathbf{x}_t$ is to be made at various future horizons, using the model (2), either from an IMS or an h-step DMS.

First, the 1-step estimator is defined by:[2]

$$\widehat{\boldsymbol{\Psi}} = \underset{\boldsymbol{\Psi} \subseteq \mathbb{R}^{(n+1)\times(n+1)}}{\operatorname{argmin}} \left| \sum_{t=1}^{T} \left(\mathbf{X}_t - \boldsymbol{\Psi}\mathbf{X}_{t-1}\right)\left(\mathbf{X}_t - \boldsymbol{\Psi}\mathbf{X}_{t-1}\right)' \right|. \tag{4}$$

[2] We denote by $\tilde{}$ the estimators for varying horizon h, and reserve $\hat{}$ for one-step ahead estimation. The same notation is used for the corresponding forecasts: $\hat{}$ for IMS and $\tilde{}$ for DMS.

The corresponding h-step ahead IMS forecasts are given by $\widehat{\mathbf{X}}_{T+h} = \widehat{\mathbf{\Psi}}^h \mathbf{X}_T$, with average conditional error:

$$\mathsf{E}\left[\left(\mathbf{X}_{T+h} - \widehat{\mathbf{X}}_{T+h}\right) \mid \mathbf{X}_T\right] = \left(\mathbf{\Psi}^h - \mathsf{E}\left[\widehat{\mathbf{\Psi}}^h\right]\right)\mathbf{X}_T,$$

where we neglect the dependence of the estimators on the latest observations. We refer to the deviation $\mathbf{\Psi}^h - \mathsf{E}[\widehat{\mathbf{\Psi}}^h]$ as the multi-step bias of the IMS estimator.[3]

The h-step DMS estimator is defined by the least-squares projection:

$$\widetilde{\mathbf{\Psi}}_h = \underset{\mathbf{\Psi}_h \subseteq \mathbb{R}^{(n+1)\times(n+1)}}{\operatorname{argmin}} \left| \sum_{t=h}^{T} \left(\mathbf{X}_t - \mathbf{\Psi}_h \mathbf{X}_{t-h}\right)\left(\mathbf{X}_t - \mathbf{\Psi}_h \mathbf{X}_{t-h}\right)' \right|, \tag{5}$$

with h-step DMS forecasts $\widetilde{\mathbf{X}}_{T+h} = \widetilde{\mathbf{\Psi}}_h \mathbf{X}_T$, and average conditional error:

$$\mathsf{E}\left[\left(\mathbf{X}_{T+h} - \widetilde{\mathbf{X}}_{T+h}\right) \mid \mathbf{X}_T\right] = \left(\mathbf{\Psi}^h - \mathsf{E}\left[\widetilde{\mathbf{\Psi}}_h\right]\right)\mathbf{X}_T.$$

Thus, the relative forecast accuracy depends on how accurate $\widehat{\mathbf{\Psi}}^h$ and $\widetilde{\mathbf{\Psi}}_h$ are as estimators of $\mathbf{\Psi}^h$ (i.e., the powered estimate *versus* the estimated power). When $\widehat{\mathbf{\Psi}}$ is badly biased, its powered values diverge increasingly from $\mathbf{\Psi}^h$, whereas, ideally, if the biases in $\widetilde{\mathbf{\Psi}}_h$ are independent of h, the dynamic estimator will produce 'good' forecasts.

3 Forecasting by well-specified stationary models

When the model is well-specified for the DGP, one-step estimation is more efficient than multi-step, essentially because of the smaller sample size available for DMS and the autocorrelation of the multi-step disturbances. Nevertheless, DMS might still be more efficient for estimating the parameters of interest, namely the *multi-step parameters*, so we compare the asymptotic distributions of the DMS and IMS parameters when model and DGP coincide for a VAR(1). As noted by Johnston (1974), in such a framework, estimation efficiency is a reliable indicator of forecast accuracy based on the mean-square forecast error (MSFE) criterion.[4]

3.1 Direct multi-step GMM estimation

Under the conditions of (1), the relationship between $\mathbf{x}_t$ and its hth lag is given by:

$$\mathbf{x}_t = \mathbf{\Gamma}^{\{h\}}\boldsymbol{\tau} + \mathbf{\Gamma}^h \mathbf{x}_{t-h} + \sum_{i=0}^{h-1} \mathbf{\Gamma}^i \boldsymbol{\epsilon}_{t-i}, \quad \text{for} \quad t \geq h, \tag{6}$$

[3] We only condition the expectation on $\mathbf{X}_T$, not on $\{\mathbf{X}_t\}_T^0 = \{\mathbf{X}_0, ..., \mathbf{X}_T\}$, because the expectation conditional on the whole sample can be difficult to establish in practice, although $\mathsf{E}[\widehat{\mathbf{\Psi}} | \{\mathbf{X}_t\}_T^0] = \widehat{\mathbf{\Psi}}$.

[4] The multivariate proofs are available on request from the first author.

where for any integer h and stable scalar or square matrix ρ:

$$\rho^{\{h\}} = \sum_{i=0}^{h-1} \rho^i = (1-\rho)^{-1}(1-\rho^h).$$

DMS focuses on:

$$\mathbf{x}_t = \boldsymbol{\tau}_h + \boldsymbol{\Gamma}_h \mathbf{x}_{t-h} + \mathbf{u}_{h,t}, \tag{7}$$

where, if $\boldsymbol{\epsilon}_t \sim \mathsf{IN}_n [\mathbf{0}, \boldsymbol{\Omega}_\epsilon]$, $\mathbf{u}_{h,t}$ is an MA$(h-1)$ process. Then, in univariate cases, generalized least-squares (GLS) or maximum likelihood (ML) should be preferred to OLS as estimation methods; in a multivariate framework, generalized method of moments (GMM), full-information maximum likelihood (FIML) or non-linear techniques are required. We rewrite (7) as:

$$\mathbf{x}_t = \boldsymbol{\beta}_h \mathbf{X}_{t-h} + \mathbf{u}_{h,t} = \left(\mathbf{X}'_{t-h} \otimes \mathbf{I}_n\right) \ vec\left(\boldsymbol{\beta}_h\right) + \mathbf{u}_{h,t},$$

where $\mathbf{X}'_t = (1 : \mathbf{x}'_t)$, $\boldsymbol{\beta}_h = (\boldsymbol{\tau}_h : \boldsymbol{\Gamma}_h)$, $\mathsf{V}[\mathbf{u}_{h,t}] = \sum_{i=0}^{h-1} \boldsymbol{\Gamma}^i \boldsymbol{\Omega}_\epsilon \boldsymbol{\Gamma}^{i\prime}$ and '$\otimes$' represents the Kronecker product. Below, we compute the GMM estimator of $vec(\boldsymbol{\beta}_h)$ and its asymptotic distribution.

Since $\mathsf{E}[(\mathbf{X}_{t-h} \otimes \mathbf{I}_n)\, \mathbf{u}_{h,t}] = \mathbf{0}$, we use the objective function:

$$\min_{\widetilde{\boldsymbol{\beta}}_h} \left(\mathbf{m}_T\left(\widetilde{\boldsymbol{\beta}}_h\right)' \widetilde{\mathbf{W}}_{h,T}\ \mathbf{m}_T\left(\widetilde{\boldsymbol{\beta}}_h\right) \right), \tag{8}$$

where, for $T_h = T - h + 1$:

$$\mathbf{m}_T\left(\widetilde{\boldsymbol{\beta}}_h\right) = T_h^{-1} \sum_{t=h}^{T} \left(\mathbf{X}_{t-h} \otimes \mathbf{I}_n\right) \left(\mathbf{x}_t - \left(\mathbf{X}'_{t-h} \otimes \mathbf{I}_n\right)\ vec\left(\boldsymbol{\beta}_h\right)\right) = T_h^{-1} \sum_{t=h}^{T} \mathbf{m}_t\left(\widetilde{\boldsymbol{\beta}}_h\right).$$

The actual estimator depends on how $\widetilde{\mathbf{W}}_{h,T}$ is computed. Following Newey and West (1987), we use the heteroscedasticity- and autocorrelation-consistent covariance matrix estimator (HAC) given by:

$$\widetilde{\mathbf{W}}_{h,T}^{-1} = \widehat{\boldsymbol{\Upsilon}}_{0,T} + \sum_{i=1}^{h-1} \left[\left(1 - \frac{i}{h}\right) \left(\widehat{\boldsymbol{\Upsilon}}_{i,T} + \widehat{\boldsymbol{\Upsilon}}'_{i,T}\right) \right],$$

where $\widehat{\boldsymbol{\Upsilon}}_{i,T}$ is the estimator of the autocovariance of $(\mathbf{X}_{t-h} \otimes \mathbf{I}_n)\, \mathbf{u}_{h,t}$ defined as:

$$\widehat{\boldsymbol{\Upsilon}}_{i,T} = T_h^{-1} \sum_{t=h+i}^{T} \left(\mathbf{X}_{t-h} \otimes \mathbf{I}_n\right) \widehat{\mathbf{u}}_{h,t} \widehat{\mathbf{u}}'_{h,t-i} \left(\mathbf{X}'_{t-h-i} \otimes \mathbf{I}_n\right).$$

Two options are available for computing $\widehat{\mathbf{u}}_{h,t}$: either as the residual from OLS estimation of (7) or defined as $\mathbf{x}_t - \widetilde{\boldsymbol{\beta}}_h \mathbf{X}_{t-h}$ when a non-linear minimizing algorithm is used for (8). Here, we assume that $\widehat{\mathbf{u}}_{h,t}$ is the OLS residual, defined as:

$$\widehat{\mathbf{u}}_{h,t} = \mathbf{x}_t - \widehat{\boldsymbol{\beta}}_h \mathbf{X}_{t-h}, \tag{9}$$

with:

$$\widehat{\boldsymbol{\beta}}_h = \left(\sum_{t=h}^{T} \mathbf{x}_t \mathbf{X}'_{t-h}\right)\left(\sum_{t=h}^{T} \mathbf{X}_{t-h}\mathbf{X}'_{t-h}\right)^{-1}.$$

Here, the weighting matrix $\widetilde{\mathbf{W}}_{h,T}$ does not depend on the GMM estimator so $vec\left(\widetilde{\boldsymbol{\beta}}_h\right)$ is obtained by differentiating (8):

$$\frac{\partial \mathbf{m}_T\left(\widetilde{\boldsymbol{\beta}}_h\right)}{\partial vec\left(\widetilde{\boldsymbol{\beta}}_h\right)'}\widetilde{\mathbf{W}}_{h,T}\left[\mathbf{m}_T\left(\widetilde{\boldsymbol{\beta}}_h\right)\right] = \mathbf{0},$$

and GMM and OLS coincide. In vectorized form:

$$vec\left(\widetilde{\boldsymbol{\beta}}_h\right) = \left[\left(\sum_{t=h}^{T} \mathbf{X}_{t-h}\mathbf{X}'_{t-h}\right)^{-1} \otimes \mathbf{I}_n\right]\sum_{t=h}^{T}\left(\mathbf{X}_{t-h}\otimes \mathbf{I}_n\right)\mathbf{x}_t.$$

However, as the disturbances are autocorrelated, the implied asymptotic distribution of $\widetilde{\boldsymbol{\beta}}_h$ depends on the limit of $\widetilde{\mathbf{W}}_{h,T}$, which, in turn, varies with the stochastic properties of the data. For stationary processes, the GMM estimator is consistent and:

$$\sqrt{T_h}\ vec\left(\widetilde{\boldsymbol{\beta}}_h - \boldsymbol{\beta}_h\right) \xrightarrow{\mathsf{L}} \mathsf{N}_{n(n+1)}\left[\mathbf{0}, \mathbf{V}_{\widetilde{\boldsymbol{\beta}}_h}\right],$$

where, under stationarity and ergodicity:

$$\mathbf{V}_{\widetilde{\boldsymbol{\beta}}_h} = \left(\mathsf{E}\left[\mathbf{X}_t\mathbf{X}'_t\right]^{-1}\otimes \mathbf{I}_n\right)\left(\operatorname*{plim}_{T\to\infty}\widetilde{\mathbf{W}}^{-1}_{h,T}\right)\left(\mathsf{E}\left[\mathbf{X}_t\mathbf{X}'_t\right]^{-1}\otimes \mathbf{I}_n\right),$$

which in the univariate case, is given by (using ρ for $\boldsymbol{\Gamma}$ in the scalar case):

$$\begin{aligned}\mathbf{V}_{\widetilde{\boldsymbol{\beta}}_h} &= -\frac{1}{(1-\rho)^2}\left(h\rho^{2h} - \frac{1+\rho^2}{1-\rho^2} + \frac{2\rho^2\left(1-\rho^{2h}\right)}{h\left(1-\rho^2\right)^2}\right)\left[\begin{array}{cc}\tau^2 & -\tau(1-\rho)\\ -\tau(1-\rho) & (1-\rho)^2\end{array}\right]\\ &\quad+\frac{1}{(1-\rho)^2}\left(1+\rho^{2h} - \frac{2\rho\left(1-\rho^{2h}\right)}{h\left(1-\rho^2\right)}\right)\left[\begin{array}{cc}\sigma^2_\epsilon & 0\\ 0 & 0\end{array}\right].\end{aligned} \tag{10}$$

3.2 Iterated multi-step estimators

We now compare the asymptotic distributions of DMS to those obtained by IMS. If the data are stationary, the asymptotic distribution of $vec\left(\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}\right)$ can be derived, following Hamilton (1994, pp 298–9), as:

$$\sqrt{T}\left(\begin{array}{c}\widehat{\boldsymbol{\tau}}_{OLS} - \boldsymbol{\tau}\\ vec\left(\widehat{\boldsymbol{\Gamma}}_{OLS} - \boldsymbol{\Gamma}\right)\end{array}\right) \xrightarrow{\mathsf{L}} \mathsf{N}\left[\left(\begin{array}{c}\mathbf{0}_{n\times 1}\\ \mathbf{0}_{n^2\times 1}\end{array}\right), \left(\begin{array}{cc}\boldsymbol{\Sigma}_{\boldsymbol{\tau},\boldsymbol{\tau}} & \boldsymbol{\Sigma}_{\boldsymbol{\Gamma},\boldsymbol{\tau}}\\ \boldsymbol{\Sigma}_{\boldsymbol{\tau},\boldsymbol{\Gamma}} & \boldsymbol{\Sigma}_{\boldsymbol{\Gamma},\boldsymbol{\Gamma}}\end{array}\right)\right], \tag{11}$$

where:

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{\tau},\boldsymbol{\tau}} &= \left(1 + \boldsymbol{\tau}'\left(\mathbf{I}_n - \boldsymbol{\Gamma}'\right)^{-1}\left(\sum_{i=0}^{\infty}\boldsymbol{\Gamma}^i\boldsymbol{\Omega}_\epsilon\boldsymbol{\Gamma}^{i\prime}\right)^{-1}\left(\mathbf{I}_n - \boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\tau}\right)\boldsymbol{\Omega}_\epsilon; \\
\boldsymbol{\Sigma}_{\boldsymbol{\tau},\boldsymbol{\Gamma}} &= \boldsymbol{\Sigma}'_{\boldsymbol{\Gamma},\boldsymbol{\tau}} = -\left(\left(\sum_{i=0}^{\infty}\boldsymbol{\Gamma}^i\boldsymbol{\Omega}_\epsilon\boldsymbol{\Gamma}^{i\prime}\right)^{-1}\left(\mathbf{I}_n - \boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\tau}\otimes\boldsymbol{\Omega}_\epsilon\right); \\
\boldsymbol{\Sigma}_{\boldsymbol{\Gamma},\boldsymbol{\Gamma}} &= \left(\sum_{i=0}^{\infty}\boldsymbol{\Gamma}^i\boldsymbol{\Omega}_\epsilon\boldsymbol{\Gamma}^{i\prime}\right)^{-1}\otimes\boldsymbol{\Omega}_\epsilon.
\end{aligned}$$

Using the delta method, we derive their asymptotic distributions in the multivariate case, which in a univariate framework becomes:

$$\sqrt{T}\begin{pmatrix}\widehat{\rho^{\{h\}}\tau} - \rho^{\{h\}}\tau \\ \widehat{\rho}^h - \rho^h\end{pmatrix} \xrightarrow{\mathsf{L}} \mathsf{N}_2\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\widehat{\sigma}_{\tau\tau}^{IMS} & \widehat{\sigma}_{\tau\rho}^{IMS} \\ \widehat{\sigma}_{\tau\rho}^{IMS} & \widehat{\sigma}_{\rho\rho}^{IMS}\end{pmatrix}\right] = \mathsf{N}_2\left[\mathbf{0}, \boldsymbol{\Sigma}_{IMS}^{\{h\}}\right], \qquad (12)$$

defined as:

$$\begin{aligned}
\widehat{\sigma}_{\tau\tau}^{IMS} &= \left(\frac{1-\rho^h}{1-\rho}\right)^2\sigma_\epsilon^2 + \tau^2 h^2\frac{1+\rho}{1-\rho}\rho^{2(h-1)}; \\
\widehat{\sigma}_{\tau\rho}^{IMS} &= \widehat{\sigma}_{\tau\rho}^{IMS} = -\tau h^2\left(1+\rho\right)\rho^{2(h-1)}; \\
\widehat{\sigma}_{\rho\rho}^{IMS} &= h^2\left(1-\rho^2\right)\rho^{2(h-1)}.
\end{aligned}$$

We compare both methods in the next subsection by means of a numerical analysis.

3.3 Numerical comparisons of efficiency

The above results allow us to proceed to a comparison of forecast accuracy as measured by the two methods' asymptotic MSFEs, given that the DMS forecast errors are:

$$\widetilde{e}_{T+h|T} = \tau_h - \widetilde{\tau}_h + \left(\rho^h - \widetilde{\rho}_h\right)x_T + \sum_{i=0}^{h-1}\rho^i\epsilon_{T+h-i},$$

so that, taking account of the asymptotic independence of x_T and $\left(\rho^h - \widetilde{\rho}_h\right)$ under the condition of stationarity, the MSFE converges to:

$$\begin{aligned}
\mathsf{E}\left[\widetilde{e}_h^2\right] &= \mathsf{Avar}\left[\tau_h - \widetilde{\tau}_h\right] + \frac{\tau^2}{\left(1-\rho\right)^2}\mathsf{Avar}\left[\rho^h - \widetilde{\rho}_h\right] \\
&\quad +2\frac{\tau}{1-\rho}\mathsf{Acov}\left[\tau_h - \widetilde{\tau}_h, \rho^h - \widetilde{\rho}_h\right] + \frac{1-\rho^{2h}}{1-\rho}\sigma_\epsilon^2 \\
&= \frac{\sigma_\epsilon^2}{\left(1-\rho\right)^2}\left\{2 - \rho\left(1-\rho^{2h}\right) - \frac{2\rho\left(1-\rho^{2h}\right)}{h\left(1-\rho^2\right)}\right\},
\end{aligned}$$

where we notice that the dependence in τ has disappeared. Similarly:

$$\mathsf{E}\left[\widehat{e}_h^2\right] = \frac{\sigma_\epsilon^2}{\left(1-\rho\right)^2}\left(1-\rho^h\right)\left(2-\rho\left(1+\rho^h\right)\right).$$

Hence the ratio of MSFEs:

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

which is independent of τ, but as $h \to \infty$, when $|\rho| < 1$:

$$h\left(\mathsf{E}\left[\widehat{e}_h^2\right]/\mathsf{E}\left[\widetilde{e}_h^2\right]-1\right) \underset{h\to\infty}{\to} \frac{2\rho}{\left(1-\rho^2\right)\left(2-\rho\right)}. \tag{13}$$

When the horizon tends to infinity both methods yield the same MSFEs, but with DMS dominating IMS in terms of forecast accuracy when the slope is positive, $\rho > 0$, and the converse being true for $\rho < 0$.

At finite horizons, the same results hold: namely that positive slopes benefit DMS and negative ones IMS. Figure 1 presents the logarithm of the ratio of MSFEs as a function of both the horizon and the—positive—slope ($0 \leq \rho \leq 0.95$). The gain from direct estimation is increasing in the slope parameter but, for fixed ρ, it first increases to some slope-dependent horizon, then decreases. The value of h for which the maximum of the log ratio of MSFEs is reached is itself increasing in ρ, and from (13), the log ratio tends to zero as h tends to infinity. We have not presented the cases corresponding to negative slopes as they are the converse of figure 1, but with the respective roles of IMS and DMS reversed and some additional non-linearities related to whether the horizon is odd or even (which renders the graphs difficult to read). To observe the variances of the estimators, figure 2 exhibits the ratios of the IMS and DMS slope and intercept estimator variances. For ρ significantly different from unity, as the horizon increases, multi-step slope estimation becomes much more accurate using IMS, whereas both methods yield similar results for the multi-step intercept. By contrast, a slope ρ closer to unity benefits DMS (until some slope-dependent horizon).

Thus, although the gain from using IMS or DMS varies with the horizon and the stochastic properties of the data, 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 mis-specification 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, the estimators are, here, asymptotically unbiased and the gain from DMS is achieved in terms of MSFEs via a reduction in the variances of the estimators. Since the MSFE consists of the sum of the squared expectation of the forecast error and of its variance, if DMS obtains parameter estimates closer to their true values in finite samples, this method can outperform IMS, even if the latter achieves a lower variance of the estimators, and hence of the forecast errors. Nevertheless, the simulations in Clements and Hendry (1996) were unfavourable to DMS for stationary data and well-specified models, so we next compare these estimation methods for integrated data.

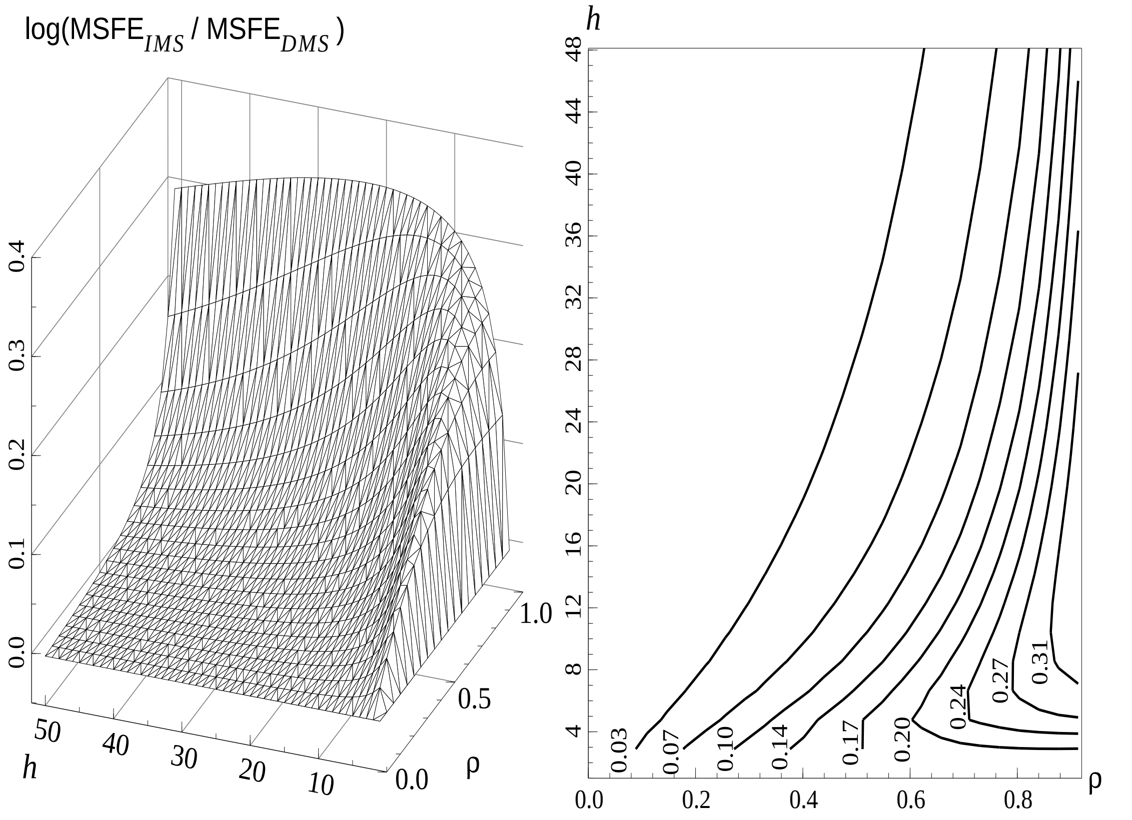


Figure 1: Ratio of the asymptotic MSFEs of the IMS over the DMS as fonction of the horizon h and the slope ρ when $\tau = 0$. The right-hand side panel represents sets of same-altitude contours of the left-hand side graphs.

4 Integrated processes

When $\mathbf{\Gamma} = \mathbf{I}_n$ so $\mathbf{x}_t \sim \mathsf{I}(1)$, the distributions differ from those presented above. We first assume that $\boldsymbol{\tau} = \mathbf{0}$ and define the n-variate Brownian motion $\mathbf{B}(r)$ such that $T^{-1/2} \sum_{t=0}^{T} \boldsymbol{\epsilon}_t \Rightarrow \mathbf{B}(r)$, where '$\Rightarrow$' denotes weak convergence of the associated probability measure. Let $\mathbf{P}$ be a square matrix of order n such that $\mathbf{P}\mathbf{P}' = \mathbf{\Omega}_\epsilon$ (e.g., the Choleski decomposition of $\mathbf{\Omega}_\epsilon$), and $\mathbf{U}(r)$ represent n-dimensional standard Brownian motion, then:

$$\mathbf{B}(r) = \mathbf{P}\mathbf{U}(r) \quad \text{where} \quad T^{-1/2} \sum_{t=0}^{T} \mathbf{u}_t \Rightarrow h\mathbf{P}\mathbf{U}(r). \tag{14}$$

The estimators are consistent with the asymptotic variance-covariance matrix:

$$\mathsf{Avar}\left[\begin{array}{c} \sqrt{T_h}\left(\widetilde{\boldsymbol{\tau}}_h - \mathbf{0}\right) \\ T_h \ vec\left(\widetilde{\mathbf{\Gamma}}_h - \mathbf{I}_n\right) \end{array}\right] = \left[h^2 + 2(h-1)\right] \mathbf{D} \left(\begin{array}{cc} \left(\mathbf{I}_n \otimes \mathbf{\Omega}_\epsilon\right) & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\ vec\left(\mathbf{\Omega}_\epsilon\right)\ vec\left(\mathbf{\Omega}_\epsilon\right)' \end{array}\right) \mathbf{D}'$$

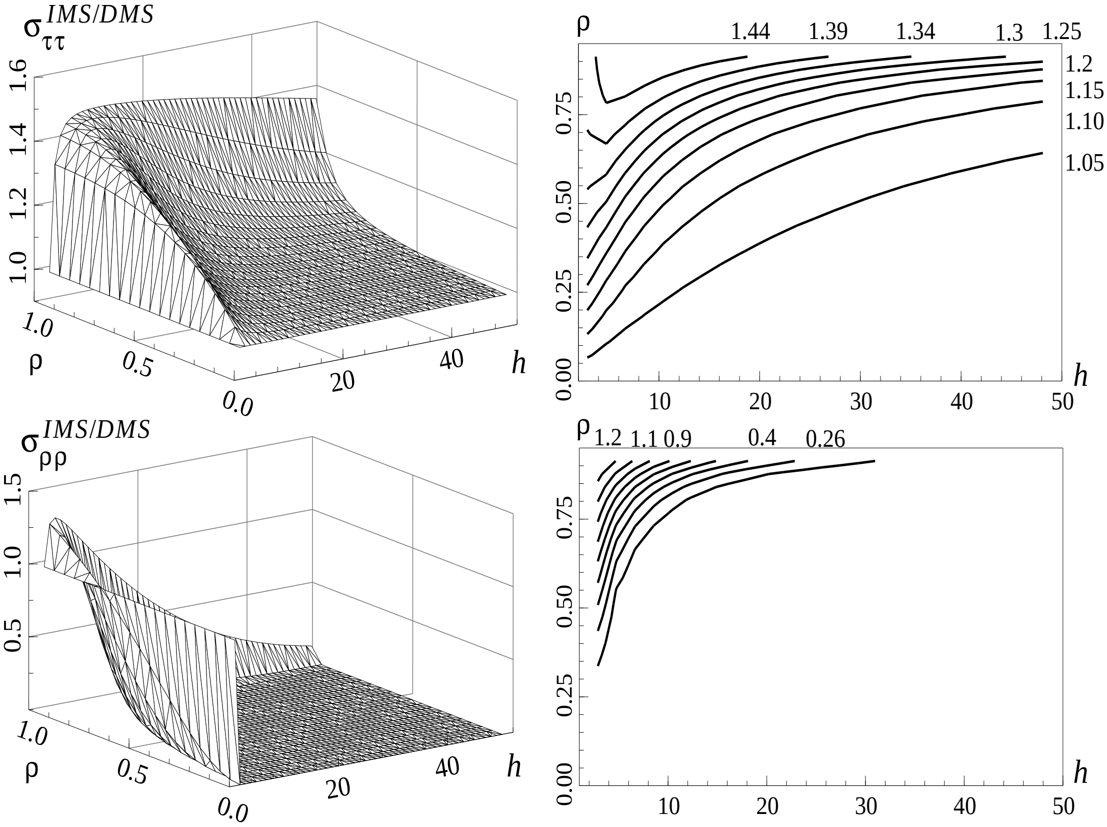


Figure 2: Ratio of the asymptotic variances of the IMS over the DMS intercept and slope estimators as fonction of the horizon h and the slope ρ when $\tau = 0$. The right-hand side panels represent sets of same-altitude contours of the left-hand side graphs.

(15)

where:

$$\mathbf{D} = \left(\begin{pmatrix} 1 & \int_0^1 \mathbf{B}(r)' \, \mathrm{d}r \\ \int_0^1 \mathbf{B}(r) \, \mathrm{d}r & \int_0^1 \mathbf{B}(r) \mathbf{B}'(r) \, \mathrm{d}r \end{pmatrix}^{-1} \otimes \mathbf{I}_n \right).$$

By contrast, the IMS estimators are such that:

$$\begin{pmatrix} T^{-1/2}\mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & T^{-1}\mathbf{I}_{n^2} \end{pmatrix} \begin{pmatrix} \sqrt{T}\left(\widehat{\boldsymbol{\tau}}_{\{h\},OLS} - h\boldsymbol{\tau}\right) \\ T\ vec\left(\widehat{\boldsymbol{\Gamma}}^h_{OLS} - \mathbf{I}_n\right) \end{pmatrix}$$

$$\Rightarrow h\left(\begin{pmatrix} 1 & \int_0^1 \mathbf{B}(r)' \,\mathrm{d}r \\ \int_0^1 \mathbf{B}(r)\,\mathrm{d}r & \int_0^1 \mathbf{B}(r)\,\mathbf{B}'(r)\,\mathrm{d}r \end{pmatrix}^{-1} \otimes \mathbf{I}_n\right)$$

$$\times \begin{pmatrix} (\mathbf{I}_n \otimes \mathbf{P})\ \ vec\,(\mathbf{U}(1)) \\ (\mathbf{P} \otimes \mathbf{P})\ \ vec\left(\left[\int_0^1 \mathbf{U}(r)\,d\mathbf{U}'(r)\right]'\right) \end{pmatrix}.$$

The asymptotic variances of the estimators differ by a ratio of $1+2\,(h-1)\,/h^2$. This ratio is unity for $h=1$, and is always larger than unity at longer horizons. Consequently, the asymptotic variance of DMS is always above that of the IMS.

So far, we have assumed that no variable in $\{\mathbf{x}_t\}$ drifts. If this condition is not satisfied, the results are substantially modified, since several I(1) regressors create deterministic trends so estimation suffers, as there are up to n perfectly-correlated variables (the $\boldsymbol{\tau}\,(t-h)$) on the right-hand side. It is, therefore, necessary to rotate the data. Although the presence of a non-zero drift radically alters the rates of convergence of the estimators, we do not derive the results for that case here, but will consider it in the next section, in a univariate framework to clarify the key features.

5 Forecasting from mis-specified non-stationary models

We now show that it is possible for DMS to lead to a substantial improvement in forecast accuracy at all horizons relative to IMS in non-stationary processes, and for the gain from DMS in terms of MSFE to increase with the lead time. The DGP, as in Clements and Hendry (1996), is an integrated process with a negative moving average which is omitted in the models, and we show that the maximum gain arises in the case of a low drift with a magnitude often found in empirical studies.

5.1 Motivation

We focus on the special case of $\boldsymbol{\Gamma} = \mathbf{I}_n$ in (1) with disturbances given by:

$$\boldsymbol{\epsilon}_t = \boldsymbol{\theta}(L)\boldsymbol{\zeta}_t,$$

where $\boldsymbol{\zeta}_t \sim \mathsf{IN}_n\left[\mathbf{0}, \boldsymbol{\Sigma}_\zeta\right]$, L denotes the lag operator and $\boldsymbol{\theta}(L) = \mathbf{I}_n + \boldsymbol{\Theta}L$. This formulation is motivated by the 'structural time-series models' of Harvey (1993, pp. 120–125) for a univariate integrated DGP represented by:

$$\begin{aligned} x_t &= \mu_t + \epsilon_t, &\text{(16a)}\\ \mu_t &= \tau_t + \mu_{t-1}, &\text{(16b)}\\ \tau_t &= \tau + \eta_t, &\text{(16c)} \end{aligned}$$

where $\epsilon_t \sim \mathsf{N}\left[0, \sigma_\epsilon^2\right]$ and $\{\eta_t\}$ is an *iid* stochastic process with zero expectation and constant variance σ_η^2. Then $\{\eta_t\}$ is a source of stochastic variation in the deterministic trend of $\{x_t\}$ and, together with τ, forms the 'local' trend of the process. The process $\{x_t\}$ can alternatively be modelled and estimated in the following congruent form:

$$x_t = \alpha + x_{t-1} + \xi_t + \theta\xi_{t-1}, \quad \text{where} \quad \xi_t \sim \mathsf{N}\left[0, \sigma_\xi^2\right]. \tag{17}$$

These models are observationally equivalent provided that $\alpha = \tau$, $\theta = \frac{1}{2}(\sqrt{q^2 + 4q} - 2 - q)$ and $\sigma_\xi^2 = -\sigma_\epsilon^2/\theta$, with noise-to-signal ratio $q = \sigma_\eta^2/\sigma_\epsilon^2$ (keeping only the invertible MA parameter). Under these conditions, $0 \leq q \leq \infty$ corresponds to $-1 \leq \theta \leq 0$, and as shown below, are favourable to DMS forecasting, which performs best (relative to IMS) when θ is close to -1 (as happens for low noise q).

The error $\{\eta_t\}$ is a shock to the level of the otherwise trend-stationary process $\{x_t\}$. Owing to the integrated nature of $\{\mu_t\}$, an occasional blip η_t entails a shift in the trend of x_t. The varying trend hence reflects variations in the growth rate of the series. Thus, any such shifts can also be regarded as the turning points of a business cycle. The fact that a modeller may be liable to omit such a variable as η_t justifies the use of DMS: as seen in Hendry (2000), location shifts often occur in macro-econometric analyses, but growth rate changes can be difficult to detect.

5.2 DGP and models

We consider a scalar process for simplicity and use the IMA(1,1), with $x_0 = 0$ and $\theta \in (-1, 1)$ in:

$$\begin{aligned} \mathsf{DGP} \;:\; & x_t = \tau + x_{t-1} + \epsilon_t, &\text{(18)}\\ & \epsilon_t = \zeta_t + \theta\zeta_{t-1}, \end{aligned}$$

and $\zeta_t \sim \mathsf{IN}[0, \sigma_\zeta^2]$. We postulate the following models (the estimators coincide for $h = 1$):

$$\begin{aligned} \mathsf{M}_{IMS} \;:\; & x_t = \alpha + \rho x_{t-1} + \varepsilon_t, &\text{(19)}\\ \text{where} \quad & \varepsilon_t \underset{c}{\sim} \mathsf{IN}\left[0, \sigma_\varepsilon^2\right] \text{ (conjectured)},\\ \mathsf{M}_{DMS} \;:\; & x_t = \tau_h + \rho_h x_{t-h} + v_t, &\text{(20)}\\ \text{where} \quad & v_t \text{ is not modelled and } h \geq 1. \end{aligned}$$

Notice that M_{IMS} is a model of the DGP and thus, mis-specified whereas M_{DMS} does not claim to reflect the DGP: it is merely a procedure, although we still use the word 'model'.

Then, (3) becomes $x_{T+h} = x_T + h\tau + \sum_{i=0}^{h-1} \epsilon_{T+h-i}$, with corresponding forecasts:

$$\mathsf{M}_{IMS} \quad : \quad \widehat{x}_{T+h} = \widehat{\rho}^{\{h\}}\widehat{\alpha} + \widehat{\rho}^{h} x_T, \qquad \text{and} \tag{21}$$

$$\mathsf{M}_{DMS} \quad : \quad \widetilde{x}_{T+h} = \widetilde{\tau}_h + \widetilde{\rho}_h x_T. \tag{22}$$

When $\tau = 0$ and the intercept is not estimated, Banerjee, Hendry, and Mizon (1996) have shown that $\widetilde{\rho}_h$ is asymptotically more (respectively less) accurate than $\widehat{\rho}^h$ if θ is negative (resp. positive). By contrast, the presence of a non-zero drift means that IMS and DMS estimators share the same asymptotic distribution: both $(T^{1/2}(\widetilde{\tau}_h - h\tau), T^{3/2}(\widetilde{\rho}_h - 1))'$ and $(T^{1/2}\left(\widehat{\rho}^{\{h\}}\widehat{\alpha} - h\tau\right), T^{3/2}\left(\widehat{\rho}^{h} - 1\right))'$ converge in law towards

$$\mathsf{N}_2\left[\mathbf{0}, h^2\begin{pmatrix} 4 & -6/\tau \\ -6/\tau & 12/\tau^2 \end{pmatrix}(1+\theta)^2\sigma_\zeta^2\right] \tag{23}$$

The distributions differ in finite samples, though since the conditional moments are non-constant, owing to the presence of a stochastic trend and an omitted moving-average component:

$$\mathsf{E}\left[x_{t+h} \mid x_t\right] = x_t + h\tau + \theta\mathsf{E}\left[\zeta_t \mid x_t\right].$$

Since:

$$\zeta_T \mid x_T \sim \mathsf{N}\left[\frac{x_T - T\tau}{1 + (1+\theta)^2(T-1)}, \sigma_\zeta^2\frac{(1+\theta)^2(T-1)}{1+(1+\theta)^2(T-1)}\right], \tag{24}$$

this leads to the non-stationary conditional moments:

$$\mathsf{E}\left[x_{t+h} \mid x_t\right] = \tau\left(h - \frac{\theta}{\dfrac{1}{t} + \dfrac{t-1}{t}(1+\theta)^2}\right) + \left(1 + \frac{\theta}{1+(1+\theta)^2(t-1)}\right)x_t, \tag{25}$$

$$\mathsf{V}\left[x_{t+h} \mid x_t\right] = \sigma_\zeta^2\left(1 + \theta^2 + (1+\theta)^2(h-1) - \frac{\theta^2}{1+(1+\theta)^2(t-1)}\right). \tag{26}$$

These converge asymptotically since, for $\theta > -1$:

$$\begin{aligned} \lim_{t\to\infty}\mathsf{E}\left[(x_{t+h} - x_t) \mid x_t\right] &= \tau\left(h - \frac{\theta}{(1+\theta)^2}\right) \\ \lim_{t\to\infty}\mathsf{V}\left[x_{t+h} \mid x_t\right] &= \left(h(1+\theta)^2 - 2\theta\right)\sigma_\zeta^2. \end{aligned}$$

The interaction between the stochastic and deterministic trends in small samples—where their influences have similar magnitudes—therefore affects estimation. Given that the estimated parameters correspond to (25), we expect that, when θ is negative, one should under-estimate the unit-root and over-estimate the intercept. The negative asymptotic covariance of the biases in (23) reinforces this. Such a mis-estimation converts the intercept from a 'drift' term to an 'equilibrium mean' of the (pseudo-) stationary estimated process. The behaviours of the estimators for the two methods are therefore non-linear and non-monotonic in the parameters of the DGP and the horizon. In such a setting, DMS is more robust to unmodelled residual autocorrelation, as in Hall (1989). The next section analyses the relation between estimation and multi-period forecasting.

Table 1: h–step ahead forecast errors for IMS estimators.

$\widehat{e}_{T+h}$	$=$	$-\sum_{i=0}^{h-1} C_i\tau - C_h x_T$	*(slope estimation)*
		$-h\delta$	*(intercept estimation)*
		$-\delta\sum_{i=0}^{h-1} C_i$	*(second-order error)*
		$+\sum_{i=0}^{h-1} \epsilon_{T+h-i}$	*(error accumulation)*

Table 2: h–step forecast errors for DMS estimators.

$\widetilde{e}_{T+h}$	$=$	$-\ \lambda_h x_T$	*(slope estimation)*
		$-\ \delta_h$	*(intercept estimation)*
		$+\ \zeta_{T+h} + \sum_{i=1}^{i=h-1}(1+\theta)\zeta_{t+h-i} + \theta\zeta_T$	*(error accumulation)*

5.3 Forecast accuracy

We consider the gains from matching the criterion used for estimating the forecasting model with that for evaluating forecast accuracy (which coincide for MSFE). A drawback from using MSFEs lies in the non-invariance of the resulting rankings to transforms of the variables in the models (see Clements and Hendry, 1998, pp. 68–73), but since DMS seeks different estimators for each combination, any results hold, anyway, only for the specific data formulation. Denoting the estimation errors, $C_i = \widehat{\rho}^i - 1$, $\lambda = \widehat{\rho} - 1$, and $\delta = \widehat{\alpha} - \tau$, we define the IMS forecast error:

$$\widehat{e}_{T+h} = x_{T+h} - \widehat{x}_{T+h}, \tag{27}$$

where $\widehat{x}_{T+h}$ is given by (21). A simplified forecast-error taxonomy for IMS follows, as shown in table 1 (see Clements and Hendry, 1998, pp. 248–250).

A similar analysis for the direct multi-step estimation procedure provides its forecast error taxonomy as reported in Table 2, with $\lambda_h = \widetilde{\rho}_h - 1$, and $\delta_h = \widetilde{\tau}_h - h\tau$.

The difference between DMS and IMS lies in the absence, for the former, of a second-order error, although it would appear in both conditional MSFEs. The interactions between estimation and error mis-specification are also similar for the two models. However, to emphasize the differences between the two methods, consider the following crude approximation to the components in $O_p\left(T^{-1}\right)$ (i.e., the slope and intercept estimators, and their direct interaction). Consider first:

$$\begin{aligned}\sum_{i=0}^{h-1} C_i &= \sum_{i=0}^{h-1} (1+\lambda)^i - h = \sum_{i=0}^{h-1} (1+i\lambda) - h + O_p\left(\lambda^2\right) \\ &= \frac{h(h-1)}{2}\lambda + O_p\left(\lambda^2\right).\end{aligned}$$

Thus, an approximation to $\mathsf{O}_p(\lambda)$ leads to:

$$\begin{aligned}\widehat{e}_{T+h} &\simeq -h\left(\lambda\left[x_T + \frac{h-1}{2}\tau\right] + \delta\left[1 + \frac{h-1}{2}\lambda\right]\right) + \sum_{i=0}^{h-1} \epsilon_{T+h-i}, \\ \widetilde{e}_{T+h} &\simeq -\left(\lambda_h x_T + \delta_h\right) + \sum_{i=0}^{h-1} \epsilon_{T+h-i}.\end{aligned}$$

Notice three main differences between the methods: first, there is a potential gain from multi-step if the corresponding estimation biases are lower than h times their one-step counterparts, i.e., $\delta_h < h\delta$ and $\lambda_h < h\lambda$ (in absolute value for the MSFE); and, more significantly, IMS estimation is equivalent to mis-estimating the forecast origin, and this effect increases with the forecast horizon (because of the additional $\frac{h-1}{2}\tau$ for IMS). Moreover, a third difference arises in the interaction between the estimated slope and intercept, via the correction in $\frac{h-1}{2}\delta\lambda$ for IMS, whose absolute value is also increasing in h. Hence, there are many channels through which DMS can be more accurate than IMS. However, this approximation may be simplistic, so the next section compares the two methods in greater detail, by numerically analyzing the accuracy of the forecasts from M_{IMS} and M_{DMS}.

6 Monte Carlo

One set of Monte Carlo simulations illustrates the preceding analyses for $\sigma_\zeta^2 = 1$ (without loss of generality) based on $10,000$ replications.[5] The following parameters are allowed to vary: τ between -3 and 3; $\theta \in (-1,1)$; with the horizon $h = 1,...,4$. For each set of parameters, all model coefficients are estimated, but the results are not presented for negative intercepts when the sign of τ has no impact. We also only report simulations for $T = 25$ as the gains from using DMS in mis-specified models fade as the sample size increases. The notation $\mathsf{E}_{MC}[\cdot]$ refers to the Monte Carlo mean. When presenting Monte Carlo results, the usual assumptions have to be made concerning the existence of any moments simulated. In the case of a well-specified AR(1) model with an estimated intercept, Magnus and Pesaran (1989) show that the forecast horizon h must not exceed $(T-3)/2$ for the MSFE to exist. We could potentially estimate these moments up to a forecast horizon of $h = 11$ when the sample is $T \geq 25$. The possibility that the slope

[5] Computations were performed using GiveWin (for graphs), OxEdit, OxRun and the Ox programming language. Figure panels are referred to as a to d left to right, top to bottom.

estimate is above unity, causing the forecast error to explode, could lead to tails too thick for the variance to exist. However, as seen in Sargan (1982) and Hendry (1991), the non-existence of the moments does not preclude sensible Monte Carlo results, which can be seen as analogous to Nagar (1959) approximations. When the error term exhibits negative autocorrelation, the probability that estimation might generate an explosive root is reduced, so we can more confidently make use of simulation.

6.1 Model Estimation

Since the aim of this section is to analyse the relation between estimation and forecast accuracy, we first study the performance of IMS for estimation of the one-step parameters, and then show how this translates into multi-step IMS and DMS.

6.1.1 One-step ahead estimation bias

Figure 3 presents the simulation results for one-step ahead estimation biases. Pattern similarities can be observed for the slope and intercept biases: first, when the true drift is zero, negative residual serial correlation leads to large negative biases. Secondly, the biases do not seem to depend much on θ when this is positive. The major difference lies in the influence of low values of the drift when θ is close to -1: whereas $\mathsf{E}_{MC}\left[\widehat{\rho}-1\right]$ is strictly increasing in both θ and τ, $\mathsf{E}_{MC}\left[\widehat{\tau}-\tau\right]$ is first increasing then decreasing in τ, and the maximum is achieved for a value of the drift which increases with θ. Notice, also, that a drift close to zero is under-estimated whereas it is over-estimated for values higher than 0.1. Panel d in fig. 3 exhibits quasi-circular contour plots in the neighborhood of $\{\tau, \theta\} = \{0, -1\}$, implying that the bias is close to a quadratic function of τ and $(1+\theta)$ in this neighbourhood.

6.1.2 Iterated and direct multi-step estimation biases

Figure 4 presents the Monte Carlo means of the IMS and DMS estimates of the slope at horizon $h = 4$. For all θ, higher drifts lead to better estimation. However, when $\{\tau, \theta\}$ is close to $\{0, -1\}$, the estimated slope is close to zero, so the estimated models imply that the observed variable is almost white noise; this is the case analyzed by Hall (1989). Figure 5 shows a pattern similar to that of 1-step estimates, except that the biases are comparatively much higher for DMS when $\theta \geq 0$, and for IMS when $\theta \leq 0$. Also, a maximum (local in the case of DMS, global for IMS) is achieved for θ largely negative and τ ranging between 0.15 and 0.40.

Thus, in the case of IMS, the non-linearity of the multi-step intercept bias results from a combination of one-step estimation uncertainties in both the slope and the drift. As we show below, this translates directly into a larger MSFE.

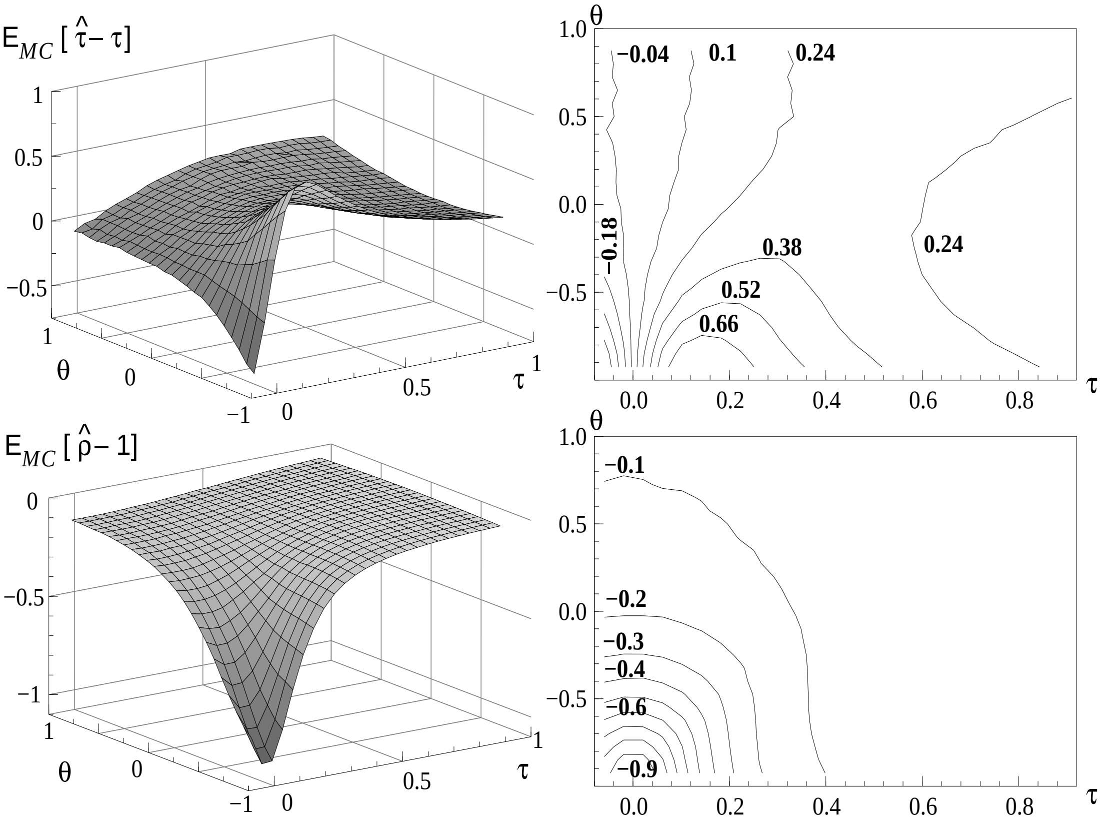


Figure 3: Monte Carlo estimates of the slope and intercept estimator biases for 1-step estimation for a sample of $T = 25$ observations, 10,000 replications and varying drift and moving average coefficient. The right-hand-side panels (*b* and *d*) exhibit a set of contours for the panels on their left. The lines join points at the same altitude (z-axis) on the 3D graphs.

6.2 Mean Square Forecast Errors

Given the non-linearities in the estimation biases, we now analyze how they affect the forecasting properties of the models. Figure 6 exhibits the Monte Carlo means of the 4-step ahead unconditional MSFEs for the two models and response surfaces for the parameters. First, in panels *b* and *d*, for non-zero θ and τ, DMS entails a lower MSFE than IMS. The striking feature is that MSFE is generally increasing in θ for DMS, which means that a more mis-specified model will forecast better, but this is not true for IMS when the drift is greater than 0.1, as the IMS MSFE surface is saddle shaped: it is increasing in θ for $\theta \geq -0.5$, decreasing elsewhere; increasing in τ for value smaller than about 0.3, but decreasing for higher values. Further, comparing fig. 6*a*, to fig. 5*a*–*c* reveals a similar pattern for IMS: for a moving-average coefficient largely negative and a low non-zero

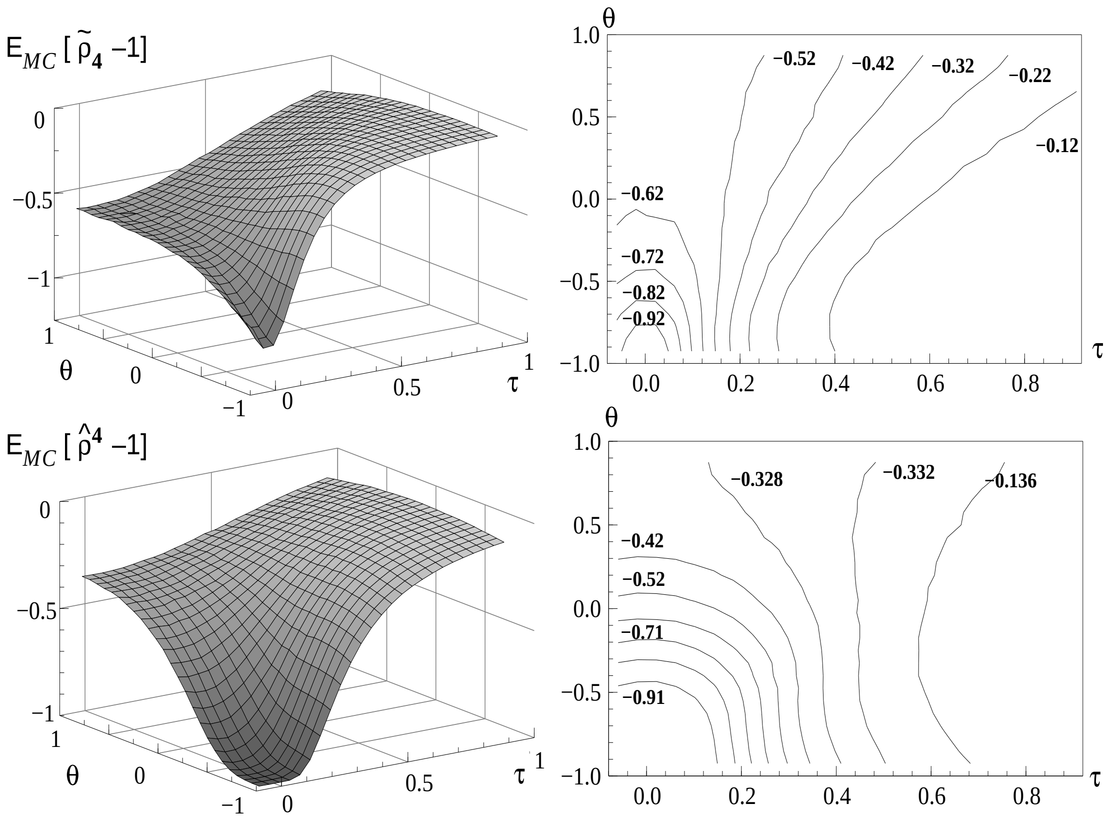


Figure 4: Monte Carlo estimates of the slope estimator biases for 4-step IMS and DMS for a sample of $T = 25$ observations, $10,000$ replications and varying drift and moving average coefficient. The right-hand-side panels (*b* and *d*) exhibit a set of contours for the panels on their left. The lines join points at the same altitude (z-axis) on the 3D graphs.

drift, the impact of iterating 1-step ahead estimates becomes large compared to direct estimation as the horizon increases. Finally, for any given θ, the MSFEs for a zero-mean process (i.e., $\tau = 0$) are lower than for any non-zero value of the intercept. Figure 7*c*–*d* record the Monte Carlo 1-step ahead MSFE for positive as well as negative values of the moving-average coefficient: notice how similar its behaviour is to the 4-step MSFE_{DMS} with respect to τ (i.e., barely dependent on the intercept, when the latter remains low), but decreasing in θ for negative values thereof, whereas the second moments are increasing at higher horizons. The MSFE seems to behave as an increasing function of θ^2, except for a zero intercept, where it hardly varies with negative values of the moving-average coefficient.

Figure 7*a*–*b*, summarizes some of the previous remarks: it shows the logarithm of the ratio of the MSFEs, positive values corresponding to a gain from using DMS. Large

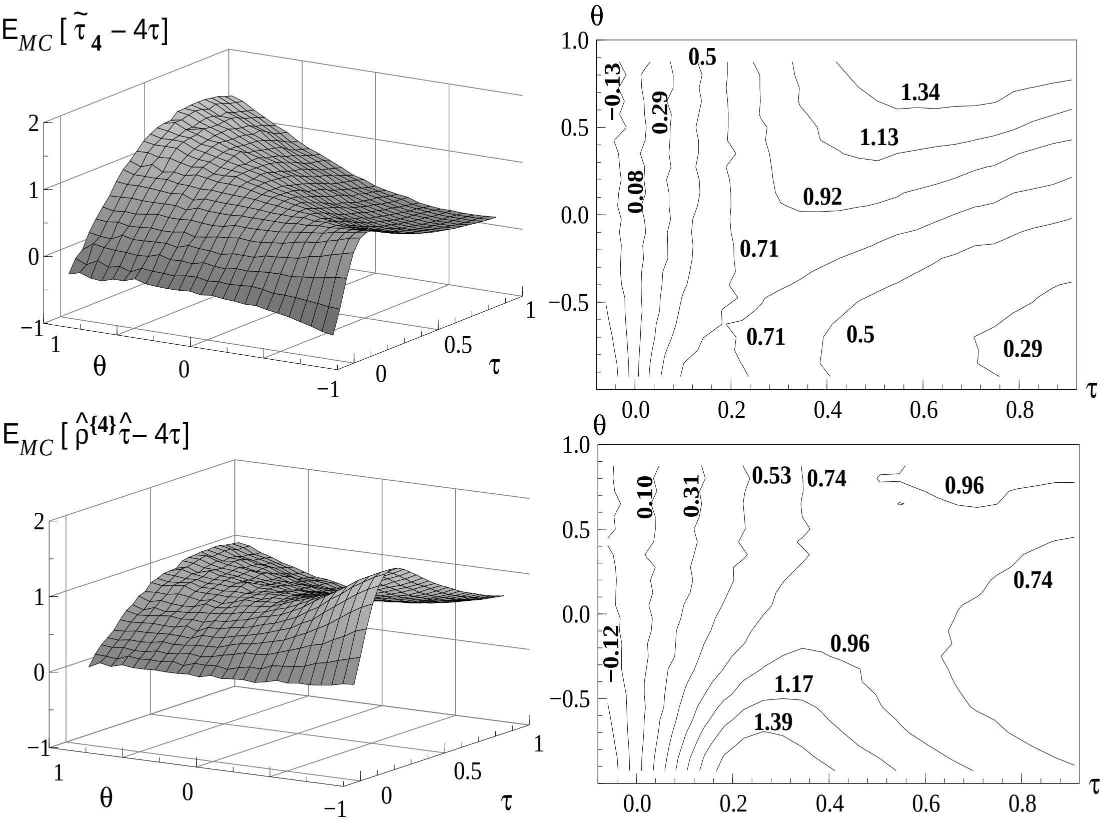


Figure 5: Monte Carlo estimates of the intercept estimator biases for 4-step IMS and DMS for a sample of $T = 25$ observations, $10,000$ replications and varying drift and moving average coefficient. The IMS implied bias (panels *c*–*d*) is that which a modeller would obtain, given that the estimated slope is strictly less than unity. The right-hand-side panels (*b* and *d*) exhibit a set of contours for the panels on their left. The lines join points at the same altitude (z-axis) on the 3D graph.

negative moving-average coefficients strongly favour DMS, even more so if combined with a low non-zero intercept. The gain can be substantial for the largest values of the log-ratio: as much as 0.8, which means that MSFE_{IMS} is 220% of MSFE_{DMS}, and is increasing in the forecast horizon. As seen on fig. 7*b*, the gain from using DMS at horizon $h = 4$, is obtained for θ lower than about -0.3.

6.3 Simulation conclusions

The conclusion that can be drawn from the Monte Carlo simulations is that when the model is reasonably well-specified, estimation of the parameters of the mean of $x_t|x_{t-h}$ is more accurate by IMS than by DMS. By contrast, a large negative moving average (here, with $\sigma_\zeta^2 = 1$, $\theta \leq -0.3$) is beneficial to DMS, and even more so when the drift is low,

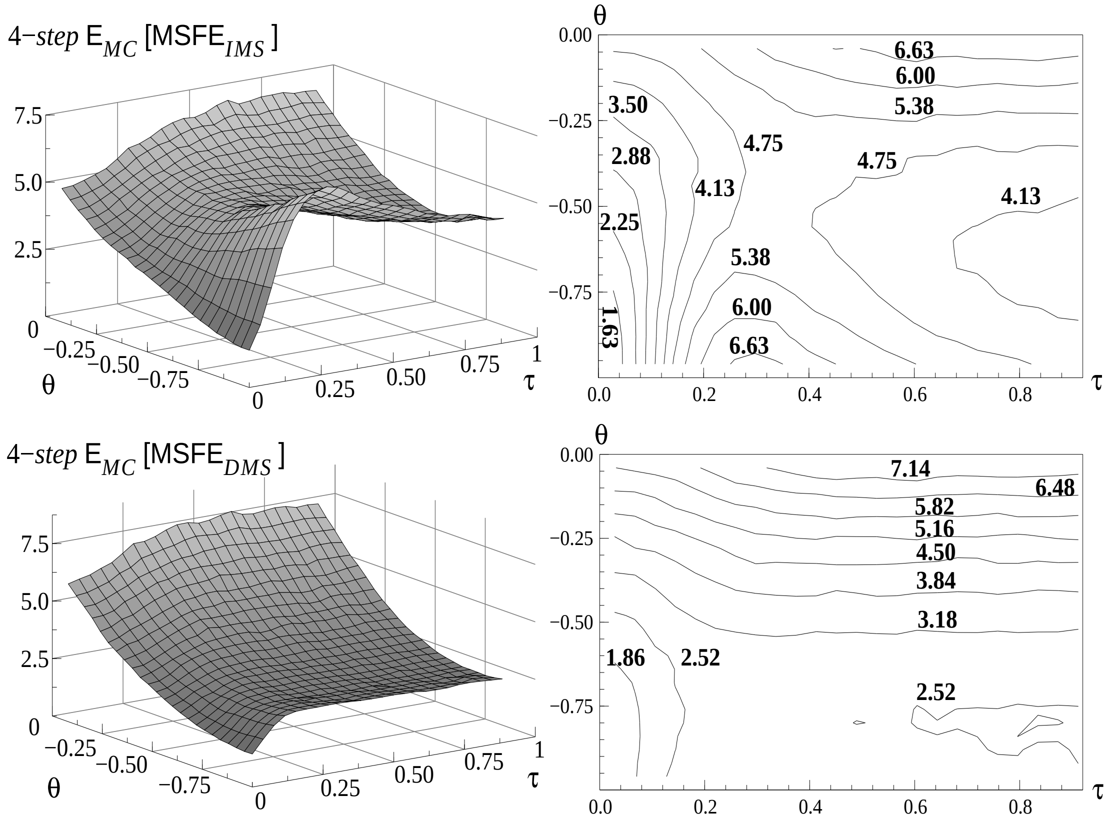
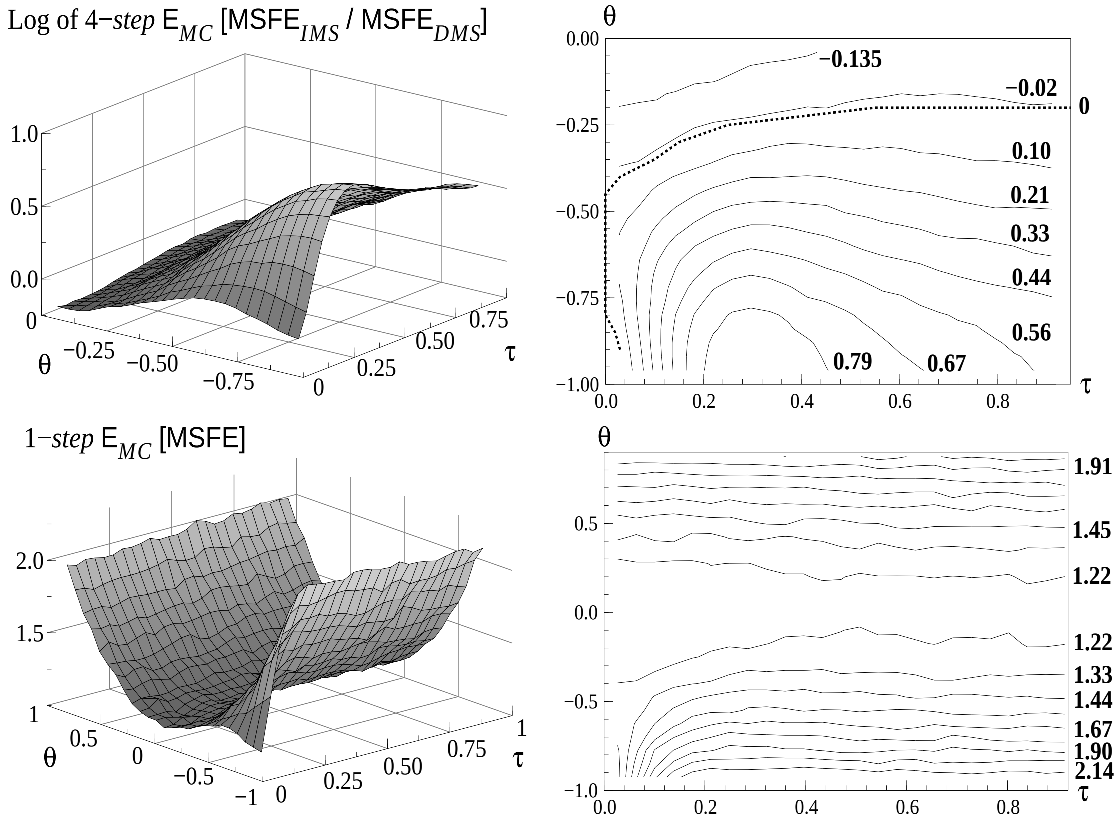


Figure 6: Monte Carlo estimates of the 4-step IMS and DMS MSFEs for a sample of $T = 25$ observations, $10,000$ replications and varying parameters.

yet non-zero, so that the h-step ahead IMS intercept estimator is badly biased (see fig. 5). The essential feature here is the deterministic trend induced by the drift: then DMS noticeably improves upon IMS. This feature combines in the multi-step intercept with the powers of the estimated slope in the implied distribution of $x_t|x_{t-h}$: whether it is well or badly estimated translates directly into the larger or lower h-step MSFE (fig. 6), and yields the results in fig. 7*a*–*b*.

7 Explaining the literature

Having shown that direct estimation can improve forecast accuracy when the model omits a negative moving average, the data are integrated and, potentially, exhibit a small deterministic trend, we are now able to understand some of results found in the literature. Findley (1983) finds that series C (226 observations) and E (100) from Box and Jenkins (1976) exhibit some accuracy gain when using DMS methods where the lag length is chosen



Log of 4−*step* E_{MC} [MSFE_{IMS} / MSFE_{DMS}]

1−*step* E_{MC} [MSFE]

Figure 7: Monte Carlo estimates of the logarithm of the ratios of 4-step IMS and DMS MSFEs (panels a and b); and Monte Carlo estimates of the one-step MSFE (panels c and d), for a sample of $T = 25$ observations, $10,000$ replications and varying parameters.

using an h-step AIC criterion. Both series are approximately integrated (with negative serial autocorrelation for E). The series C can be represented by an AR(2):

$$\begin{aligned} \mathrm{C}_t &= \underset{(0.103)}{0.2645} + \underset{(0.0387)}{1.809}\,\mathrm{C}_{t-1} - \underset{(0.0387)}{0.821}\,\mathrm{C}_{t-2} + \varepsilon_t \\ \mathsf{R}^2 &= 0.99, \quad \widehat{\sigma} = 0.133, \end{aligned}$$

whereas series E exhibits longer memory, and is more difficult to model. Findley (1983) uses larger samples than the ones here, namely 150 to 200 for series C, and 80 to 90 for series E. He finds that at longer horizons, multi-step estimation can improve forecast accuracy, and this corroborates our analysis since series C is indeed integrated with a small drift. Stoica and Nehorai (1989) simulate the forecasting properties of an AR(1) model applied to an ARMA$(2, 2)$ process where the first lag has zero coefficient:

$$y_t = 0.98y_{t-2} + \epsilon_t - 0.87\epsilon_{t-1} - 0.775\epsilon_{t-2}.$$

Statistically, $\{y_t\}$ behaves as if it were integrated, so corresponds to forecasting an I(1) process with a model which omits a negative moving average. In light of our previous analysis, it is natural that Stoica and Nehorai (1989) find that DMS outperforms IMS at horizon $h = 4$ with a sample of $T = 200$ observations. In a simulation study of DMS where the data is generated by autoregressive processes, with potential additional regressors, Weiss (1991) finds that DMS is preferred when the model is an AR(1) which omits negative serial correlation, and the data exhibit long memory (via either integratedness or a large autoregressive coefficient). This is again in line with the results above. Weiss (1991) also finds that, when the DGP is:

$$\begin{aligned} y_t &= 0.8y_{t-1} + 0.5z_t + \alpha z_{t-1} + \epsilon_t && \text{(28a)} \\ z_t &= 0.5z_{t-1} + u_t, && \text{(28b)} \end{aligned}$$

where ϵ_t and u_t are both standard normal and are independent, and the model is $y_t = \phi_1 y_{t-1} + \phi_2 z_t + v_t$, DMS outperforms IMS for $\alpha = 1/2$, 2 and 5 at forecast horizon $h = 4$ for a sample size of 1000 observations. Notice that in this case the DGP can be re-formulated as:

$$y_t = 0.8y_{t-1} + (0.5 + 2\alpha)\, z_t + (\epsilon_t - 2u_t)\,,$$

so that both the implied coefficient on z_t and the variance of the disturbances are larger than in (28a). Given that the autoregressive coefficient of y_t is large, this series exhibits some long memory, so corresponds to neglecting some serial correlation in the case of a near-integrated processes. Since Weiss (1991) had shown that DMS can be asymptotically efficient, he uses very large samples for his simulations. In an empirical application of multivariate multi-step estimation methods (which are similar to using the GFESM as a forecast accuracy criterion), Liu (1996) finds that DMS can be more accurate when forecasting the log of real US GNP and the US consumer price index for food in levels, but when using the first-differences of the data, IMS is preferred. In both cases, the series—in levels—are integrated and trending which could be modelled either as a linear deterministic trend, or as an I(2) stochastic trend. This is close to the situation presented above when DMS fares well. Finally, Clements and Hendry (1996) found that using DMS was beneficial for forecast accuracy when the series were integrated and negative moving averages were neglected. They derived their findings from models with small drifts, but only compared these to the cases of no deterministic terms when the differences, not the levels, of the variables were to be forecast, so the impact of the joint occurrence of deterministic and stochastic trends was not obvious. Thus, our results corroborate those obtained by these various authors. However, none of the previous findings stressed the major influence that non-zero drift has in the case of integrated data, although positive drifts were present in the simulations by Clements and Hendry (1996), and also for series C in Findley (1983) and the seasonal model in Haywood and Tunnicliffe-Wilson (1997).

8 Conclusions

We have shown that when the one-step model is well specified and estimation takes into account the dynamic properties of the disturbances, DMS can asymptotically prove a more accurate forecasting technique than IMS when it leads to more efficient estimation of the *multi-step* intercept.

Moreover, when the DGP is unknown, and least-squares estimation entails an inaccurate dynamic specification of the disturbances, direct multi-step estimation may be preferred as a forecasting procedure in finite samples. Model mis-specification is then necessary, but not sufficient, to justify the use of DMS. When the process is integrated, 'large' negative moving-average errors amplify the bias of the one-step ahead estimators. Even so, omitting negative autocorrelation of the errors is not sufficient for DMS to prove more accurate in small-sample forecasting. Processes which exhibit a deterministic trend, and which are mis-specified for the errors, present the conditions for a successful use of DMS. Unfortunately, the resulting gain is highly non-linear in the various parameters of the series and cannot be easily modelled. The stylized results are that the nearer (but not equal to) zero the trend—with a maximum for certain low values which depend on the other parameters—and the larger the negative residual autocorrelation, the higher the relative accuracy from DMS. The class of structural time-series models allows us to understand the role played by the negative autocorrelation of the residuals, and shows that potential gains from forecasting with DMS arise when economic variables exhibit varying trends, or are subject to cyclical patterns. Thus, DMS-based forecasts exhibit some robustness compared to IMS.

When should a modeller decide to use direct multi-step estimation rather than iterate the forecasts? Our analysis shows that DMS can prove either more efficient asymptotically, or more precise in finite samples. A practitioner does not know *ex ante* how mis-specified her model might be, or what shocks the economic series may encounter. By observing the situations most beneficial here to DMS, she would be advised to resort to direct multi-step estimation whenever the data she wishes to forecast exhibit either stochastic or deterministic non-stationarity (unit-root and breaks) and the available sample is too small for reliable inferences.

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.

Banerjee, A., D. F. Hendry, and G. E. Mizon (1996). The econometric analysis of economic policy. *Oxford Bulletin of Economics and Statistics 58*, 573–600.

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. (1997). Direct autoregressive predictors for multistep prediction: order selection and performance relative to the plug-in predictors. *Statistica Sinica 7*, 425–449.

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.

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.

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 (1998). Forecasting economic processes. *International Journal of Forecasting 14*, 111–131.

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.

Fildes, R. and K. Ord (2002). Forecasting competitions – their role in improving forecasting practice and research. In M. P. Clements and D. F. Hendry (Eds.), *A Companion to Economic Forecasting*, pp. 322–253. Oxford: Blackwells.

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.

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

Hall, A. (1989). Testing for a unit root in the presence of moving average errors. *Biometrika 76*, 49–56.

Hamilton, J. D. (1994). *Time Series Analysis.* Princeton: Princeton University Press.

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

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

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. (1991). Using PC-NAIVE in teaching econometrics. *Oxford Bulletin of Economics and Statistics 53*, 199–223.

Hendry, D. F. (2000). On detectable and non-detectable structural change. *Structural Change and Economic Dynamics 11*, 45–65.

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.

Kang, I.-B. (2003). Multi-period forecasting using different models for different horizons: an application to U.S. economic time series data. *International Journal of Forecasting 19*, 387–400.

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.

Magnus, J. R. and B. Pesaran (1989). The exact multiperiod mean square forecast error of the first-order autoregressive model with an intercept. *Journal of Econometrics 42*, 157–179.

Marcellino, M., J. H. Stock, and M. W. Watson (2004). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. Mimeo, Harvard University.

Nagar, A. L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica 27*, 575–595.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocoreelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

Sargan, J. D. (1982). On monte carlo estimates of moments that are infinite. *Advances in Econometrics 1*, 267–299.

Stock, J. H. and M. W. Watson (1999). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time-series. In R. F. Engle and H. White

(Eds.), *Cointegration, Causality and Forecasting: A Festschrift in honour of Clive W. J. Granger*, pp. 1–44. Oxford: Oxford University Press.

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

Tsay, R. S. (1993). Comment: Adpative forecasting. *Journal of Business and Economic statistics 11*(2), 140–142.

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

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.

Wold, H. O. A. (1938). *A study in the analysis of stationary time series.* Stockolm: Almqvist and Wicksell.