Bayesian forecasting with the structural damped trend model

Abstract

In an important study, Sbrana and Silvestrini (2020) [Sbrana, Giacomo, Silvestrini, Andrea, 2020. Forecasting with the damped trend model using the structural approach. International Journal of Production Economics, to appear] consider the structural damped trend model which is standard in the arsenal of forecasting analysis. The authors compare its performance to the "single source of error" model. In this paper, we consider Bayesian analysis of the damped trend model, which is a well known benchmark in forecasting, using the structural approach without the need to compute the Kalman filter or evaluate the likelihood function. Monte Carlo and empirical applications show the superior performance of the structural model versus the "single source of errors" model. Additional evidence is provided by a Bayesian optimal model pool approach.

JEL classifications: C11, C13, C22.

Keywords: Damped trend model; Bayesian analysis; Out-of-sample forecasting; Forecast accuracy.

1 Introduction

In this paper we consider the damped trend model recently considered in an important study by Sbrana and Silvestrini (2020). Specifically, they consider the "multiple sources of error" (MSOE) formulation (or "structural approach") which seems to be relatively unpopular in forecasting, probably due to the complexities of likelihood estimation, which requires the use of Kalman filter (Harvey, 1989; West and Harrison, 2006; Durbin and Koopman, 2012).

Hyndman et al. (2008) show that the damped trend's parameters can be estimated by maximum likelihood using the Kalman filter. Of course, this is true for the whole class of exponential smoothing models in the "single source of error" (SSOE) form, in which a single source of disturbances drives the whole system (Ord et al., 1997).

Gardner and McKenzie (1985) propose a modification of Holt's linear method to dampen the trend when the length of the forecast horizon is large. The method, called "damped trend exponential smoothing", is useful when the time series has a trend but one believes that its growth rate is unlikely to continue at a constant pace during the forecast horizon. Since the pioneering work of Gardner and McKenzie (1985), the damped trend model has been applied extensively in empirical applications owing to its good forecasting record (Gardner and McKenzie, 1988, 1989, 2011; McKenzie and Gardner, 2010). In a survey of the most popular forecasting methods in operational research, Fildes et al. (2008) mention that the damped trend provides a "benchmark forecasting method for all others to beat" (p. 1154). It is the consensus of the literature that the damped trend appears is an accurate forecasting method (McKenzie and Gardner, 2010).

As Sbrana and Silvestrini (2020) write: "in a Monte Carlo experiment, we compare the out-of-sample forecasting performance of the damped trend model in MSOE form, estimated using our new method, with that of the damped trend model in the SSOE framework, estimated using the standard innovations approach. Simulation results show that the forecasting performances of the two models are rather similar, even though some forecasting gains from the MSOE over the SSOE are possible when time series are short. Furthermore, in two empirical applications using annual data from the M3-competition dataset (Makridakis and Hibon, 2000) and quarterly credit-to-GDP data published by the Bank for International Settlements (BIS), we provide further evidence that the damped trend model in the MSOE form competes well with its SSOE counterpart, confirming the Monte Carlo results."

In this paper we propose a Bayesian approach to the MSOE which relies on an efficient implementation of the Gibbs sampler with data augmentation (Tanner and Wong, 1987; Gelfand and Smith, 1990). The Bayesian MSOE model performs better than SMOE in a decisive way, resolving an ambiguity regarding forecasting performance relative to SSOE in Sbran and Silverstrini (2020) who employ sampling-theory based techniques. This illustrates the fact that it makes a substantive difference whether one uses a Bayesian or a sampling-theory approach.

2 Model

The model is as follows:

$$y_t = l_{t-1} + \phi b_{t-1} + \varepsilon_t,\ \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2),\ t = 1, \ldots, T, \tag{1}$$

$$l_t = l_{t-1} + \phi b_{t-1} + \eta_t,\ \eta_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2),\ t = 1, \ldots, T, \tag{2}$$

$$b_t = \phi b_{t-1} + \xi_t,\ \xi_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\xi^2),\ t = 1, \ldots, T. \tag{3}$$

In this formulation, “iid” means “independent and identically distributed, $\mathcal{N}(0, \sigma^2)$ denotes the normal distribution with zero mean and variance σ^2,$0 < \phi < 1$, and $\varepsilon_t, \eta_t, \xi_t$ are pairwise uncorrelated. The data is $y = \{y_t\}_{t=1}^T$ where T denotes the number of observations. We assume $\sigma_\varepsilon > 0$, and $\sigma_\eta, \sigma_\xi \geq 0$. To proceed we consider the complete data likelihood as if we knew the latent $\{l_t, b_t\}_{t=1}^T$ which is given by:

$$\begin{gathered} L(\theta, \{l_t, b_t\}_{t=1}^T; y) \propto \\ \sigma_\varepsilon^{-n} \sigma_\eta^{-n} \sigma_\xi^{-n} \cdot \exp\left\{ -\tfrac{1}{2\sigma_\varepsilon^2} \textstyle\sum_{t=1}^T \left(y_t - l_{t-1} - \phi b_{t-1} \right)^2 \right\} \\ \exp\left\{ -\tfrac{1}{2\sigma_\eta^2} \textstyle\sum_{t=1}^T \left(l_t - l_{t-1} - \phi b_{t-1} \right)^2 \right\} \cdot \exp\left\{ -\tfrac{1}{2\sigma_\xi^2} \textstyle\sum_{t=1}^T \left(b_t - \phi b_{t-1} \right)^2 \right\}, \end{gathered} \quad . \tag{4}$$

where $\theta = [\phi, \sigma_\varepsilon, \sigma_\eta, \sigma_\xi]'$ is the vector of unknown parameters.

To proceed we use a benchmark prior for the coefficients given as follows:

$$p(\theta) = p(\phi) p(\sigma_\varepsilon) p(\sigma_\eta) p(\sigma_\xi), \tag{5}$$

$$p(\phi) = 1,\ \phi \in (0, 1), \tag{6}$$

$$p(\sigma_j) \propto \sigma_j^{-(\bar{N}_j + 1)} e^{-\bar{Q}_j / (2\sigma_j^2)},\ j \in \{\varepsilon, \eta, \xi\}, \tag{7}$$

where $\bar{N}_j, \bar{Q}_j \geq 0$ are parameters of the priors. The prior for ϕ is flat in $(0, 1)$. The priors for the scale parameters are equivalent to

$$\frac{\bar{Q}_j}{\sigma_j^2} \sim \chi^2_{\bar{N}_j},\ j \in \{\varepsilon, \eta, \xi\}, \tag{8}$$

where χ^2_N denotes the chi-square distribution with N degrees of freedom.

3 Bayesian analysis by MCMC

The posterior distribution is given by Bayes' theorem:

$$
\begin{array}{c}
p(\theta, \{l_t, b_t\}_{t=1}^T | y) \propto \\
\sigma_\varepsilon^{-(T+\bar{N}_\varepsilon+1)} \sigma_\eta^{-(T+\bar{N}_\eta+1)} \sigma_\xi^{-(T+\bar{N}_\xi+1)} \cdot \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left[\bar{Q}_\varepsilon + \sum_{t=1}^T (y_t - l_{t-1} - \phi b_{t-1})^2\right]\right\} \quad \cdot \\
\exp\left\{-\frac{1}{2\sigma_\eta^2}\left[Q_\eta + \sum_{t=1}^T (l_t - l_{t-1} - \phi b_{t-1})^2\right]\right\} \cdot \exp\left\{-\frac{1}{2\sigma_\xi^2}\left[\bar{Q}_\xi + \sum_{t=1}^T (b_t - \phi b_{t-1})^2\right]\right\}.
\end{array} \tag{9}
$$

To provide access to the posterior we use Gibbs sampling with data augmentation (Tanner and Wong, 1987; Gelfand and Smith, 1990) which means that we have to draw random numbers recursively from the following conditional posterior distributions:

$$\phi | \sigma_\varepsilon, \sigma_\eta, \sigma_\xi, \{l_t\}_{t=1}^T, \{b_t\}_{t=1}^T, y, \tag{10}$$

$$\sigma_\varepsilon | \phi, \sigma_\eta, \sigma_\xi, \{l_t\}_{t=1}^T, \{b_t\}_{t=1}^T, y, \tag{11}$$

$$\sigma_\eta | \sigma_\varepsilon, \phi, \sigma_\xi, \{l_t\}_{t=1}^T, \{b_t\}_{t=1}^T, y, \tag{12}$$

$$\sigma_\xi | \sigma_\varepsilon, \sigma_\eta, \phi, \{l_t\}_{t=1}^T, \{b_t\}_{t=1}^T, y, \tag{13}$$

$$l_t | \phi, \sigma_\varepsilon, \sigma_\eta, \sigma_\xi, \{l_{-t}\}, \{b_t\}_{t=1}^T, y, \ t = 1, \ldots, T, \tag{14}$$

$$b_t | \phi, \sigma_\varepsilon, \sigma_\eta, \sigma_\xi, \{l_t\}_{t=1}^T, \{b_{-t}\}, y, \ t = 1, \ldots, T, \tag{15}$$

where $\{l_{-t}\}$ and $\{b_{-t}\}$ are, respectively, the sequences of $\{l_t\}$ and $\{b_t\}$without the elements l_t and b_t.

For the scale parameters we have the following conditional posteriors:

$$\tfrac{\bar{Q}_\varepsilon + \sum_{t=1}^T (y_t - l_{t-1} - \phi b_{t-1})^2}{\sigma_\varepsilon^2} | \phi, \sigma_\eta, \sigma_\xi, \{l_t\}_{t=1}^T, \{b_t\}_{t=1}^T, y \sim \chi^2_{T+\bar{N}_\varepsilon}, \tag{16}$$

$$\tfrac{\bar{Q}_\eta + \sum_{t=1}^T (l_t - l_{t-1} - \phi b_{t-1})^2}{\sigma_\eta^2} | \phi, \sigma_\varepsilon, \sigma_\xi, \{l_t\}_{t=1}^T, \{b_t\}_{t=1}^T, y \sim \chi^2_{T+\bar{N}_\eta}, \tag{17}$$

$$\tfrac{\bar{Q}_\eta + \sum_{t=1}^T (l_t - l_{t-1} - \phi b_{t-1})^2}{\sigma_\xi^2} | \phi, \sigma_\varepsilon, \sigma_\eta, \{l_t\}_{t=1}^T, \{b_t\}_{t=1}^T, y \sim \chi^2_{T+\bar{N}_\eta}. \tag{18}$$

This involves only drawing random numbers from *chi*-square distributions. Let us define b, b_{-1}, l, l_{-1} that contain current and lagged values of the sequences $\{l_t\}_{t=1}^T$ and $\{b_t\}_{t=1}^T$. It is not difficult to show that the conditional posterior distribution of ϕ is:

$$\phi | \sigma_\varepsilon, \sigma_\eta, \sigma_\xi, \{l_t\}_{t=1}^T, \{b_t\}_{t=1}^T, y \sim \mathcal{N}\left(\hat{\phi}, \ s_\phi^2\right), \tag{19}$$

where $\phi = \left(\frac{1}{\sigma_\varepsilon^2} + \frac{1}{\sigma_\eta^2} + \frac{1}{\sigma_\xi^2}\right)^{-1} b'_{-1}b_{-1} + \left(\frac{b'_{-1}(y-l_{-1})+b'_{-1}(l-l_{-1})}{\sigma_\varepsilon^2} + \frac{b'_{-1}b}{\sigma_\xi^2}\right)$, and $s_\phi^2 = \left(\frac{1}{\sigma_\varepsilon^2} + \frac{1}{\sigma_\eta^2} + \frac{1}{\sigma_\xi^2}\right)^{-1} b'_{-1}b_{-1}$. The conditional posterior distribution of l_t is as follows.

$$l_t|\phi, \sigma_\varepsilon, \sigma_\eta, \sigma_\xi, \{l_{-t}\}, \{b_t\}_{t=1}^T, y \sim \mathcal{N}\left(\hat{l}_t,\ s_l^2\right),\ t = 1, \ldots, T, \tag{20}$$

where $\hat{l}_t = \frac{\sigma_\eta^2(y_{t+1}-\phi b_t)+\sigma_\varepsilon^2[l_{t-1}-\phi(b_t+b_{t-1})-l_{t+1}]}{\sigma_\varepsilon^2+2\sigma_\eta^2}$, $s_l^2 = \frac{2\sigma_\varepsilon^2\sigma_\eta^2}{\sigma_\varepsilon^2+2\sigma_\eta^2}$, with obvious modifications for the end points. Finally, the conditional posterior distribution of b_t is the following.

$$b_t|\phi, \sigma_\varepsilon, \sigma_\eta, \sigma_\xi, \{l_t\}_{t=1}^T, \{b_{-t}\}, y \sim \mathcal{N}(\hat{b},\ s_b^2),\ t = 1, \ldots, T, \tag{21}$$

where $\hat{b} = \phi\left(\frac{\phi^2}{\sigma_\varepsilon^2} + \frac{\phi^2}{\sigma_\eta^2} + \frac{1+\phi^2}{\sigma_\xi^2}\right)^{-1}\left[\frac{y_{t+1}-l_t}{\sigma_\varepsilon^2} + \frac{l_{t+1}-l_t}{\sigma_\eta^2} + \frac{b_{t-1}+b_{t+1}}{\sigma_\xi^2}\right]$, and $s_b^2 = \left(\frac{\phi^2}{\sigma_\varepsilon^2} + \frac{\phi^2}{\sigma_\eta^2} + \frac{1+\phi^2}{\sigma_\xi^2}\right)^{-1}$, with obvious modifications for the end points.

To see how these expressions can be derived easily, let us focus on the terms of the posterior that involve b_t. These terms can be written as follows.

$$\begin{aligned} y_{t+1} - l_t &= \phi b_t + v_{b,t}^{(1)}, \\ l_{t+1} - l_t &= \phi b_t + v_{b,t}^{(2)}, \\ \phi b_{t-1} &= b_t + v_{b,t}^{(3)}, \\ b_{t+1} &= \phi b_t + v_{b,t}^{(4)} \end{aligned} \tag{22}$$

where $v_{b,t}^{(1)} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$, $v_{b,t}^{(2)} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2)$, $v_{b,t}^{(3)}, v_{b,t}^{(4)} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\xi^2)$. This is a system of four observations with one parameter (b_t) so application of generalized least squares yields the expressions in (21). Similarly, we can derive all other conditional posterior distributions. The expressions in (16), (17), and (18) can be derived using the result that if $p(\sigma) \propto \sigma^{-(N+1)}e^{-Q/(2\sigma^2)}$ then $\frac{Q}{\sigma^2} \sim \chi_N^2$. In turn, the conditional posterior distributions in (19), (20), and (21) and normal from which random number generation is commonly available in statistical software.

4 Technical details, Monte Carlo simulation and empirical results

4.1 Technical details

To implement Gibbs sampling with data augmentation we use 150,000 iterations omitting the first 50,000 in the burn-in phase to mitigate possible start up effects.[1] Convergence is checked using the standard Geweke (1992) diagnostics. Sbrana and Silvestrini (2020) recommend the use of the initial conditions $\begin{bmatrix} l_0 \\ b_0 \end{bmatrix} = \begin{bmatrix} y_1 \\ 0 \end{bmatrix}$ although in principle these can be treated as unknown parameters and estimated. Our benchmark prior has $\bar{N}_j = 1$ and

[1] Our implementation relies on `fortran77` (`gnu` compiler) making extensive use of commonly available subroutines including `linpack` in `netlib`. Computations were carried out in the High End Computing cluster (HEC) of Lancaster University. The High End Computing (HEC) Cluster offers 8,800 cores, 40 TB of aggregate memory, 70TB of high performance filestore for general use and 4PB of medium performance filestore for GridPP data. The cluster operating system is `CentOS Linux`, with job submission handled by `Son of Grid Engine` (SGE).

$Q_j = 10^{-6}$ $(j \in \{\varepsilon, \eta, \xi)$ which is minimally informative, if at all.

For the endpoints in (20) and (21) to draw $l_1^{(s)}$ and $b_1^{(s)}$(where s indexes the particular MCMC draw) we use $l_T^{(s-1)}$ and $b_T^{(s-1)}$ for the missing l_0 and b_0 while for $l_T^{(s)}$ and $b_T^{(s)}$ we use $l_1^{(s-1)}$ and $b_1^{(s-1)}$ for the missing l_{T+1} and b_{T+1}. This is justified by stationarity. Finally, the SSOE model can be obtained by setting $\rho_{\varepsilon\eta} = \rho_{\varepsilon\xi} = \rho_{\eta\xi}$=1, where $\rho_{\varepsilon\eta}, \rho_{\varepsilon\xi}, \rho_{\eta\xi}$ are defined below.

4.2 Monte Carlo experiment

For the Monte Carlo simulation exercise, we use the same set up as in Sbrana and Silvestrini (2020). Model (1) – (2) is misspecified by assuming that

$$\begin{bmatrix} \varepsilon_t \\ \eta_t \\ \xi_t \end{bmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho_{\varepsilon\eta}\alpha & \rho_{\varepsilon\xi}\beta \\ & \alpha^2 & \rho_{\eta\xi}\alpha\beta \\ & & \beta^2 \end{bmatrix} \right), \tag{23}$$

where the correlation coefficients $\rho_{\varepsilon\eta}, \rho_{\varepsilon\xi}, \rho_{\eta\xi} \in (-1, 1)$.

The number of observations is T = 50, 100, 200 and we consider 5,000 replications or artificial data sets. For each replication, the parameter ϕ is drawn from a uniform distribution in the interval (0.01, 0.99). The innovation variance σ_ε^2 and parameters α and β are independently drawn from a uniform in the interval $(0.01, 1)$. Finally, $\rho_{\varepsilon\eta}$, $\rho_{\varepsilon\xi}$ and $\rho_{\eta\xi}$ are drawn from a uniform in the interval $(–1, 1)$. The use of random rather than fixed parameters allows the results of the Monte Carlo simulation experiment to be general rather than case-specific. We evaluate the forecast performance of each model (h = 1, 2, 3, 4, 5, 6-steps ahead) and results are compared using the Mean Absolute Scaled Error (MASE) as proposed by Hyndman and Koehler (2006). Median MASEs are reported in Table 1 and compared with the results of Tables 1 – 3 of Sbrana and Silvestrini (2020).

Table 1: Out of sample forecast comparisons

Forecast horizon	$T = 50$	$T = 100$	$T = 200$
$h = 1$	$0.702^{(a)}$	$0.703^{(b)}$	$0.702^{(c)}$
	0.689	*0.694*	*0.698*
$h = 2$	0.775	0.783	0.765
	0.770	*0.778*	*0.750*
$h = 3$	0.818	0.813	0.805
	0.805	*0.810*	*0.800*
$h = 4$	0.836	0.828	0.827
	0.827	*0.820*	*0.814*
$h = 5$	0.855	0.840	0.840
	0.832	*0.837*	*0.838*
$h = 6$	0.871	0.861	0.846
	0.862	*0.855*	*0.830*

Notes: $^{(a)}$ Taken from Table 1 of Sbrana and Silvestrini (2020). $^{(b)}$ Taken from Table 2 of Sbrana and Silvestrini (2020). $^{(c)}$ Taken from Table 3 of Sbrana and Silvestrini (2020). Reported numbers are medians of Mean Absolute Scaled Error (MASE) as proposed by Hyndman and Koehler (2006) and Sbrana and Silvestrini (2020). The results of the Monte Carlo simulations in this paper are reported in italics.

Notably, the forecast errors are smaller than the ones using minimizing the sum of squared prediction errors in samples as large as 200. Thus, it is essential, as it seems, to use a Bayesian approach as it takes better account of "small samples".

4.3 Empirical application

As in Sbrana and Silvestrini (2020) we use quarterly data of time series on total credit to the non-financial sector relative to GDP using data published by the Bank for International Settlements. It covers almost 50 economies with, on average, more than 45 years of data, reaching back to the 1940s and the 1950s in some cases and it is taken from the same source as in Sbrana and Silvestrini (2020).[2] several studies of past financial crises found that the credit-to-GDP ratio is an early warning indicator of systemic banking crises (see, for instance, De Bonis and Silvestrini, 2014). This motivates our interest in forecasting this variable. Each time series (271 in total) is divided into a training sample and into a test sample, which corresponds to the last six observations. The predictive performance of the models is evaluated based on h =1, 2, 3, 4, 5, 6-step ahead forecasts, using the MASE. Starting in 2008:Q1 all time series have 45 observations. This allows us to better appreciate the influence of the sample size on the forecasting performance of different estimation and forecasting techniques.

Our empirical results are reported in Table 2 and we compare with SSOE rather than MSOE as it behaved better in Sbrana and Silvestrini (2020).

[2] `https://www.bis.org/ statistics/about_credit_stats.htm`

Table 2: Out-of-sample forecast comparison of Bayesian MSOE

	Annual M3 data		Quarterly BIS credit data		quarterly BIS credit data, shorter sample (2008:Q1-2019:Q1)	
	Mean	Median	Mean	Median	Mean	Median
$h = 1$	$1.117^{(a)}$	0.745	$0.636^{(b)}$	0.499	$0.667^{(c)}$	0.536
	1.103	*0.710*	*0.625*	*0.445*	*0.610*	*0.528*
$h = 2$	1.468	0.991	0.781	0.653	0.814	0.688
	1.320	*0.977*	*0.693*	*0.610*	*0.792*	*0.665*
$h = 3$	1.884	1.214	0.988	0.799	1.040	0.872
	1.720	*1.128*	*0.970*	*0.782*	*0.986*	*0.864*
$h = 4$	2.250	1.405	1.197	0.981	1.278	1.115
	2.130	*1.317*	*1.182*	*0.977*	*1.265*	*1.042*
$h = 5$	2.581	1.646	1.373	1.137	1.460	1.308
	2.410	*1.630*	*1.355*	*1.124*	*1.432*	*1.265*
$h = 6$	2.873	1.878	1.528	1.294	1.612	1.421
	2.773	*1.853*	*1.510*	*1.286*	*1.602*	*1.389*

Notes: [a] Taken from Table 4 of Sbrana and Silvestrini (2020). [b] Taken from Table 5 of Sbrana and Silvestrini (2020). [c] Taken from Table 6 of Sbrana and Silvestrini (2020). All these results are for the SSOE model. Reported numbers are medians of Mean Absolute Scaled Error (MASE) as proposed by Hyndman and Koehler (2006) and Sbrana and Silvestrini (2020). The results of the Monte Carlo simulation in this paper are reported in italics. For the data, see Sbrana and Silvestrini (2020).

From the evidence in Table 2 it turns out that Bayesian MASEs are smaller for the MSOE model relative to the SSOE (and, therefore, the sampling-theory based MSOE) model.

To assess the out-of-sample behavior of the Bayesian MSOE we use predictive Bayes factors (PBF). First, we need the predictive likelihood $p(y_{T+h}|y_{1:T})$ where $y_{1:T}$ is the estimation sample, and y_{T+h} is the observation in period $T + h$ where h is the forecast horizon. Suppose $\Lambda_{1:T} = \{l_t, b_t\}_{t=1}^{T}$ denotes our latent variables. In turn, we have:

$$p(y_{T+h}|y_{1:T}) = \int p(y_{T+h}|\theta, \Lambda_{1:T}, y_{1:T})p(\theta, \Lambda_{1:T}|y_{1:T})\,\mathrm{d}\theta\,\mathrm{d}\Lambda_{1:T}. \tag{24}$$

Provided we have a set of draws $\left\{\theta^{(s)}, \Lambda_{1:T}^{(s)}\right\}_{s=1}^{S}$ which converges to the distribution whose density is $p(\theta, \Lambda_{1:T}|y_{1:T})$ (a task carried out by MCMC), we have

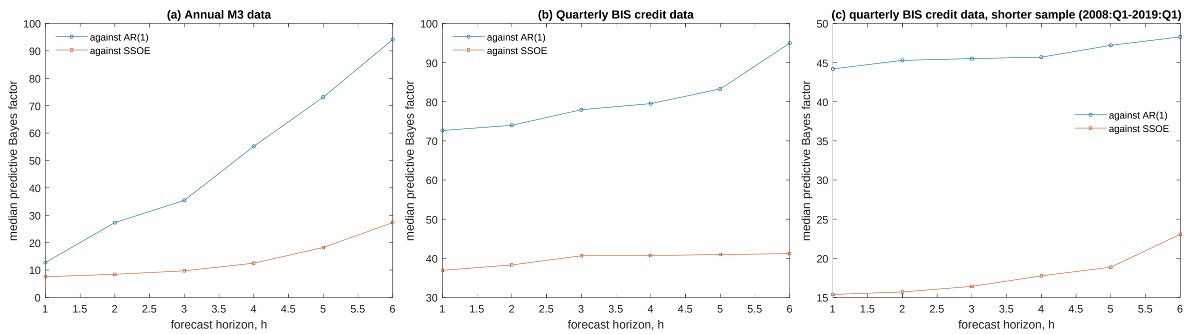
$$p(y_{T+h}|y_{1:T}) \simeq S^{-1}\sum_{s=1}^{S} p(y_{T+h}|\theta^{(s)}, \Lambda_{1:T}^{(s)}, y_{1:T}), \tag{25}$$

where the quality of approximation depends only on S and $y_{T+h}|\theta^{(s)}, \Lambda_{1:T}^{(s)}, y_{1:T}$ is available in closed form from (1) – (3). If we have different models indexed by $m \in \mathscr{M} = \{0, 1, \ldots, M\}$ then we can obtain different predictive densities denoted $p_m(y_{T+h}|y_{1:T})$. The PBF in favor of model m and against model "0" is simply

$$PBF_{m:1} = \frac{p_m(y_{T+h}|y_{1:T})}{p_0(y_{T+h}|y_{1:T})},\ m \in \mathscr{M}. \tag{26}$$

For our empirical application, median PBFs (across all time series in a given data set) are reported in Figure 1. In panel (a) we report PBFs in favor of the Bayesian SSOE model and against MSOE as well as AR(1) when the benchmark model "0" is a random walk with drift.

Figure 1: Predictive Bayes factors



The evidence in Figure 1 shows that the against both AR(1) and SSOE are quite large, ranging from 10 to 100 in panel (a), $70-100$ in panel (b) and nearly 45 in panel (c). This behavior is consistent across all forecast horizons so it resolves, in a sense, the ambiguity in forecasting performance of MSOE versus SSOE which comes out of the evidence in Sbrana and Silvestrini (2020). This ambiguity arises in a sampling-theory context when models are estimated using minimizing the sum of squared prediction errors. The Bayesian evidence is more decisive and shows the superior performance of MSOE in terms of out-of-sample forecasting.

4.4 Optimal model pools

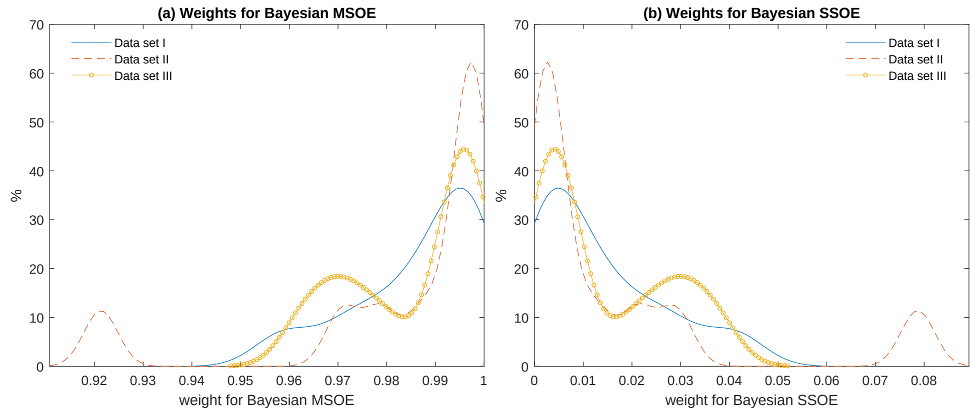
Suppose y_o is an out-of-sample observation. Given a set of models whose posterior predictive densities are $p_m(y_o|y_{1:T})$, we consider predictive densities of the form:

$$\sum_{m\in\mathscr{M}} w_m p_m(y_o|y_{1:T});\ \sum_{m\in\mathscr{M}} w_m = 1;\ w_m \geq 0\,\forall m \in \mathscr{M}, \tag{27}$$

which are known as *linear opinion pools.* We consider using the log predictive score function:

$$\sum_{t=1}^{T} \log\left[\sum_{m=1}^{M} w_m p_m(y_o|y_{1:t})\right], \tag{28}$$

Figure 2: Optimal model pools



Notes: Data set I is Annual M3 data. Data set II is Quarterly BIS credit data quarterly BIS credit data, and Data set III is the shorter sample (2008:Q1-2019:Q1). For the data, see Sbrana and Silvestrini (2020). Results are reported for all time series in a given data set.

as in Geweke and Amisano (2011 a, b). The log predictive score function is a measure of the out-of-sample prediction record of the model.

Bernardo (1979) showed that *the only proper local scoring rule takes the form of* (28). Maximizing (28) subject to (27) is quite feasible given nonlinear programming software. Therefore, the problem is:

$$\max_{\{w_m, m\in\mathscr{M}\}} \sum_{t=1}^{T} \log\left[\sum_{m=1}^{M} w_m p_m(y_o|y_{1:t})\right], \\ \text{subject to} \\ \sum_{m\in\mathscr{M}} w_m = 1;\ w_m \geq 0\, \forall m\in\mathscr{M}. \tag{29}$$

The purpose of this optimization program is to maximize weighted log predictive scoring (or, equivalently, maximize out-of-sample performance) subject to the usual portfolio-like constraints on the weights attached to different models. *It must be emphasized that this approach is quite general in that it allows for different, general models not necessarily sharing the same parameters although, apparently, all models are conditioned on (that is use) the same data.*[3]

For each time series in a data set, we consider a random walk model with drift, an AR(1), the SSOE and the MSOE models. We solve the problem in (29) for each different time series and we report the optimal weights in Figure 2.

In all three data sets, from Figure 2, the weight of Bayesian MSOE model (panel (a)) is higher than 0.92 and fairly close to unity for the majority of time series. The remaining weight is allocated to SSOE (panel (b)) with the random walk and AR(1) models receiving a zero weight in all cases examined here (results not reported here but

[3]In our computations we used the `fortran 77` version of library `lbfgs` in `netlib`. The problem can be converted to unconstrained optimization by reparametrizing $w_m = \frac{\exp\left(-\omega_m^2\right)}{1+\sum_{m'>1}\exp\left(-\omega_{m'}^2\right)}$, and the ω_is are defined on the real line. Convergence is quite fast given any initial conditions and no other local optima were found.

available on request). The evidence from optimal weight pools shows that MSOE definitely has better predictive better than the SSOE model in a substantive sense.

Concluding Remarks

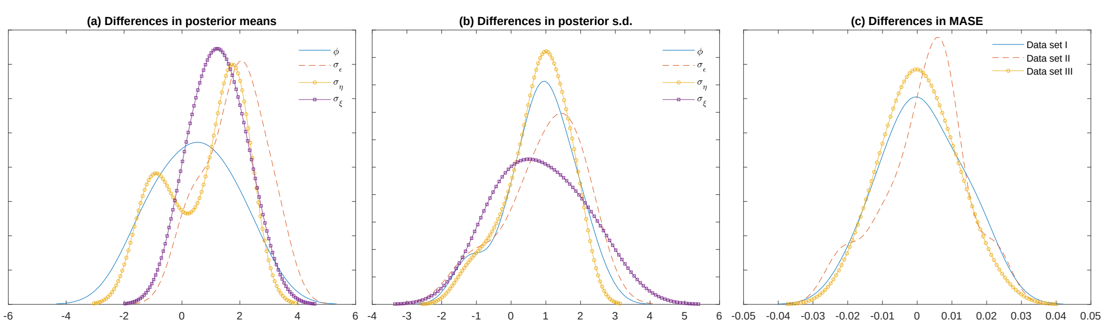
In this paper we have considered Bayesian inference and out-of-sample forecasting using the multiple sources of error (MSOE) damped trend model. Monte Carlo simulations and empirical evidence point to the direction that the Bayesian MSOE model performs better than the Bayesian single source of error (SSOE) damped trend model.

Technical Appendix

A.1 Prior sensitivity analysis

We deviate from our benchmark prior by assuming that $\bar{N}_j$ and $\bar{Q}_j$ are drawn from a uniform distribution in the interval (0.001, 10). We draw 10,000 values for these prior parameters and we redo posterior analysis using the Sampling-Importance-Resampling (SIR) approach of Rubin (1987, 1988), see also Smith and Gelfand (1992). The size of the re-sample in the SIR algorithm is set to 20% of the original MCMC sample. Differences of posterior means and posterior standard deviations of the model (across all time series in the three data sets) are reported in Figure A.1.

Figure A.1: Prior sensitivity analysis

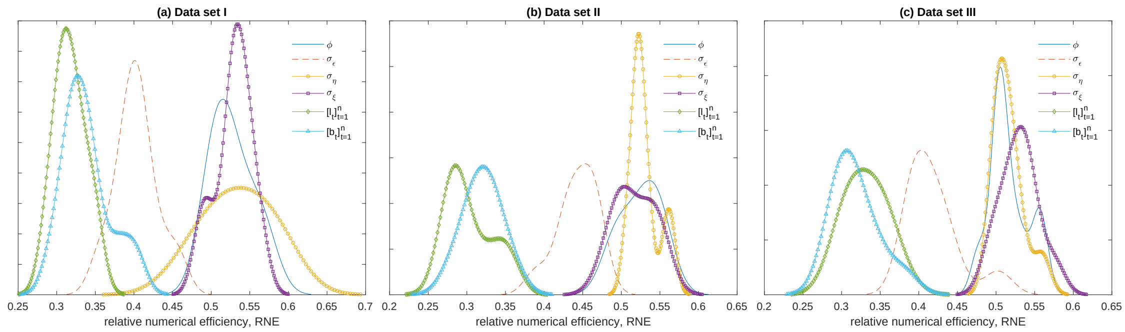


Notes: Data set I is Annual M3 data. Data set II is Quarterly BIS credit data quarterly BIS credit data, and Data set III is the shorter sample (2008:Q1-2019:Q1). MASE is Mean Absolute Scaled Error as proposed by Hyndman and Koehler (2006) and Sbrana and Silvestrini (2020). For the data, see Sbrana and Silvestrini (2020). Results in panels (a), (b), and (c) are across 50 randomly selected time series in a given data set.

A.2 MCMC autocorrelation

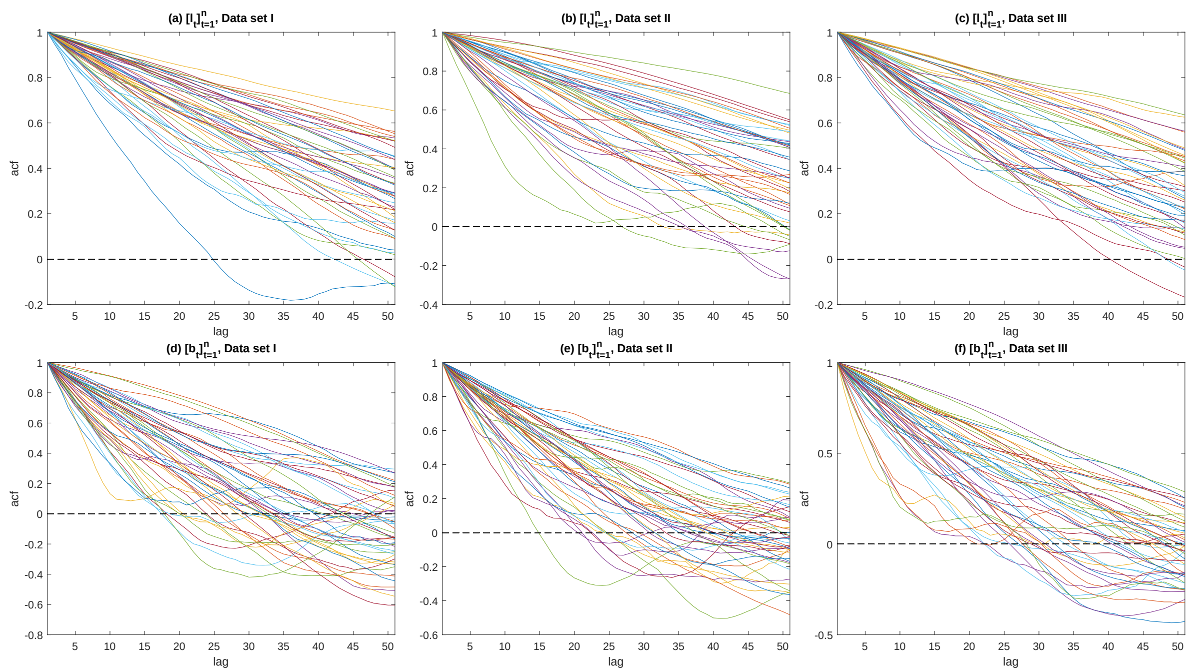
In Figure A.3 we report median autocorrelation functions across 50 (randomly selected) time series in a given data set.

Figure A.3: Relative Numerical Efficiency (RNE)



Notes: Data set I is Annual M3 data. Data set II is Quarterly BIS credit data quarterly BIS credit data, and Data set III is the shorter sample (2008:Q1-2019:Q1). MASE is Mean Absolute Scaled Error as proposed by Hyndman and Koehler (2006) and Sbrana and Silvestrini (2020). For the data, see Sbrana and Silvestrini (2020). Results in panels (a), (b), and (c) are medians across 50 randomly selected time series in a given data set.

Figure A.2: Autocorrelation functions



Notes: Data set I is Annual M3 data. Data set II is Quarterly BIS credit data quarterly BIS credit data, and Data set III is the shorter sample (2008:Q1-2019:Q1). MASE is Mean Absolute Scaled Error as proposed by Hyndman and Koehler (2006) and Sbrana and Silvestrini (2020). For the data, see Sbrana and Silvestrini (2020). Results in panels (a), (b), and (c) are across 50 randomly selected time series in a given data set.

In Figure A.4 we report median relative numerical efficiency (RNE) indices (Geweke, 1992) for 50 (randomly selected) time series in a given data set. If iid sampling from the posterior were possible, RNEs would be equal to one. In MCMC we have to settle, of course, for lower values of RNEs, so that autocorrelation is not destructively large (defined as a situation where RNEs are close to zero).

References

[1] Bernardo, Jose, M., 1979. Expected information as expected utility. The Annals of Statistics 7, 686–690.

[2] De Bonis, Riccardo, Silvestrini, Andrea, 2014. The Italian financial cycle: 1861–2011. Cliometrica J. Hist. Econ. Econom. Hist. 8 (3), 301–334.

[3] Durbin, James, Koopman, Siem Jan, 2012. Time Series Analysis by State Space Methods. Oxford University Press, Oxford.

[4] Fildes, Robert, Nikolopoulos, Konstantinos, Crone, Sven, Syntetos, Aris A., 2008. Forecasting and operational research: A review. J. Oper. Res. Soc. 59 (9), 1150–1172

[5] Gardner, Jr., Everette, McKenzie, Eddie, 1988. Model identification in exponential smoothing. J. Oper. Res. Soc. 39 (9), 863–867.

[6] Gardner, Jr., Everette, McKenzie, Eddie, 1989. Seasonal exponential smoothing with damped trends. Manage. Sci. 35 (3), 372–376.

[7] Gardner, Jr., Everette, McKenzie, Eddie, 2011. Why the damped trend works. J. Oper. Res. Soc. 62 (6), 1177–1180.

[8] Gelfand, Alan, E., Smith, Adrian, F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. J. Amer. Statist. Assoc. 85 (410), 398–409.

[9] Geweke, John, 1992. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4* (J. M. Bernado, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Clarendon Press, Oxford, UK, 169–193.

[10] Geweke, John, Amisano, Gianni, 2011a. Optimal prediction pools. Journal of Econometrics 164, 130–141.

[11] Geweke, John, Amisano, Gianni, 2011b. Hierarchical Markov normal mixture models with applications to financial asset returns. Journal of Applied Econometrics 26, 1–29.

[12] Harvey, Andrew C., 1989. Forecasting, Structural Time Series Analysis, and the Kalman Filter. Cambridge.

[13] Hyndman, Rob J., Koehler, Anne B., Ord, John Keith, Snyder, Ralph D., 2008. Forecasting with Exponential Smoothing: The State Space Approach. Springer Science & Business Media.

[14] Makridakis, Spyros, Hibon, Michele, 2000. The M3-Competition: results, conclusions and implications. Int. J. Forecast. 16 (4), 451–476.

[15] Ord, John Keith, Koehler, Anne B., Snyder, Ralph D., 1997. Estimation and prediction for a class of dynamic nonlinear statistical models. J. Amer. Statist. Assoc. 92 (440), 1621–1629.

[16] Rubin, Donald, B., 1987. Comment on 'The calculation of posterior distributions by data augmentation', by M. A. Tanner and W. H. Wong", J. Amer. Statist. Assoc. 82, 543–546.

[17] Rubin, Donald,. B., 1988). Using the SIR Algorithm to Simulate Posterior Distributions", In Bayesian Statistics 3, ed. J. M. Bernardo, M. H. DeGroot , D. V. Lindley, and A. F. M. Smith, 395–402. Oxford: Oxford University Press.

[18] Sbrana, Giacomo, Silvestrini, Andrea, 2020. Forecasting with the damped trend model using the structural approach. International Journal of Production Economics, to appear.

[19] Smith, Adrian, F. M., Gelfand, Alan, 1992. Bayesian statistics without tears: A sampling–resampling perspective. The American Statistician 46, 84–88.

[20] Tanner, Martin A., Wong, Wing Hung (1987). The Calculation of Posterior Distributions by Data Augmentation. J. Amer. Statist. Assoc. 82 (398), 528–540.

[21] West, Mike, Harrison, Jeff, 2006. Bayesian Forecasting and Dynamic Models. In: Springer Series in Statistics, Springer, New York.