Stochastic Modeling and Applications
Vol.20, No. 1 (June 2016), 1-16

MUK PUBLICATIONS
Open Access Publisher

SEQUENTIAL MODELS FOR TIME-EVOLVING REGRESSION PROBLEMS WITH AN APPLICATION TO ENERGY DEMAND PREDICTION

ROBERT G. AYKROYD AND NADA ALFAER

ABSTRACT. In recent years, there has been a dramatic increase in the use of data that are collected over time and hence models with a temporal component, leading to dynamic models, have received increasing attention. The proposed approach uses a general framework which permits many special cases to be considered. Put simply, for each time a parametric observation model is defined with a conditional auto-regressive type model defined relating the parameters at one time to previous parameter values, this is called the evolution equation. Simulation results will be presented investigating estimator properties considering a temporally changing regression problem with results demonstrating improved estimation. The technique will also be applied to a real dataset examining the changing relationship between ambient temperature and electricity consumption in the UK. The fitted model can then be used to predict future demand based on easily obtained temperature forecast information.

1. Introduction

Statistical modelling is being used in an ever increasing range of applications from molecular human biology to the search for exoplanets. These new areas typically provide richer datasets which need correspondingly more complex and specialized statistical techniques to be created. One exemplary situation is in economics and finance were "big data" problems arise due to high-frequency sampling and to the linking of data collected from diverse sources. The term *dynamic* indicates that temporal change occurs in a process as it evolves over time. It has been claimed that "process modelling is the single technology that has had the biggest impact on business in the last decade", and hence it is clearly a valuable and important area of statistical research. Dynamic model analysis is based on groups of models which can be define as a sequence changing over time and allows the detailed properties of the process to be studies using a rigorous framework. As with all models, a dynamic model is a simplified representation of reality but it aims to capture the salient features of the temporal behaviour of the process. In particular, the aim in the dynamic model is to explain as much of the relationship between the variables which are observed over time as possible. It is usual to use standard statistical models for each time point, but to also describe changes in model parameters through Markov or auto-regressive type processes as discrete time approximations to differential equations.

Dynamic models, sometimes called state space models, date back to the Kalman filter [13] but were first used in time series analysis in [11], then further modelling was studied in [12], [19] and [10]. Good reviews can also be found in, for example, [14], [17] and [26].

2000 *Mathematics Subject Classification.* Primary 62L12; Secondary 62J05.

Key words and phrases. Bayesian estimation, big data, dynamic models, electricity consumption, hierarchical modelling, maximum likelihood, regression, supply and demand, temporal models.

The linear dynamic model was extended to the nonlinear case in [27]. Popular examples of dynamic approaches involve ARCH [6] and GARCH models which use autoregressive processes to describe changing variance and general correlation structures in time series. For a general review of the family of ARCH models see [7].

Many applications of dynamic modelling exist, for example in finance applications it is not possible to perform a designed experiment with replication but instead the data can be observed only from repeated observations as the system is evolving and hence there is no choice other than using a dynamic model to analyse the data. For an application to financial mathematics see [5]. In biology, dynamic models have been used to look at how the behaviour of different species, which share a common habitat, changes through time. For further examples in biology see, for example, [3] and [28], and for examples in geophysics see [1] and [4].

Here we adopt a statistical approach with a Bayesian perspective giving careful attention to the general approach, but also showing useful specific examples. Key theoretical results, using the general linear dynamic model, are derived, for two novel sequential approaches. The first of these used conditional updating with estimation for the current time based on the estimates at the previous time, whereas the second performs simultaneous estimation for the current and several previous times. The results presented allow readers to apply the same techniques, not just on the examples given here, but also for their own specific linear model and hence can have widespread impact. The theory is investigated through simulation and illustrated using a real data example. In particular, this paper is structured as follows. Section 2 provides background to dynamic modelling and details of the proposed linear dynamic model approach focusing on general linear dynamic problems. Section 3 presents simulation studies to investigate estimation properties. The methods are then applied to real data in Section 4. The final summary and conclusions are presented in Section 5.

2. Dynamic modelling and parameter estimation

2.1. General definitions. Consider a dynamic process with a relationship between data and model parameters where the parameters change through time, and even where the relationship itself might vary with time. For background information see, for example, [10], [12] and [19]. Suppose there are T times and that these occur at distinct times $\boldsymbol{T} = \{t_k : k = 1, \ldots, T\}$. Further suppose that a dataset of n_k measurements, $\mathbf{y}_k = \{y_{ik} : i = 1, \ldots, n_k\}$, is recorded at time t_k, $k = 1, \ldots, T$, which depends on a parameter vector $\boldsymbol{\theta}_k = \{\theta_{jk} : j = 1, \ldots, p\}$ through a known function $F_k(\cdot)$ and random noise $\boldsymbol{\epsilon}_k = \{\epsilon_{ik} : i = 1, \ldots, n_k\}$, that is

$$\mathbf{y}_k = F_k(\boldsymbol{\theta}_k) + \boldsymbol{\epsilon}_k, \qquad k = 1, \ldots, T, \tag{2.1}$$

which is known as the *observation equation*. The second equation is the *evolution equation* which can be written as

$$\boldsymbol{\theta}_k = G_k(\boldsymbol{\theta}_{k-1}, \ldots, \boldsymbol{\theta}_1) + \boldsymbol{\nu}_k, \qquad k = 2, \ldots, T \tag{2.2}$$

where $G(\cdot)$ is a known function and $\boldsymbol{\nu}_k = \{\nu_{jk} : j = 1, \ldots, p\}$ is random noise. Although this general case allows greater flexibility, it is usual to assume only a first-order dependency in which case $G_k(\boldsymbol{\theta}_{k-1}, \ldots, \boldsymbol{\theta}_1) \equiv G_k(\boldsymbol{\theta}_{k-1})$. Of course in the most general setting there is a very wide choice of error model, and even the errors do not need to be

additive. Perhaps the most flexible is to introduce a *generalized dynamic model* incorporating a link function, $\mu(\cdot)$, leading to the definition $\mu(E[\boldsymbol{y}_k]) = F_k(\boldsymbol{\theta}_k)$. The most usual assumption, however, is that the errors are additive and well described by a multivariate normal distribution, with a possible further assumption of independent variables with identical distributions. Background to regression modelling can be found in, for example, [8], [18], [21] and [25].

Now to move to a specific setting, consider the linear model with Gaussian distributed errors which is defined by observation equation

$$\mathbf{y}_k = F_k\,\boldsymbol{\theta}_k + \boldsymbol{\epsilon}_k, \qquad\qquad \boldsymbol{\epsilon}_k \sim N_{n_k}(\mathbf{0}, \Sigma_\epsilon), \tag{2.3}$$

where $F_k = \{F_{ijk}, i = 1, \ldots, n_k, j = 1, \ldots, p\}$ is an $n_k \times p$ design matrix of explanatory variables, $\boldsymbol{\theta}_k$ is, as above, a p-dimensional vector of regression parameters, and $\boldsymbol{\epsilon}_k$ is a Gaussian random vector with mean zero and known covariance matrix Σ_ϵ. The second equation, the linear evolution equation, is given by

$$\boldsymbol{\theta}_k = G_k\,\boldsymbol{\theta}_{k-1} + \boldsymbol{\nu}_k, \qquad\qquad \boldsymbol{\nu}_k \sim N_p(\mathbf{0}, \Sigma_\nu), \tag{2.4}$$

where $G_k = \{G_{jj'k}, j = 1, \ldots, p, j' = 1, \ldots, p\}$ is a $p \times p$ evolution matrix and $\boldsymbol{\nu}_k$ is a Gaussian random vector with mean zero and known covariance matrix Σ_ν. In the case where G_k is the identity matrix and $\boldsymbol{\nu}_k$ is identically zero at each time point then, this becomes a non-dynamic model.

In the terminology of state-space models, for a discrete time process, $\mathbf{y}_k$ is the output vector, F_k is the output matrix, $\boldsymbol{\theta}_k$ is the state vector, and G_k is the state, system or transition matrix. Further, especially in financial applications, the distribution of the errors $\boldsymbol{\nu}_k$ is known as the shock distribution and the covariance matrix defines a volatility matrix. If F_k and G_k do not depend on time, then the process is said to be time-invariant.

In the dynamic linear model the $\mathbf{y}_k$ are independent given $\boldsymbol{\theta}_k$, and then $\mathbf{y}_k$ depends only on $\boldsymbol{\theta}_k$. Further, the sequence $\boldsymbol{\theta}_k, k = 1, \ldots, T$ forms a first-order p-dimensional vector auto-regressive process, that is a Markov chain random walk. Under these assumptions the model can be written as

$$\mathbf{y}_k \mid \boldsymbol{\theta}_k \sim N_{n_k}(F_k\,\boldsymbol{\theta}_k, \sigma^2 I_{n_k}), \qquad k = 1, \ldots, T \tag{2.5}$$

and

$$\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{k-1} \sim N_p(G_k\,\boldsymbol{\theta}_{k-1}, \tau^2 I_p), \qquad k = 2, \ldots, T. \tag{2.6}$$

These equations describe the general and specific system which is a dynamic model.

2.2. Step-wise conditional parameter estimation. When the observation and evolution equations are written in the forms of equations (2.5) and (2.6), the hierarchical structure is more evident and hence these can be usefully thought of, in a Bayesian setting, as defining likelihood and prior respectively. For a general introduction to Bayesian methods see, for example, [15] and for a more in-depth and theoretical perspective see [2]. That is with likelihood, $\ell(\boldsymbol{\theta}_k) \equiv f(\mathbf{y}_k|\boldsymbol{\theta}_k)$ defined by the density function

$$f(\mathbf{y}_k|\boldsymbol{\theta}_k) = \frac{1}{(2\pi\sigma^2)^{n_k/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_k - F_k\boldsymbol{\theta}_k)^T(\mathbf{y}_k - F_k\boldsymbol{\theta}_k)\right\}, \tag{2.7}$$

and prior distribution, $\pi(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})$, defined by the density function

$$\pi(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}) = \frac{1}{(2\pi\tau^2)^{p/2}} \exp\left\{-\frac{1}{2\tau^2}(\boldsymbol{\theta}_k - G_k\boldsymbol{\theta}_{k-1})^T(\boldsymbol{\theta}_k - G_k\boldsymbol{\theta}_{k-1})\right\}. \tag{2.8}$$

In the Bayesian approach the likelihood and the prior are combined into the posterior distribution, $\pi(\boldsymbol{\theta}_k|\mathbf{y}_k, \boldsymbol{\theta}_{k-1})$, with density function defined, through Bayes's theorem by

$$\begin{aligned}\pi(\boldsymbol{\theta}_k|\mathbf{y}_k, \boldsymbol{\theta}_{k-1}) &= \frac{f(\mathbf{y}_k|\boldsymbol{\theta}_k)\pi(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})}{f(\mathbf{y}_k)} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_k - F_k\boldsymbol{\theta}_k)^T(\mathbf{y}_k - F_k\boldsymbol{\theta}_k)\right. \\ &\qquad \left. -\frac{1}{2\tau^2}(\boldsymbol{\theta}_k - G_k\boldsymbol{\theta}_{k-1})^T(\boldsymbol{\theta}_k - G_k\boldsymbol{\theta}_{k-1})\right\}. \end{aligned} \tag{2.9}$$

Estimation and inference is then based on this distribution which balances evidence from data with information from temporal constraints.

For the estimation of $\boldsymbol{\theta}_k$, consider finding the maximum a posterior (MAP) estimate by differentiation the log-posterior distribution

$$\begin{aligned}\log \pi(\boldsymbol{\theta}_k|\mathbf{y}_k, \boldsymbol{\theta}_{k-1}) &= -\frac{1}{2\sigma^2}(\mathbf{y}_k - F_k\boldsymbol{\theta}_k)^T(\mathbf{y}_k - F_k\boldsymbol{\theta}_k) \\ &\quad - \frac{1}{2\tau^2}(\boldsymbol{\theta}_k - G_k\boldsymbol{\theta}_{k-1})^T(\boldsymbol{\theta}_k - G_k\boldsymbol{\theta}_{k-1}) + \mathcal{C}, \end{aligned} \tag{2.10}$$

where $\mathcal{C}$ contains the accumulated constant terms, to give

$$\frac{d}{d\boldsymbol{\theta}_k}\log \pi(\boldsymbol{\theta}_k|\mathbf{y}_k, \boldsymbol{\theta}_{k-1}) = \frac{1}{\sigma^2}F_k^T(\mathbf{y}_k - F_k\boldsymbol{\theta}_k) - \frac{1}{\tau^2}(\boldsymbol{\theta}_k - G_k\boldsymbol{\theta}_{k-1}) \tag{2.11}$$

and setting to zero. Hence the MAP estimate, $\widehat{\boldsymbol{\theta}}_k$ of $\boldsymbol{\theta}_k$, is the given by

$$\widehat{\boldsymbol{\theta}}_k = \left(F_k^TF_k + \kappa I_p\right)^{-1}\left(F_k^T\mathbf{y}_k + \kappa G_k\boldsymbol{\theta}_{k-1}\right) \tag{2.12}$$

where $\kappa = \sigma^2/\tau^2$. Then the posterior expectation and covariance of $\widehat{\boldsymbol{\theta}}_k$ can be shown to be

$$E(\widehat{\boldsymbol{\theta}}_k) = \left(F_k^TF_k + \kappa I_p\right)^{-1}\left(F_k^TF_k\boldsymbol{\theta}_k + \kappa G_k\boldsymbol{\theta}_{k-1}\right) \tag{2.13}$$

and

$$\text{var}(\widehat{\boldsymbol{\theta}}_k) = \sigma^2(F_k^TF_k + \kappa I_p)^{-T}F_k^TF_k(F_k^TF_k + \kappa I_p)^{-1}. \tag{2.14}$$

Further, the Hessian matrix is given by

$$\frac{d^2}{d\boldsymbol{\theta}_k^2}\log \pi(\boldsymbol{\theta}_k|\mathbf{y}_k, \boldsymbol{\theta}_{k-1}) = -\frac{1}{\sigma^2}\left(F_k^TF_k + \kappa I_p\right) \tag{2.15}$$

and hence an asymptotic approximation to the posterior covariance matrix

$$\text{var}(\widehat{\boldsymbol{\theta}}_k) = \sigma^2\left(F_k^TF_k + \kappa I_p\right)^{-1}. \tag{2.16}$$

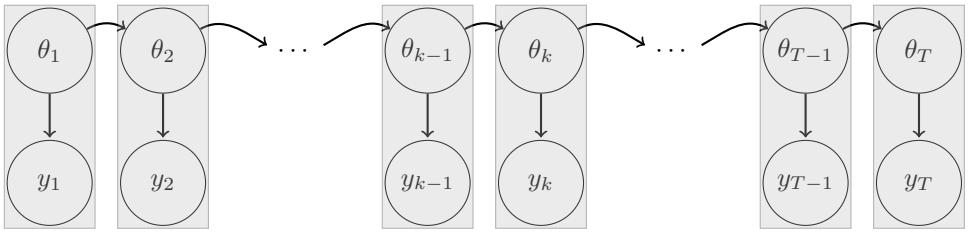
As special estimation cases note that as $\kappa \to 0$, that is as $\tau \to \infty$, these become $\widehat{\boldsymbol{\theta}}_k = \left(F_k^TF_k\right)^{-1}F_k^T\mathbf{y}_k$ which is the usual regression estimate, and hence $E(\widehat{\boldsymbol{\theta}}_k) = \boldsymbol{\theta}_k$, $\text{var}(\widehat{\boldsymbol{\theta}}_k) = \sigma^2(F_k^TF_k)^{-1}$. Then, similarly, as $\kappa \to \infty$, that is as $\tau \to 0$, the estimate becomes $\widehat{\boldsymbol{\theta}}_k = G_k\boldsymbol{\theta}_{k-1}$, which is simple the projection of the previous parameter estimates forward to the new time. In particular, this is non-random and hence $E(\widehat{\boldsymbol{\theta}}_k) = G_k\boldsymbol{\theta}_{k-1}$ and $\text{var}(\widehat{\boldsymbol{\theta}}_k) = 0$. This estimation has only considered a single time for given value at the previous time. This is illustrated in Figure 1 and in Table 1.

In practice, these given values are in fact also estimates and hence have an attached uncertainty. If a sequence of step-wise estimates are calculated, then these accumulated

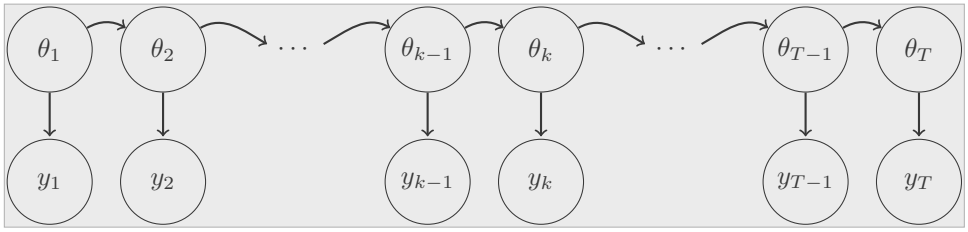
Time	$k=1$	$k=2$	$\cdots$	$k=T$
Estimation	$\hat{\boldsymbol{\theta}}_1 \vert \mathbf{y}_1$	$\hat{\boldsymbol{\theta}}_2 \vert \hat{\boldsymbol{\theta}}_1, \mathbf{y}_2$		$\hat{\boldsymbol{\theta}}_T \vert \hat{\boldsymbol{\theta}}_{T-1}, \mathbf{y}_T$

TABLE 1. Conditional structure of step-wise conditional estimation.

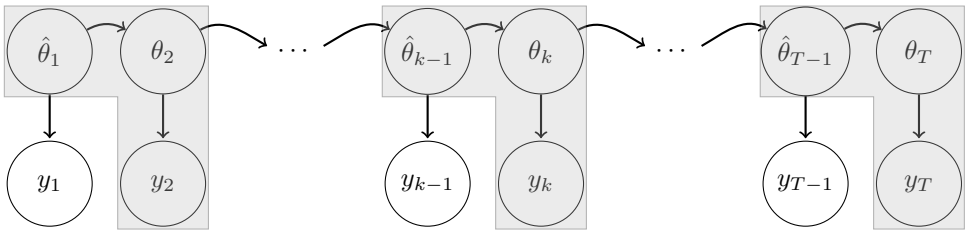
uncertainties could be substantial. Although it might be possible to derive expressions for the total uncertainty it is beyond the scope of this paper.



(A) Independent parameter estimation ignoring dynamic component, where the grey area shows that each parameter is estimated using the corresponding single dataset.



(B) Full parameter estimation with joint dynamic component, where the grey area shows that all parameters are estimated using all datasets.



(C) Conditional parameter estimation with step-wise dynamic component, where the grey area shows that a single parameter is estimated using the corresponding single dataset and the previous parameter estimate.

FIGURE 1. *Directed graphs showing the hierarchical relationships between model parameters and datasets for (A) independent estimation, (B) joint estimation and (C) conditional estimation.*

2.3. Joint modelling. The aim is now to estimate all parameters up to and including a particular time, $k = K$, with $1 \leq K \leq T$, using all available data. Let the complete set of parameters be labelled $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$, and the available data $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_K\}$. Also, for ease of notation, define two extra partial parameter vectors $\boldsymbol{\theta}_{-1} = \{\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K\}$ which has the parameters from the first time removed, and $\boldsymbol{\theta}_{-K} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{K-1}\}$ which has the parameters from the K-th time removed.

Now the likelihood, $\ell(\boldsymbol{\theta}) \equiv f(\mathbf{y}|\boldsymbol{\theta})$ is defined by

$$f(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - F\boldsymbol{\theta})^T(\mathbf{y} - F\boldsymbol{\theta})\right\}, \tag{2.17}$$

where $N = n_1 + \cdots + n_K$, and F is a block diagonal matrix formed from $F_1, \ldots, F_K$. Similarly, a block matrix G_{-K} is formed from $G_1, \ldots, G_{K-1}$. These two matrices can be written as

$$F = \begin{bmatrix} F_1 & 0 & \ldots & 0 \\ 0 & F_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & F_K \end{bmatrix}, \quad \text{and} \quad G_{-K} = \begin{bmatrix} G_1 & 0 & \ldots & 0 \\ 0 & G_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & G_{K-1} \end{bmatrix}. \tag{2.18}$$

Next the prior distribution, $\pi(\boldsymbol{\theta})$, defined by the density function

$$\pi(\boldsymbol{\theta}) = \frac{1}{(2\pi\tau^2)^{P/2}} \exp\left\{-\frac{1}{2\tau^2}(\boldsymbol{\theta}_{-1} - G_{-K}\boldsymbol{\theta}_{-K})^T(\boldsymbol{\theta}_{-1} - G_{-K}\boldsymbol{\theta}_{-K})\right\}, \tag{2.19}$$

where $P = (K-1) \times p$. In the Bayesian approach these are combined into the posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{y})$, with density function defined by

$$\begin{aligned} \pi(\boldsymbol{\theta}|\mathbf{y}) &= f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \big/ f(\mathbf{y}) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - F\boldsymbol{\theta})^T(\mathbf{y} - F\boldsymbol{\theta})\right. \\ &\qquad \left. -\frac{1}{2\tau^2}(\boldsymbol{\theta}_{-1} - G_{-K}\boldsymbol{\theta}_{-K})^T(\boldsymbol{\theta}_{-1} - G_{-K}\boldsymbol{\theta}_{-K})\right\}. \end{aligned} \tag{2.20}$$

Again, this distribution is the basis for estimation and inference.

For the estimation of $\boldsymbol{\theta}$, consider finding the maximum a posterior (MAP) estimate by differentiation the log-posterior distribution

$$\log \pi(\boldsymbol{\theta}|\mathbf{y}) = -\frac{1}{2\sigma^2}(\mathbf{y} - F\boldsymbol{\theta})^T(\mathbf{y} - F\boldsymbol{\theta}) - \frac{1}{2\tau^2}(\boldsymbol{\theta}_{-1} - G_{-K}\boldsymbol{\theta}_{-K})^T(\boldsymbol{\theta}_{-1} - G_{-K}\boldsymbol{\theta}_{-K}) + \mathcal{C}, \tag{2.21}$$

where $\mathcal{C}$ is again a constant term. This time, however, greater care must be used as not all parameters appear in the prior component in the same way. Instead three cases must be considered leading to the following derivatives

$$\frac{d}{d\boldsymbol{\theta}_1} \log \pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{\sigma^2} F_1^T(\mathbf{y}_1 - F_1\boldsymbol{\theta}_1) - \frac{1}{\tau^2} G_1^T(\boldsymbol{\theta}_2 - G_1\boldsymbol{\theta}_1), \tag{2.22}$$

for the parameters from the first time, then

$$\frac{d}{d\boldsymbol{\theta}_k} \log \pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{\sigma^2} F_k^T(\mathbf{y}_k - F_k\boldsymbol{\theta}_k) - \frac{1}{\tau^2} \begin{bmatrix} 1 \\ -G_k \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\theta}_k - G_{k-1}\boldsymbol{\theta}_{k-1} \\ \boldsymbol{\theta}_{k+1} - G_k\boldsymbol{\theta}_k \end{bmatrix}, \tag{2.23}$$

for the parameters from each time, $2 \leq k \leq K-1$. Then, for the parameters from the final time

$$\frac{d}{d\boldsymbol{\theta}_K} \log \pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{\sigma^2} F_K^T(\mathbf{y}_K - F_K \boldsymbol{\theta}_K) - \frac{1}{\tau^2} 1_p^T(\boldsymbol{\theta}_K - G_{K-1}\boldsymbol{\theta}_{K-1}). \tag{2.24}$$

Similarly, there are three parts to the MAP estimate, $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, given by

$$\widehat{\boldsymbol{\theta}}_1 = \left(F_1^T F_1 - \kappa G_1^T G_1\right)^{-1} \left(F_1^T \mathbf{y}_1 + \kappa G_1^T \boldsymbol{\theta}_2\right), \tag{2.25}$$

for the parameters from the first time, then

$$\widehat{\boldsymbol{\theta}}_k = \left(F_k^T F_k + \kappa I_p\right)^{-1} \left(F_k^T \mathbf{y}_k + \kappa G_{k-1}^T \boldsymbol{\theta}_{k-1}\right), \qquad 2 \leq k \leq K-1, \tag{2.26}$$

then, for the parameters from the final time

$$\widehat{\boldsymbol{\theta}}_K = \left(F_K^T F_K + \kappa I_p\right)^{-1} \left(F_K^T \mathbf{y}_K + \kappa\, G_{K-1}\boldsymbol{\theta}_{K-1}\right). \tag{2.27}$$

For ease of solution, these equations can be formed into a linear system

$$A\,\widehat{\boldsymbol{\theta}} = F^T \mathbf{y}, \tag{2.28}$$

where the $p \times p$ matrix A is given by

$$\begin{bmatrix} F_1^T F_1 + \kappa G_1^T G_1 & -\kappa G_1^T & 0 & 0 & \dots & 0 \\ -\kappa G_2 & F_2^T F_2 + \kappa(G_2^T G_2 + I_p) & -\kappa G_2^T & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & -\kappa\, G_{K-2} & F_{K-2}^T F_{K-2} + \kappa\,(G_{K-2}^T G_{K-2} + I_p) & -\kappa\, G_{K-2}^T \\ 0 & \dots & 0 & 0 & -\kappa\, G_{K-1} & F_K^T F_K + \kappa I_p \end{bmatrix}, \tag{2.29}$$

and hence the estimates are given by

$$\widehat{\boldsymbol{\theta}} = A^{-1}\, F^T \mathbf{y}. \tag{2.30}$$

Then, for completeness, the expectation and the covariance are given by

$$E(\widehat{\boldsymbol{\theta}}) = A^{-1}\, F^T F \boldsymbol{\theta} \quad \text{and} \quad \text{var}(\widehat{\boldsymbol{\theta}}) = \sigma^2 A^{-1}\, F^T F A^{-T}. \tag{2.31}$$

The above estimation considers data from all previous times simultaneously and estimates all corresponding parameters. This is illustrated in Figure 1 and in Table 2.

Time	$k=1$	$k=2$	$\cdots$	$k=K$
Estimation	$\hat{\boldsymbol{\theta}}_1 \vert \mathbf{y}_1$	$\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2 \vert \mathbf{y}_1, \mathbf{y}_2$		$\hat{\boldsymbol{\theta}}_1, \cdots, \hat{\boldsymbol{\theta}}_K \vert \mathbf{y}_1, \cdots, \mathbf{y}_K$

TABLE 2. Conditional structure of simultaneous joint estimation

3. Estimator properties through simulation

3.1. Dynamic linear models. The purpose of this section is to demonstrate the performance of step-wise conditional and simultaneous joint estimation of parameters in dynamic linear model. In particular, consider the dynamic simple linear regression model, see [14], [17] and [19], which is fully defined by the observation equation, describing the data $\mathbf{y}_k$, with $n_k = n$ for all $k = 1, \ldots, T$, in terms of explanatory variables $\mathbf{x}_k$, given by

$$\mathbf{y}_k = \alpha_k + \beta_k \mathbf{x}_k + \boldsymbol{\epsilon}_k, \qquad k = 1, \ldots, T, \tag{3.1}$$

where α_k and β_k are the regression parameters and $\boldsymbol{\epsilon}_k \sim N_n(\mathbf{0}, \sigma^2 I_n)$. Hence, $\boldsymbol{\theta}_k = [\alpha_t, \beta_k]^T$ and $F_t = [\mathbf{1}, \mathbf{x}_k]$. The second equation, the evolution equation, describes the change in α_k and β_k over time, and here is given by

$$\begin{bmatrix} \alpha_k \\ \beta_k \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_{k-1} \\ \beta_{k-1} \end{bmatrix} + \begin{bmatrix} u_k \\ v_k \end{bmatrix}, \qquad k = 2, \ldots, T, \tag{3.2}$$

where u_k and v_k are independent Gaussian random variables with zero mean and standard deviation τ. Here the evolution matrix G defined in equation (3.2) is the identity matrix.

The simulation experiment considers fixed sample size $n = 10$ and $T = 10$ times. In the observation equation the error standard deviation is set at $\sigma = 10$, which gives substantial variation around the line, and in the evolution equation $\tau = 2$, which produces moderate changes across the times. To start the parameter evolution initial information $\alpha_0 = 15$ and $\beta_0 = 25$ is used. Further, a set of $M = 1000$ replicates are considered to allow a reliable comparison of the different modelling approaches.

In this experiment interest is in estimating α_k and β_k over time. The procedure is to consider the usual linear regression to estimate the parameters individually, called independent estimation, and then sequentially using the simultaneous joint estimation and the step-wise conditional estimation. First in the individual estimation, represented in Figure 1(A), for each time point the parameters are estimate by fitting the usual linear regression model. This means that there is a single set of parameters estimates for each time. Second, for the simultaneous joint estimation, which is shown diagrammatically in Figure1(B) and in Table 2, the procedure starts at time $k = 1$ by simply estimating α_1 and β_1 from $\mathbf{y}_1$ as with the independent estimation. Then, at time $k = 2$, when new data are available, all data so far available are used to estimate α_2 and β_2 and also to re-estimate α_1 and β_1. For the remaining time points we repeat the process using all available data to estimate or re-estimate all corresponding parameters. Finally, for the last time point, $k = T$, all the data is used to re-estimate all the parameters from $k = 1$ up to $k = T - 1$ and estimate the final set of parameters at $k = T$. Third, the step-wise conditional estimation, which is shown diagrammatically in Figure1(C) and Table 1, is applied. The procedure is again to start at time $k = 1$ to estimate α_1 and β_1 using data $\mathbf{y}_1$. Then, at time $k = 2$ the parameters α_2 and β_2 are estimated using the just calculated estimates $\hat{\alpha}_1$ and $\hat{\beta}_1$, along with data $\mathbf{y}_2$. At the other time points the procedure is repeated to estimate α_k and β_k given $\hat{\alpha}_{k-1}$ and $\hat{\beta}_{k-1}$, along with data $\mathbf{y}_k$.

3.2. Output summary. The simulation experiment has produced $M = 1000$ replicate results with parameter estimates $\hat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{\theta}}_{jm}, j = 1, \ldots, T, m = 1, \ldots, M)$ where each estimate is made-up of the intercept and the slope of the linear regression, that is $\hat{\boldsymbol{\theta}}_{jm} = (\hat{\alpha}_{jm}, \hat{\beta}_{jm})$. All calculations were performed in R [20] using standard functions to allow

clearer understanding of the model structure – the corresponding code is available from the authors.

To summarise the results two output measures will be considered, the MSE and the bias which, for $\hat{\alpha}_j$, are defined as

$$\text{MSE}(\hat{\alpha}_j) = \frac{1}{M}\sum_{m=1}^{M}(\hat{\alpha}_{jm} - \alpha_{jm})^2 \qquad \text{and} \qquad \text{Bias}(\hat{\alpha}_j) = \frac{1}{M}\sum_{m=1}^{M}(\hat{\alpha}_{jm} - \alpha_{jm}), \tag{3.3}$$

with corresponding definitions for $\hat{\beta}_j$. Similarly, an average MSE and average bias combining the values over all times can be defined as

$$\text{AMSE}(\hat{\alpha}) = \frac{1}{T}\sum_{j}\text{MSE}(\hat{\alpha}_j) \qquad \text{and} \qquad \text{ABias}(\hat{\alpha}) = \frac{1}{T}\sum_{j}\text{Bias}(\hat{\alpha}_j), \tag{3.4}$$

with corresponding definitions for $\hat{\beta}$.

3.3. Numerical results. To start the investigation, values of τ were picked and the corresponding average MSE and the average bias of $\hat{\alpha}$ and $\hat{\beta}$ calculated. Figure 2 illustrates the performance of the joint estimation process at the final time, that is with the simultaneous estimation of all parameters, $\alpha_1, \ldots, \alpha_T$ and $\beta_1, \ldots, \beta_T$ from all the data $\mathbf{y}_1, \ldots, \mathbf{y}_T$. Notice that in these figures the scale in (A) is 10 times that in (B), that is the average MSE for $\hat{\alpha}$ is much greater than that for $\hat{\beta}$. Also, the scales for average bias cover a very narrow range and hence each estimator is essentially unbiased.

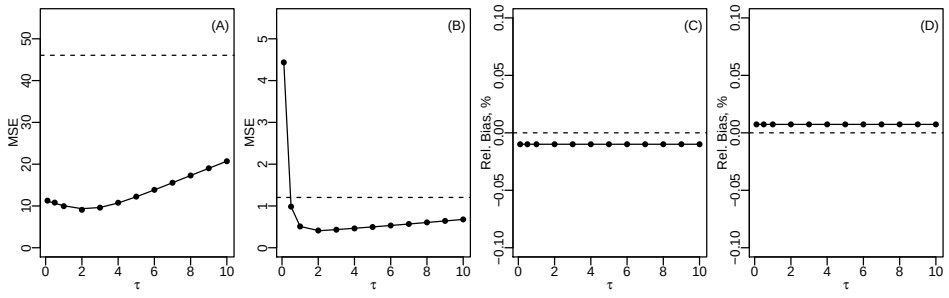


FIGURE 2. *Assessing the influence of τ on estimation using simultaneous joint estimation, with (A) MSE of $\hat{\alpha}$, (B) MSE of $\hat{\beta}$, (C) bias of $\hat{\alpha}$ and (D) bias of $\hat{\beta}$.*

We can see from Figure 2(A) and (B) how the MSE for both parameters initially decreases as τ increases, reaching a minimum MSE at $\tau = 2$, and then increases again towards the value corresponding to the independent estimation – which is shown as the horizontal dotted line. Hence, the MSE for all τ values is, almost always, lower than the MSE for the individual estimation. Looking now at the bias shown in Figure 2(C) and (D). There is no substantial change in the average bias, though it very gradually moves towards zero as τ increases. The limiting value, corresponding to independent estimation, is zero to 2 decimal places in each case. Hence, the estimators for all values of τ can be considered to be unbiased. Overall, it can be said that for very high values of τ, there is

no clear advantage of using the joint estimation, but for small and moderate values there is a clear advantage. In particular, the minimum in the average MSE occurs at the value of τ used in the simulation, and in the remainder of the study this is the case.

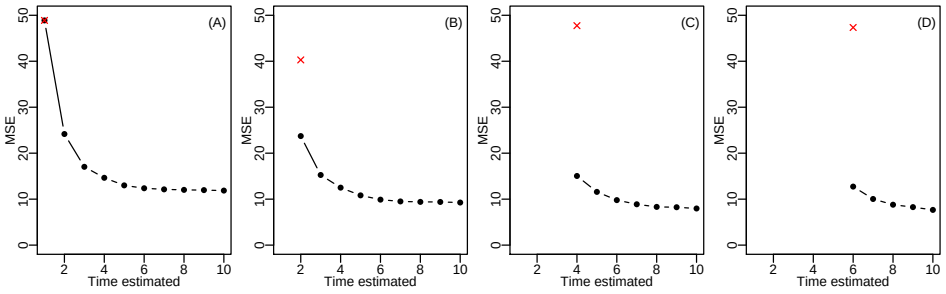


FIGURE 3. *The MSE for the joint estimation of α with (A) MSE of $\hat{\alpha}_1$, (B) MSE of $\hat{\alpha}_2$, (C) MSE of $\hat{\alpha}_4$ and (D) MSE of $\hat{\alpha}_6$.*

Consider now the effect of using the joint estimation sequentially. That is at time K, for $2 \leq K \leq T$, the data so far collected, $\mathbf{y}_1, \ldots, \mathbf{y}_K$, are used to estimation all of the corresponding parameters, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$. Clearly, this can be repeated at each time with the amount of available data and the number of parameters steadily growing. Figures 3 and 4 show the MSE of $\hat{\alpha}_k$ and $\hat{\beta}_k$, respectively, for $k = 1, 2, 4, 6$. In each panel the red cross represents the MSE from the individual independent estimation and the black line with dots shows the MSE from the simultaneous joint estimation. Note that it is not possible to estimate a parameter until the corresponding data has been collected. It is clear from the figures that MSE for the joint estimation decreases over time and in both cases there is a big jumping from the independent estimation to the first joint estimate, this is certainly due to the use of the previous information at each time point. For the later times, as more and more data is available there is only a small further decrease in MSE. Hence, this suggests that there is a great advantage in performing joint estimation with three or four previous times, but the further gain may not be worthwhile. This latter point, is most important if we imagine that the model assumptions may not be true in all examples and limiting the number of previous times will limit any biassing due to unexpected abrupt changes in the parameters.

Figures 5 and 6 show the bias of $\hat{\alpha}_k$ and $\hat{\beta}_k$, respectively, again for $k = 1, 2, 4, 6$ estimated sequentially. In all cases the bias when using joint estimation is less than for independent estimation, but there is only a small change as more data is used to estimate the larger number of parameters. Also, all bias values are very small and hence it can be concluded that estimation is essentially unbiased and conclusions can be based on the MSE pattern.

The final part of the investigation is to look at step-wise conditional estimation. In this case at each time k, for $2 \leq k \leq T$, the estimation involves the previous estimate $\hat{\boldsymbol{\theta}}_{k-1}$ and current data $\mathbf{y}_k$, to estimate one set of parameters $\boldsymbol{\theta}_k$. Figure 7 shows the MSE and bias for all times. Note that at the first time point the only estimation possible is independent estimation and hence all methods perform equally. In all panels, the circles represent the independent estimation and the crosses represent the step-wise conditional

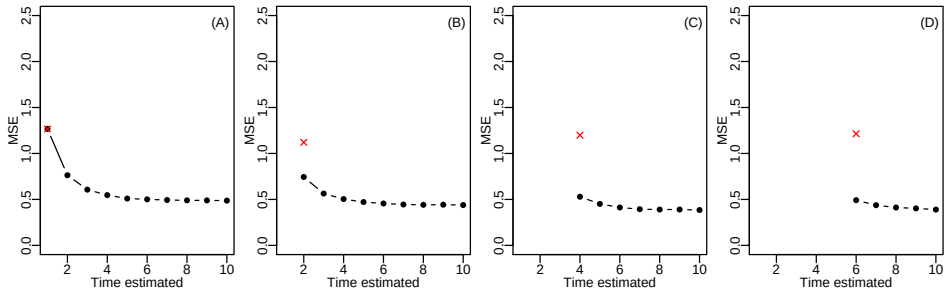


FIGURE 4. *The MSE for the joint estimation of β with (A) MSE of $\hat{\beta}_1$, (B) MSE of $\hat{\beta}_2$, (C) MSE of $\hat{\beta}_4$ and (D) MSE of $\hat{\beta}_6$.*

estimation. For the MSE, panels (A) and (B), the conditional estimates gradually reduces while the independent randomly varies around a constant level. The reduction is, however, not as dramatic as in the joint estimation. This shows the positive reinforcement of basing

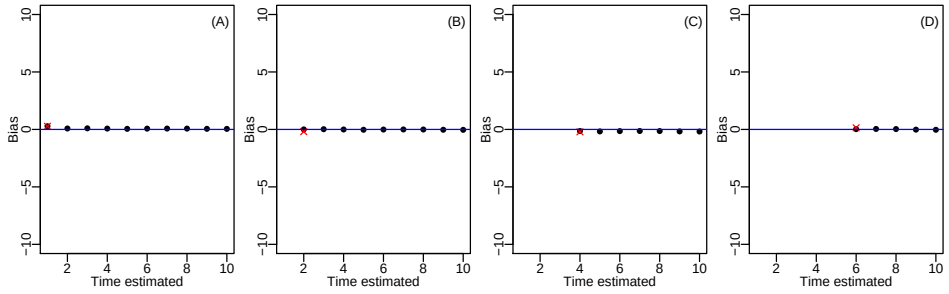


FIGURE 5. *The bias for the joint estimation of α with (A) bias of $\hat{\alpha}_1$, (B) bias of $\hat{\alpha}_2$, (C) bias of $\hat{\alpha}_4$ and (D) bias of $\hat{\alpha}_6$.*

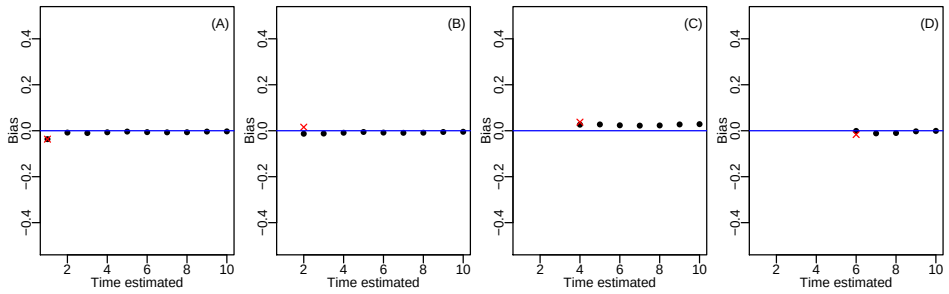


FIGURE 6. *The bias for the joint estimation of β with (A) bias of $\hat{\beta}_1$, (B) bias of $\hat{\beta}_2$, (C) bias of $\hat{\beta}_4$ and (D) bias of $\hat{\beta}_6$.*

current estimation on a good previous estimate. For the estimator bias, there is no clear pattern, but all values are small and hence compatible with being unbiased.

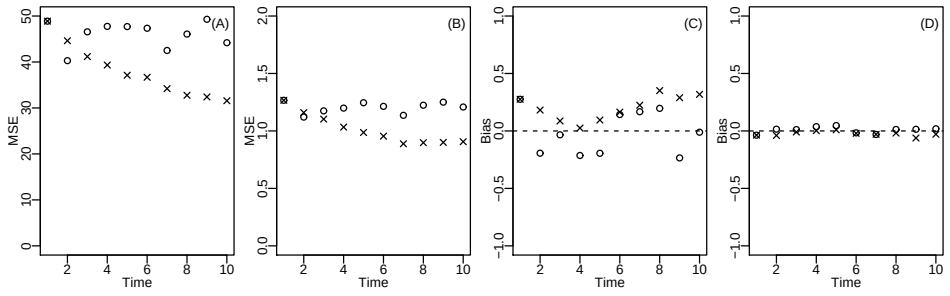


FIGURE 7. *A comparison of step-wise conditional estimation (crosses) and independent estimation (circles), with (A) MSE of $\hat{\alpha}$, (B) MSE of $\hat{\beta}$, (C) bias of $\hat{\alpha}$ and (D) bias of $\hat{\beta}$.*

4. Modelling temperature related electricity demand

There are many potential applications of dynamic modelling as almost every relationship will have a potentially hidden dependency on time. One such application is the use of temperature values to predict future electricity demand. Several approaches incorporate climatic variables, see for example [23], but most use time series based approaches, see for example [22] and [24]. In contrast to these approaches, the proposed method models the underlying dependency of demand on temperature. However, rather than treating successive years as replicate information, the relationship is allowed to vary. There are many mechanisms which will cause the relationship to change, such as socio-economic factors, but which would be very difficult to model in any useful fashion. Instead the use of dynamic models allows the changing relationship to be studied and provides a natural framework for prediction.

Only recently has detailed electricity demand data been made avaialble to the public with one such source, [9], giving the instantaneous demand at 5 minute intervals since mid-2011. This is a very large, and ever growing, dataset containing more than half a million records with total demand broken-down into different energy sources, for example coal, nuclear, gas and renewable. From these figures the total monthly demand is calculated and then presented as an hourly average. The temperature data, obtained from the UK Meteorological Office website [16] is part of a much bigger record of daily temperatures dating back to 1772. In particular, the Central England Temperature readings from the Met Office Hadley Centre were used and the median monthly temperature was calculated as a representative value for the whole month over the period 2012-2015. Figure 8 shows the temperature-demand relationship over the four years being considered. There is clear similarity in these graphs, but equally they are not identical, and hence an analysis based on dynamic modelling is appropriate.

First, individual linear regression is used to produce independent parameter estimates, see Table 3 and the dotted line in Figures 8, with a pooled estimate of variance used to give

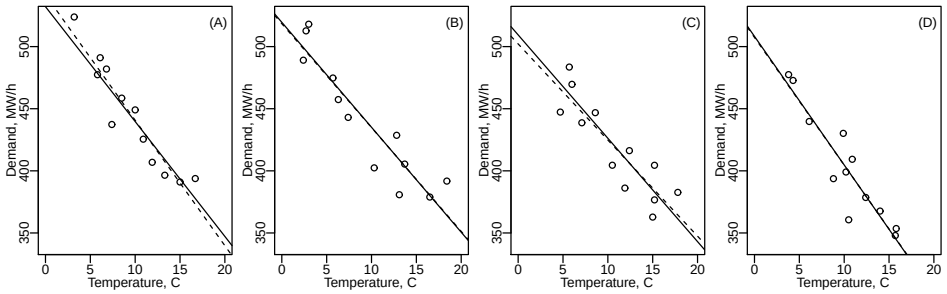


FIGURE 8. *Temperature-related electricity demand based on 12 monthly values for the years, (A) 2012, (B) 2013, (C) 2014 and (D) 2015. Temperature is the monthly median Central England Temperature and electricity demand is total UK demand. The solid line is the jointly fitted model with the separate fits shows as dotted lines.*

	2012	2013	2014	2015
$\hat{\alpha}$	541.61	518.46	502.16	507.37
$\hat{\beta}$	-10.09	-8.36	-7.74	-10.28

TABLE 3. Electricity demand analysis: parameter estimates using independent estimation.

an estimate of the standard deviation in the observational equation, leading to $\hat{\sigma} = 22.45$. From the results it can be seen that there is a small increase in the temperature values and a corresponding decrease in electricity demand. This has produced a general decrease in linear regression intercept, but a decrease and then increase in the slope parameter.

Second, the simultaneous joint estimation procedure was applied for a range of evolution equation parameter values. The smallest value which did not substantially effect the residual sum of squares was selected, giving $\hat{\tau} = 20$. The corresponding parameter estimates are given in Table 4 and corresponding fitted regression equations as solid lines in Figure 8. It is important to note that moderate changes in τ have little effect on the resulting parameter estimates and hence the approach is robust. Although the changes have only been small, non-the-less they have the potential of making valuable impact. Further, this example has illustrated the dynamic modelling framework and show that it can be successfully applied to real data problems.

	2012	2013	2014	2015
$\hat{\alpha}$	532.14	519.55	509.42	508.23
$\hat{\beta}$	-9.25	-8.45	-8.32	-10.35

TABLE 4. Electricity demand analysis: parameter estimates using the dynamic model with $\tau = 20$.

5. Conclusions

This paper has presented the fundamentals of dynamic modelling for the general case and taken a step-by-step approach to deriving the estimation equations for the Gaussian linear model. This has considered both a simultaneous joint model and a step-wise conditional approach. It is hoped that this will provide a starting point for other researchers and those looking to perform a similar analysis on their own data. Basic properties were derived for completeness, and which could be the basis of hypothesis testing or the construction of confidence intervals.

A simulation study was performed to assess the performance of the modelling framework on the specific example of linear regression. This was easy to perform and showed good results. In particular, the mean squared error is substantially lower for dynamic models compared to independent estimation. It was shown that after about four previous times have been included there is little further gain. It was also shown that the smallest MSE was achieve when the same value was used in the analysis as was used to simulate the data. However, a smaller value was nearly as good in this simulation. It is worth noting, however, that to retain some robustness to abrupt temporal changes, it is better not to go beyond 3 or 4 previous times and not to use too small a value of τ as bias could be introduced. In a brief study of the step-wise conditional estimation, again an improvement in MSE was seen, but this was not as dramatic as with the simultaneous joint estimation. Further, there is a danger in that a bad estimate at one time will lead to poor estimation at the next, and later times. Of course the joint estimation can be adapted to only consider the most resent 3 or 4 previous times, as a kind of running windowing method.

The electricity demand example was considered to demonstrate the approach on real data. Two separate data sources were combined and manipulated before a dynamic linear regression model was examined. The procedure worked well, with error parameters estimated as part of the procedure, as well as the regression parameters. This dataset only covered a few years and so only produced four times meaning that there was limited scope for comment on the use of many time points. Although there was only slight change in the parameters, and this was not consistent, the framework is clear. Short-term predictions can be based on the final fitted linear observation equation, and those for future years would use parameters obtained by projecting into the future using the fitted evolution model, that is $\hat{\boldsymbol{\theta}}^p_{T+1} = G_T\hat{\boldsymbol{\theta}}_T + \boldsymbol{\nu}_T$ and then $\hat{\mathbf{y}}^p_{T+1} = F_{T+1}\hat{\boldsymbol{\theta}}^p_{T+1}$. Repeating the steps for different errors $\boldsymbol{\nu}_K$ will then produce a distribution of future predictions.

Although the generally sequential approach proposed is fully defined in Section 2, there is infinite scope for variations. Hence, clearly, it is of interest to see how the methods will perform in other applications. There are many uses where the assumed Gaussian distribution error model is not appropriate. Hence following the same derivation but for binomial and Poisson discrete distributions, gamma as a skew continuous distribution or Student-t to give a heavier tailed distribution will provide future work. The approach also allows richly-structured and high-dimensional problems to be considered within a unified framework. Similarly, there is still many questions to be addressed for the linear dynamic model considered here. For example, the structure need not be time-invariant, nor does it need to involve independent or heteroskedastic errors – linking back to ARCH/GARCH models. All these would be worthy areas for future research.

Dynamic modelling has a great potential impact for many real world applications where data are collected from evolving processes, such as in biology, economics, environmental science and engineering. The approach is valid for any situation were the relationship between input and output can change with time. As with the majority of statistical analyses, the linear model has the ability to capture most phenomenon using the recorded variables or transformations of the original variables, hence the study of linear dynamic models will cover most commonly encountered situations. Dynamic modelling should become a regular topic in undergraduate courses as much as standard linear regression is now and might become commonplace in the toolkit of applied statisticians.

Acknowledgment. The authors are grateful for the comments of two anonymous referees and for the efficiency of the Editor in dealing with this paper.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. S. I. Aanonsen, G. Nævdal, D. S. Oliver, A. C. Reynolds, B. Vallès, et al., *The ensemble Kalman filter in reservoir engineering–a review*, SPE Journal **14** (2009), No. 3, 393–412.
2. J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, Wiley Series in Probability and Statistics, 2000.
3. C. Bretó, D. He, E. L. Ionides, and A. A. King, *Time series analysis via mechanistic models*, The Annals of Applied Statistics **3** (2009), 319–348.
4. N. Cressie and C. K. Wikle, *Statistics for Spatio-temporal Data*, John Wiley & Sons, 2015.
5. D. Crisan and B. Rozovskii, *The Oxford Handbook of Nonlinear Filtering*, Oxford University Press, 2011.
6. R. F. Engle, *Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation*, Econometrica **50** (1982), 987–1008.
7. ———, *Garch 101: The use of ARCH/GARCH models in applied econometrics*, Journal of Economic Perspectives **15** (2001), 157–168.
8. M. A. Golberg and H. A. Cho, *Introduction to Regression Analysis*, Wit Press, 2004.
9. Gridwatch, *Gridwatch - GB National Electrcity Demand*, 2016. Accessed on 12 July 2016. `http://www.gridwatch.templar.co.uk`.
10. J. Harrison and M. West, *Bayesian Forecasting & Dynamic Models*, Springer, 1999.
11. P. J. Harrison and C. F. Stevens, *Bayesian forecasting*, Journal of the Royal Statistical Society (B). **38** (1976), 205–247.
12. A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, 1990.
13. R. E. Kalman, *A new approach to linear filtering and prediction problem*, Journal of Basic Engineering, series D **82** (1960), 34–45.
14. H. R. Kunsch, *State space and hidden Markov models*, Monographs on Statistics and Applied Probability **87** (2001), 109–174.
15. P. M. Lee, *Bayesian Statistics: An Introduction*, Second ed., Arnold, 1989.
16. Metoffice, *Hadley Centre Central England Temperature Data*, 2016. Accessed on 12 July 2016. `www.metoffice.gov.uk/hadobs/hadcet/data/download.html`.
17. H. S. Migon, D. Gamerman, H. F. Lopes, and M. A. R. Ferreira, *Dynamic models*, Handbook of Statistics **25** (2005), 553–588.
18. R. H. Myers, *Classical and Modern Regression with Applications*, Duxbury Press, Pacific Grove (2000).
19. A. Pole, M. West, and J. Harrison, *Applied Bayesian Forecasting and Time Series Analysis*, CRC Press, 1994.
20. R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. `www.R-project.org`.
21. T. P. Ryan, *Modern Regression Methods*, vol. 655, John Wiley & Sons, 2008.

22. I. Shah and F. Lisi, *Day-ahead electricity demand forecasting with nonparametric functional models*, 12th International Conference on the European Energy Market (EEM), Institute of Electrical and Electronics Engineers, 2015, pp. 1 – 5.
23. J. W. Taylor and R. Buizza, *Using weather ensemble predictions in electricity demand forecasting*, International Journal of Forecasting **19** (2003), 5770.
24. W. J. Taylor, *Short-term electricity demand forecasting using double seasonal exponential smoothing*, Journal of the Operational Research Society **54** (2003), 799–805.
25. S. Weisberg, *Applied Linear Regression*, Vol. 528, John Wiley & Sons, 2005.
26. M. West, *Bayesian dynamic modelling*, Bayesian Inference and Markov Chain Monte Carlo: In Honour of Adrian FM Smith (2013), 145–166.
27. M. West and P. J. Harrison, *Monitoring and adaptation in Bayesian forecasting models*, Journal of the American Statistical Association **81** (1986), 741–750.
28. D. J. Wilkinson, *Stochastic Modelling for Systems Biology*, CRC press, 2011.

ROBERT G. AYKROYD, DEPARTMENT OF STATISTICS, UNIVERSITY OF LEEDS, LEEDS, LS2 9JT, UK.
E-mail address: r.g.aykroyd@leeds.ac.uk

NADA ALFAER, DEPARTMENT OF STATISTICS, UNIVERSITY OF LEEDS, LEEDS, LS2 9JT, UK.
E-mail address: mmna@leeds.ac.uk