Multivariate Time Series Forecasting in Incomplete Environments

Technical Report PARG 08-03

Seung Min Lee and Stephen J. Roberts

Robotics Research Group
Department of Engineering Science
University of Oxford

December 2008
Multivariate Time Series Forecasting in Incomplete Environments

Summary

We consider the problem of predicting missing observations and forecasting future values in incomplete multivariate time series data. We study three forecasting models: a dynamic multivariate autoregressive model, a multivariate local trend model and a Gaussian process model. To evaluate each of the models we analyse air temperature data collected by a network of weather sensors. We discuss advantages and disadvantages of the models.

Acknowledgements

This work was undertaken as part of the ALADDIN (Autonomous Learning Agents for Decentralised Data and Information Systems) project and is jointly funded by a BAE Systems and EPSRC (Engineering and Physical Research Council) strategic partnership.
## Contents

1 Introduction .................................................. 1

2 Dynamic Linear Models ...................................... 1

3 Learning Dynamic Linear Models .......................... 6

4 Multivariate Time Series Models ......................... 10

5 Results .......................................................... 14

6 Discussions and Future Work ............................... 27
1 Introduction

A time series is a sequence of data points collected over time. Analysis of time series is required in many different scientific fields from economics to engineering. In the engineering literature state space methods have been developed for the sequential analysis of data. Several researchers have tried to bridge the gap between engineering methods and statistics (e.g. [Harvey, 1990, Durbin and Koopman, 2001]).

We adopt classes of dynamic models [West and Harrison, 1997] to model time series processes. In particular we focus on the linear Gaussian dynamic model. If all components of the dynamic model are known, then the Kalman filter [Kalman, 1960, Kalman and Bucy, 1961] produces optimal (in a square error sense) estimates of hidden state variables. It is, however, common that these components are not available. Shumway and Stoffer [1982] and Ghahramani and Hinton [1996] propose techniques for estimating the unknown components of the dynamic model using the expectation-maximisation (EM) algorithm [Dempster et al., 1977].

Our main interest in this report is the forecasting of observation values in incomplete time series data. We are particularly interested in a multi-sensors system that give rise to time series data with missing values owing to the intermittent failure of the sensors. In order to handle this highly dynamic and incomplete set of time series, we shall implement three forecasting models: (1) a dynamic multivariate autoregressive model, which can be formulated as a linear Gaussian dynamic model, (2) a multivariate local trend model whose parameters are estimated using the EM algorithm, and (3) a Gaussian process model, which is a semi-parametric model. We shall also discuss how to make a dynamic linear model able to deal with data having a periodic cycle.

In Section 2 we introduce a linear Gaussian model and demonstrate estimation methods, namely filtering, smoothing and forecasting. In Section 3 we outline the EM algorithm, which finds the maximum likelihood estimates of hidden variables. We also show how the algorithm can be applied to estimating unknown components of the dynamic linear model. In Section 4 we study three multivariate forecasting models. In Section 5 we apply the forecasting models to air temperature data collected by a network of weather sensors, and discuss their performances and characteristics. Section 6 concludes the report with discussions of future research directions.

2 Dynamic Linear Models

There are several reasons why a Gaussian distribution enjoys huge popularity. Firstly, many physical processes are approximately Gaussian; it is associated with the central limit theorem. Secondly, it is analytically elegant and convenient to use. Thirdly, it can be easily manipulated; that is, if a Gaussian variable is propagated through
a linear function, an output variable is also Gaussian. As a result, the most widely used class of dynamic linear model is a dynamic linear Gaussian model, which we will focus on throughout this report. It may be represented by two equations: one defining an observation variable \( y_t \), the other a state variable \( w_t \). We denote index \( t \) for discrete time steps, that is, \( t = \{1, 2, 3, \ldots \} \). The observation equation is of form

\[
y_t = H_t w_t + n_t, \tag{1}
\]

where \( n_t \) is a Gaussian observation error term with mean 0 and covariance \( R_t \). The matrix \( H_t \) is called the observation or design matrix. The state variable \( w_t \) in Equation (1) evolves over time according to the following evolution equation:

\[
w_t = F_t w_{t-1} + v_t, \tag{2}
\]

where \( v_t \) is a Gaussian diffusion term with mean 0 and covariance \( Q_t \). The matrix \( F_t \) governs the linear dynamic evolution of the state variable and is referred to as the evolution, or transition, matrix. Note that \( H_t, F_t, R_t \) and \( Q_t \) form the parameterisation of the dynamic model and they are typically assumed to be known unless otherwise stated.

We here define notations that will be used for the rest of the report. An observation set available at time \( t \) is denoted by \( D_t \), i.e., \( D_t = \{y_1, y_2, \ldots, y_t\} \). The mean and covariance of a state variable at time \( t \) given an observation set \( D_t \) are represented by

\[
\hat{w}_{t|t} = E(w_t|D_t),
\]

\[
P_{t|t} = \text{Cov}(w_t|D_t). \tag{4}
\]

In addition, the expected value of a one-step ahead observation forecast is denoted by

\[
\hat{y}_{t+1|t} = E(y_t|w_t, D_{t-1}). \tag{5}
\]

Broadly speaking, there are three main inference and prediction tasks in the dynamic model: Filtering, Smoothing and Forecasting. The following probability densities define concisely what we shall discussed shortly.

- State Filtering: \( p(w_t|D_t) \)
- State Smoothing: \( p(w_{t-k}|D_t) \) for \( k \geq 1 \)
- Observation Forecasting: \( p(y_{t+k}|D_t) \) for \( k \geq 1 \)
2.1 State Filtering

The Kalman filter [Kalman, 1960, Kalman and Bucy, 1961] is a recursive method for finding an optimal estimate of the state variable that minimises the minimum mean-squared error. Although the original idea behind the Kalman filter was not limited to any Gaussian assumptions on the state or observation variables, we show here a recursive algorithm for computing filtered states of the dynamic linear Gaussian model given in Equations (1) and (2).

Suppose that the posterior probability distribution of the state variable at time $t-1$ given $D_{t-1}$ is Gaussian with mean $\hat{w}_{t-1|t-1}$ and covariance $P_{t-1|t-1}$:

$$p(w_{t-1}|D_{t-1}) = N(\hat{w}_{t-1|t-1}, P_{t-1|t-1}).$$ (6)

The prior distribution of $w_t$ given $D_{t-1}$ is also Gaussian because the evolution equation of Equation (2) is linear and the state evolution noise $v_t$ is Gaussian. Thus, the mean and covariance of the posterior distribution can be readily computed, namely

$$p(w_t|D_{t-1}) = N(\hat{w}_{t|t-1}, P_{t|t-1}),$$ (7)

where

$$\hat{w}_{t|t-1} = F_t \hat{w}_{t-1|t-1},$$ (8)

$$P_{t|t-1} = F_t P_{t-1|t-1} F_t^T + Q_t.$$ (9)

Likewise, the computation of the one-step ahead forecast distribution is also straightforward due to the linear observation equation and the Gaussian observation noise given in Equation (1); it takes the form

$$p(y_t|w_t, D_{t-1}) = N(\hat{y}_{t|t-1}, S_{t|t-1}),$$ (10)

where

$$\hat{y}_{t|t-1} = H_t \hat{w}_{t-1},$$ (11)

$$S_{t|t-1} = H_t P_{t-1|t-1} H_t^T + R_t.$$ (12)

Via the Bayes’ rule the posterior distribution of $w_t$ can be represented as follows:

$$p(w_t|D_t) = p(w_t|y_t, D_{t-1})$$

$$\propto p(w_t, y_t|D_{t-1})$$

$$\propto p(w_t|D_{t-1})p(y_t|w_t, D_{t-1}).$$ (13)
Note that the two densities in the last line of Equation (13) are given in Equations (7) and (10) respectively. If we use the fact that the product of two Gaussian densities is also Gaussian (e.g. [Petersen and Pedersen, 2008]), we can obtain the posterior distribution of $w_t$, which is proportional to the following Gaussian distribution:

$$ p(w_t|D_t) \propto N(\hat{w}_t|t, P_t|t), \quad (14) $$

where

$$ \hat{w}_t|t = \left( P_{t|t-1}^{-1} + S_{t|t-1}^{-1} \right)^{-1} \left( P_{t|t-1}^{-1} \hat{w}_t|t-1 + S_{t|t-1}^{-1} \hat{y}_t|t-1 \right), \quad (15) $$

$$ P_t|t = \left( P_{t|t-1}^{-1} + S_{t|t-1}^{-1} \right)^{-1}. \quad (16) $$

The mean and covariance of the posterior distribution of $w_t$ in Equation (15)and (16) can be rearranged to be the familiar Kalman filter equations of form

$$ \hat{w}_t|t = \hat{w}_t|t-1 + K_t(y_t - \hat{y}_t|t-1), \quad (17) $$

$$ P_t|t = P_t|t-1 - K_t H_t P_t|t-1, \quad (18) $$

where $K_t$ is referred to as a Kalman gain and defined by

$$ K_t = P_{t|t-1} H_t^T S_{t|t-1}^{-1}. \quad (19) $$

### 2.2 State Smoothing

Filtering is concerned with the inference of a state at time $t$ given all observations up to time $t$. Meanwhile, smoothing considers the picture of what happened in the past. In other words, smoothing infers past states given all observations up to the present.

Rauch et al. [1965] proposed a retrospective estimation technique known as the Kalman-Rauch smoother. It consists of two steps: one is a forward step and the other a backward step. Suppose that we have a set of observations up to time $t = N$, i.e. $D_N = \{y_1, \ldots, y_N\}$. In the forward step a process runs forward in time from $t = 1$ to $t = N$. In other words, we compute the posterior mean and covariance of the state variable, $\hat{w}_t|t$ and $P_t|t$ (for $t = 1, 2, \ldots, N$), via the filtering algorithm detailed in the previous subsection.

In the backward step we infer the past states (for $t = N - 1, \ldots, 1$), on the basis of the forward filtered estimates of the states. The backward equations are given by

$$ \hat{w}_{t-1}|N = \hat{w}_{t-1}|t-1 + J_{t-1} (\hat{w}_{t|N} - F_t \hat{w}_{t|t-1}), \quad (20) $$

$$ P_{t-1}|N = P_{t-1}|t-1 + J_{t-1} (P_{t|N} - P_{t|t-1}) J_{t-1}^T, \quad (21) $$
where
\[
J_{t-1} = P_{t-1|t-1}F_t^T(P_{t|t-1})^{-1}.
\] (22)

If we recursively run the backward equations, the smoothed estimates of the states, \( \hat{w}_{t|N} \) and \( P_{t|N} \), can be computed from \( t = N - 1 \) to \( t = 1 \).

Additionally, the covariance \( P_{t,t-1|N} \) between two states adjacent in time given \( D_N \) can be computed using the backward equations. For \( t = N, N - 1, \ldots, 2 \),
\[
P_{t-1,t-2|N} = P_{t-1|t-1}J_{t-2}^T + J_{t-1}(F_tP_{t-1|t-1}F_t^T)J_{t-2}^T,
\] (23)
where
\[
P_{N,N-1|N} = (I - K_N H_N) F_N P_{N-1|N-1}.
\] (24)

This quantity will be re-considered when we discuss estimating unknown parameters of a dynamic linear model in Section 3.

2.3 Observation Forecasting

In the state filtering process we compute the one-step ahead forecast distribution of the observation variable, \( p(y_t|D_{t-1}) \). We are often interested in forecasting not only a one-step ahead observation but also a multiple-step ahead observation with observations up to time \( t \), i.e. \( D_t \); we would like to obtain
\[
p(y_{t+k}|D_t), \quad \text{for } k \geq 1.
\] (25)

According to Equation (1), the observation variable at \( t + k \) can be represented as follows:
\[
y_{t+k} = H_{t+k}w_{t+k} + n_{t+k}.
\] (26)

Hence, computing the forecast distribution of Equation (25) is closely related to inferring the distribution of future states, \( p(w_{t+k}|D_t) \). The inference of the future state given \( D_t \) can be thought of as running the filtering process recursively without new observations being received. This means that the mean and covariance of the state variable at \( t + k \) can be computed from those of the state at \( t + k - 1 \) as follows:
\[
\hat{w}_{t+k|t} = F_{t+k}\hat{w}_{t+k-1|t},
\] (27)
\[
P_{t+k|t} = F_{t+k}P_{t+k-1|t}F_{t+k}^T + Q_{t+k}.
\] (28)

Using the observation process of Equation (26) and the two forecast moments of the state variable given in Equations (27) and (28), we can readily compute the forecast distribution of the observation at \( t + k \):
\[
p(y_{t+k}|D_t) = N(\hat{y}_{t+k|t}, S_{t+k|t}),
\] (29)
Learning Dynamic Linear Models

where

\[
\hat{y}_{t+k|t} = \mathbf{H}_{t+k} \hat{\mathbf{w}}_{t+k|t},
\]

\[
\mathbf{S}_{t+k|t} = \mathbf{H}_{t+k} \mathbf{P}_{t+k|t} \mathbf{H}_{t+k}^T + \mathbf{R}_{t+k}.
\]

In the case of a constant evolution matrix, i.e., \( \mathbf{F}_t = \mathbf{F} \), inferring the distribution of the future state \( \mathbf{w}_{t+k} \) is significantly simplified without recursive computation. In other words, it can be computed directly from the filtered distribution at \( t \), \( p(\mathbf{w}_t|\mathbf{D}_t) \). The first two moments of the distribution are given by

\[
\hat{\mathbf{w}}_{t+k|t} = \mathbf{F}^k \hat{\mathbf{w}}_{t|t},
\]

\[
\mathbf{P}_{t+k|t} = \mathbf{F}^k \mathbf{P}_{t|t} \left( \mathbf{F}^k \right)^T + \sum_{i=0}^{k-1} \mathbf{F}^i \mathbf{Q}_{t+k-i} \left( \mathbf{F}^i \right)^T.
\]

Accordingly, the mean and covariance of the forecast distribution of the observation variable are

\[
\hat{y}_{t+k|t} = \mathbf{H}_{t+k} \mathbf{F}^k \hat{\mathbf{w}}_{t|t},
\]

\[
\mathbf{S}_{t+k|t} = \mathbf{H}_{t+k} \mathbf{P}_{t+k|t} \mathbf{H}_{t+k}^T + \mathbf{R}_{t+k}.
\]

3 Learning Dynamic Linear Models

The filtering, smoothing and forecasting processes described in the previous section have all assumed that the set of quadruples \( \{ \mathbf{F}_t, \mathbf{H}_t, \mathbf{Q}_t, \mathbf{R}_t \} \) of a dynamic linear model is known to a data-modeller. However, it is common that a part of, or the whole of, this set is not available. This makes it necessary to estimate simultaneously the state variable and the unknown components (called parameters) of the model. In this section we briefly introduce the expectation-maximisation (EM) algorithm [Dempster et al., 1977] as a method of estimating latent variables. In addition, we discuss how the EM algorithm is applied to learning a dynamic linear model, which is first proposed by Shumway and Stoffer [1982].

3.1 The EM Algorithm

The expectation-maximisation algorithm is an iterative method of finding maximum likelihood estimates of latent variables or unknown parameters when data available is incomplete. We first consider a standard maximum likelihood estimation method. Suppose that we have a data set \( \mathbf{Y} = \{ \mathbf{y}_1, \ldots, \mathbf{y}_T \} \) drawn from a probability density function \( p(\mathbf{y}; \Theta) \), which is governed by a set of parameters, \( \Theta \). Assuming that each observation \( \mathbf{y}_i \) is independent and identically distributed, the joint probability density is represented by

\[
p(\mathbf{Y}; \Theta) = \prod_{i=1}^{T} p(\mathbf{y}_i; \Theta).
\]
This can be viewed as a function of the parameter set given the data set, and it is referred to as a likelihood function denoted by $L(\Theta|Y)$. The maximum likelihood estimate of $\Theta$ is one that maximises the likelihood function; that is,

$$
\hat{\Theta} = \arg\max_\Theta L(\Theta|Y).
$$

(37)

If $p(y; \Theta)$ is a multivariate Gaussian with $\Theta = \{\mu, \Sigma\}$, then we can readily compute the maximum likelihood estimate of $\Theta$ by setting to zero the derivative of the logarithm of the likelihood function and solving the resultant equation with respect to $\mu$ and $\Sigma$.

We now suppose to have a latent variable $w$ in addition to the observable variable $y$. Hence, a complete data set is $Z = \{Y, W\}$. We wish to find an estimate of $\Theta$ that maximises the following likelihood function

$$
L(\Theta|Z) = p(Z; \Theta) = p(Y, W; \Theta).
$$

(38)

However, as $w$ is a latent variable, it is not so straightforward to find the maximum likelihood estimate of $\Theta$. In other words, we are not given the complete data set $Z$, but only the observable part $Y$. The basic concept of the expectation-maximisation algorithm is elegantly simple. Since we cannot directly evaluate $p(Y, Z; \Theta)$, we compute its expectation over the unobserved variable $w$ (expectation step), and then we obtain the maximum likelihood estimate of $\Theta$ by maximising the expectation (maximisation step).

The probability density of the complete set can be rewritten as

$$
p(Z; \Theta) = p(Y, W; \Theta) = p(W|Y; \Theta)p(Y; \Theta).
$$

(39)

In other words, we know the values of the latent variables in $W$ through the conditional density of $W$ given $Y$, $p(W|Y; \Theta)$. In the expectation step, assuming that the observation set $\{y_1, \ldots, y_T\}$ and an estimate of the parameter set $\hat{\Theta}$ are given, we take expectations over $p(Y, W; \Theta)$ with regard to the latent variables $W$ as follows:

$$
E_W \left[ p(Y, W; \Theta) \bigg| Y, \hat{\Theta} \right] = \int p(Y, W; \Theta)p(W|Y, \hat{\Theta})dW.
$$

(40)

In the maximisation step we maximise the above expectation term with respect to $\Theta$. Hence, the given estimate $\hat{\Theta}$ is updated according to

$$
\hat{\Theta}_{(new)} := \arg\max_\Theta E_W \left[ p(Y, W; \Theta) \bigg| Y, \hat{\Theta}_{(old)} \right].
$$

(41)

The expectation and maximisation steps alternate until the value of the log-likelihood converges. It is known that the log-likelihood is guaranteed to increase at each iteration and thus the algorithm is guaranteed to reach a local maximum of the likelihood function [Dempster et al., 1977].
3.2 Estimating Parameters of a Dynamic Linear Model

Shumway and Stoffer [1982] suggested a method of learning unknown components of a dynamic linear model using the expectation-maximisation algorithm. This subsection summarises their approach. Suppose that we have the following dynamic linear model:

\[ y_t = H_t w_t + n_t, \]
\[ w_t = F w_{t-1} + v_t, \]

where \( n_t \sim N(0, R) \) and \( v_t \sim N(0, Q) \). Note that time index \( t \) is not present on \( F, R \) and \( Q \) because they are assumed to be static, not because of brevity of notation. In the dynamic linear model we assume that the observation matrix, \( H_t \), is known, but the state evolution matrix, \( F \), and the state and observation noise covariances, \( Q \) and \( R \), are not given. In addition, we assume that the initial state \( w_0 \) is Gaussian distributed with mean \( \mu_0 \) and covariance \( \Sigma_0 \), which are also unknown components of the model. If we regard the unknown components of the model as a set of parameters, i.e. \( \Theta = \{ F, R, Q, \mu_0, \Sigma_0 \} \), and the state variable \( w \) as a latent variable, the we can use the expectation-maximisation algorithm to compute the maximum likelihood estimate of \( \Theta \) with a set of observations, \( D_T = \{ y_1, \ldots, y_T \} \), given.

Since the state and observation variables are all Gaussian distributed, the likelihood function of the complete data is of form

\[
L(\Theta | w_0, \{y_i\}_{i=1}^{T}, \{w_i\}_{i=1}^{T}) = \frac{1}{(2\pi)^{\alpha/2} |\Sigma_0|^{1/2}} \exp \left( -\frac{1}{2} (w_0 - \mu_0)^T \Sigma_0^{-1} (w_0 - \mu_0) \right) \\
\frac{1}{(2\pi)^{\beta T/2} |Q|^{1/2}} \exp \left( -\frac{1}{2} \sum_{t=1}^{T} (w_t - F w_{t-1})^T Q^{-1} (w_t - F w_{t-1}) \right) \\
\frac{1}{(2\pi)^{\alpha T/2} |R|^{1/2}} \exp \left( -\frac{1}{2} \sum_{t=1}^{T} (y_t - H_t w_t)^T R^{-1} (y_t - H_t w_t) \right). \tag{44}
\]

Here \( \alpha \) and \( \beta \) denote the dimension of \( w \) and \( y \) respectively. Ignoring constants, hence, the log likelihood function of the complete data is

\[
\log L = -\frac{1}{2} \log |\Sigma_0| - \frac{1}{2} (w_0 - \mu_0)^T \Sigma_0^{-1} (w_0 - \mu_0) \\
- \frac{T}{2} \log |Q| - \frac{1}{2} \sum_{t=1}^{T} (w_t - F w_{t-1})^T Q^{-1} (w_t - F w_{t-1}) \\
- \frac{T}{2} \log |R| - \frac{1}{2} \sum_{t=1}^{T} (y_t - H_t w_t)^T R^{-1} (y_t - H_t w_t). \tag{45}
\]
In the expectation step we compute the conditional expectation of the log likelihood with regard to the latent states, given \( D_T \) and the \((j - 1)\)th iterate estimate of the parameter set denoted by \( \hat{\Theta}^{(j-1)} \). We can compute this conditional expectation [Shumway and Stoffer, 1982]:

\[
E[\log L|D_T, \hat{\Theta}^{(j-1)}] = \\
-\frac{1}{2} \log |\hat{\Sigma}_0^{(j-1)}| \\
-\frac{1}{2} \text{tr} \left\{ \left( \hat{\Sigma}_0^{(j-1)} \right)^{-1} \left( P_0|T + (\hat{\omega}_0|T - \mu_0^{(j-1)})(\hat{\omega}_0|T - \mu_0^{(j-1)})^T \right) \right\} \\
-\frac{T}{2} \log |Q| - \frac{1}{2} \text{tr} \left\{ Q^{-1} (C - BF^T - FB^T + FAA^T) \right\} \\
-\frac{T}{2} \log |R| - \frac{1}{2} \text{tr} \left\{ R^{-1} \sum_{t=1}^{T} \left( (y_t - H_t \hat{\omega}_t|T)(y_t - H_t \hat{\omega}_t|T)^T + H_t P_t|T H_t^T \right) \right\}.
\]

(46)

Here \( \text{tr} \) denotes the trace of a matrix and

\[
A = \sum_{t=1}^{T} \left( P_{t-1|T} + \hat{\omega}_{t-1|T} \hat{\omega}_{t-1|T}^T \right), \\
B = \sum_{t=1}^{T} \left( P_{t,t-1|T} + \hat{\omega}_{t|T} \hat{\omega}_{t-1|T}^T \right), \\
C = \sum_{t=1}^{T} \left( P_{t|T} + \hat{\omega}_{t|T} \hat{\omega}_{t|T}^T \right).
\]

(47) \quad (48) \quad (49)

The conditional mean and covariance terms (i.e. \( \hat{\omega}_{t|T}, \hat{\mu}_{t|T} \) and \( \hat{\Sigma}_{t|T} \)) can be computed by the Kalman-Rauch smoother detailed in Section 2.2. In the maximisation step the parameter set is re-estimated by setting the corresponding partial derivative of the expected log likelihood to zero and solving it with respect to each element of the parameter set. The \((j)\)th iterate estimates for the parameters are given by

\[
\hat{F}^{(j)} = BA^{-1}, \\
\hat{Q}^{(j)} = \frac{1}{T}(C - BA^{-1}B^T), \\
\hat{R}^{(j)} = \frac{1}{T} \sum_{t=1}^{T} \left( (y_t - H_t \hat{\omega}_t|T)(y_t - H_t \hat{\omega}_t|T)^T + H_t P_t|T H_t^T \right), \\
\hat{\mu}^{(j)} = \hat{\omega}_0|T, \\
\hat{\Sigma}^{(j)} = P_0|T - \hat{\omega}_0|T \hat{\omega}_0|T^T.
\]

(50) \quad (51) \quad (52) \quad (53) \quad (54)

We iterate the estimation and maximisation steps until the value of the log likelihood function converges.

Ghahramani and Hinton [1996] proposed a generalised algorithm in which the observation matrix \( H \) is also regarded as a parameter. In addition, for a non-linear dynamic model Roweis and Ghahramani [2001] suggested a method that employs extended Kalman smoothing to estimate parameters of the model. For time-varying parameters, it is possible to provide estimate using observations over a time window.
4 Multivariate Time Series Models

A need for dealing with multivariate time series data is common in many problem domains. For example, we can consider a network of sensors measuring weather conditions, or multiple financial indices reflecting global and domestic economic states. Since such data are essentially non-stationary, correlated and often have periodic patterns, it is a hard challenge to analyse and forecast a multivariate time series. Moreover, such a time series can be occasionally incomplete: some sensor readings or financial instruments may not be available. In this section we discuss forecasting models that deal with an incomplete multivariate time series.

4.1 Dynamic Multivariate Autoregressive Model

An autoregressive (AR) model is one of the most widely used time series models. It is a type of multiple linear regression model in which a dependent variable is regressed not on independent variables but on past values of the dependent variable. For a scalar observation variable $y_t$ an autoregressive model of order $p$ is defined by

$$y_t = w_0 + w_1 y_{t-1} + \cdots + w_p y_{t-p} + n_t,$$  \hspace{1cm} (55)

where $\{w_0, w_1, \ldots, w_p\}$ are parameters of the model and $n_t$ is a Gaussian noise term with zero mean and finite variance. This model is abbreviated to AR($p$). Here the order $p$ determines the number of past time steps required to predict the current value.

If there are multiple observation variables and they interrelate, we can consider a multivariate autoregressive (MAR) model. Suppose that there are $M$ different sensors, each of which has a time series. These sensors are correlated in a way that the value of a sensor at $t$ is related to the $p$ previous values of all other sensors including its own past values. We can represent each sensor’s value at $t$ as follows:

$$y_t^{(1)} = w_0^{(1)} + w_{11}^{(1)} y_{t-1}^{(1)} + \cdots + w_{D1}^{(1)} y_{t-1}^{(M)} + \cdots + w_{1p}^{(1)} y_{t-p}^{(1)} + \cdots + w_{Dp}^{(1)} y_{t-p}^{(M)} + n_t^{(1)},$$

$$\vdots$$

$$y_t^{(M)} = w_0^{(M)} + w_{11}^{(M)} y_{t-1}^{(1)} + \cdots + w_{D1}^{(M)} y_{t-1}^{(M)} + \cdots + w_{1p}^{(M)} y_{t-p}^{(1)} + \cdots + w_{Dp}^{(M)} y_{t-p}^{(M)} + n_t^{(M)}.$$  \hspace{1cm} (56)

These equations can be rewritten as a single equation: a MAR($p$) model is defined by

$$y_t = w_0 + W_1 y_{t-1} + \cdots + W_p y_{t-p} + n_t,$$  \hspace{1cm} (57)

where $y_t = [y_t^{(1)} \ldots y_t^{(M)}]^T$, $w_0 = [w_0^{(1)} \ldots w_0^{(M)}]^T$, $W_i = \begin{bmatrix} w_{1i}^{(1)} & \cdots & w_{1i}^{(M)} \\ \vdots & \ddots & \vdots \\ w_{Di}^{(M)} & \cdots & w_{Di}^{(M)} \end{bmatrix}$ and $n_t = [n_t^{(1)} \ldots n_t^{(M)}]^T$. 
For the MAR\( (p) \) model to deal with a non-stationary property of a time series we consider a \textit{dynamic} version of the model by allowing the parameters (i.e. \( w_0 \) and \( W_i \)) to vary over time. Re-arranging Equation (57) we have

\[
y_t = H_t w + n_t,
\]

where \( w = [w_0^{(1)} \ w_0^{(1)} \ \ldots \ w_0^{(M)} \ w_0^{(M)} \ w_D^{(1)} \ \ldots \ w_D^{(M)} \ w_p^{(1)} \ \ldots \ w_p^{(M)}]^T \). The observation matrix \( H_t \) consists of \( p \) previous values of the observation variable \( y_t \):

\[
H_t = [1, y_{t-1}^T, \ldots, y_{t-p}^T] \otimes I_M.
\]

Here \( I_M \) is the identity matrix of dimension \( M \), and \( \otimes \) denotes the Kronecker product. We assume that the state variable evolves according to a random walk. Therefore, a dynamic multivariate autoregressive (DMAR) model of order \( p \) is defined by

\[
y_t = H_t w_t + n_t,
\]

\[
w_t = w_{t-1} + v_t,
\]

where \( n_t \sim N(0, R_t = r_t I) \) and \( v_t \sim N(0, Q_t = q_t I) \).

We estimate the unknown noise variances, \( r_t \) and \( q_t \), adaptively by covariance-matching techniques [Jazwinski, 1970, Mehra, 1972]. The key idea underlying the technique is to use a residual defined by

\[
e_{t|t-1} = y_t - \hat{y}_{t|t-1}.
\]

The covariance matrix of the residual error is

\[
S_{t-1} \triangleq \text{E}[(y_t - \hat{y}_{t|t-1})(y_t - \hat{y}_{t|t-1})^T] = r_t I + H_t P_{t|t-1} H_t^T + q_t H_t H_t^T.
\]

This covariance can be approximated as \( S_{t-1} \approx e_{t|t-1} e_{t|t-1}^T \), which we denote by \( C_{t-1} \). If \( q_t \) is assumed to be zero, \( S_{t-1} \) reduces to \( r_t I + H_t P_{t|t-1} H_t^T \), which is denoted by \( S_{t-1}^{q=0} \). We can, hence, argue that if \( C_{t-1} \leq S_{t-1}^{q=0} \), then \( \hat{q}_t = 0 \). Otherwise,

\[
\hat{q}_t = \frac{C_{t-1} - S_{t-1}^{q=0}}{H_t H_t^T}.
\]

We smooth the estimate, \( \hat{q}_t \), with its one-step previous estimate, assuming that \( q_t \) change smoothly:

\[
\hat{q}_t := (1 - \alpha) \hat{q}_{t-1} + \alpha \hat{q}_t,
\]
where $0 < \alpha < 1$ is a smoothing parameter.

From Equation (63) we can see that if $r_t$ is assumed to be zero, then $S_{t-1}$ is simply $H_t P_{t|t-1} H_t^T$. As with the estimation for $q_t$, this value is compared with $C_{t-1}$: if $C_{t-1} \leq H_t P_{t|t-1} H_t^T$, then $\hat{r}_t = 0$. Otherwise,

$$\hat{r}_t = C_{t-1} - H_t P_{t|t-1} H_t^T.$$  \hspace{1cm} (66)

The smoothing scheme is also applied to this case:

$$\hat{r}_t := (1 - \beta) \hat{r}_{t-1} + \beta \hat{r}_t,$$  \hspace{1cm} (67)

where $0 < \beta < 1$ is another smoothing parameter. Note that we smooth the estimates of $q$ and $r$ because they are based on only one residual, which implies that the estimates have little statistical significance. Hence, a small value for $\alpha$ or $\beta$ is undesirable in order to ensure a smooth change of the estimates. We used $\alpha = 0.9$ and $\beta = 0.7$ in experiments conducted in the next section.

We can consider an advanced model in which the state variable varies according to unknown dynamics, not according to a random walk; that is, it is of form

$$y_t = H_t w_t + n_t,$$  \hspace{1cm} (68)

$$w_t = F w_{t-1} + v_t,$$  \hspace{1cm} (69)

where the unknown matrix $F$ governs the movement of the state variable. In addition to the noise covariances we could estimate this matrix via the expectation-maximisation algorithm detailed in Section 3.2. However, it should be noted that the dimension of $w_t$ is $(1 + pM)M \times 1$, and accordingly, that of $F$ is $(1 + pM)M \times (1 + pM)M$. If the AR order, $p$, and/or the dimension of $y_t$, $M$, are large, the number of parameters to estimate is innumerably many. For this reason it would be unrealistic to use this approach in practice.

### 4.2 Multivariate Local Trend Model

As a multivariate time series model we consider the following simple dynamic linear model

$$y_t = w_t + n_t,$$  \hspace{1cm} (70)

$$w_t = F_t w_{t-1} + v_t,$$  \hspace{1cm} (71)

where $n_t \sim N(0, R_t)$ and $v_t \sim N(0, Q_t)$. A major difference from the DMAR model in Equations (60) and (61) depends on how we treat the state variable $w_t$. In the DMAR model $w_t$ represents a time-varying parameter vector of a multivariate autoregressive model. Meanwhile, in the above model $w_t$ is treated as unknown true
values of observations. We assume that the true values of \( M \) sensors at \( t \) are related to the true values at \( t-1 \) as follows:

\[
\mathbf{w}_t = \begin{bmatrix} w_{t(1)}^{(1)} \\ \vdots \\ w_{t(M)}^{(M)} \end{bmatrix} = \begin{bmatrix} f_{11} & \cdots & f_{1M} \\ \vdots & \ddots & \vdots \\ f_{M1} & \cdots & f_{MM} \end{bmatrix} \begin{bmatrix} w_{t-1(1)}^{(1)} \\ \vdots \\ w_{t-1(M)}^{(M)} \end{bmatrix} + \begin{bmatrix} v_{t(1)}^{(1)} \\ \vdots \\ v_{t(M)}^{(M)} \end{bmatrix},
\]

\[= \mathbf{F}_t \mathbf{w}_{t-1} + \mathbf{v}_t. \tag{72} \]

In this model we think of \( y_t \) as a noisy observation vector of the sensors’ true values.

The diagonal elements of the state evolution matrix, \( \mathbf{F}_t \), depicts unknown local trends of the sensors; the off-diagonal elements represent the relationships between the sensors. We refer to this model as a multivariate local trend (MLT) model. We can estimate the unknown components of the model, i.e. \( \mathbf{F}_t, \mathbf{Q}_t \) and \( \mathbf{R}_t \), using the expectation-maximisation algorithm. Note that this MLT model is similar to the DMAR\((p = 1)\) model in that they are both based on a first-order Markov model. However, for the DMAR\((p = 1)\) model, a one-step-ahead forecast of the sensors is computed by \( \hat{y}_{t|t-1} \), whereas for the MLT model by \( \hat{w}_{t|t-1} \). In other words, the DMAR model’s forecast is a linear function of noisy past observations, \( y_{t-1}, \ldots, y_{t-p} \), whereas that of the MLT model is a filtered value, thereby being more smooth.

### 4.3 Gaussian Process Model

Gaussian processes [Rasmussen and Williams, 2006] represent a method to perform Bayesian inference about functions. A Gaussian process (GP) is defined as being a probability distribution over a (possibly infinite) number of variables, such that the distribution over any finite subset of them is a multivariate Gaussian.

We consider a function \( y(t) \). We are principally concerned with prediction and hence functions in time. For a single sensor case we use \( y = \{y_1, y_2, \ldots, y_n\} \) to refer to a possible vector of function outputs and \( t = \{t_1, t_2, \ldots, t_n\} \), function inputs. We define a GP as the following multivariate Gaussian distribution

\[
p(y|t) = N(\mu, \mathbf{C}). \tag{73} \]

Here \( \mathbf{C} \) is a square matrix indicating the strength of correlations amongst the entries of \( y \). It is generated by a covariance function, \( c(t, t') \):

\[
\mathbf{C} = \begin{bmatrix} c(t_1, t_1) & c(t_1, t_2) & \cdots & c(t_1, t_n) \\ c(t_2, t_1) & c(t_2, t_2) & \cdots & c(t_2, t_n) \\ \vdots & \vdots & \ddots & \vdots \\ c(t_n, t_1) & c(t_n, t_2) & \cdots & c(t_n, t_n) \end{bmatrix}. \tag{74} \]
Almost all functions of interest, we expect, possess some degree of smoothness. That is, the value of a function at $t$ is strongly correlated with the values close to $t$, these correlations becoming weaker further away. A prototypical choice is the squared exponential

$$c(t, t') = h^2 \exp \left( -\frac{1}{2m} \left| t - t' \right|^2 \right).$$

(75)

Here $h > 0$ and $m > 0$ specify the expected length scales of the function in output and input spaces respectively.

Consider knowing the predictor data $y$ and $t$ and being interested in the value of the predictant $y_*$ at known $t_*$. Assuming $\mu = 0$ we can write a conditional distribution of $y_*$ as follows:

$$p(y_*|t_*, y, t) = N(C_*C_*^{-1}y, c(t_*, t_*) - C_*C_*^{-1}C_*^T),$$

(76)

where $C_* = [c(t_*, t_1) \ c(t_*, t_2) \ \cdots \ c(t_*, t_n)]$. Our best estimate for $y_*$ is the mean of this distribution and the uncertainty in our estimate is captured in its variance. Osborne and Roberts [2007] presented an on-line formulation of a multi-dimensional GP that allows us to model the correlations between sensor readings, and to update this model on-line as new observations are sequentially available.

5 Results

In this section we apply the multivariate time series models detailed in the previous section to real-world data. We are particularly interested in the case in which sensor readings are partially or entirely missing. We examine two data sets: (1) annual surface temperature series (single sensor problem) and (2) air temperature series measured at three locations (multiple sensor problem). To obtain the GP results given in this section we used a set of MATLAB code written by Osborne and Roberts [2007].

5.1 Single Sensor Case

We consider a global temperature series record from 1880 to 2007 (Figure 1). The data is a combination of air-sea temperature anomalies; the temperature anomalies are computed relative to the base period 1951~1980. For further information about the data refer to http://data.giss.nasa.gov/gistemp/. It appears that there exists an overall upward trend in the series with a level-off for 1940~1980 followed by a rapid upward trend. It is arguable whether the overall trend is natural or artificial, or whether there even exists any upward trend. In this experiment we are primarily interested in examining how our forecasting models discover a hidden trend in the time series and forecast future values. Hence, we will focus on their performances not on historical data but on future data.
Figure 1: Annual surface temperature anomaly relative to 1951-1980 mean, based on surface air measurements at meteorological stations and ship and satellite measurements of sea surface temperature.

In the dynamic linear model framework the forecasting problem can be treated as the problem of missing observations. In other words, if we are interested in forecasting a series of future observations from \( t + 2 \) to \( t + T \), then we can compute them via the Kalman filter merely by treating \( \{ y_{t+1}, \ldots, y_{t+T-1} \} \) as missing values. A GP model does not need such replacement because it forecasts future values as a function of time.

We first apply to this data dynamic multivariate autoregressive (DMAR) models of various AR orders – as this time series is one-dimensional the term ‘dynamic autoregressive’ would be more suitable, but for consistency we shall stick to the term ‘DMAR’. To demonstrate how these models perform, refer to Figure 2. We can see that when an AR order is small (i.e. (a) \( p = 1 \) or (b) \( p = 5 \)), the DMAR models failed to forecast any temperature trends because with such a small AR order the models are able to detect no more than noise. However, if we consider more past observations (i.e. (c) \( p = 10 \) or (d) \( p = 20 \)), the models predicted temperature forecasts showing an upward trend, with a steeper increase rate for DMAR(\( p = 20 \)). This improvement results from the fact that looking further backwards allows the models to find better any directional movement of temperature.

We now consider a multivariate local trend (MLT) model in which a hidden local trend is estimated within a window of recent observations (again, it is more appropriate to drop the term ‘multivariate’, but we retain it for consistency). We used two different sizes of a data window, namely \( n = 50 \) and \( n = 80 \); we sequentially build a MLT model with 50 or 80 past data points from 1880 to year 2007, and forecast values for the next 33 years using the trend found at 2007. Figure 3 illustrates forecasting results of the models and corresponding local trend estimates. When 50 past data points were used, an estimated local trend (i.e. \( \hat{F}_t \) in Equation (71)) at 2007 was 1.0598, which gave rise to a steep upward trend (Figure 3 (a)). Meanwhile, when
Figure 2: Predicted values for annual surface temperature anomaly by a DMAR model of an AR order: (a) $p = 1$, (b) $p = 5$, (c) $p = 10$ and (d) $p = 20$. Observed values are represented by a black line, and predictions by a red line. Shaded areas depict a 95% confidence region of the predictions. DMAR models with a higher AR order predicted temperature forecasts showing an upward trend.
more past data points were considered, the estimated local trend at 2007 was 1.0375 and this model produced less steeply increasing forecasts (Figure 3 (c)). Comparing between the two series of local trend estimates by MLT($n = 50$) and the MLT($n = 80$), we can find that the former is more erratic, ranging from 0.2628 to 1.0818, whereas the latter is more stable, varying between 0.9139 and 1.0461, which seems more realistic (Figure 3 (b) and (d)). Note that the predicted values of the MLT models are smooth compared with that of a DMAR model displayed Figure 2, in which future temperature forecasts wobble despite no such features being noticed from past observations.

Figure 3: Predicted values for annual surface temperature anomaly by a MLT model with a data window of size (a) $n = 50$ and (c) $n = 80$. (b) and (d) depict time plots of estimated local trends (i.e. the annual rate of temperature change) for MLT($n = 50$) and MLT($n = 80$) respectively.

We discuss an interesting feature that differentiates the two dynamic linear models, namely the DMAR and the MLT. The former is an autoregressive model with time-varying weights; its predicted values are not smooth as they are directly related to noisy observations. Accordingly, its future forecasts shown in Figure 2 wobble despite no such features being noticed from past observations. Meanwhile, the MLT treats true temperature as a state variable; its predicted values are filtered and thus they are more smooth as displayed in Figure 3.

As the third forecaster we consider a GP model, in which a covariance function determines the relationship between values at different times. We used a squared exponential covariance function given in Equation
(75) because it is a prototypical choice when no prior information about data is available. When a GP model used one squared exponential term that represents local disturbances in temperature, the model could not forecast an upward trend (Figure 4 (a)). To allow a GP model to deal with a trend, we considered a covariance function that is the sum of two squared exponential terms: one for modelling short term irregularities and the other for modelling a long-term trend, i.e. \( c(t, t') = \tau_1 c_1(t, t') + \tau_2 c_2(t, t') \). Here \( c_1 \) and \( c_2 \) are a squared exponential covariance function, and \( \tau_1 \) and \( \tau_2 \) are weights. A GP model using this covariance function predicted a mildly increasing trend (Figure 4 (b)). It should be noted that the predicted values by the GP model are smooth as it models underlying true temperature, not noisy observations. This is a nice property of a GP model in comparison to a dynamic autoregressive model.

In this section we applied three forecasting models to annual surface temperature data. Each model provided us with a different forecasting result on a future trend of the temperature. As discussed earlier, no one could argue with certainty that an upward trend of the global surface temperature exists, or how rapidly it grows if any. The question of how to make use of a model and how to interpret it is a challenge faced by a data modeller.

### 5.2 Multiple Sensor Case

We consider air temperature data collected by a network of weather sensors located off the coast near Southampton [Chan, 2000]. Three sensors located at Chichester Bar, Southampton Dockhead and Bramble Pile read the weather condition every second, and every five minutes the readings are averaged and uploaded to the internet with free-to-the-public access (Figure 5). Figure 6 (a) illustrates an example of such data of size \( t = 1,493 \) (i.e. about 5.2 days). We can see that there are many missing observations; such occurrences of missing data are due to a failure of transmitting collected data to the internet. Moreover, the sensor readings display a strong degree
of correlation; as the three sensor locations are a few tens of kilometers apart from each other, the air temperature readings are expected to be reasonably similar. These conditions are common in real-world applications and what we are most interested in studying. An important feature of this data is a periodic cycle that describes daytime ups and nightly downs in air temperature. For analysis of this data we use three time series models detailed in Section 4. Missing values are recovered after a certain amount of time has passed and thus true air temperature readings are available (Figure 6 (b)). As a result, we can evaluate forecasting performances of the models. Five grey regions in the figures are of our particular interest as these are periods when much data was missing.

We consider two DMAR models of AR order, \( p = 1 \) and \( p = 15 \). To demonstrate how these models perform in sparse and correlated data, refer to Figure 7. First, the DMAR(\( p = 1 \)) model performed very poorly: except for the first grey area it failed to predict any upward or downward trends when sensor readings are partially or completely missing. On the contrary, the DMAR(\( p = 15 \)) model predicted reasonably well an upward trend of the sensors in the second interest region. However, the model could not capture the daily cyclical pattern of the data and predicted poor forecasts in the third and fourth grey areas where all sensors were missing.

We now consider a MLT model in which a state evolution matrix is sequentially estimated from a set of observations up to date via the expectation-maximisation algorithm. Figure 8 illustrates the predicted values of the three sensors computed by the MLT model. In the second region of interest, long sequences of missing values of the red and blue sensors were well predicted when the green sensor readings were available. However, when all sensors were missing, the model could not predict a daily cyclical pattern of the data. We can see this
Figure 6: (a) Air temperature data collected in real time by a network of weather sensors located at Chichester Bar (red), Southampton Dockhead (green) and Bramble Pile (blue). (b) True air temperature readings recovered afterwards. Five grey regions are of our particular interest as these are periods when missing values regularly occurred.
Figure 7: Predicted values for three sensors by a DMAR model of AR order (a) $p = 1$ and (b) $p = 15$. Pink areas represent a 95% confidence region of the predictions. A DMAR model with a higher AR order performed better particularly in the second interest region. However, both models failed to predict a daily cyclical pattern of air temperature as shown in the third and fourth grey areas.
failure in the third and fourth grey areas as well. This is an expected failure as the MLT model is a first-order Markov model and it only considers local trends of the sensors. We will shortly discuss a method that allows DMAR and MLT models to deal with a periodic cycle as well as a local trend.

![Multivariate Local Trend Model](image)

Figure 8: Predicted values for three sensors by a MLT model. Despite partial missing values being well-predicted, the MLT model, which is based on a first-order Markov model, failed to capture a daily cyclical pattern of air temperature.

We also consider a GP model, in which a covariance function is constructed so as to account for two main characteristics of the air temperature data: correlated sensor readings and a daily cycle of temperature. We used a covariance function that consists of three terms: (1) a periodic term whose period is found to be almost exactly 1 day, (2) a longer-term drift term having a time scale of 3 days and (3) a higher frequency disturbance term possessing a time scale of 4 hours [Osborne and Roberts, 2007]. To illustrate how the GP model performed, refer to Figure 9. We can see that partial missing values were well-predicted on the basis of observed information. Moreover, the daily periodic feature was captured by the model, and accordingly it predicted well the missing values when all sensors were unavailable. This outstanding result is not surprising as we explicitly incorporated prior knowledge about the data into the model in the form of a covariance function.

5.2.1 Cycle-Augmented Model

Dynamic-linear-model-based models (i.e. DMAR and MLT) were found to be unsuitable for data possessing a periodic cycle. There are various ways of allowing for a periodic term in traditional statistical time series analysis (e.g. Chatfield [2004]). We here discuss an idea of incorporating a periodic cycle term into a dynamic linear model.

We know that in our air temperature data, sensor readings are observed every 5 minutes, thereby giving
Figure 9: Predicted values for three sensors by a GP model. Since the GP model incorporated prior knowledge about the data such as a daily cyclical pattern into a covariance function, it successfully predicted missing values even when all sensors were missing.

rise to a daily cycle having a period of 288 time steps. To allow a dynamic linear model to deal with the cycle we introduce a variable for such a cyclic pattern and treat its values as if they were observations. In other words, we create a time series, denoted by $s_t$, having a sine wave with a period of a day (Figure 10), and use the following ‘augmented’ observation vector:

$$\tilde{y}_t = \begin{bmatrix} y_t \\ s_t \end{bmatrix}.$$  \hfill (77)

Figure 10: A time series having a daily periodic cycle.

The cycle variable $s_t$ can be thought of as a ‘time stamp’, which provides information about which time of a day to forecast. In other words, even when all sensors are missing, the information is still available to a forecasting model, which allows it to predict missing values possessing a cyclical period. In the analysis of our air temperature data it is sufficient to consider a single periodic term because all sensors measure the same...
object. However, we may need to consider more than one periodic terms; for example, when each sensor has a cycle of different period. This case could be solved by adding several “artificial” observation variables for the cycles to an original observation variable.

To illustrate how DMAR and MLT models perform when using the above new observations, refer to Figure 11: (a) and (b) depict the results of DMAR models of AR order $p = 1$ and $p = 15$ respectively; (c) shows that of a MLT model. In comparison to their corresponding results when using original observations, which are displayed in Figures 7 and 8, we can easily see improvements made by augmenting the cycle variable. In particular, the DMAR($p = 15$) model correctly captures the daily cyclical pattern and predicted well missing values even when all sensors were unavailable. It is interesting to note that the first-order Markov models (i.e. DMAR($p = 1$) and MLT) underperformed compared to the higher-order model, DMAR($p = 15$). This result indicates that a high-order Markov model is more suitable for discovering an underlying trend as it looks further backward.

5.2.2 Comparison

To evaluate performances of the three multivariate forecasting models studied so far, we computed the root mean square error of each model for the three sensors (Table 1). For the DMAR and MLT models we used cycle-augmented observations, and the AR order of the DMAR was $p = 15$. We can see that the high-order Markov model, DMAR($p = 15$), and the GP model outperformed the first-order Markov model, MLT; the MLT model stood out in Area 2 but it performed poorly in other areas, which can be visually checked in Figure 11 (c). From the bottom part of Table 1 we can see that the GP model was better for the blue and red sensors, and the DMAR model for the green sensor. To check if these performance differences between the two models are statistically significant, we performed a two-sided t-test; the null hypothesis is that the differences between the two models’ prediction errors are a random sample from a Gaussian distribution with mean 0 and unknown variance. For the blue sensor the null hypothesis is not rejected with the $p$-value being 0.7677. This indicates that DMAR and GP did not perform differently for the blue sensor. For the green and red sensors the $p$-value is 0.0005 and 0.003 respectively, thereby rejecting the null hypothesis. These results confirm that DMAR performed better for the green sensor than GP, and for the red sensor GP better than DMAR.

In this section we analysed real-world air temperature data, in which multiple (correlated) sensors measured noisy and periodic air temperature, and frequent (partially or completely) missing observations occurred. These features of the data provided a nice testbed for our multivariate forecasting models. By using cycle-augmented observations we could improve DMAR and MLT models in a way that they deal with a daily
Figure 11: Predicted values for three sensors by (a) the DMAR\((p = 1)\), (b) the DMAR\((p = 15)\) and (c) the MLT when cycle-augmented observations are used. By allowing the models to be informed of a daily cyclical pattern, we obtained much improved performances, particularly for the DMAR\((p = 15)\) model. It predicted very well the periodic pattern of the air temperature even when all sensors were missing.
Table 1: Root mean square error of predictions by DMAR, MLT and GP models.

<table>
<thead>
<tr>
<th>Area</th>
<th>Sensor 1 (Blue)</th>
<th>Sensor 2 (Green)</th>
<th>Sensor 3 (Red)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area 1</td>
<td>1.3302</td>
<td>2.3824</td>
<td>0.9999</td>
</tr>
<tr>
<td></td>
<td>1.1178</td>
<td>1.3243</td>
<td>0.6791</td>
</tr>
<tr>
<td></td>
<td>0.8458</td>
<td>3.7949</td>
<td>1.4998</td>
</tr>
<tr>
<td>Area 2</td>
<td>1.1664</td>
<td>0.7465</td>
<td>1.5141</td>
</tr>
<tr>
<td></td>
<td>0.5920</td>
<td>0.6200</td>
<td>0.5647</td>
</tr>
<tr>
<td></td>
<td>1.0728</td>
<td>0.7783</td>
<td>0.9065</td>
</tr>
<tr>
<td>Area 3</td>
<td>0.5928</td>
<td>1.9748</td>
<td>0.4213</td>
</tr>
<tr>
<td></td>
<td>2.7788</td>
<td>4.2150</td>
<td>2.7883</td>
</tr>
<tr>
<td></td>
<td>0.5451</td>
<td>2.0685</td>
<td>0.5575</td>
</tr>
<tr>
<td>Area 4</td>
<td>0.2009</td>
<td>1.6594</td>
<td>0.3262</td>
</tr>
<tr>
<td></td>
<td>2.3582</td>
<td>3.9596</td>
<td>2.4805</td>
</tr>
<tr>
<td></td>
<td>0.7000</td>
<td>1.7338</td>
<td>0.9191</td>
</tr>
<tr>
<td>Area 5</td>
<td>0.4141</td>
<td>0.8549</td>
<td>0.4667</td>
</tr>
<tr>
<td></td>
<td>0.4253</td>
<td>0.9313</td>
<td>0.2312</td>
</tr>
<tr>
<td></td>
<td>0.6731</td>
<td>0.9191</td>
<td>0.1970</td>
</tr>
<tr>
<td>Total</td>
<td>0.7966</td>
<td>1.0294</td>
<td>0.9437</td>
</tr>
<tr>
<td></td>
<td>1.0141</td>
<td>1.5556</td>
<td>1.0237</td>
</tr>
<tr>
<td></td>
<td>0.7267</td>
<td>1.0439</td>
<td>0.7005</td>
</tr>
</tbody>
</table>

| Sensor 2 (Green) | 1.0294 | 1.5556 | 1.0439 |
| Sensor 3 (Red)   | 0.9437 | 1.0237 | 0.7005 |

periodic cycle in air temperature. We summarise characteristics of these forecasting models (Table 2). All the models are suitable for predicting a trend. However, the GP model accounts for periodic or seasonal terms in a more elegant manner as a covariance function encapsulates these effects. This feature gives rise to more flexibility to the GP model in comparison to the DMAR and MLT models. In return for the flexibility the GP model has more hyper-parameters to adjust. This is not a trivial problem, but Osborne et al. [2008a] proposed a Bayesian method that allows the GP model to infer the values of the hyper-parameters in an automated sequential manner. These useful properties of the GP model come at the cost of very high computation. Meanwhile, the DMAR model is computationally simple and rapid in its estimations.

For the air temperature data analysed here, the dynamic multivariate autoregressive model using cycle-augmented observations performed as well as the Gaussian process model. Moreover, the former is computationally considerably faster than the latter. We cannot conclude which model is universally superior. Generally speaking, there exists no such ‘blind black-box’ model in which we put in a data point and receive a best-possible prediction. It should be a data-modeller’s wise decision which model to use in what circumstances.
6 Discussions and Future Work

In this report we have demonstrated how an incomplete multivariate time series can be dealt with. We have studied three forecasting models; two of them are based on the dynamic linear model framework, namely a dynamic multivariate autoregressive (DMAR) model and a multivariate local trend (MLT) model; thirdly, a Gaussian process (GP) model that is a semi-parametric approach. In comparison to traditional statistical time series analysis, an important feature of a dynamic linear model is that it can cope easily with incomplete and/or missing observations [Harvey, 1990, Durbin and Koopman, 2001]. In addition, since in the time series analysis the GP model predicts outputs as a function of time, missing observations can be easily dealt with.

We have analysed two real-world weather data sets: (1) annual surface temperature anomalies and (2) air temperature measured at three locations. They represent the problems of forecasting a single sensor and multiple sensors respectively. In particular, the second data set is a good example of an incomplete multivariate time series, in which the sensors are correlated and intermittently generate missing observations. We have found that all of the forecasting models were good at discovering local trends of the surface temperature data. Moreover, for the air temperature data the models successfully predicted partially missing observations when the sensors are correlated. However, when all sensors are missing the standard DMAR and MLT models could not capture a periodic cycle. Meanwhile, the GP model performed very well as it makes use of prior knowledge about the periodic movement of the temperature, which is formulated as a covariance function. For the DMAR and MLT models to deal with such data having a daily periodic cycle, we have created an extra observation variable having a sine wave with a 1 day period and added it to an original observation vector. This extra information has allowed the models to forecast the periodic pattern even when no sensors were observed. Overall, the MLT has underperformed the DMAR and the GP as it is a first-order Markov model. As the DMAR and GP models have different features and advantages, we cannot conclude which one of them is superior; a decision on which model to use depends on circumstances and resources available. Interesting work related to this forecasting problem is that of Girard et al. [2003] in which they tried to find underlying dynamics of an

<table>
<thead>
<tr>
<th></th>
<th>DMAR</th>
<th>MLT</th>
<th>GP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trend</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Cycle</td>
<td>Partial yes</td>
<td>Partial yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Flexibility</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Complexity</td>
<td>Low</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td>Computation</td>
<td>Very cheap</td>
<td>Expensive</td>
<td>Very expensive</td>
</tr>
</tbody>
</table>

Table 2: Comparison between three multivariate forecasting models.
autoregressive model using the GP method.

In the financial markets it is common to meet this type of multivariate forecasting problem in incomplete environments; the markets constantly change over time; they are highly volatile; there are many different types of factors that affect the markets; these factors interrelate; and information is often incomplete. These features of the financial markets provide a huge challenge. For this reason we shall be interested in applying multivariate forecasting models to analysing such financial data.

Moreover, the forecasting problem is closely related to the decision making problem. As described in Lee and Roberts [2008], we are often interested in making a future decision with information to date. A forecasting model can be used to predict future values of inputs and a decision model can make a future decision with the predicted values. A relevant decision making problem is that of ‘Reinforcement learning’ [Sutton and Barto, 1998], in which of interest is an optimal decision that maximises the sum of future rewards. Ambitious future research would be to bridge between the forecasting and the decision making.

References


