Bayesian Model Averaging of Dynamic Linear Models

E. Mubwandarikwa, P. H. Garthwaite and A. E. Faria.

November 7, 2005

Abstract

In this paper we aim to compare the performance of three different Bayesian model averaging (or mixture) methods applied to regression dynamic linear models for beverage data from Zimbabwe. The models are chosen to reflect different plausible causal structures of association between beverage sales and other variables, thought to influence beverage sales, such as prices, temperature and maize crop production. A model averaging method based on Akaike weights performed on average 4% better than a Monte Carlo Markov Chain (MCMC) simulation method judged by the predictive one step ahead percentage root mean square error (\textit{prmse}) for forecasting a test set. The Akaike method also performed on average 73% better than a Quasi Bayes method on the same \textit{prmse} measure on the test set.

Keywords: Bayesian Model Averaging, Regression Dynamic Linear Models, model weights, model density.

1 Overview

In this paper we investigate the Bayesian Model Averaging (BMA) approach, Hoeting \textit{et al} (1999) when the models are Regression Dynamic Linear Models (RDLMs) as defined by West and Harrison (1997). We define a number of plausible RDLMs for the forecasting of beer sales in the volatile Zimbabwean market for which we are interested in obtaining a single predictive density.

It is well known that model uncertainty is usually ignored in model selection and that this leads to the underestimation of the uncertainty associated with the underlying quantities, see e.g. Draper (1995). Therefore, good practise requires that further to the uncertainty associated with model parameters, the uncertainty associated with model

\footnote{Address for Correspondence: Department of Statistics, Faculty of Mathematics and Computing, The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom. Email: e.mubwandarikwa@open.ac.uk}
selection should also be considered as part of the modelling process. In many cases, a model selection process involves the search for the “best” model according to some criteria. During this process other relevant models are simply discarded with the winning model being treated as the “true” model and used for inference.

In practice, however, finding the “true” model is not straightforward and can be a daunting task for various reasons. Identifying the regressor variables that directly influence the variable of interest at all periods of time is perhaps the most difficult task. For example, structural changes in the underlying process may have been caused by factors that were initially unforeseen. In our beer sales case, unanticipated political, economical and environmental factors have caused at least three structural changes in the volume sales of beer in Zimbabwe, as we shall see.

A way of dealing with the difficulty of selecting an individual model is to entertain multiple models. These models are then “averaged” to give a model that takes them all into consideration. Specifically the individual models become components that are linearly combined to give a new model or model mixture. We propose to carry out the model averaging or mixtures, see West and Harrison (1997) multi-process class II models for further detail. The multi-process class II models are very general and under certain assumptions and restrictions shown later in the paper we obtain the standard Bayesian Model Averaging.

Bayesian Model Averaging (BMA), like multi-process class II models, linearly combine the predictive densities generated by individual models. In the combination, weights are assigned to each model and these weights can be seen as the probability of the associated model being the “true” model. In the time series context, the model mixture method has a sequential nature with the weights being calculated as posterior probabilities in the Bayesian paradigm.

BMA has several advantages over model selection. The mixture explicitly accounts for model uncertainty thus producing more accurate and realistic inferences. After the model formulation stage, model averaging resolves the issue of having to select a model that could potentially be a “wrong” or inadequate model. Note that model selection can be seen as a particular (and rather extreme) case of model averaging when the chosen model is given full weight in the average and the remaining models are given zero weight. Therefore, model averaging is a more general approach in which model selection is a particular case.

There are some general difficulties associated with the model averaging/mixture method. First, if there are a large number of parameters this results in a large number of possible models which could be impractical and computationally costly. A pure Bayesian approach would require that we specify the prior distribution of all the relevant
parameters conditional on each possible model. This means that for $p$ parameters there will be $\sum_{j=0}^{p} \binom{p}{j} = 2^p$ possible prior models. The sequential nature in time series can lead to an ever increasing number of components, of order $O(p^t)$, in the linear combinations. This increase in the number of components leads to increasing complexity and computational costs. This can make the specification of prior distributions an impractical task even for values of $p$ that are not very large.

However there are a number of ways of dealing with these difficulties, either directly via simulation or through approximation that will be discussed later in the paper. The possible solutions include using simulation methods such as the Markov Chain Monte Carlo (MCMC) methods to estimate the posteriors directly and also using other approximation methods to stop the explosion of possible models and the effect that the passage of time has on probabilities associated with the models.

There is some existing literature on analysis of beverage data such as that of Johnston and Harrison (1980) who apply Bayesian dynamic linear model methods to sales of cider in the United Kingdom. Goldstein et al (1999) applied belief nets in model building for prediction and decision-making at a United Kingdom brewery. These models were on data coming from the more stable UK economy and have emphasis on model selection. We extend the Dynamic Linear models of Johnston and Harrison (1980) to consider data which originates from a volatile economy. Alternatively to the model selection implemented by Johnston and Harrison (1980), we perform model averaging on our data. As we are interested in investigating how different causal structures for the process that generates the data influence the structural changes in the data, we entertain a set of distinct regression DLMs for which model averaging is performed. This approach is suitable for the Zimbabwean beverage data as successive structural changes during the period of analysis means that straight forward model selection may not be sufficient.


The Dynamic Linear Models, their properties and application, including multi-process class II models which generalise the standard Bayesian Model Averaging, are discussed extensively by West & Harrison (1997). Bayesian Model Averaging and its application, via approximation and simulation methods, to linear models are discussed in Hoeting et al (1999). Estimating the mixture densities are discussed
2 THE REGRESSION DYNAMIC LINEAR MODEL

in West & Harrison (1997), and conditions for multi-modality for a normal mixture are discussed by Robertson and Fryer (1969).

In Section 2 we describe the structure of the regression Dynamic Linear Model (RDLM) and the updating of DLMs. In Section 3 we discuss the Class II multi-process models and how they generalise Bayesian Model Averaging. We also elaborate the different methods of calculating the weights applied in the model mixtures. Section 4 shows the results obtained from the model averaging. Finally we discuss our conclusions.

2 The Regression Dynamic Linear Model

The Dynamic Linear Model (DLM), West and Harrison (1997), can be represented as a system of equations specifying how observations of a process are stochastically dependent on the current process state and can be represented by how the process parameters evolve in time. The system of equations captures the inherent process dynamics with the stochastic elements modelled by random shocks or disturbances. The DLM is a general class of linear models flexible enough to represent most real time series processes. The model is stated in terms of discrete, equally spaced intervals of time although it is possible to extend it to unequal intervals. Cases with missing values can also be dealt with within the DLM formulation, West and Harrison (1997).

The formulation of the DLM must follow certain principles for Bayesian forecasting and dynamic modelling. Suppose that at a certain time \( t - 1 \) all the relevant information available up to that time is denoted as \( D_{t-1} \), i.e. \( D_{t-1} \) is the set of all observations up to time \( t - 1 \). From the modellers/forecaster point of view interest lies in the forecast value of some scalar quantity, the beer sales in our case, denoted by \( Y_t \) with observed values of this quantity as \( y_t \). It follows that \( D_t = \{ Y_t, D_{t-1} \} \). At time \( t - 1 \) a meaningful parameterisation such that all historical information relevant to predicting future observations is contained in a vector denoted \( \theta_{t-1} \). This relevant information is represented in terms of a probability distribution \( (\theta_{t-1} \mid D_{t-1}) \) such that given \( D_{t-1} \), \( (\theta_{t-1} \mid D_{t-1}) \) is sufficient for predicting the future. The parameter vector \( (\theta_{t-1}) \) must be meaningful, dynamic and changeable to allow incorporation of expert information from the decision makers and from other influential factors outside the system. Current information can then be related to the future via some derived predictive distribution \( (Y_{t+k} \mid D_{t-1}) \), \( k = 0, 1, 2, \ldots \). This derivation is via specification of a sequential parametric relation \( (\theta_t \mid \theta_{t-1}, D_{t-1}) \) together with \( (Y_t \mid \theta_{t-1}, D_{t-1}) \). In combination with \( (\theta_{t-1} \mid D_{t-1}) \) these distributions enable derivation of a full joint forecast Student \( t \)-distribution. The main property enabling effective dynamic modelling is conditional independence, which can be stated generally as follows: given the present state \( \theta_t \), the present observation \( y_t \) and the future observation \( y_{t+m} \) for \( m \geq 1 \) are independent of the past observation \( y_{t-1} \).
2.1 The Normal Regression Dynamic Linear Model

For a time point \( t \), \( t = 1, 2, \ldots \), let \( M_j \), \( j = 1, 2, \ldots \), represent a model for the underlying time series \( Y_t \) with observation \( y_t \). Let \( \theta_t \) be a \((p \times 1)\) vector of regression parameters at time \( t \), with \( p \) being the number of parameters in the model. The distribution of the underlying time series conditional on the regression parameters and the distribution of the parameters conditional on the previous parameter state can be expressed as:

\[
(Y_t \mid \theta_t) \sim N(F_t^' \theta_t, V_t) \tag{1}
\]

and

\[
(\theta_t \mid \theta_{t-1}) \sim N(G_t \theta_{t-1}, W_t) \tag{2}
\]

where \( F_t \) is an known design vector, \( G_t \) is the system matrix, \( V_t \) and \( W_t \) are the respective variances. There are two main interpretations of the process parameters. A process parameter may be expressed as a relationship between dependent variables \( y_t \) and a matrix of independent variables \( F_t \) in which case it extends the classical parametric interpretation of the static model. The second interpretation is if \( F_t \) were constant the model is a state space representation of the time series in which the parameters may be thought of as process growth and so on. The DLM may represent linear combinations of models and hierarchical models.

**Definition 2.1.1** For a time series \( Y_t \), at time \( t \) the general normal Regression Dynamic Linear Model (RDLM) is defined as:

**Observation Equation:**

\( y_t = F_t^' \theta_t + v_t \quad \{v_t \sim N(0, V_t)\} \)

**System Equation:**

\( \theta_t = G_t \theta_{t-1} + w_t \quad \{w_t \sim N(0, W_t)\} \)

**Initial Information:**

\( (\theta_0 \mid D_0) \sim N(\mu_0, \Sigma_0) \) \tag{3}

The DLM can alternatively be specified by the quadruple:

\[
M_t(j_i) = \{F_t, G_t, V_t, W_t\} \quad j = 1, 2, \ldots \tag{4}
\]

where at time \( t \),

\[
F_t' = (X_{t1}, X_{t2}, \ldots, X_{tp}) \text{ is a } (1 \times p) \text{ design vector of known values of independent variables or regressors at time } t,
\]

\[
G_t \text{ is a } (p \times p) \text{ known evolution, transfer, state or system matrix at time } t,
\]
\(v_t\) and \(w_t\) are independent normally distributed random vectors with zero mean and time varying variance \(V_t = E[v_t^2]\) and variance-covariance matrices \(W_t = E[w_t w_t']\) respectively at time \(t\).

Suppose there are \(p\) regressor/parameter variables. The variables can be identified and labelled \(X_{t1}, \ldots, X_{tp}\) for time \(t = 1, 2, \ldots\). If the regression DLM has an intercept this can be represented in the models by setting the first term \(X_{t1} = 1\). The quadruple, equation 4 defines the model relating volume sales \(Y_t\) to the \((p \times 1)\) parameter vector \(\theta_t\) at time \(t\), and the \(\theta_t\) sequence through time, via the sequentially specified distributions. It is possible for the number of known parameters in the selected DLM to vary between time points \(t - 1\) and \(t\) unless there is a unique underlying model generating the data, which is not always the case in practice. The evolution matrix \(G_t\) and evolution variance \(w_t\) describe the changes in the parameter vector \(\theta_t\) between time points \(t - 1\) and \(t\).

The usual linear regression model formulated by the classical statistical theory corresponds to the static DLM which has \(\theta_t = \theta\) and \(V_t = V\) are known and constant in time. The system equation is redundant as the state parameter vector \(\theta_t = \theta_{t-1}\) is constant. The classical linear regression model assumes that there is a unique static vector of parameters for all the time points.

Dynamic linear modelling assumes that, at each time point \(t\), we have knowledge of the observation variance matrices \(V_t\) and the system evolution variance matrices \(W_t\). However these matrices can be difficult to specify and quantify, especially in practice. The system variance matrices, \(W_t\), determines and controls the extent and nature of stochastic changes in state vectors over time. A way around this is to estimate the matrix using information discounting. Information discounting from Time Series methods has led to Component discounting in dynamic linear modelling. Discounting captures the essence of data diminishing in value as time evolves, (Pole, West and Harrison 1994). Using discount factors has the advantages of being easy to understand, are independent of any scale measures (being percentages) and take away the need to implement prior to posterior updating of individual parameters in the model.

### 2.2 Sequential Updating for the normal RDLM

In Dynamic Linear Models updating of the vector of regression parameters, \((\theta_{t-1})\), at time \(t - 1\) to \(\theta_t\), at time \(t\) occurs via the Kalman Filter, West and Harrison (1997). Let \(D_t\) represent all the knowledge from the past up to time \(t\). Inferences about \(\theta_t\), also known as the State of Nature, given all the knowledge available, \(D_t\), can be carried out in terms of the densities, \(p(\cdot)\), through application of Bayes Theorem as follows:
\[ p(\text{State of Nature} \mid \text{Data}) \propto p(\text{Data} \mid \text{State of Nature}) \times p(\text{State of Nature}), \]

\[ p(\theta_t \mid D_t) \propto p(Y_t \mid \theta_t, D_{t-1}) \times p(\theta_t \mid D_{t-1}). \]  \hspace{1cm} (5)

This expresses of the relationship between the posterior and the prior distributions. At time point \( t - 1 \), having observed all the data up to the current time \( D_{t-1} \) our state of knowledge (or the posterior distribution for \( \theta_{t-1} \)) of \( \theta_{t-1} \) can be written as:

\[ (\theta_{t-1} \mid D_{t-1}) \sim f(\theta_{t-1} \mid D_{t-1}), \]  \hspace{1cm} (6)

where \( f(\theta_{t-1} \mid D_{t-1}) \) is a distribution function for \( \theta_{t-1} \). The distribution function can be any of the family of statistical distributions. If, for example, \( (\theta_{t-1} \mid D_{t-1}) \) were Normally distributed then \( (\theta_{t-1} \mid D_{t-1}) \sim N(m_{t-1}, C_{t-1}) \), with \( m_{t-1} \) being the expectation and \( C_{t-1} \) the variance.

At time \( t = 0 \), the initial distribution \( (\theta_0 \mid D_0) \) and its properties are specified by the modeller. Depending on how much knowledge is available this distribution may be known explicitly or is estimated using a priori knowledge. In general, updating from time \( t - 1 \) to time \( t \) occurs in two stages. First, prior to observing \( Y_t \), we have our prior distribution \( (\theta_t \mid Y_{t-1}) \). The one-step ahead forecast for \( Y_t \) can be estimated via the distribution

\[ (Y_t \mid D_{t-1}) \sim f(Y_t \mid D_{t-1}). \]  \hspace{1cm} (7)

Second, after to observing \( Y_t \). The aim now is to compute the posterior of \( \theta_t \) via equation 5. First the likelihood \( L(\theta_t \mid D_t) \) equivalently \( p(Y_t \mid \theta_t, D_{t-1}) \) has to be calculated. Let \( e_t = Y_t - \hat{Y}_t \) be the error in predicting \( Y_t \), with \( \hat{Y}_t \) being the forecast value. Thus observing \( Y_t \) is equivalent to observing \( e_t \). Eq 5 can be re-written as:

\[ p(\theta_t \mid D_t) = p(\theta_t \mid e_t, D_{t-1}) \propto p(e_t \mid \theta_t, D_{t-1}) \times p(\theta_t \mid D_{t-1}), \]  \hspace{1cm} (8)

where \( p(e_t \mid \theta_t, D_{t-1}) \) is the likelihood. From Bayes Theorem, we can describe our current state of knowledge as:

\[ p(\theta_t \mid D_t) = \frac{p(e_t \mid \theta_t, D_{t-1}) \times p(\theta_t \mid D_{t-1})}{\int_{\theta_t} p(e_t \mid \theta_t, D_{t-1}) \, d\theta_t}. \]  \hspace{1cm} (9)

Once \( p(\theta_t \mid D_t) \) is computed updating can be done recursively for time point \( t + 1 \) and so forth.
Estimating the observational variance can be difficult. One reason is that the available sample data is not very large and thus sample estimates may not be reliable enough to be thought of as the true population values. In this paper we adopt classes of RDLMs for which the observational variance, \( V_t \), is unknown but constant. The evolution/updating cycle is based on the following distributions:

\[
\begin{align*}
(\theta_{t-1} \mid D_{t-1}) & \sim T_{n_{t-1}}[m_{t-1}, C_{t-1}], \\
(\theta_t \mid D_{t-1}) & \sim T_{n_{t-1}}[a_t, R_t], \\
(\phi \mid D_{t-1}) & \sim G[n_{t-1}/2, d_{t-1}/2] \text{ with } S_{t-1} = d_{t-1}/n_{t-1}, \\
(Y_t \mid D_{t-1}) & \sim T[f_t, Q_t], \\
(\theta_t \mid D_t) & \sim T_n[m_t, C_t], \\
(\phi \mid D_t) & \sim G[n_t/2, d_t/2] \text{ with } S_t = d_t/n_t, 
\end{align*}
\]

where \( a_t = m_{t-1}, R_t = C_{t-1} + W_t, f_t = F_t'm_{t-1}, Q_t = F_t'R_{t-1}F_t + S_{t-1} \) and the remaining elements are defined by the usual updating equations of Section 2.2. \( n_t = n_{t-1} + 1, d_t = d_{t-1} + S_{t-1}e_t^2/Q_t, m_t = m_{t-1} + A_te_t \) and \( C_t = [R_t - A_tA_t']S_t/S_{t-1} \) where \( e_t = Y_t - f_t \) and \( A_t = R_tF_t/Q_t. \) The multivariate Student \( t \)-distribution for \( \theta_t \) with \( p \) degrees of freedom, mode \( m_t \) and a positive-definite scale matrix \( C \) is

\[
P(\theta_t \mid D_t) = C_p \frac{|C_t|^{-1/2}}{[n_t + (\theta_t - m_t)'C_t^{-1}(\theta_t - m_t)]^{(n_t+p)/2}}
\]

where

\[
C_p = \frac{n_p^{n_t/2} \Gamma((n_t+p)/2)}{\pi^{p/2} \Gamma(n_t/2)}
\]

For forecasting to \( t + k \) the distributions are:

\[
\begin{align*}
(\theta_{t+k} \mid D_t) & \sim T_n[m_t, R_t(k)], \\
(Y_{t+k} \mid D_t) & \sim T_n[f_t(k), Q_t(k)],
\end{align*}
\]

where \( R_t(k) = C_t + \sum_{r=1}^k W_{t+r}, f_t(k) = F_t'R_t(k)m_t \) and \( Q_t(k) = S_t + F_t'R_t(k)F_{t+k}. \)

### 2.3 Similar RDLMs

We have so far defined the types of models, the RDLMs, that will be used for model averaging. However before performing the model averaging, it is important to introduce the concept of model similarity. This is particularly important by entertaining only a set of non-similar models, we avoid the problem of explosion of the number of component models in the model averaging as shown later. In
3 Bayesian Averaging of RDLMs

One of the main advantages of DLMs is that at any time, \( t - 1 \), the only information required to forecast \( y_t \) at time \( t \) is a DLM which describes the evolution, of whatever process is generating the data, in the interval \((t - 1, t)\) together with the probability description of the relevant parameters \( (\theta_{t-1} \mid y_{t-1}) \). It is possible that the DLM used to explain the evolution of the process at one time can be different, in either its specification or parametric structure, to that used at another time.
time.

In model selection, a single model is selected from a set of candidate models, $M_t = \{ M_t(j), j = 1, 2, 3, \ldots, N \}$, to represent the underlying time series process. This is analogous to parameter estimation and model identification in classical statistics. Only in our case we are doing this for DLMs rather than static models. Forecast and decision making based on information relating to the whole set of models are the main drivers of this work and as such more emphasis will be put on them in later sections through model averaging rather than model selection. Model selection suffices if we can be sure that we have landed on the “best” model to describe the data. However this is not always the case and the disadvantages are discussed in the next section.

3.1 Multi-process Class II RDLM

The most general class of Bayesian Model Averaging (BMA), Hoeting et al (1999) is the multi-process class II models of West and Harrison (1997). Contrary to model selection, model averaging allows the explicit estimation of model uncertainty (see, e.g. Clyde and George (2003)). In fact fitting a model to the data there is some uncertainty associated with selecting the parameters in the model. By selecting a single, unique “best” fitting model to the data there is further uncertainty about which model to select and whether it suffices to account for all the information in the data.

Bayesian Model Averaging is a class of model averaging where the uncertainty associated with each model is modelled within the Bayesian paradigm. Unlike model selection, model averaging methods explicitly handle the uncertainty associated with the models themselves. Model uncertainty is assessed by prior distributions and is incorporated in the BMA through averaging over the model predictive distributions of $Y_t$ according to each model.

The main idea behind model averaging is to linearly combine the predictive densities generated by individual models. Each model's predictive density in the combination is associated with a weight, this can be interpreted as the probability of the model being the “true” model. Uncertainty about a model being the true model can be statistically assessed in a Bayesian framework work prior-to-posterior updating of the associated weight. If the weights are made equal for all the models this becomes simple forecast averaging.

In the use of DLMs it is implicitly assumed that though a certain DLM may be an appropriate representation of the underlying process during a period of time, it may not (and usually is not in practice) be appropriate at all times. Certainly interventions can be carried out in a model during or after a structural change in the process however those will be limited to the models parameter set. This is clearly
not satisfactory when structural changes in the underlying process are caused by changes in the factors that govern (or drive) that process. An example of this on the Zimbabwean beverage sales was the effect of rapid currency devaluation towards the end of the 1990s which was not present at the beginning of the decade. Alternative models can be used to counter the inadequacy of using an individual model on the whole data set (West and Harrison, 1997). A single DLM defines a process model and linear combinations of DLMs form multi-process Class II models.

Suppose that any defining parameters of a DLM that are possibly subject to uncertainty are denoted by \( \alpha \). At each time \( t \), \( \alpha \) takes a value in the discrete set \( \mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_k\} \), the values possibly differing over time. Each DLM is indexed by \( \alpha_j \), but for simplicity as before we index the models by setting \( \alpha_j = j \), \( (j = 1, 2, \ldots, k) \), \( \mathcal{A} = \{1, 2, \ldots, k\} \). The \( j \)-th model at time \( t \) will be defined as \( M_t(j) \).

**Definition 3.1.1** Under the framework described, if a model \( M_t(j) \) is selected with known probability

\[
P(M_t(j_t)) = P[M_t(j_t) \mid D_{t-1}],
\]

then the series \( Y_t \) follows a multi-process class II model.

The combination of DLMs in a class provides an overall model for the series. As the combination uses discrete mixtures of DLMs, multi-process models are also known as mixture models.

**Definition 3.1.2** For each time point \( t \) and integer \( h \), \( 0 \leq h < t \), define the probabilities

\[
P(M_t(j_t), M_t(j_{t-1}), \ldots, M_t(j_{t-h})) = P[M_t(j_t), M_{t-1}(j_{t-1}), \ldots, M_{t-h}(j_{t-h}) \mid D_t].
\]

West and Harrison (1997) show that the posterior density, for \( \theta_t \mid D_t \), can be written as

\[
p(\theta_t \mid M_t(j_t), D_t) = \sum_{j_t=1}^{k} p(\theta_t \mid M_t(j_t), D_t) P(M_t(j_t)),
\]

or equivalently as

\[
\sum_{j_t=1}^{k} \sum_{j_{t-1}=1}^{k} p(\theta_t \mid M_t(j_t), M_{t-1}(j_{t-1}), D_t) P(M_t(j_t), M_t(j_{t-1}))
\]

substituting and so on down to the final stage

\[
\sum_{j_t=1}^{k} \sum_{j_{t-1}=1}^{k} \ldots \sum_{j_1=1}^{k} p(\theta_t \mid M_t(j_t), M_{t-1}(j_{t-1}), \ldots, M_t(j_1), D_t)
\]

\[
\times P(M_t(j_t), M_t(j_{t-1}), \ldots, M_t(j_1)).
\]
Only at the final stage of elaboration equation 16 where the sequence of models obtaining at each of the times 1, 2, \ldots, t are assumed, are the posteriors

\[ p(\theta_t | M_t(j_t), M_{t-1}(j_{t-1}), \ldots, M_1(j_1), D_t) \]  

of standard DLM form. Thus to obtain the marginal posterior as a mixture, it is necessary to consider all \( k^t \) possible combinations that may apply. For any combination of models the DLM analysis applies. The posterior combination of each combination \( P(M_t(j_t), M_{t}(j_{t-1}), \ldots, M_t(j_t)) \) weight the combination of the overall mixture.

Without loss of generality, moving between any two time points \( t-1 \) to time \( t \) can be done via the Kalman Filter for each possible current state to each component of the time \( t \) posterior, thus generating a distribution comprising \( k^t \) distinct normal or Student-\( t \) components. As mentioned earlier, we end up with \( k^t \) possible combinations for all the time points 1, 2, \ldots, t. A method to inhibit this explosive growth must be used for practical purposes, hence our limiting the set of models to be averaged over to non-similar models as described earlier.

### 3.1.1 Approximations for Mixtures

West and Harrison, (1997), propose an approximation for practical implementation to curtail the explosion of terms in Equations. 14, 15 and 16. The concept of information becoming less relevant with the passage of time, also applies to model mixtures. Therefore it is expected that the full conditional posterior \( p(\theta_t | M_t(j_t), M_{t-1}(j_{t-1}), \ldots, M_1(j_1), D_t) \) will depend negligibly on models from early times for large values of time, \( t \). For some fixed integer \( h \geq 1 \) the full posterior will depend essentially on those models applying only up to \( h \)-steps back in time,

\[ p(\theta_t | M_t(j_t), M_{t-1}(j_{t-1}), \ldots, M_t(j_t), D_t) \approx p(\theta_t | M_t(j_t), M_{t-1}(j_{t-1}), \ldots, M_{t-h}(j_{t-h}), D_t) \]  

On this assumption, the number of components of the mixture posterior at any time will not exceed \( k^{h+1} \) therefore there is no need to consider models up to \( h \)-steps back in time for the analysis hence approximating the full mixture by

\[ \sum_{j_{t-1}=1}^{k} \sum_{j_{t-2}=1}^{k} \ldots \sum_{j_{t-h}=1}^{k} p(\theta_t | M_t(j_t), M_{t-1}(j_{t-1}), \ldots, M_{t-h}(j_{t-h}), D_t) \times P(M_t(j_t), M_{t}(j_{t-1}), \ldots, M_t(j_t)). \]  

This mixture is still in DLM form but now contains a fixed number \( k^{h+1} \) components compared to the \( k^t \) components in the full mixture. Bayes Theorem then gives the posteriors as
The second term is given by

$$p(Y_t \mid M_t(j_t), \ldots, M_{t-h}(j_{t-h}), D_{t-1}) = \sum_{j_{t-h-1}=1}^k p(Y_t \mid M_t(j_t), \ldots, M_{t-h}(j_{t-h}), M_{t-h-1}(j_{t-h-1}), D_{t-1})$$

$$\times P[M_{t-h-1}(j_{t-h-1}) \mid D_{t-1}]$$  (21)

an average of the normal or Student $t$-distribution one step ahead predictive densities for $Y_t$ under each of the models in the conditionings. The average is with respect to models $(h+1)$-steps back, those at time $t-(h+1)$. The probabilities weighting these term are available from the identity

$$P[M_{t-(h+1)}(j_{t-(h+1)}) \mid D_{t-1}] = \sum_{j_{t-1}=1}^k \ldots \sum_{j_{t-h}=1}^k p_{t-1}(M_t(j_{t-1}), \ldots, M_t(j_{t-h}), M_t(j_{t-(h+1)})$$  (22)

The first term is calculated by,

$$P[M_t(j_t), \ldots, M_{t-h}(j_{t-h}) \mid D_{t-1}] = P[M_t(j_t) \mid M_{t-1}(j_{t-1}), \ldots, M_{t-h}(j_{t-h}), D_{t-1}]$$

$$\times p_{t-1}(M_t(j_{t-1}), \ldots, M_t(j_{t-h}))$$

$$= \pi(M_t(j_t)) p_{t-1}(M_t(j_{t-1}), \ldots, M_t(j_{t-h}))$$

$$= \pi(M_t(j_t)) \sum_{j_{t-(h+1)}=1}^k p_{t-1}(M_t(j_{t-1}), \ldots, M_t(j_{t-(h+1)}))$$  (23)

This can be found directly since the summands here are just the $k^{h+1}$ posterior model probabilities at time $t-1$. Further approximations can be done by taking into consideration,

(i) ignoring components with small posterior probabilities,

(ii) combining components that are roughly equal into a single component and also combining the probabilities and

(iii) replacing collections of components by a single component that roughly estimates the contribution of the collection.

If we put a further restriction:
3 BAYESIAN AVERAGING OF RDLMS

(iv) Setting \( h = 0 \) means that the averaging at time \( t \) is only carried out on the current \( k \) non-similar models only. The selection of non-similar models avoids duplication as similar models are intrinsically the same model as one can be obtained through an appropriate transformation of the other.

This restriction removes the need for collapsing different models over time as it fixes the number of available models in the average to only those at current time. The last restriction, (iv), leads to the standard Bayesian Model Averaging which is a special case of the multi-process class II models. In general Bayesian Model Averaging is not limited to non-similar models. At time \( t - 1 \), the one step ahead predictive mixture distribution, Hoeting et al (1999), for \( Y_t \) given data \( D \) is:

\[
p(Y_t | M_t(j_t), D_{t-1}) = \sum_{j_t=1}^{k} p(Y_t | M_t(j_t), D_{t-1}) P[M_t(j_t) | D_{t-1}], \tag{24}
\]

where \( p(Y_t | M_t(j_t), D_{t-1}) \) is the predictive density of the RDLM \( M_t(j_t) = \{F_t, G_t, V_t, W_t\} \), \( k \) is the total number of possible models under consideration and \( P[M_t(j_t) | D_{t-1}] \) is the prior probability of \( M_t(j_t) \) being the true model at time \( t \). This is the average of the posterior distributions under each of the models considered weighted by their posterior model probabilities. In equation 24 the posterior probabilities for each model are given by,

\[
P(M_t(j_t) | D_t) = \frac{P(D_t | M_t(j_t))p(M_t(j_t))}{\sum_{j=1}^{k} P(D_t | M_t(j))p(M_t(j))}, \tag{25}
\]

where

\[
P(D_t | M_t(j_t)) = \int P(D_t | \theta_t, M_t(j_t))p(\theta_t | M_t(j_t))d\theta_t \tag{26}
\]
is the integrated likelihood of model \( M_t(j_t) \), \( \theta_t \) is the vector of parameters of model \( M_t(j) \), \( p(\theta_t | M_t(j_t)) \) is the prior density of \( \theta_t \) under \( M_t(j_t) \), \( P(D_t | \theta_t, M_t(j_t)) \) is the likelihood and \( p(M_t(j)) \) is the prior probability that \( M_t(j) \) is the true model, under the assumption that one of the models is the true one.

Which models to average over can be difficult as there can potentially be a large number of models. Hoeting et al (1999) suggest several methods for this. One of them is to select a subset of models that belong to a class:

\[
A = \left\{ M_t(j_t) : \frac{\max\{P(M_t(j^*) | D)\}}{P(M_t(j_t) | D)} \leq C \right\}, \tag{27}
\]
where $C$ is a constant specified by the modeller. This method is based on the argument that if a model does not predict the data as well as the model, $M_t(j^*)$, which provides the best predictions then it is effectively disregarded.

One alternative to the problem of specifying the priors in Eq 25 can be overcome by assuming that all models are equally likely. As such each model is given equal weight. Another alternative to this method is to use a method that finds a “good” model and then selecting several models that are close to this model under some performance measure. The models are then averaged around this model. This is the approach used in this paper. Various selection criteria, namely the percentage root mean square error ($prmse$), mean absolute deviation ($mad$), the log Likelihood ($log-L$), Akaike Information Criteria ($AIC$), Bayesian Information Criteria ($BIC$), Fractional Bayes Factors ($FBF$) and geometric mean relative absolute error ($gmrae$) are used to select the models. The models are then averaged using Bayesian Model Averaging.

3.2 Methods for the Posterior Model Probabilities

Bayesian Model Averaging (BMA) methods linearly combine predictive densities generated by individual models. In the combination, weights are assigned to each model. A weight can be interpreted as the probability of the associated model being the “true” model. The BMA method requires the prior and posterior probabilities for its application. Weights based on the posterior probabilities have the advantage of occurring naturally and also allow ranking of the models. At time $t$, the weights $w_{jt}$ associated with the model $M_t(j_t)$ have the condition that they are non-negative and sum up to one, $\sum_{j=1}^{k} w_{jt} = 1$, $w_{jt} \geq 0$.

The vector for the weights is $\omega = (w_{t1}, \ldots, w_{tk})'$. Note that if any of the weights is equal to 1 then automatically all the others are zero. This case with $w_{jt} = 1$ for any $j$, is a special case of model averaging because it yields the standard model selection. Two main methods of obtaining the weights (which is in effect specifying the priors as the weights are based on the posterior probabilities) used in the model mixtures will be explored. The methods of obtaining the weights are either via approximations or exact results (via simulation) for the component models in the resulting mixture.

As a mixture model is a combination of various densities, interest lies in finding out the resulting density of this combination or at least approximating it. On one hand mixtures can be multi-modal and thus approximating them by uni-modal densities may be misleading though on the other hand, uni-modality may be desirable from a decision making stand point. If $\theta_t$ has a density described by a mixture:
with known probability $P(M_t(j_t))$, the true density of the mixture can be approximated by combining component densities in the mixture. A possible approximation is to minimise the Kullback-Leibler divergence measure between the resulting mixture and the individual components. Note this approximation can be considered to be optimal in the Kullback-Leibler, see West and Harrison (1997) for further details.

### 3.2.1 Method 1: Akaike Weights Approximation

The Akaike Information Criteria (AIC) by Akaike, (1974), originates from the maximum (log-)likelihood estimate (MLE) of the error variance of a Gaussian linear regression model. The maximum (log-)likelihood method can be used to estimate the values of parameters in models in classical linear regression. AIC advocates that from a class of competing models select the model that minimizes:

$$AIC = -2 \log f_j(y|\hat{\theta}_i) + 2p = -2 \log L_j + 2p$$

where for the $j$-th model $L_j$ is the likelihood, $p$ is the number of parameters in the model and $\hat{\theta}_i$ the maximum likelihood estimate of $\theta_i$.

The method picks the model that gives the best approximation asymptotically (Akaike, (1974)) in the Kullback-Leibler sense. The Akaike weights are calculated by first defining

$$\Delta_j = AIC_j - \min(AIC).$$

where $\min AIC$ is the smallest value of AIC in the model set. The likelihood $L_j$ of each model $M_t(j)$ is conditional on the data and the set of models. The Akaike weights, $w_j$, are then calculated as:

$$w_j = \frac{e^{-\delta_j}}{\sum_k e^{-\delta_k}}$$

where $k$ is the number of possible models under consideration and the rest of the models components as defined, Turkheimer, Hinz and Cunningham (2003).

### 3.2.2 Method 2: Quasi-Bayes Approximation

A method to estimate the weights for the model averaging called quasi-Bayes is proposed by Smith and Makov (1978) and reviewed
by Faria and Souza (1995). Let \( M_t(j_t) \), \( (j_t) = (1, 2, \ldots, k) \) be the models generating one step-ahead forecasts for \( Y_t \) with observations as \( y_t = (y_1, \ldots, y_t)' \). Let the predictive densities (denoted \( p(y_t \mid \omega) \)) for the forecast be assumed known. At time \( t \) the weights \( w_{j,t} \) associated with the model \( M_t(j_t) \) have the conditions that they are non-negative and sum to one, \( \omega = (w_{1t}, w_{2t}, \ldots, w_{kt}) \). By Bayes Theorem, for \( t \geq 1 \),

\[
p(\omega \mid y_t) \propto p(y_t \mid \omega)p(\omega \mid y_{t-1}). \tag{32}
\]

If we now define random variables \( x_1, x_2, \ldots, x_t \), such that \( x_t = j \) if and only if \( x_t \) belongs to model \( M_t(j) \) then,

\[
p(\omega \mid y_t) = \sum_{j=1}^{k} p(x_t = j \mid y_t)p_j(\omega \mid y_t). \tag{33}
\]

where for \( j = 1, 2, \ldots, k \)

\[
p(x_t = j \mid y_t) \propto p(y_t)\hat{w}_{j,n-1}(y_{t-1}) \tag{34}
\]

and

\[
\hat{w}_{j,n-1} = \int_{0}^{1} \int_{0}^{1} w_{j}p(\omega \mid y_{t-1})d\omega \tag{35}
\]

Sequential learning about the weights, \( \omega \), takes place through \( p(\omega \mid y_t) \) and classification of the successive observations on the basis of \( p(x_t = j \mid y_{t-1}) \), \( j = 1, 2, \ldots, k \). Successive computation of equation 32 or equation 33 introduces an ever expanding linear combination of component posterior densities, each of which corresponds to an updating based on a particular choice of previous classifications. At the \( t \)-th stage there are \( k^t \) possible classifications and computation becomes quickly prohibitive. The quasi-Bayes is an approximation method to overcome this.

To initialise the approximation, at time \( t = 0 \), assume that the prior density for the weights \( p(\omega) \) has the form of the Dirichlet density,

\[
p(\omega) = \frac{\Gamma(\alpha_{1,0} + \alpha_{2,0}, \ldots, + \alpha_{k,0})}{\Gamma(\alpha_{1,0})\Gamma(\alpha_{2,0}), \ldots, \Gamma(\alpha_{k,0})} \prod_{j=1}^{k} \theta_{j,0}^{\alpha_{j,0} - 1} \tag{36}
\]

which is denoted as \( D(\omega; \alpha_{1,0}, \alpha_{2,0}, \ldots, \alpha_{k,0}) \) where \( \alpha_{j,0} > 0 \), \( j = 1, 2, \ldots, k \). After observing the first value, at time \( t = 1 \), in the test set \( Y_1 = y_1 \) and letting \( x_1, x_2, \ldots, x_t \), be a random variable such that \( x_t = j \) if and only if \( x_t \) belongs to model \( M_t(j) \),

\[
p(\omega \mid y_1) = \sum_{j=1}^{k} p(x_1 = j \mid y_1)D(\omega; \alpha_{1,0} + \delta_{j1}, \alpha_{2,0} + \delta_{j2}, \ldots, \alpha_{k,0} + \delta_{jk}) \tag{37}
\]

where \( p(x_1 = j \mid y_1) \propto p(y_1)\alpha_{j,0} \) and
\[
\delta_{ji} = \begin{cases} 
0 & \text{if } j \neq i \\
1 & \text{if } j = i
\end{cases}
\]

Equation 37 can be approximated by a Dirichlet density. Note that if the model generating \( y_1 \) was known, with realisations for \( \delta_{ji} = \Delta_{ji} \), we can obtain

\[
p(\omega \mid y_1) = D(\omega; \alpha_{1,0} + \Delta_{11}, \alpha_{2,0} + \Delta_{12}, \ldots, \alpha_{k,0} + \Delta_{1k})
\]

(38)

where

\[
\Delta_{j1} = \begin{cases} 
1 & \text{if } y_1 \text{ belongs to } M_t(j) \\
0 & \text{otherwise.}
\end{cases}
\]

However as the true underlying density generating the data is not known we have \( p(x_1 = j \mid y_1) \) the expectation (given \( y_1 \)) of \( \Delta_{j1} \) when the latter is regarded as known. This provides the starting point for the proposed approximation replacing equation 37

\[
p(\omega \mid y_1) = D(\omega; \alpha_{1,0} + p(x_1 = 1 \mid y_1), \alpha_{2,0} + p(x_1 = 2 \mid y_1), \\
\ldots, \alpha_{k,0} + p(x_1 = k \mid y_1)).
\]

(39)

Subsequent updating is then done within the Dirichlet family of distributions: \( p(\omega \mid y_t) \) is Dirichlet with parameters \( \alpha_{t,j} = \alpha_{(t-1),j} + p(x_t = j \mid y_t) \), \( j = 1, 2, \ldots, k \) where the \( \alpha_{(t-1),j} \) are the parameters of \( p(\omega \mid y_{t-1}) \) and the calculation of \( p(x_t = j \mid y_t) \) proceed through equation 34 and equation 35.

For application at the realisation of a time point \( t = T \) this simplifies to estimating sequentially from a Dirichlet distribution with initial parameters \( \alpha_{j,0} \) assumes the following quasi-Bayes simplification:

\[
p(\omega \mid y_T) = D(\omega; \alpha_{1,T}, \alpha_{2,T}, \ldots, \alpha_{k,T})
\]

(40)

where the parameters are updates as

\[
\alpha_{j,T} = \alpha_{j,T-1} + \frac{f_j(y_T)\alpha_{j,T-1}}{\sum_{j=1}^{k} f_j(y_T)\alpha_{j,T-1}}; j = 1, 2, \ldots, k,
\]

(41)

where the parameters are as defined and \( f_j(y_T) \) is the point value of the density function after observing \( (y_T) \). The mean of the posterior distribution for \( w_j \) at \( t = T \) is then

\[
\hat{w}_{j,T} = \bar{w}_{j,T} = \frac{\alpha_{j,T}}{\sum_{j=1}^{k} \alpha_{j,T}}; j = 1, 2, \ldots, k.
\]

(42)

Hence at each time \( t \), the weights of each individual model in the quasi-Bayes model mixture are estimated from equation 42. In our application we initialise the analysis by setting the values of \( \alpha_{1,0}, \alpha_{2,0}, \ldots, \alpha_{k,0} = 0.2 \), giving each of the 5 models an equal chance.
3.3 Method 3: Markov Chain Monte Carlo (MCMC)

Simulation methods to estimate the posterior probabilities, equation 24, exactly are carried out via the Markov Chain Monte Carlo (MCMC) method. There are several methods for carrying out MCMC simulation. Here implementation was via Reversible Jump Markov Chain Monte Carlo (RJ-MCMC). Suppose that $({\theta}_k, M_t(k))$ is the current state. The aim of RJ-MCMC is to construct efficient model jumping proposals and appropriate mapping functions to move to a new state $({\theta}_j, M_t(j))$. Reversible Jump Markov Chain Monte Carlo (RJ-MCMC) samples over the model space and parameter space and does not require exhaustive enumeration of the model space, (Clyde and George (2003)). This property theoretically makes it possible to use RJ-MCMC methods for moderate and large dimensional problems.

The simulation method was implementing using the statistical package $R$, (Ihaka and Gentleman (1996)), by modification and application of Markov chain Monte Carlo model composition (MC3 hereafter) software written by Hoeting et al (1999). In the MC3 method the model $M_j$ is determined by picking one of the $p$ variables at random and considering whether to delete it, if it is currently in the model, or add it, if it is not currently in the model.

The method’s aim is to estimate equation 24, (Hoeting et al, (1999)), by constructing a Markov chain $\{M(t), t = 1, 2, 3, \ldots\}$ with state space $\mathcal{M}$ and equilibrium distribution $P(M_t(j) \mid D)$. This Markov Chain can be simulated to obtain observations $M(1), \ldots, M(N)$. Then for any function $g(M_t(j))$ defined on $\mathcal{M}$, the average

$$\hat{G} = \frac{1}{N} \sum_{t=1}^{N} g(M(t)) \quad (43)$$

is an estimate of $E(g(M))$. This results in:

$$\hat{G} \to E(g(M)) \text{ as } N \to \infty \quad (44)$$

To compute equation 24 in this way we set

$$g(M) = p(\hat{Y}_j \mid M_t(j), D). \quad (45)$$

Suppose that at the current time the model is $M_t(k)$. The Markov chain is constructed by defining a neighbourhood for each model $M_t(k) \in \mathcal{M}$ denoted $\text{nbd}(M_t(k))$. A neighbourhood for the DLMs might be models that have one less or one more parameter than the model in question $M_t(k)$. Obviously $M_t(k)$ itself is in the neighbourhood.

Define a transition matrix $q$ by setting
\[ q(M_t(k) \rightarrow M_t(j)) = 0 \forall M_t(j) \notin nbd(M_t(k)) \]  
\[ \text{(46)} \]

and

\[ q(M_t(k) \rightarrow M_t(j)) \neq 0 \forall M_t(j) \in nbd(M_t(k)). \]  
\[ \text{(47)} \]

If the chain is currently in state \( M_t(k) \), proceed by drawing \( M_t(j) \) from \( q(M_t(k) \rightarrow M_t(j)) \). Accept \( M_t(j) \) with probability:

\[ \alpha = \min\{1, \frac{P(M_t(j) \mid D)}{P(M_t(k) \mid D)}\}. \]  
\[ \text{(48)} \]

Otherwise the chain remains in state \( M_t(k) \).
4 Application: BMA for the Beer Sales Data

In this section, we apply the model averaging methods described in Section 3 to a time series of beer sales in Zimbabwe. The time series plot of the 120 month sales data can be seen in Fig. 1. The plot clearly shows an element of seasonality, which is not unexpected. Beer sales are higher during the festive period with peaks in December coinciding with the peak of summer in the Southern Hemisphere. The troughs bottom out in July, which coincides with winter. There is a downward trend in the period 1991-95, an upward trend in the 1995-99 period and a downward trend in the period from 1999 onwards.

The data were split into two sets, a training set and a test set. The training set comprises the first 108 observations and will be used for fitting the models. The second set, with time points 109 to 120, is used for checking the performance of the fitted models. The beer sales data showed 3 clear distinct periods, showing clear linear growth or decline. Models were fitted on each of the 3 periods and on the whole global data set. The sales data were transformed by taking the natural logarithm and then deseasonalised to account for the seasonality in the data. The regressors were transformed by taking the natural logarithm and forecasted using simple DLMs (with or without seasonality depending on whether the regressor exhibited seasonality or not) for
The analysis is carried out in line with the principles for Bayesian forecasting and dynamic modelling outlined in Section 2.1. We propose in line with Johnston and Harrison (1980), a class of Bayesian dynamic linear models which account for trend, seasonal components, factors such as price, weather and economic indicators amongst others which the company believes are influential on sales. There are potentially $4 \times 2^7 = 512$ models, found from each of the 4 data sets (the 3 distinct segments of either constant linear growth or decline within the data and the global set) and the 7 possible regressors (with a possible $2^7$ model combinations). A step-wise method to find the best models quickly at each time point for each set of models was used. The “best” fitting models were then selected.

4.1 The Plausible RDLMs

The models selected from each of the 3 periods were then extrapolated on to the whole data set and compared with the global models. The models selected are labelled (depending on how many variables were in each model see later in this Section) and put in the set of models to be averaged $\mathcal{A} = \{M0, M1, M2, M3, M7\}$. The full model $M7$ was not among the “best” models but is added as the full set of regressors is of interest. During the model averaging the effect of the full model being outside the original “best” set should not be very significant as we expect it to be assigned small weight. Note that the final models selected are non-similar and non-equivalent. The main objective of our application is to determine the averaging method that performs better in terms of their predictive densities producing better forecasts.

Discount Factors were determined by fitting the models using varying discount factors to determine the optimal values. The derived optimal trend component discount $\delta_T = 0.9$, the seasonal component discount $\delta_S = 0.95$, the variance component discount $\delta_W = 0.99$ and the regression component discount $\delta_R = 0.98$ were used in all the modelling.

Writing the models in the defined DLM quadruple form, $\{F_t, G_t, V_t, W_t\}$, the model MA0 which has an intercept and a linear growth, is as follows:

$$MA0 = \left\{ \begin{pmatrix} 1 \\ t \end{pmatrix}, I_2, v_t, \begin{pmatrix} W_{0t} & 0 \\ 0 & W_{1t} \end{pmatrix} \right\}.$$  

Likewise the other models can be written in their respective DLM form and quadruple forms. Model MA1 has a level, a linear growth and one regression variable $LogDcrop$. It is defined as follows:
The Models Regressor Structure

<table>
<thead>
<tr>
<th>Models</th>
<th>MA0</th>
<th>MA1</th>
<th>MA2</th>
<th>MB3</th>
<th>MG7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Trend Components, $\delta_T = 0.9$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Linear</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Quadratic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model Regression Components, $\delta_R = 0.98$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$LogDcrop$</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$LogMaize1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$LogRain1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$LogTemp$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$LogDprice$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$LogEmpl$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* shows the parameter present. There are $n=108$ observations in the training set and 12 more in the test set.

Table 1: The “best” models as selected for use in Bayesian Model Averaging

$MA1 = \left\{ \begin{pmatrix} 1 \\ t \\ LogDcrop \end{pmatrix}, I_3, v_t, \begin{pmatrix} W_{0t} & 0 & 0 \\ 0 & W_{1t} & 0 \\ 0 & 0 & W_{2t} \end{pmatrix} \right\}$

$MA2 = \left\{ \begin{pmatrix} 1 \\ t \\ LogDcrop \\ LogMaize1 \end{pmatrix}, I_4, v_t, \begin{pmatrix} W_{0t} & 0 & 0 & 0 \\ 0 & W_{1t} & 0 & 0 \\ 0 & 0 & W_{2t} & 0 \\ 0 & 0 & 0 & W_{3t} \end{pmatrix} \right\}$

$MB3 = \left\{ \begin{pmatrix} 1 \\ t \\ LogDcrop \\ LogMaize1 \\ LogRain1 \end{pmatrix}, I_5, v_t, \begin{pmatrix} W_{0t} & 0 & 0 & 0 & 0 \\ 0 & W_{1t} & 0 & 0 & 0 \\ 0 & 0 & W_{2t} & 0 & 0 \\ 0 & 0 & 0 & W_{3t} & 0 \\ 0 & 0 & 0 & 0 & W_{4t} \end{pmatrix} \right\}$

MG7 has a level, a linear growth, a quadratic growth and all seven regression variables, $LogDcrop$, $LogMaize1$, $LogRain1$, $LogTemp$, $LogDprice$, $LogEmpl$ and CPI.
4 APPLICATION: BMA FOR THE BEER SALES DATA

4.2 Implementing the BMA Methods

4.2.1 The Akaike Method

The Akaike model average Section 3.2.1 is calculated on models $MA_0$, $MA_1$, $MA_2$, $MB_3$ and $MG_7$, (see Table 1). The density at each time 109 to 120 for each component model and the Akaike model average of the models were calculated and plots for the time points 109 and 120 are given Fig. 2. Overall the mixture model shows a peaked Student $t$-distribution type shape. The averaged density appears uni-modal but that is because the component models are not significantly different.

$$MG_7 = \begin{pmatrix} 1 \\ \frac{t}{2} \\ LogDcrop \\ LogMaize_1 \\ LogRain_1 \\ LogTemp \\ LogDprice \\ LogEmpl \\ CPI \end{pmatrix}, \quad I_{10}, \quad v_t,$$

$$\begin{pmatrix} W_{0t} & 0 & 0 & \ldots & 0 \\ 0 & W_{1t} & 0 & \ldots & 0 \\ 0 & 0 & W_{2t} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & W_{0t} \end{pmatrix}$$
The Akaike averaged model has the weights for model MA1 close to one. This results in the model MA1 being the most influential and it dominates the model average. The effect of the remaining models is negligible. The more complex models MA2, MB3 and MG7 have distributions that are more diffuse compared to the simpler models MA0 and MA1.

The plot, Fig. 2, only shows time points 109 and 120. However the densities for the model components and the mixture model densities are similar for the rest of the time points. The Akaike model has its peak lying in the range 10 to 13 on the log sales scale. The full model MG7, is very diffuse and in fact looks more like a uniform density than the regular Student t-distribution shape. This does not affect the mixture as the models associated weight is negligible.

4.2.2 The Quasi-Bayes (QB) Method

The Quasi-Bayes model average were calculated for the component model set \( \mathcal{A} = \{MA0, MA1, MA2, MB3, MG7\} \). The weights were calculated by estimating the Dirichlet distribution as described in Sec 3.2.2.

Fig. 3 shows the component models and the resulting mixture model densities at time 109 and 120, again only two time points are plotted.
to highlight the test set densities. The plot shows that at time $t = 109$ the mixture is influenced by the more complex models $MA_2$, $MB_3$ and $MG_7$ almost as much as the simpler $MA_0$ and $MA_1$ models. This influence of the less complex $MA_0$ and $MA_1$ increases over the time interval till at time $t = 120$ it is clear that the Quasi Bayes mixture density has a high peak that is close to that of $MA_0$ and $MA_1$. This shows graphically how the weights for $MA_0$ are influential as they increase from about 0.27 at time $t = 109$ to around 0.6 at time $t = 120$. At all time points the mixture appears uni-modal with a peak and Student $t$-distribution shape averaging the component densities. The range for the positions of the model peaks is roughly between 10 and 13.

### 4.2.3 The Reversible Jump Markov Chain Monte Carlo Method

At each time point reversible jump MCMC was run to simulate the models and calculate their associated posteriors. Hoeting et al (1999) used the reversible-jump MCMC simulation method to estimate the posteriors. They consider only the first 10 models with the highest number of visits. Basically, the models are ranked with the model with highest number of visits corresponding to the highest posterior probability estimate being assigned rank 1, and so on for subsequent models. This approach was used with our data. The ranking is repeated for each of the time points 109 to 120, in the test set. An alternative method to select which models to average over is to use a scree plot. Scree plots are popularly used in factor analysis to choose the number of factors to use.

The plots, Fig. 4, show the posteriors estimated by simulation against the rank of each model. The plots show that the points are steeper on the left hand side of the plot and less steep on the right. The idea to find a subset is to find a cut-off point for which models to use, (Jolliffe (2002)). This cut-off point or “elbow” in the plot is the point at which the plot changes from “steep” to “shallow” The models retained are all those at the cut-off and backwards. For our data the elbows for the different times are roughly at 5 models or less. Thus we select to average over 5 models for the times 109 to 120. This should give a set of models that reasonably make the most contribution. The inclusion of more models than the 2 models the scree plot seems to suggest at, say, time 109 in Fig. 4 will not affect or skew our results as the less significant models will be given smaller weights. Thus their net effect will also be much smaller.

Table 2 shows the 5 component models with highest posterior probabilities at each time point in the test set, 109 to 120. The component models for times 111, 112 and 113 have the same parametric structure. Times 114 and 115 share between them the same parametric models. The times 117, 119 and 120 also share models with the same paramet-
Figure 4: Scree Plots for Posterior Probabilities of the first 20 Models from MCMC method. Time Points 109 to 120
### Models Regressor Structure for models selected to be averaged by Reversible Jump MCMC Method.

<table>
<thead>
<tr>
<th>MODEL</th>
<th>6A</th>
<th>4A</th>
<th>5A</th>
<th>5B</th>
<th>7A</th>
<th>7B</th>
<th>5C</th>
<th>3A</th>
<th>4B</th>
<th>5D</th>
<th>4C</th>
<th>5E</th>
<th>3C</th>
<th>6B</th>
<th>3B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Linear</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Quadratic</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$CP1$</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Log$Dcrop$</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log$Dprice$</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Log$Empl$</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Log$Maize1$</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Log$Rain1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log$Temp$</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>109</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>110</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>111</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>112</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>113</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>114</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>115</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>116</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>117</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>118</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>119</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>120</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

* shows the parameter present. There are $n=108$ observations in the training set and 12 more in the test set.

Table 2: The “best” models as selected for use in Bayesian Model Averaging. Models selected by MCMC Reversible Jump Method.
Figure 5: Model Mixture and Component Predictive Forecast Student-t Densities with MCMC weights. Time Points 109 and 120

ric structure. The model components and the mixture models in Fig. 5 show that the mean for deseasonalised logged volumes is in the range 10 to 14. The mixtures look uni-modal with a Student $t$-distribution type shape.

4.3 Comparing The BMA Methods

The averaged densities all show a uni-modal Student $t$-distribution type shape. There is uni-modality in the averages here as the differences in the component means and variances were not large enough to introduce multi-modality. The plot Fig. 6 shows the averaged predictive densities superimposed on each other.

The forecasting performances of the methods (Akaike, MCMC and Quasi Bayes) are compared with the component models by analysing the forecasted values compared with the actual observations, the percentage root Mean Square Error ($prmse$) and the Geometric Mean Relative Absolute Error ($gmrae$). The $prmse$ measure can be sensitive to large errors. However $prmse$ is relatively simple to calculate, understand and use. The Geometric Mean Relative Absolute Error ($gmrae$) is the mean absolute error of the current model versus the absolute error of a naive model.
Figure 6: Model Mixture and Component Predictive Forecast Student-t Densities with the Akaike, MCMC and Quasi Bayes weights. Time Points 109 and 120
Figure 7: Actual Values and One Step Ahead Forecast Values from the Component Models MA0, MA1, MA2 and MB3. Fig(b) shows the One Step ahead Forecasts from the Model Averages with Akaike, MCMC and Quasi Bayes weights.
\[ g_{mrae} = \frac{\prod_{t=1}^{n} |y_t - \hat{y}_t|}{\prod_{t=1}^{n} |y_t - y_{t-1}|}^{\frac{1}{n}}, \]  

where \( n \) is the sample size, \( y_{t-1} \) the observed values at time \( t - 1 \), \( y_t \) the observed values at time \( t \) and \( \hat{y}_t \) is the fitted or forecast value all this is assuming that model \( M_j \) is the model being compared with the naive model.

For test set, \( \{109, \ldots, 120\} \), we plot the one step ahead forecasts in Fig. 7, the \textit{prmse} in Fig. 8 and the \textit{gmrae} in Fig. 9. For the component models the one step ahead forecast values, Fig. 7, were plotted excluding those of the full model which were clearly very varied in range. The simple model \( MA0 \) is just a linearly descending line whilst the other models were closer to the actual values. The one step ahead forecasts from the Akaike and MCMC averaged models are closer to the actual observed values with an average deviation of 0.12 and 0.13, respectively. The values from the Quasi Bayes method are more dispersed and have an average absolute deviation of 0.87, making them 7 times further from the actual values compared with the Akaike one step ahead forecast values. However, the one step ahead forecast values from the Quasi Bayes average get closer to the actual values towards the end of the test series, suggesting that the method improves with the passage of time. The large variations in the one step ahead forecast values from the Quasi Bayes average result from the method tending to initially assign the full model \( MG7 \) significant weight.

The plot Fig. 8, shows the \textit{prmse} values for the model components and the model averages (Akaike, MCMC and Quasi Bayes). The component model plot shows that the models \( MB3 \) and \( MA2 \) are very close, and in fact the \( MB3 \) is only 0.08\% better than the \( MA2 \) model. The Akaike model is consistently better than the MCMC and Quasi Bayes averages. The Akaike model performs within 1\% of that of model \( MB3 \) which had the smallest \textit{prmse} for the component models. The Akaike model performed on average 4\% better than the MCMC model average. The MCMC model initially does not perform as well as the Akaike but with the passage of time its \textit{prmse} values improve. The Akaike has a \textit{prmse} that is 73\% smaller than that for the Quasi Bayes model on average.

The plot of Fig. 9, shows the component and model average \textit{gmrae} plots. The \textit{gmrae} measures the performance compared to the naive model. The component models show that the model that performs closest to the naive model is model \( MB3 \), within 3\%. The \textit{gmrae} for one step ahead forecasts from the MCMC model average are on average 63\% higher than the naive model whilst those for the Akaike and Quasi Bayes model averages are 115\% and 1969\% respectively. Thus the MCMC model average performs better than the Akaike or Quasi
Figure 8: The \textit{prmse} from Akaike, BIC and MCMC models.

Figure 9: The \textit{gmrae} from Akaike, BIC and MCMC models.
Bayes average on the \textit{gmrae} criteria.

The component models in the methods used earlier all show that the mixture or model average achieved is unimodal. This should not mislead and give the impression that unimodality is guaranteed in all cases. In fact, it is easily shown that multi-modality could occur in our data by simply tweaking the one of the means of the component densities. The mixture models as shown by Fig. 10 has all the parameters in all the component models unchanged except for an increase in the location parameter for model $\text{MA0}$ from \( \mu = 11.68, \tau = f(\hat{\sigma} = 0.0234) \) to \( \mu = 16, \tau = f(\hat{\sigma} = 0.0234) \). Immediately the mixture model becomes bi-modal showing the dependence of the number of models in the mixture on the location parameter. Conditions for bi-modality for a pair of normal distributions are illustrated in Robertson and Fryer (1969).

\section{Conclusions}

We have compared the performance of different methods for model averaging using different methods to calculate the associated weights.
The posterior probability based weights were calculated by approximation methods, namely the Akaike and the Quasi Bayes method and by simulation to estimate exact values of the posteriors through MCMC methods.

The results showed that the Akaike method was the best performing method, compared to the MCMC and Quasi Bayes methods, under the \textit{prmse} performance measure. However it should be noted that because the minimum AIC values for $MA1$ were very far from the rest of the models, the weights tended to heavily favour model $MA1$. Thus the Akaike model has the weights for $MA1$ very close to one. Under the \textit{gmrae} performance measure the MCMC method performed better than the Akaike and Quasi Bayes methods.

The Akaike and MCMC model averaging methods did much better than the Quasi Bayes method. The two methods seemed to do consistently well though they were out-performed under both the \textit{prmse} and \textit{gmrae} measures by model $MB3$. This does not mean the averaging methods are bad because they bring with them all the advantages of model averaging including removing the need to make a choice for an individual model, incorporating the uncertainty associated with individual models and robustness associated with model averaging. There is also scope to increase the sample size and in particular the test set size.

We have shown that in our case that models averaging using MCMC and Akaike methods performed better than the Quasi Bayes method. This is a way to compare performance of various statistical model averaging methods via the different ways of calculating the associated weights.
References


REFERENCES


