The Geometric Combination of Bayesian Forecasting Models

A. E. Faria and E. Mubwandarikwa

Department of Statistics,

The Open University

Walton Hall,

Milton Keynes,

MK7 6AA, UK

November 1, 2007

Abstract

A non-linear geometric combination of statistical models is proposed as an alternative approach to the usual linear combination or mixture. Contrary to the linear, the geometric model is closed under the regular exponential family of distributions as we show. As a consequence, the distribution which results from the combination is uni-modal and a single location parameter can be chosen for decision making. In the case of Student t-distributions (of particular interest in forecasting) the geometric combination can be uni-modal under a sufficient condition we have established.

A comparative analysis between the geometric and the linear combinations of predictive distributions from three Bayesian regression dynamic linear models, in a case of beer sales forecasting in Zimbabwe, shows the geometric model to consistently outperform its linear counterpart as well as its component models.

Keywords*: Non-linear model combination, regression dynamic linear models, multi-modality.*

1 Introduction.

In this paper we consider the situation where an analyst entertains $k \geq 2$ statistical forecasting models $(\mathcal{M}_1, \ldots, \mathcal{M}_k)$ *she* believes are good plausible models for a time series process Y_t $(t = 1, 2, \ldots)$, but is interested in determining a single model she can use for forecasting. There are basically two courses of action that the analyst could take. She could either:

(i) choose a single model from the set $\underline{\mathcal{M}} = \{\mathcal{M}_1, \ldots \mathcal{M}_k\}$ based on some model selection criteria; or

(ii) combine the models together according to some model combination approach.

Within the action course (i), model selection, there are a number of different approaches that could be adopted. Some are based on comparing the models predictive performances and choosing the best according to a defined performance measure. Others, like the Bayes factor methods for example, are based on calculating the models predictive likelihood ratios and selecting the one with the largest likelihood. Also, in case of regression models, the analyst could choose a model associated with the belief net she thinks is the most appropriate for the forecasting period in consideration. In any case, a single model from the set of plausible models is selected and used for forecasting.

Under option (ii) above, model combination, there is a main class of available methods, all linear, that are usually referred to as *mixture models* (see e.g. Titterington et.al., 1985). In the Bayesian context, Raftery et. al. (1997) refer to the combination of predictive distributions weighted by the component models probabilities as *Bayesian model averaging* (BMA). West and Harrison (1997) considered classes of *dynamic linear models* (DLMs) with their mixtures referred to as *multi-process* models.

Perhaps the main advantage of adopting a model combination over a model selection approach is that the uncertainty about the models can be explicitly accounted for in the analysis. Conditioning on a single selected model ignores model uncertainty, and thus leads to the underestimation of uncertainty when making inferences about quantities of interest. Please see for example Raftery et. al. (1997) for more details. Also, the combination of models can be interpreted as a form of aggregating information from different sources (models) with obvious advantages.

However, *linear combinations* also have their disadvantages. The main drawback in many cases being the difficulty of obtaining the exact resulting combination. this may be remedied (as in the BMA) by adopting approximate solutions like Markov Chain Monte Carlo (MCMC) simulation. Further to that, even when the resulting combination can be determined, its distribution will typically present multi-modality which may be undesirable in decision making situations when a single estimate of the location parameter of the predictive distribution is required.

Here, we propose an alternative approach to the mixture, the *geometric combination*, which resolves some of the main disadvantages of the linear methods. The main reasons we propose the geometric combination of probability distribution models are two-fold. First, geometric combinations preserve the distribution form of the combined models for many distributions. In particular we show that the geometric combination of exponential family distributions is also a member of the exponential family. This result means that the geometric combination of exponential family densities is uni-modal. This is not the case with linear combinations where, in general, uni-modality occurs only under certain conditions, see e.g. Titterington et. al. (1985). Further, whilst linear combinations of Student's t-distributions (which play an important role in Bayesian forecasting) are typically multi-modal under linear combination, it is not always so under the geometric rule as we also shall see.

The second reason, is that geometric combinations are externally Bayesian thus possessing the advantages of such types of combinations (e.g. immunity of influence on decision making). External Bayesianity, Madansky (1964), ensures that the combination rule will give the same result *a posteriori*, independently of being obtained before or after the individual distributions are updated by new data. In the case where the individual distributions are subjective probability distributions provided by experts, Raiffa (1968) illustrated how the relevance over the order in which the combination and updating are done can lead to subjects trying

to increase their influence on the "consensus" or resulting combination of distributions by insisting that their opinions be computed before the outcome of an experiment is known. External Bayesianity makes such argument pointless.

To exemplify some of the results mentioned above, we apply the geometric and the linear combination approaches to the predictive distributions of three models formulated for a quarterly time series of beer sales by a brewery in Zimbabwe. In particular, we will consider a class of Bayesian time series forecasting models, the *regression dynamic linear models* proposed by West and Harrison (1997). In that application, each formulated model is individually characterised by its individual set of (economic and environmental) regressor variables which represents a historical scenario. The weight of each model in the combination is then interpreted as a subjective uncertainty measure of the represented scenario being the one occurring during the forecasting horizon. In practice, an analyst may also include in the combination a model for a scenario that did not happen in the past but for which she is prepared to formulate a forecasting model.

The remainder of this paper is structured as follows. In Section 2 we introduce both the linear and the geometric combinations of probability distributions function in general, and show the conditions for uni-modality for models from the exponential and the Student t families. Section 3 reviews the regression dynamic linear models that are used in the application section. Section 4 compares the predictive performances of the two combination methods applied to three regression dynamic linear models formulated for a quarterly series of beverage sales in Zimbabwe. Section 5 concludes the paper.

2 Model combination approaches

Let (Ω, μ) be a measure space. Also, let Δ be the class of all μ-measurable functions $p : \Omega \to [0, \infty)$ such that $\int p d\mu = 1$ with μ almost everywhere (a.e.). A (generic) *combination* function $P : \Delta^k \to \Delta$, is one which maps a vector of probability density functions $(p_1, ..., p_k)$, where $p_j \in \Delta$ is a density associated with a statistical model $\mathcal{M}_j$ (for $j = 1, ..., k$), into a single density $p(\cdot)$ also in Δ.

In the following subsections we define both the linear and the geometric combinations for univarite distributions. We also show the uni-modality conditions for models within the exponential family as well as for models with (Student) t-distributions.

2.1 The linear combination

The linear combination $P_L : \Delta^k \to \Delta$ of k densities for a process $Y \in \Omega$, has the following general form:

$$P_L(p_1, ..., p_k)(Y) = \sum_{j=1}^{k} w_j p_j(Y) \tag{1}$$

where $p_j(Y)$ is the probability density function associated with model $\mathcal{M}_j$ and w_j is an arbitrary weight (in general, not necessarily nonnegative) given to $\mathcal{M}_j$ in the combination $(j = 1, ..., k)$ such that $\sum_{j=1}^{k} w_j = 1$. As $P_L(Y)$ is a probability density function itself, it means $P_L \geq 0$ for all $Y \in \Omega$ and care must be taken when choosing negative weights.

The weights could be elicited by the analyst based on her knowledge about the relative predictive capabilities of the individual models for the period of interest. There are a number of methods -including Bayesian, such as Bunn's (1975, 1978) *outperformance*- available for determining the weights based on the models past predictive performances. In the Bayesian model averaging framework, the weight w_j is treated as the posterior probability for model $\mathcal{M}_j$, that is $w_j = p(\mathcal{M}_j|Y)$ obtained by Bayes theorem via MCMC. Please see Raftery et. al. (1997) for further details.

In the application section of this paper, we have considered a time series process Y_t $(t = 1, 2, \ldots)$ for which predictive densities $p_j(Y_t|D_t)$, where $D_t = \{y_t, D_{t-1}\}$, are formulated for each model $\mathcal{M}_{jt}$ $(j = 1, \ldots, k)$ and adopted a probabilistic interpretation for the weights, with w_{jt} being the probability that the economic and environmental scenario influencing the process Y_t is that represented by model $\mathcal{M}_{jt}$ as we shall see.

2.1.1 Uni-modality in linear combinations of Gaussian models

It is well known that the linear combinations of Gaussian densities are uni-modal only under rather restrictive conditions, see e.g. Titterington et. al. (1985). For example, for the linear combination of two Gaussian densities $p_j(y|\mu_j, \sigma_j^2)$, with means μ_j and variances σ_j^2, $(j = 1, 2)$, Eisenberger (1964) showed that independently of the combining weights, a sufficient condition for uni-modality of the combined density is that

$$(\mu_2 - \mu_1)^2 < \frac{27}{4} \frac{\sigma_1^2 \sigma_2^2}{(\sigma_1^2 + \sigma_2^2)}. \tag{2}$$

This condition means that to obtain uni-modality, the distance between the location parameters of the components' densities must be small enough relative to a ratio of their spread parameters relative to the total spread. Otherwise, the combined density will present bi-modality.

This result when extended for more than two densities yields a similar interpretation, that is, uni-modality of the combined density will result when the distances between the components' means are small enough relative to a certain ratio of their standard deviations. Otherwise, the combined density will have between two and k modes where $k > 2$ is the number of components.

In the case where the combining densities in (1) are t-distributions, Faria and Mubwandarikwa (2006) have shown that the resulting combination can also be uni-modal under conditions which are more restrictive than (2) above. In fact, uni-modality in this case would require the distance between the means of the two t-distributed components (relative to their spreads) to be smaller than that required in the Gaussian case.

At the end of the following section we show an example where the linear combination of three t-distributed models results in a bi-modal density (see Figure 1).

2.2 The geometric combination

The geometric combination $P_G : \Delta^k \to \Delta$ of k predictive densities for a time series $Y \in \Omega$ has the following general form:

$$P_G(p_1, ..., p_k)(Y) = c \prod_{j=1}^{k} p_j^{w_j}(Y) \ , \tag{3}$$

where $c^{-1} = \int \prod_{j=1}^{k} p_j^{w_j}(Y)dY$, $w_j \in \mathbb{R}$, $j = 1, ..., k$, is the weight associated with the density $p_j(Y)$ associated with model $\mathcal{M}_j$) such that $\sum_{j=1}^{k} w_j = 1$.

As mentioned in the introduction, one of the advantages of adopting the geometric over the linear combination method is that geometric combinations are externally Bayesian. The advantage proved by Raiffa (1968), being that of immunity of influence over decision making in cases where component distributions reflect subjective opinions from experts (or interested parties).

In general terms, an externally Bayesian (EB) combination function P is characterised as one satisfying the following condition:

$$P\left(\frac{p_1}{\int \mathcal{L}p_1 d\mu}, \ldots, \frac{p_k}{\int \mathcal{L}p_k d\mu}\right) = \frac{\mathcal{L}P(p_1, \ldots, p_k)}{\int \mathcal{L}P(p_1, \ldots, p_k)d\mu} \ , \quad \mu \ a.e. \ , \tag{4}$$

where $\mathcal{L} : \Omega \to (0, \infty)$ is a likelihood function for the actually observed data, such that $0 < \int \mathcal{L}p_j d\mu < \infty$ $(j = 1, \ldots, k)$.

Briefly, an EB combination policy ensures that the combination rule will give the same result a posteriori, independently of being obtained before or after each individual combining density is updated when new data is observed.

It can be easily seen that the geometric combination $P_G \in P$ in (3) satisfies (4) and therefore is EB. This is not the case for P_L in (1). The reader can refer to Faria (1996) or Genest et. al. (1986) for more details.

2.2.1 The geometric combination of exponential family densities

In this section we prove that the geometric combination of densities from the regular exponential family of distributions has a density which is strongly uni-modal. A probability measure is said to be strongly uni-modal if it is log-concave (i.e. its logarithm is a concave function) over its parameter space. In fact, the strong uni-modality of the geometric combination comes from the fact, shown below, that the geometric combination of strongly uni-modal densities from the regular exponential family also belongs to that family.

Recall that a density $p(y|\underline{\eta})$ with $\underline{\eta} = (\eta_1, \ldots, \eta_n) \in \Omega$ belonging to the n-parameter exponential family has the natural representation

$$p(y|\underline{\eta}) = h(y)c(\underline{\eta}) \exp[\underline{\eta}'\underline{d}(y)]$$

where $h(y) \geq 0$ does not depend on $\underline{\eta}$ and $\underline{d}(y) = (d_1(y), \ldots, d_n(y))$ with $d_i(y) : \Omega \to \mathbb{R}$ not depending on $\underline{\eta}$. The natural parameter space Ω is the set where the kernel function has a finite integral (or sum):

$$\Omega = \{(\eta_1, \ldots, \eta_n) : \frac{1}{c(\underline{\eta})} = \int_{-\infty}^{+\infty} h(y) \exp[\underline{\eta}'\underline{d}(y)]dy < \infty\} \ .$$

The exponential family is said to be regular if (i) the elements of $\underline{\eta}$ and those of $\underline{d}$ are linearly independent, and (ii) Ω is a n-dimensional open set. Also, recall that the elements of $\underline{d}(y)$ are linearly independent if $\sum_{i=1}^{n} a_i d_i(y) = b$ for all y if and only if $a_1 = \ldots = a_n = 0$ where a_i and b are constants.

Now, assume that a random variable y whose density under a model $\mathcal{M}$, $p(y|\underline{\eta})$ belongs to the n-parameter strongly uni-modal regular exponential family. In this case, $p(y|\underline{\eta})$ raised

to any constant power $w \in (0,1)$ (not a function of y or $\underline{\eta}$) is a density which also belongs to the n-parameter strongly uni-modal regular exponential family.

In fact, for a fixed $w \in \mathbb{R}$ we can write:

$$\begin{aligned} p^w(y|\underline{\eta}) &= [h(y)]^w [c(\underline{\eta})]^w \exp[w\underline{\eta}'\underline{d}(y)] \\ &= h^*(y)c^*(y)\exp[\underline{\eta}'\underline{d}(y)] \end{aligned}$$

where $h^*(y) = h^w(y)\exp(w)$ and $c^*(\underline{\eta}) = c^w(\underline{\eta})$. Therefore p^w is from the exponential family.

Note that because $\ln p(y|\underline{\eta})$ is concave, $\ln p^w(y|\underline{\eta}) = w[\ln h(y) + \ln c(\underline{\eta})] + \underline{\eta}\underline{d}(y)$ is also concave, and thus p^w is strongly uni-modal.

Also note that the product of densities from the strongly uni-modal regular exponential family also belongs to the same family. In fact, for $j = 1, \ldots, k$, let $p_j(y|\underline{\eta}_j)$ belong to the n_j-parameter strongly uni-modal regular exponential family as above. Thus, given $uw = \{w_1, \ldots, w_k\}$ with $w_j \in (0,1) : \sum_{j=1}^k w_j = 1$, we can write:

$$\begin{aligned} P_G(y|\underline{\tilde{\eta}}) &= a(\underline{\tilde{\eta}}) \prod_{j=1}^k [h_j(y)]^{w_j} [c_j(\underline{\eta})]^{w_j} \exp[w_j \underline{\eta}_j' \underline{d}_j(y)] \\ &= \tilde{h}(y)\tilde{c}(\underline{\eta}) \exp[\sum_{j=1}^k \underline{\eta}_j' \underline{d}_j(y)] \ , \end{aligned}$$

where $\underline{\tilde{\eta}}$ is the parameter set of $P(y|\underline{\tilde{\eta}})$, $a^{-1}(\underline{\tilde{\eta}}) = \int \prod_{j=1}^k [h_j(y)]^{w_j} [c_j(\underline{\eta})]^{w_j} \exp[w_j \underline{\eta}_j' \underline{d}_j(y)] dy$, $\tilde{h}(y) = \prod_{j=1}^k [h_j(y)]^{w_j}$ and $\tilde{c}(\underline{\tilde{\eta}}) = a(\underline{\tilde{\eta}}) \prod_{j=1}^k [c_j(\underline{\eta}_j)]^{w_j}$.

Therefore, $P_G(y|\underline{\tilde{\eta}})$ is a nk-parameter density from the regular exponential family, where $n = \sum_{j=1}^k n_j$. Notice that

$$\ln P_G(y|\underline{\tilde{\eta}}) = \ln a(\underline{\tilde{\eta}}) + \sum_{j=1}^k w_j \ln h_j(y) + \sum_{j=1}^k w_j \ln c_j(\underline{\eta}_j)) + \sum_{j=1}^k \underline{\eta}_j' \underline{d}_j(y)$$

is concave and thus P_G is uni-modal.

We have just shown that when $p_1(Y), \ldots, p_k(Y)$ are strongly uni-modal densities from the regular exponential family, the density of the geometric combination $P_G(p_1, \ldots, p_k)(Y)$ in (3) is also a strongly uni-modal density from the regular exponential family.

This result is rather interesting from a decision analysis viewpoint. It may in many cases give a decision maker a reason to adopt the geometric rather than the linear combination of models.

In the particular case where $p_j(y|\mu_j, \tau_j)$ is a normal density with mean μ_j and precision $\tau_j = \sigma_j^{-2}$ $(j = 1, \ldots, k)$, we have that $P_G(y)$ is also Gaussian with mean $\tilde{\mu} = \frac{\sum_{j=1}^k \tau_i \mu_i}{\sum_{j=1}^K \tau_i}$ and precision $\tilde{\tau} = \sum_{j=1}^k w_i \tau_i$.

2.2.2 Uni-modality in geometric combinations of t-densities

In the case where the component densities have t-distributions, Faria and Mubwandarikwa (2006) showed the following uni-modality condition for the geometric combination for $k = 2$.

For Y under $\mathcal{M}_j$ having a Student t-density with n degrees of freedom, mean μ_j and variance $n/(n-2)\sigma_j^2$, that is $[Y|\mathcal{M}_j] \sim St_n(\mu_j, \sigma_j^2)$ $(j = 1, 2)$, P_G is uni-modal when

$$(\mu_2 - \mu_1)^2 < \frac{64}{27} \frac{(S^2 - \bar{T}^* \nu^*)^3}{\nu^{*2}(T_2^* - T_1^*)^2} \ . \tag{5}$$

where $\bar{\mu} = \frac{1}{2}(\mu_1 + \mu_2)$, $\bar{T}^* = \frac{1}{2}(T_1^* + T_2^*)$, $T_j^* = \frac{w}{(w\nu^2 + w\nu - \nu)} \tau_j^{-1}$, $\nu^* = w(\nu + 1) - 1$ and $S^2 = \frac{1}{2} \sum_{i=1}^{2} (\mu_i - \bar{\mu})^2$.

The geometric combination of regular exponential family densities is always uni-modal, the density of the geometric combination of two t-densities can go from having a single mode to having two modes as the distance between the means μ_1 and μ_2 increases.

As an example, the plot Figure 1 shows the densities of the linear and the geometric combinations (solid lines) of three t-distributed component models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ (dashed lines) for a variable y with 25 degrees of freedom (dof) and means and standard deviations of $(48.04, 2.881)$, $(47.49, 3.258)$ and $(39.37, 3.063)$ respectively. The combination weights were $w_1 = 0.3$, $w_2 = 0.1$ and $w_3 = 0.6$. The resulting bi-modal linear combination density has modes at $y = 39.50$ (the largest) and at $y = 47.62$. The anti-mode occurred at $y = 44.43$ (over its mean at $y = 42.78$). Note that μ_1 and μ_2 are relatively close to each other and both are relatively far from μ_3 thus the resulting bi-modality. Were μ_1 and μ_2 further apart (and away from μ_3) the linear combination could have three modes instead. On the other hand, were mu_1 and μ_2 closer to μ_3 the resulting linear combination could have a skewed uni-modal density. The geometric combination in its turn has a symmetric uni-modal density with mode at $y = 42.79$.

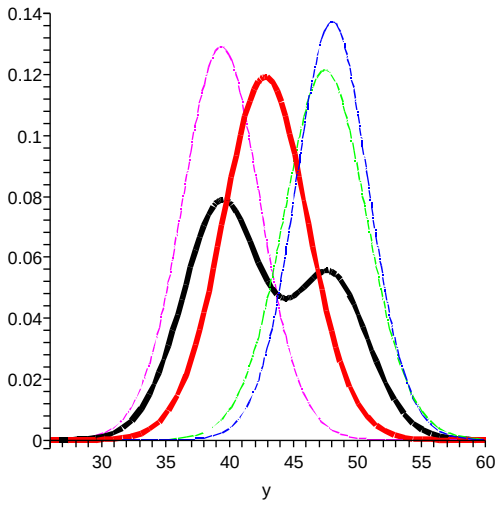

Figure 1: The linear and the geometric combinations (solid lines) of three t-distributed models $\mathcal{M}_1\mathcal{M}_2$ and $\mathcal{M}_3$ (dashed lines) with 25 dof and means and standard deviations of $(48.04, 2.881)$, $(47.49, 3.258)$ and $(39.37, 3.063)$ respectively.

The above example was obtained from the predictive densities from three regression dynamic linear models formulated and applied to the problem of beer sales forecasting in Zimbabwe. Before proceeding to the application section, we introduce in the following section

a short description of the adopted regression dynamic linear models. The reader familiar with the class of Bayesian forecasting dynamic linear models proposed by West and Harrison (1997) can proceed directly to Section 4.

3 Regression Dynamic Linear Models

At a fixed period of time t, a *regression dynamic linear model* (RDLM), $\mathcal{M}_j$, $(j = 1, 2, \ldots, k)$, for a time series, Y_t, is characterised by the quadruple $\{\underline{F}, \mathbf{G}, V, \mathbf{W}\}_j$, where $\underline{F}_j = (X_1, \ldots, X_r)'_j$ is the $(r_j \times 1)$ regression vector, X_{ij} being the i^{th} regression variable $(i = 1, 2, \ldots, r_j)$, $\mathbf{G}_j$ is the $(r_j \times r_j)$ system evolution matrix, V_j is the observational variance (i.e. the variance of Y_t), and $\mathbf{W}_j$ is the $(r_j \times r_j)$ state evolution covariance matrix. For simplicity and without loss we have omitted the subscript t in defining $\mathcal{M}_j$. Also, we have used underlined and boldfaced characters to represent vectors and matrices respectively.

Note that the regression vector $\underline{F}_j$, the evolution matrix $\mathbf{G}_j$ and the evolution covariance matrix $\mathbf{W}_j$ are defined by the analyst during the model specification stage of the modelling process. The evolution matrix $\mathbf{W}_j$ can, for instance, be chosen with the use of discount factors, $\delta_j \in (0, 1)$, interpreted as measures of how quickly the value of the current information set $D_t = \{y_t, D_{t-1}\}$ is expected to decay in time. The observational variance V_j is usually unknown (and large relative to the evolution variance $\mathbf{W}_j$). Also, it is usually the major source of uncertainty but appropriate Bayesian learning procedures can be used in its specification and estimation.

A RDLM $\mathcal{M}_{jt}$ can be represented by an observation and an evolution equation:

$$\begin{aligned} \text{Observation:} \quad & Y_t = \underline{F}'_{jt}\underline{\theta}_{jt} \quad + \ \nu_{jt} \ ; \ \nu_{jt} \sim p(\nu_{jt}) \\ \text{Evolution:} \quad & \underline{\theta}_{jt} = \mathbf{G}_{jt}\underline{\theta}_{j,t-1} + \ \underline{\omega}_{jt} \ ; \ \underline{\omega}_{jt} \sim p(\underline{\omega}_{jt}), \end{aligned}$$

where $p(\nu_{jt})$ is a density (or mass) function with zero mean and variance V_{jt} for the observational error ν_{jt}, and, $p(\underline{\omega}_{jt})$ is a joint density with zero mean and covariance matrix $\mathbf{W}_{jt}$ for the evolution error vector $\underline{\omega}_{jt}$.

In its simple form, a RDLM $\mathcal{M}_{jt}$ will assume specific known probability density functions for the observational and the evolution errors such that a sequential prior-to-posterior distribution updating can be performed in analytical closed form such that the posterior preserves the same distributional form of the prior distribution. This is called conjugate analysis. Perhaps the best known case (and the one we have adopted here) is the Gaussian RDLM where the observational error ν_{jt} is assumed to follow a normal density, i.e. $\nu_{jt} \sim N[0, V_{jt}]$. Unknown model parameters such as in some cases, the observational variance V_{jt}, can be dealt with in a Bayesian framework by assuming they follow a density function which is also sequentially updated within the Bayesian paradigm. In this case, the evolution error $\underline{\omega}_{jt}$ is assumed to follow a multivariate (Student) t-density with $n_{j,t-1}$ degrees of freedom, zero mean vector and covariance matrix $\mathbf{W}_{jt}$, i.e. $\underline{\omega}_{jt} \sim St_{n_{j,t-1}}[\underline{0}, \mathbf{W}_{jt}]$.

In cases where it is not possible to adopt a conjugate analysis, numerical integration methods can be employed to determine the posterior parametric density in the sequential updating described above. One of the most popular methods is the Markov chain Monte Carlo (MCMC).

Now, in the Gaussian case, when all the elements of $\mathcal{M}_{jt}$ are known, the forecast ahead

to time $t+h$ from time t, $p_j(Y_{t+h}|D_t)$, $(h \geq 1)$ is also Gaussian with mean

$$f_{jt}(h) = \underline{F}'_{j,t+h}\underline{a}_{jt}(h)$$

and variance

$$Q_{jt}(h) = \underline{F}'_{j,t+h}\mathbf{R}_{jt}(h)\underline{F}_{j,t+h} + V_{j,t+h}$$

can be calculated recursively for $h \geq 1$ using

$$\underline{a}_{jt}(h) = \mathbf{G}_{j,t+h}\underline{a}_{jt}(h-1)$$

and

$$\mathbf{R}_{jt}(h) = \mathbf{G}_{j,t+h}\mathbf{R}_{jt}(h-1)\mathbf{G}'_{j,t+h} + \mathbf{W}_{j,t+h} \ ,$$

with the initial values $\underline{a}_{jt}(0) = \underline{m}_t$ and $\mathbf{R}_{jt}(0) = \mathbf{C}_{jt}$. The vector $\underline{m}_{jt}$ and the matrix $\mathbf{C}_{jt}$ are the posterior location and spread parameters respectively of the state vector $\underline{\theta}_{jt}$. Please see West and Harrison (1997) for further details.

A case of particular interest occurs in situations where some of the elements of $\mathcal{M}_{jt}$ are unknown (e.g. V_{jt}) or they are known but the sample sizes are small. In such cases, the h-steps-ahead predictive density $p_j(Y_{t+h}|D_t)$ will typically be the density of a t-distribution. This density will have $n_{jt} = n_{j,t-1} + 1$ degrees of freedom, mean $f_{jt}(h)$ and variance $Q_{j,t}(h)$, that is $(Y_{t+h}|D_t) \sim St_{n_{jt}}(f_{jt}(h), Q_{jt}(h))$. The mean and variance are obtained as for the Gaussian case above but with the sample variance S_{jt} used as estimator of the unknown V_{jt}. The posterior for $\underline{\theta}_{jt}$ is $(\underline{\theta}_{jt}|D_t) \sim St_{n_{jt}}(\underline{m}_{jt}, \mathbf{C}_{jt})$ with n_{jt}, $\underline{m}_{jt}$ and $\mathbf{C}_{jt}$ determined recursively by the Kalman filter.

Note that from a pragmatic point of view, it makes sense in combining only RDLMs that represent distinct characteristics of the underlying process. Therefore, we will be interested in RDLMs which are associated with distinct causal structures of association between the underlying variables. In this paper, two RDLMs $\mathcal{M}_i$ and $\mathcal{M}_j$ $(i \neq j)$ are considered to differ from one another if they are *non-similar*, i.e. their forecast functions f_{jt} have different algebraic form. This is certainly the case when they have different sets of regressors.

4 Forecasting Beverage Sales in Zimbabwe

In this section, the predictive performances of the geometric and the linear combination methods are compared when applied to three plausible Bayesian RDLMs formulated for a quarterly series of beer sales from a Zimbabwean brewery. Each RDLM was formulated to represent a distinct economical-environmental scenario which was believed to have occurred in Zimbabwe and strongly influenced sales during three different periods of time. Therefore, each model is characterised by its particular set of regressors as we shall see.

The formulated models were combined and tested for "fitting" during a selected within-sample period and for forecasting in an out-of-sample period (the last 4 quarters of data in the time series). We have, in the role of the analyst, arbitrarily chosen the combining weights which were our subjective probabilities that their associated models represent the scenario believed to prevail during the fitting and the forecasting periods. The specific values will be described later.

The underlying series comprises of 40 quarterly seasonally adjusted observations of total beer sales volume (Y_t), in hectoliters $(\times 10^4)$, from the second quarter of 1991 (Q2 1991) to

the first quarter of 2001 (Q1 2001). The series was seasonally adjusted not only because we were interested in focusing on the trend component but also the original data presented a very strong and predictable seasonal behaviour. In fact, all seasonal series in this application were also seasonally adjusted for similar reasons. The seasonally adjusted beer sales series displayed in Figure 2(a) shows three distinct trend patterns during three different periods of time. During the first period, which we call period *A* (from Q2 1991 to Q4 1994), there was a linear but steep decline in sales. This was followed by a period B (from Q1 1995 to Q4 1997) of a positive linear trend, and a period C (from Q1 1998 to Q1 2000) of a negative linear trend. Period D (from Q2 2000 to Q1 2001) was chosen as the out-of-sample period. The period from Q2 1991 to Q1 2000 (i.e periods A, B and C together) is the within-sample period.

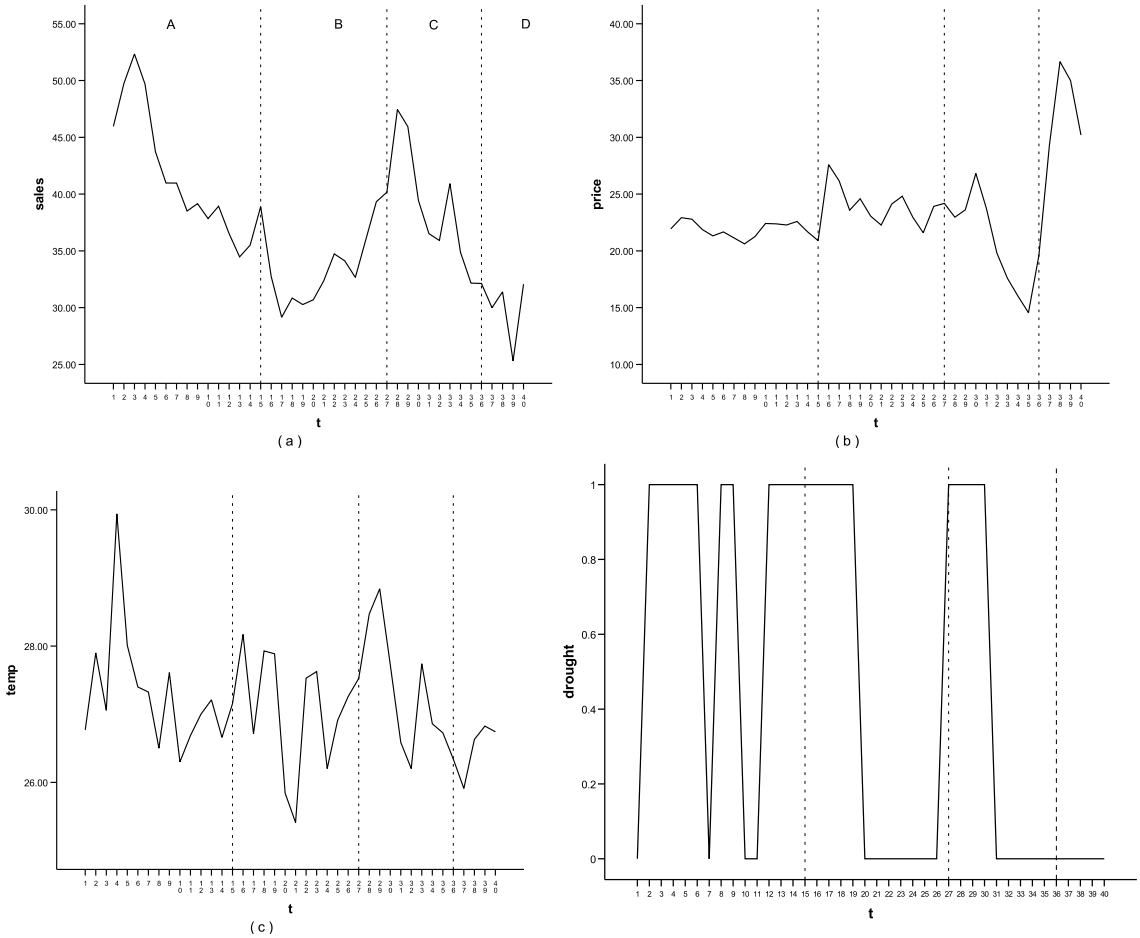

Figure 2: Seasonally adjusted series of (a) total beer sales, (b) average deflated beer price per unit and (c) average maximum temperature. Dummy variable (d) drought indicator.

Figure 2 (b), (c) and (d) shows some of the data used as regressors, they are the seasonally adjusted series of average deflated beer price per unit(in Zimbabwe dollars), the seasonally

adjusted average maximum temperature (in degrees Celsius) and drought indicator respectively.

4.1 The formulated RDLMs

A plausible RDLM were determined from the believed causal structure for each one of the periods A, B and C above. In fact, there were a number of economic and weather variables which had marked influence on sales at those periods. The average beer price (X_{1t}) as well as the average maximum temperature (X_{2t}) were explanatory variables which were believed to strongly influence the total beer sales (Y_t) at all times. However, each period A, B and C had other distinct explanatory factors believed strongly influential on sales only at those periods. Each formulated model is believed to best represent the economic and environmental scenario at the corresponding period of time it was obtained.

During period A, the decline in beer sales was heavily influenced not only by governmental policy of general price de-regulation (following the introduction in Zimbabwe of an economic structural adjustment programme, formulated by the World Bank and the International Monetary Fund, from 1990 to 1995 – see e.g. World Bank, 1996) which had a strong effect on increasing prices, but also by the drought which occurred between 1991 and 1992 and was characterised by a combination of relatively lower rainfall and higher temperature. The drought indicator series in Figure 2 (d) was obtained from the time series of temperature and rainfall simultaneously. It was set to 1 when the standardised values of both rainfall and temperature were below and above their sample averages respectively; and to zero otherwise. That series is consistent with well known drought periods in Zimbabwe.

The plausible model $\mathcal{M}_{1t}$ formulated for period A's scenario included X_{1t}, X_{2t} as well as the drought indicator (X_{3t}) as explanatory variables. The regression vector was set as $\underline{F}'_{1t} = (1, t, X_{1t}, X_{2t}, X_{3t})$ with the two initial terms (1 and t) used to model level and trend respectively. The evolution matrix $\mathbf{G}_{1,t}$ was set as an (5×5) identity matrix.

In period B, beer sales was thought to have been influenced by beer prices, which continued to rise steadily in nominal (but not in deflated) terms, compounded by the long lasting drought (started in period A) with a severe impact on agricultural seasons leading to higher temperatures, lower rainfall and hence low crop sales (X_{5t}) following poor harvests. Model $\mathcal{M}_{2t}$ formulated for period B included X_{1t}, X_{2t}, X_{3t} and total crop sales (X_{4t}). The regression vector was thus $\underline{F}'_{2t} = (1, t, X_{1t}, X_{2t}, X_{3t}, X_{5t})$ with the evolution matrix $\mathbf{G}_{2,t}$ set as an (6×6) identity matrix.

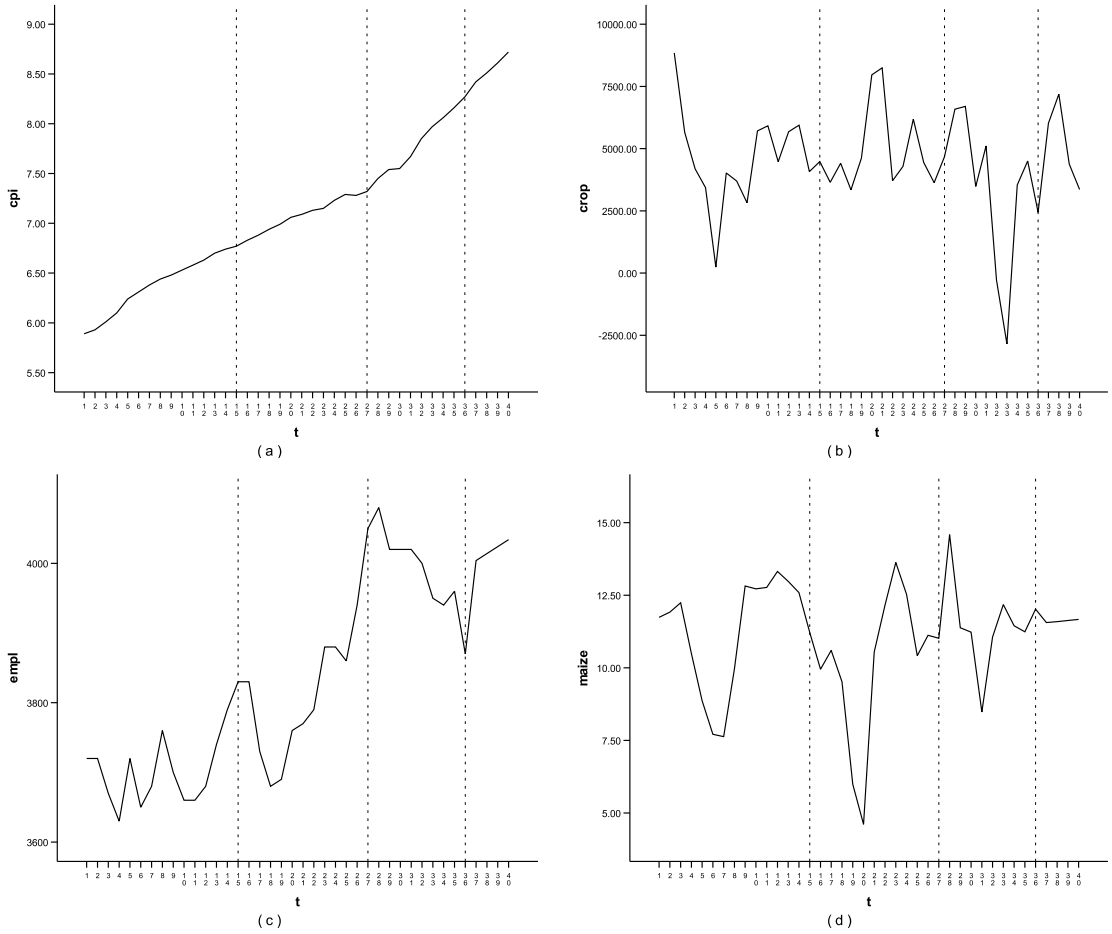

Figure 3: Quarterly series of (a) log-transformed consumer price index (CPI), (b) seasonally adjusted total crop sales, (c) employment and (d) seasonally adjusted log-transformed total maize production, from Q2 1991 to Q1 2001.

Period C had beer sales markedly influenced by the well documented social, political and economical events that occurred in Zimbabwe in the late nineties. In fact, two major political events in 1997 and 1998 were thought to be most influential for the sudden turn in the economy. First, the awarding of large grants and pensions to liberation war veterans which were paid for by an increase in sales taxes. Second, the government announced and started to implement the 1993 Land Designation Act which saw the redistribution of agricultural land with compensation covering buildings and infrastructure rather than land value. Those had a significant effect of output decrease on the commercial agricultural sector and related industries according to Bond (1999). Those changes in the economy were believed to have caused the decrease in sales of beer seen at that period. Those effects were represented in model $\mathcal{M}_{3t}$ by: (a) the linear increase in the (log-transformed) consumer price index (X_{4t}); (b) the decrease in employment levels (X_{6t}); as well as (c) the decrease in (deseasonalised log-transformed) total maize sales (X_{7t}). Those explanatory variables can be seen in Figure 3(a), (c) and (d) respectively.

The regression vector for $\mathcal{M}_{3t}$ was then set as $\underline{F}'_{3t} = (1, t, X_{1t}, X_{2t}, X_{4t}, X_{6t}, X_{7t})$. The evolution matrix $\mathbf{G}_{3t}$ was set as a (7×7) identity matrix.

4.2 Predictive performances

In this section we compare the within-sample and the out-of-sample forecasting performances of the geometric and linear combinations of the formulated models. For that, we made use of the BATS (Bayesian Analysis of Time Series) program developed by Pole, West and Harrison (1994) to run the formulated models.

The initial prior values at $t = 0$ for all models were determined from a *reference analysis* which used initial observations to assign some 'reasonable' initial values to the model parameters (see e.g. Box and Tiao, 1973, or Pole and West, 1989).

For $\mathcal{M}_{1t}$, the reference means and standard deviations for level, growth, price, temperature and drought were $(47.54, 339.4)$, $(-1.472, 5.164)$, $(2.869, 12.11)$, $(1.196, 3.495)$ and $(3.7, 21.58)$ respectively. The reference observational variance $V_{1,0}$ was 3.467. $\mathcal{M}_{2t}$ had reference prior means and standard deviations of $(13.70, 14.220)$, $(-0.846, 1.441)$, $(6.796, 5.356)$, $(1.623, 1.977)$, $(-4.547, 7.441)$ and $(-0.001, 0.0014)$ for level, trend, beer price, temperature, drought and crop sales respectively. The reference observational variance $V_{2,0}$ was 1.700. The reference prior means and standard deviations for $\mathcal{M}_{3t}$ were $(38.10, 5.935)$, $(0.537, 0.186)$, $(-1.083, 0.263)$, $(2.293, 0.721)$, $(-19.38, 3.536)$, $(0.032, 0.0007)$ and $(0.298, 0.283)$ for level, growth, price, temperature, CPI, crop sales, employment and maize respectively. The reference observational variance $V_{3,0}$ was set at 35.222.

Discount factors of 0.99 were used for all models to determine subsequent parameter values for level, growth, regressors as well as the observational and evolution variances. A discount factor, $\delta \in (0, 1]$, is a practical solution to the problem of setting values for the dynamic parameters in RDLMs. It represents the amount of information loss attributed to time evolution. In our case, $\delta = 0.99$ represents a 1% loss through the evolutionary process, for example at time t the actual variance is δ^{-1} of the variance at $t-1$. The 0.99 factors were chosen to minimize the model's forecasting performances (during the within-sample period) in terms of its *mean absolute deviation* (MAD) and the *square root of the mean squared error* (RMSE). Those values are displayed in Table 1.

No interventions were carried out during the analysis such that all models parameters were only updated by the available data. The combining weights were chosen to reflect the believed scenarios at each period A, B, C and D. They were arbitrarily chosen as $w_{1A} = 0.6, w2A = 0.3, w_{3A} = 0.1$ for period A, $w_{1B} = 0.2, w2B = 0.6, w_{3B} = 0.2$ for period B, $w_{1C} = 0.3, w2C = 0.1, w_{3C} = 0.6$ for period C and $w_{1D} = 0.15, w2D = 0.05, w_{3D} = 0.8$ for the out-of-sample period.

As in any regression model, forecasting of the response series requires that forecasts of each of the regression variables in the model be produced. In this application we have also made use of BATS (with reference prior initialisation) to obtain such forecasts for each of the component models. Those forecasts were entered to obtain the predictive densities of each model during out-of-sample period. The actual regressor values were used for obtaining the one-step-ahead forecasting densities of the models during the within-sample period.

The plot of Figure 4 shows the observed beer sales as well as the forecasts (means of Student t one-step-ahead forecasting densities) for each of the component models obtained by BATS. It can be seen that all models follow the beer sales series reasonably well most of the time during the within-sample period ($t = 1$ to $t = 36$). The main exception occurred at $t = 16$ (Q1 1995) when the forecasts by $\mathcal{M}_{1,16}$ and $\mathcal{M}_{2,16}$ were much higher than the observed y_{16} (and the forecast by $\mathcal{M}_{3,16}$).

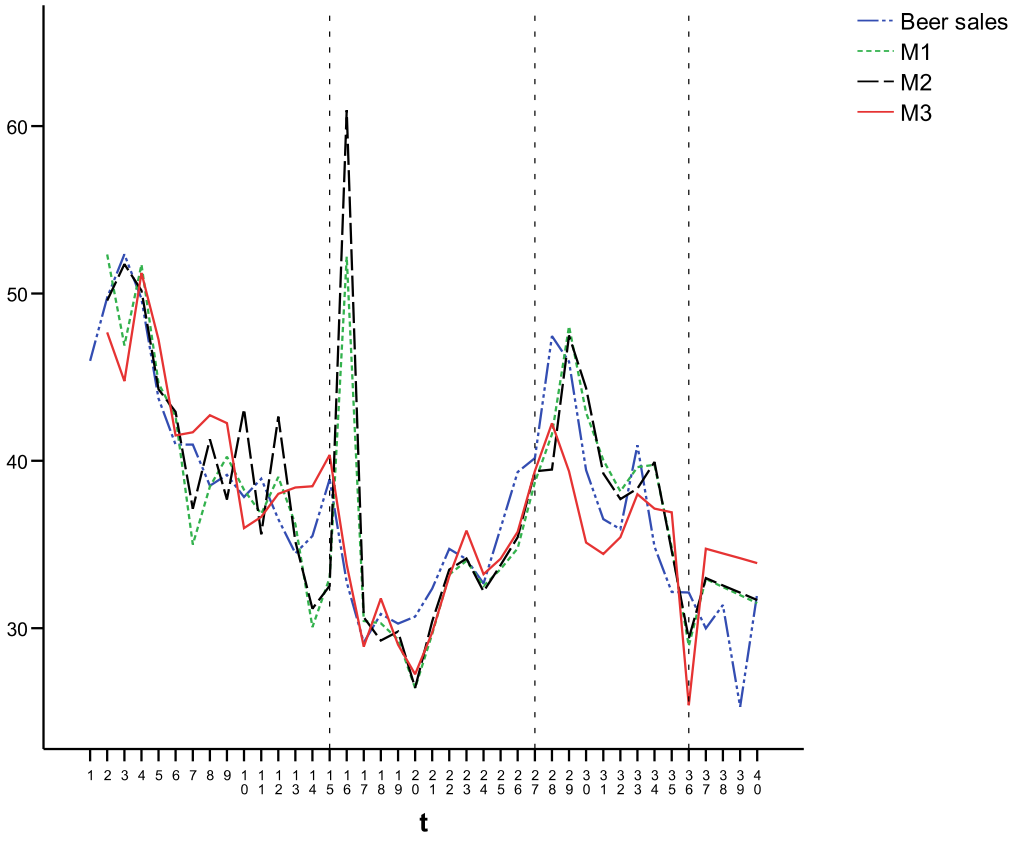

Figure 4: The observed sales (dash-dotted line) and the means of Student t forecasting densities by $\mathcal{M}_{1t}$ (dotted line), $\mathcal{M}_{2t}$ (dash line) and $\mathcal{M}_{3t}$ (solid line) from $t = 1$ (Q2 1991) to $t = 36$ (Q1 2000) (one-step forecasts) and from $t = 37$ (Q2 2000) to $t = 40$ (Q1 2001) (the h-steps forecasts made at $t = 36$, $h = 1, 2, 3, 4$).

Table 1 shows the MAD, the RMSE and the *geometric mean relative absolute error* (GMRAE) of all models during the within-sample period. The GMRAE is a measure of error relative to a naive model (that uses y_{t-1} as a one-step forecast for Y_t) and is calculated for a period of time of size h by

$$GMRAE_h = \left[\frac{1}{h} \prod_{t=1}^{h} \left| \frac{(y_t - \hat{Y}_t)}{(y_t - y_{t-1})} \right| \right]^{\frac{1}{h}}$$

where $\hat{Y}_t$ is the point forecast for Y_t made at time $t - 1$. Note that while a naive model usually produces the best performance measures amongst most models, it does not capture any characteristic patterns in the data that can be useful for forecasting. Naive models are usually suitable to represent random walk processes. So, the GMRAE value of 1.07 for $\mathcal{M}_1$ indicates that the size of $\mathcal{M}_1$'s forecasting errors during the within-sample period is only 7% larger than the size of the errors generated using the naive model data set. Note that for being a relative measure, it avoids some of the scaling problems associated with the MAD or the RMSE.

	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	P_G	P_L
MAD	3.04	3.33	**2.64**	**2.42**	3.21
RMSE	4.50	5.75	**3.22**	**3.12**	5.53
GMRAE	**1.07**	1.08	1.17	**0.95**	1.12

Table 1: The MAD, RMSE and GMRAE of the one-step within-sample forecasts by $\mathcal{M}_{1t}, \mathcal{M}_{2t}, \mathcal{M}_{3t}$ (means), the geometric (P_G's modes) and the linear (P_L's largest modes) combinations. Best performances in bold.

The geometric combination (P_G) produce the best within-sample performance of all models in all three performance measures. In particular, it outperformed the linear combination and the best performing component model. It is the only model that would outperform the naive model with a GMRAE of 0.95. Amongst the component models, $\mathcal{M}_3$ produced the smallest MAD (2.64) and RMSE (3.22) but the largest GMRAE (1.17). That result is, at least in part, due to the large errors by both $\mathcal{M}_1$ and $\mathcal{M}_2$ at $t = 16$. Those forecast errors at $t = 16$ omitted, all the performance measures would have favored $\mathcal{M}_1$ instead.

In cases where the combined density is multi-modal (or skewed), the issue of what location and spread parameters to adopt by the analyst becomes an issue. In fact, the mean (or the median) loses most of its usefulness as a descriptive statistics in such situations as it is expected to fall in the interval between the modes (in many cases near an anti-mode). The modes themselves tend to coincide with means of the component distributions of the combination. Similarly, the variance fails to describe the peakedness (or spread around the mean) of multi-modal (or skewed) distributions. Faria and Mubwanrikwa (2006) have shown, that the analyst's choice of loss function (associated with the consequences of her potential decisions) can change the optimal choice of location parameter when forecasting densities are multi-modal.

As an example, refer to the plot of Figure 1 where the one-step-ahead Student's t forecasting densities at Q2 1998 ($t = 29$) for the three component models (dashed lines) as well as those of their linear and geometric combinations (solid lines) are shown. The component model's densities obtained by BATS had 25 degrees of freedom and means and standard deviations of $(48.04, 2.881)$, $(47.49, 3.258)$ and $(39.37, 3.063)$ for $\mathcal{M}_{1,29}$, $\mathcal{M}_{2,29}$ and $\mathcal{M}_{3,29}$ respectively. The combination weights were $w_{1C} = 0.3$, $w_{2C} = 0.1$ and $w_{3C} = 0.6$. The resulting density of the linear combination is bi-modal with the largest mode at $y = 39.50$, mean at $y = 42.78$, anti-mode at $y = 44.43$ and smallest mode at $y = 47.62$. The geometric combination in its turn has a practically symmetric uni-modal density with mode at $y = 42.79$. The observed beer sales value at Q2 1998 was 45.93. So, in this instance the linear combination's anti-mode produced the forecast closest to the observed value in absolute error terms. The obvious choice of the largest mode as point forecast (before observation) would have produced the second worst result of all models at this time.

The plot of Figure 5 shows the observed values of beer sales (dash-dotted line) from $t = 1$ (Q2 1999) to $t = 40$ (Q1 2001), together with the means from the predictive densities of the best performing component model, $\mathcal{M}_{3,t}$ (dashed line), the modes from the geometric combination (solid line) and the largest modes from the linear combination (dotted line). It seems that in general all models followed the sales series reasonably well during the within-sample period. The combining models seem to have produced very close forecasts most of the

	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	P_G	P_L
MAD	**2.82**	2.84	4.63	**2.82**	**2.82**
RMSE	**3.69**	3.77	5.34	3.74	**3.72**
GMRAE	0.40	**0.37**	0.83	**0.38**	0.39

Table 2: The MAD, RMSE and GMRAE of the h-step out-of-sample forecasts ($h = 1, \ldots, 4$) by $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ (means), the geometric (P_G's modes) and the linear (P_L's largest modes) combinations. Best performances in bold.

time with few exceptions such as at time $t = 16$ when the linear model produced a much larger over-prediction of sales compared with the geometric. The combining models also produced very close short term forecasts in the out-of-sample period. Model $\mathcal{M}_{3t}$ has not reproduced its good within-sample performance in the out-of-sample period.

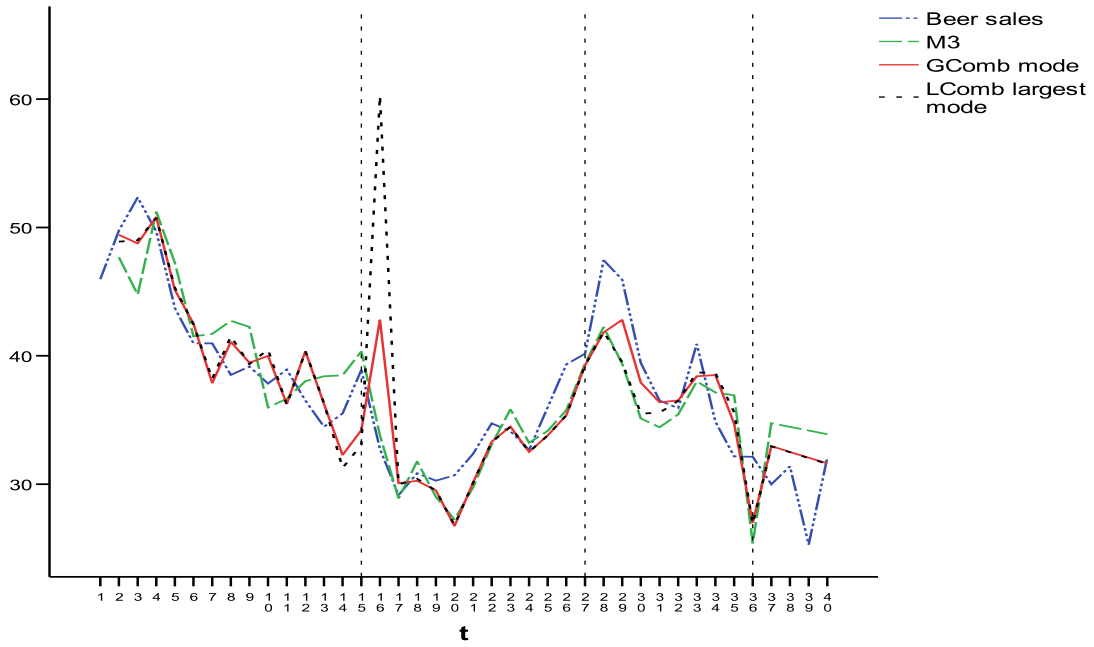

Figure 5: The observed sales (dash-dotted line) and the means of $\mathcal{M}_{3t}$ (dash line), the modes of the geometric combination (solid line) and the largest modes of the linear combination (dotted line) as one-step within-sample point forecasts, and h-steps out-of-sample point forecasts at $t = 36$ ($h = 1, 2, 3, 4$).

Table 2 shows the MAD, RMSE and GMRAE for the h-steps out-of-sample forecasts ($h = 1, 2, 3, 4$) by the means of the predictive t-densities of each component model as well as the largest modes of the linear combination densities (P_L) and the modes of the geometric combination densities (P_G).

Note that despite having the best within-sample performance of all component models, $\mathcal{M}_3$ had the worst out-of-sample performance in this instance. The best performing individual

model on the MAD and RMSE measures was $\mathcal{M}_1$ but closely followed by $\mathcal{M}_2$ which had the smallest GMRAE. The performance results for the combination models were very similar indeed with matching MAD and very close RMSE and GMRAE. The performances of the combining models were also very close to the best component models. So, no clear winner in the out-of-sample performance competition.

5 Conclusion

In this paper we have proposed a non-linear geometric approach to the combination of statistical models (including Bayesian forecasting models) as an alternative to (linear) mixture models. In situations where the models to be combined are from the regular exponential family of distributions we have shown that the resulting distribution of the geometric combination is also from the same family and, thus, uni-modal. This is not the case for mixture models which are typically multi-modal. In the Student t case, we have shown a sufficient uni-modality condition for the geometric model. uni-modality may be desirable in decision making situations where a single location parameter is required.

Geometric combinations are also externally Bayesian such that whether to combine before or after new data become available is irrelevant. In fact, for such combinations, Bayes theorem applied to the combined prior distributions will give the same result as combining the posterior distributions themselves. When component distributions are subjective (i.e. coming from expert judgements), external Bayesianity can guarantee that no individual expert will influence the decision making process by insisting the prior and not the posterior distributions be combined first.

An application of the geometric and the linear combinations to the forecasting of beverage sales in Zimbabwe showed that the geometric approach outperformed both the linear and the component regression dynamic linear models in the within-sample period. When used for forecasting it performed as well the the linear and the best component models.

Also, in this application we have adopted a probabilistic interpretation to the combining weights which allows the analyst to explicitly model her beliefs about how past economical-environmental scenarios are likely to repeat in the forecasting period. In fact, any scenario (past or not) for which the analyst is able to formulate a statistical model could in principle be included in the combination. This can be very useful in situations, similar to our beer sales forecasting application, where the underlying process is subject to structural changes caused by economic and environmental factors such as droughts and hyper-inflation.

Although we have only considered continuous distributions from the exponential and Student t family, there is certainly scope for investigating the behaviour of other types distributions under the geometric combination approach we proposed here.

References

Bond, P. (1999), Political reawakening in Zimbabwe. *M*onthly Review, Date: April 1, 1999.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Massachusetts.

Bunn, D. W. (1975), A Bayesian approach for linear combining models. *Operations Research Quartely*, **40**, 322-327.

Bunn, D. W. (1978), A simplification of the matrix beta distribution for combining estimators. *J. Opl. Res. Soc.*, **29**, 1013-1016.

Eisenberger, I. (1964), Genesis of Bimodal Distributions. *Technometrics*, **4**, 357–367.

Faria, A. E. (1996), *Graphical Bayesian Models in Multivariate Expert Judgements and Conditional External Bayesianity*, PhD Thesis, Department of Statistics, University of Warwick, U.K.

Faria, A. E. and Mubwandarikwa, E. (2006), *The Geometric Combination of Forecasting Models*, Technical Report 06/11, Department of Statistics, The Open University, U.K.

Genest, C. McConway, K. J. and Schervish, M. J. (1986), Characterization of externally Bayesian pooling operators. *Ann. Statist.*, **14**, 487–501.

Madansky, A. (1964) Externally Bayesian groups, *Rand Corporation Memo* Rm-4141-PR, Santa Monica, Rand.

Pole, A. and West, M. (1989), *Reference analysis of the DLM*, J. Time Series Analysis bf 10, 131–147.

Pole, A., West, M. and Harrison, P.J. (1994), it Applied Bayesian Forecasting and Time Series Analysis. Chapman & Hall/CRC, Boca Raton.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997), Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **92**, 179–191.

Raiffa, H. (1968), *Decision Analysis : Introductory Lectures on Choices under Uncertainty.* Random House, New York, pp. 211-226.

Titterington, D. M., Smith,A. F. M. and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions.* John Wiley and Sons, New York.

West, M. and Harrison, P. J. (1997), *Bayesian Forecasting and Dynamic Models* (2nd edition). Springer-Verlag, New York.

World Bank. (1996), *Understanding Poverty and Human Resources* : Changes in the 1990s and Directions for the Future, World Bank, Washington D.C.