# The Geometric Combination of Forecasting Models

A. E. Faria and E. Mubwandarikwa

*Department of Statistics, The Open University*

*MK7 6AA, U.K.*

**Abstract**

The most commonly used method for combining probability models is the linear combination (LComb) also known as mixture. In this report we propose an alternative approach, the *geometric combination* (GComb), which overcomes some of the main disadvantages of the linear methods. In fact, the proposed GCombs not only preserve the distributional form of the combination for many distribution types (including those of the exponential family and $t$-distributions), but are also externally Bayesian. While closedness to distribution forms can preserve unimodality, external Bayesianity will preserve immunity of influence on decision making. The classification theorem of catastrophe theory is used to establish the unimodality conditions for the GComb of densities from the Student t family.

We apply both the LComb and the GComb to a case of beverage sales forecasting in Zimbabwe. Several plausible (non-similar) regression dynamic linear models which include both economic and weather variables are entertained. The performances of the two combination

methods are compared in terms of symmetric loss functions such as the quadratic and the exponential losses as well as the non-symmetric logarithmic loss. Consequences of decisions for such losses under multimodal predictive densities are compared when the mean and the largest mode are chosen as point forecasts.

Keywords: Bayesian forecasting, non-linear model combination, regression dynamic linear models, multimodal predictive densities, loss functions

# 1 Introduction.

A statistical model like any mathematical model consists of a simple representation of a usually complex real process or phenomena. As such it is not always appropriate to assume that a model is the true representation of the underlying process. We assume here that a plausible (or appropriate) model irrespective of its truth is useful for the purposes of forecasting a time series.

In our context, we assume that a plausible model is mainly characterised by its set of regressors. In particular, we will consider a class of Bayesian time series forecasting models, the *regression dynamic linear models* (RDLMs) proposed by West and Harrison (1997).

One of the main features of the Bayesian forecasting modelling approach is to allow for model intervention to accommodate subjective information. This is particularly important at times of major changes in the time series process. Intervention allows for both parametric and structural changes to be made to a model.

Changes to existing parameters of a model allow it to adapt faster to changes in the series. For example, suppose a model is structured such that it has a specific parameter to account

for the series mean level at each period of time. Thus, by properly modifying the model's prior probability distribution for the mean level, a significant change in the level of the series can be better fitted by the model. Such a model should also produce improved predictions in the short term. Similarly, the model can be made to adapt to stochastic changes in the variance of the series by having the prior distribution for the observational variance changed accordingly.

Certain (structural) changes in the process generating the data, where prior expert intervention only is not appropriate, may require changes in the structure of the Bayesian model itself. Those changes can be implemented by modifications to either (i) the model's parameter set (by augmentation, for example, to allow an explicit modelling of a newly observed characteristic or change in the process), or (ii) the model's design matrix (one example being to include a new causal variable in the regressor set).

Further to model intervention, model monitoring for continual assessment and detection of model inadequacies, is an implicit recognition that a single model may not always be appropriate to represent the global behaviour of the series. In effect, despite a particular RDLM with a specific formulation being more appropriate for a time series during a period of time, it will not necessarily be appropriate at all times. Also, in many situations, there will be a number of models with distinct formulations which can be considered to be appropriate (or plausible) for the time series process even during a common period of time.

In this paper, we assume that a Bayesian decision maker (BDM) entertains a number of RDLMs, say $\mathcal{M}_1, \ldots, \mathcal{M}_k$ ($k \geq 2$) *she* believes are good (plausible) models for a time series process, represented by the random variable $Y_t$ ($t = 1, 2, \ldots$), but is interested in determining a single model she can use for forecasting. The BDM possesses a set of historical time series

data of dimension $T$, $\underline{y}^T = (y_1, \ldots, y_T)'$, about the underlying process $Y_t$ as well as on a set of auxiliary variables $X_{it}$ (we call influential variables) which realisations we represent by the vector $\underline{x}_i^T = (x_{i,1}, \ldots, x_{i,T})'$, $(i = 1, \ldots, r)$ she believes relate in a causal sense to $Y_t$. The notation adopted throughout, underlined characters represent vectors, boldfaced characters represent matrices and prime denotes transposition.

Each RDLM that the BDM entertains at a given period of time can be seen as coming from her belief about how the influential variables $\underline{X}_t$ relate to $Y_t$. Also, this relationship is not fixed but may change over time, giving origin to further alternative models that reflect the BDM's believed causal associations at different periods.

The BDM's task is to obtain, at time $t$, a single predictive distribution for future values of $Y_t$ to suport her decision making. There are basically two courses of action that the BDM could take to obtain a single predictive density:

(i) she can choose a single model from the set $\underline{\mathcal{M}} = \{\mathcal{M}_1, \ldots \mathcal{M}_k\}$ based on some model selection criteria and use that model for forecasting or

(ii) she could obtain a model from the combination of $\mathcal{M}_1, \ldots \mathcal{M}_k$ and use the combined model for forecasting.

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

prediction.

Under option (ii) above, model combination, there is a main class of available methods all based on the linear combination of models usually referred to as *mixture models*. In the Bayesian context, Raftery, Madigan and Hoeting (1997) refer to the combination of the predictive distributions weighted by the models probabilities as *Bayesian model averaging* (BMA). West and Harrison (1997) considered classes of *dynamic linear models* (DLMs) with their linear combinations (or mixtures) referred to as *multi-process* models. However, they only consider models whose parameter sets are identical and differ only on their variances.

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, the *linear combinations* of probability distributions such as the BMA also has its disadvantages. The exact resulting combination, in certain cases, can be difficult to obtain in practice. This is remedied by adopting approximate solutions like Markov Chain Monte Carlo (MCMC) simulation . Furthermore, in many situations the resulting distribution presents multi-modality which is an undesirable characteristic from a decision making standpoint as a decision usually requires a single estimate for the location parameter of the predictive distribution.

In this article we propose an alternative approach to the linear combination, 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 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.

The second reason, is that geometric combinations preserve the distribution form of the combined models for many distributions. In particular we shall 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 unimodal. This is not the case with linear combinations where, in general, unimodality occurs only under certain conditions, see e.g. Titterington, Smith and Makov (1985). Further, whilst linear combinations of (Student) $t$-distributions are necessarily multimodal under linear combination, it is not always so under the geometric rule as we also shall see. Student $t$-distributions play an important role in Bayesian forecasting.

The remainder of this paper is structured as follows. In Section 2 we introduce some further notation and define the structure of plausible RDLM's and their associated proper-

ties. In Section 3 we introduce both the linear the geometric combinations of probability distributions. We show their main characteristics and properties in terms of conditions for unimodality when applied to Gaussian and Student-$t$ models in particular, and more generally to the exponential family. In Section 4, we produce a comparative analysis of the performances of the two combination methods to a time series of beverage sales in Zimbabwe. Section 5 concludes the paper and points out some future research.

# 2 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 $\mathcal{M}_j$.

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 BDM 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. Please see West and Harrison (1997) for further details.

The following subsection describes how in general terms (for any assumed error distri-

bution) the sequential Bayesian (prior-to-posterior) updating of parameters of a RDLM is carried out.

## 2.1 Sequential Bayesian Model Updating

A RDLM $\mathcal{M}_{jt} = \{\underline{F}, \mathbf{G}, V, \mathbf{W}\}_{jt}$ is sequentially updated in time in the following manner. Let $\underline{\theta}_{jt} = (\theta_1, \ldots, \theta_{r_j})'_{jt}$ be the $(r_j \times 1)$ state vector associated with $\mathcal{M}_{jt}$. At each time $t$, $Y_t$ conditional on $\underline{\theta}_{jt}$, has a known observation probability density (or mass) function $p_j(Y_t|\underline{\theta}_{jt})$. Conditional on $\underline{\theta}_{jt}$, $Y_t$ is assumed to be independent of $Y_s$ and $\underline{\theta}_{js}$ for all past and future values of $s$ $(s \neq t)$. The parameter vector, $\underline{\theta}_{jt}$, evolves sequentially in time according to a known Markovian process described by an evolution density $p_j(\theta_{jt}|\underline{\theta}_{j,t-1})$. Notice that $\underline{\theta}_{jt}$ is assumed independent of $\underline{\theta}_{js}$ for $s \neq t-1$ given $\underline{\theta}_{j,t-1}$.

Like a *dynamic linear model* (DLM), a RDLM $\mathcal{M}_{jt}$ can be represented by an observation and an evolution equation:

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

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

Within the Bayesian dynamic forecasting framework we adopt here, the predictive densities for $Y_t$ are obtained sequentially for each model $\mathcal{M}_{j,t}$ $(j = 1, \ldots, k)$ from the marginalisation of the model's parameter vector $\underline{\theta}_{j,t} = (\theta_{1,t}, \ldots, \theta_{n_j,t})$ of dimension $n_j$ from the joint density

$$p_j(Y_{t+1}, \underline{\theta}_{j,t+1}|D_t) = p_j(Y_{t+1}|\underline{\theta}_{j,t+1}, D_t)\, p_j(\underline{\theta}_{j,t+1}|D_t)$$

where the prior density

$$p_j(\underline{\theta}_{j,t+1}|D_t) = \int p_j(\underline{\theta}_{j,t+1}|\underline{\theta}_{j,t}, D_t) p_j(\underline{\theta}_{j,t}|D_t) d\underline{\theta}_{j,t} \ . \tag{1}$$

The posterior density at time $t$,

$$p_j(\underline{\theta}_{j,t}|D_t) \propto p_j(\underline{\theta}_{j,t}|D_{t-1}) p_j(Y_t|\underline{\theta}_{j,t}, D_{t-1}) \tag{2}$$

is determined by Bayes' theorem. The one–step ahead predictive density is then

$$p_j(Y_{t+1}|D_t) = \int p_j(Y_{t+1}|\underline{\theta}_{j,t+1}) p_j(\underline{\theta}_{j,t+1}|D_t) d\underline{\theta}_{j,t+1}$$

where the integral $\int$ denotes multiple integration over the parameter space generated by $\underline{\theta}_{j,t+1}$.

An important property that effectively enables dynamic modeling is that of *conditional independence*, in which, given $\underline{\theta}_{j,t}$ at time $t$, $Y_{t+h}$ $(h = 0, 1, \ldots)$ is independent of $Y_{t-i}$ $(i = 1, 2, \ldots)$. That is, given the present $\underline{\theta}_{j,t}$, the past, present, and future observations are mutually independent. Also, given $D_t$, all information concerning the future is contained in the posterior density $p_j(\underline{\theta}_{j,t}|D_t)$ so that its hyperparameters are sufficient for $\{Y_{t+h}, \underline{\theta}_{j,t+h}\}$ $(h = 1, 2, \ldots)$. The predictive (or forecasting) densities can be computed sequentially for larger forecasting horizons. In general, the $h$–steps ahead predictive density for $Y_{t+h}$ calculated at time $t$ (sequentially) for $h = 1, 2, \ldots$ will be:

$$p_j(Y_{t+h}|D_t) = \int p_j(Y_{t+h}|\underline{\theta}_{j,t+h}) p_j(\underline{\theta}_{j,t+h}|D_t) d\underline{\theta}_{j,t+h}. \tag{3}$$

In its simple form, a DLM $\mathcal{M}_j$ (at a fixed time $t$), characterised by $\{\underline{F}, \mathbf{G}, V, \mathbf{W}\}_j$, will assume specific known probability density functions for the observational and the evolution

errors such that the general sequential updating above can be performed in analytical closed form. Perhaps the best known case is the Gaussian DLM where the observational error $\nu_{jt}$ 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. $\underline{\omega}_{jt} \sim St_{n_{j,t-1}}[\underline{0}, \mathbf{W}_{jt}]$.

Recall that, in general, if $\underline{\theta} \sim St_n[\underline{m}, \mathbf{C}]$ then

$$p(\underline{\theta}|n, \underline{m}, \mathbf{C}) = c\left\{ n + (\underline{\theta} - \underline{m})'\mathbf{C}^{-1}(\underline{\theta} - \underline{m}) \right\}^{-(\frac{r+n}{2})}. \tag{4}$$

where $c$ is a constant such that $\int_{-\infty}^{\infty} p(\underline{\theta}|n, \underline{m}, \mathbf{C}) d\underline{\theta} = 1$, $n$ is the number of degrees of freedom, $\underline{m}$ is the $r \times 1$ mean vector of $\underline{\theta}$, $\mathbf{C}$ is the $r \times r$ covariance matrix and $r$ is the dimension of the vector $\underline{\theta}$.

In cases where it is not possible to adopt a conjugate analysis in (2), 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), which we shall adopt in our application section.

Now, that we have defined a RDLM and showed how it can be updated in time, we shall next define how we consider two RDLMs to differ from one another.

## 2.2 Similar RDLMs

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. For that reason it is important that we characterise when two RDLMs are considered to differ from one another.

Within the theory of DLMs, any two models producing the same forecasts are said to be *equivalent models*. Similarly, any two models with the same qualitative (i.e. algebraic) form of forecast functions are said to be *similar models*.

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 *not* similar, i.e. if they do not have forecast functions of exactly the same algebraic form. Formally, $\mathcal{M}_i$ and $\mathcal{M}_j$ are similar models, $\mathcal{M}_i \sim \mathcal{M}_j$, if and only if their evolution matrices $\mathbf{G}_i$ and $\mathbf{G}_j$ have identical eigenvalues. Equivalently, $\mathbf{G}_i$ and $\mathbf{G}_j$ are similar matrices such that there exists a non-singular *similarity matrix* $\mathbf{H}$ such that $\mathbf{G}_i = \mathbf{H}\mathbf{G}_j\mathbf{H}^{-1}$. Please refer to West and Harrison (1997) for further details.

Now that we have revised how a RDLM is sequentially updated in time, as well as introduced non-similar RDLMs as the types of models the BDM will be entertaining, we can move to the problem of how to combine them together to obtain a single forecasting model.

# 3 Characterising combinations of models

Let $(\Omega, \mu)$ be a measure space. Also, let $\Delta$ be the class of all $\mu$-measurable functions $p : \Omega \to [0, \infty)$ such that $\int p d\mu = 1$ with $\mu$ 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 $p_j = p(\cdot|\mathcal{M}_j) \in \Delta$ (for $j = 1, ..., k$), into a single density $p(\cdot)$ also in $\Delta$.

In the following subsections we define both the linear and the geometric combinations for univarite distributions. We also show unimodality conditions for models within the exponential family as well as for models with t-distributions for both types of combination.

## 3.1 Linear Combination of Predictive Models

The linear combination $P_L : \Delta^k \to \Delta$ of $k$ (predictive) densities for a time series $Y_t \in \Omega$, given the actual information $D_t = (y_t, D_{t-1})$, has the following general form:

$$P_L(p_1, ..., p_k)(Y_{t+h}|D_t) = \sum_{j=1}^{k} w_{jt} p_j(Y_{t+h}|D_t) \tag{5}$$

where $p_j(Y_{t+h}|D_t)$, the $h$-steps-ahead ($h = 1, 2, \ldots$) predictive density of model $\mathcal{M}_{jt}$, is sequentially obtained from (3) and $w_{jt}$, $j = 1, ..., k$, are arbitrary weights (not necessarily nonnegative) adding up to one. As $P_L(Y_{t+h}|D_t)$ is a probability density function (pdf) it means $P_L \geq 0$ for all $Y_t \in \Omega$, care must be taken when choosing negative weights.

The weights could be elicited by the BDM 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) *outperformance*- available for determining the weights based on the models past predictive performances. In the Bayesian model averaging framework, the weight $w_{jt}$ is treated as the posterior probability for model $\mathcal{M}_{jt}$, that is $w_{jt} = p(\mathcal{M}_{jt}|D_t)$ obtained by Bayes theorem via MCMC. Please see Raftery et. al. (1997) for further details.

In the application section of this paper, we have adopted such interpretation and obtain their values sequentially in time.

Now, in the context where the component models are non-similar RDLMs, the combination (5) above is applied to the predictive densities obtained from the posterior distributions of the parameter vector $\underline{\theta}_{jt}$, i.e. $p_j(\underline{\theta}_{jt}|D_t)$. The predictive density will therefore be obtained from the combined posterior densities by the integration in (3).

Consider the function $P_L(Y_{t+h}|D_t)$ as the BDM's linear combination of all h-steps-ahead $(h = 1, 2, \ldots)$ predictive probability density functions $p_j(Y_{t+h}|D_t)$ obtained from the RDLM $\mathcal{M}_{jt}$.

Note that the future values of regressors are required when forecasting ahead. There are various possible ways in which those forecasts can be obtained. A more complex way is to use multivariate time series modelling and forecasting to obtain a joint model for forecasting the regressor variables as time series along with the variable of interest $Y$. We adopt the simpler approach here of estimating the future values of regressors by individual univariate models for those.

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 (1996) 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 $f_{j,t}(h)$ and variance $Q_{j,t}(h)$, that is $(Y_{t+h}|D_t) \sim St_{n_{j,t}}(f_{jt}(h), Q_{jt}(h))$. 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.

## 3.2 Unimodality in Linear Combinations

This section presents the unimodality conditions for the linear combination of Gaussian and t-distributions.

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

$$(\mu_2 - \mu_1)^2 < \frac{27}{4} \frac{\sigma_1^2 \sigma_2^2}{(\sigma_1^2 + \sigma_2^2)}. \tag{6}$$

This condition means that to obtain unimodality, 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, unimodality 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.

### 3.2.1 Linear Combination of $t$-distributions

In the case where the combining densities in (5) are $t$-distributions, the resulting combination can also be unimodal under conditions which are even more restrictive than that for the Gaussian case.

Without loss (for simplicity), we omit the time index $t$ in this subsection. For $k = 2$ and for some fixed weight $w$, the linear combination (5) can be written as

$$P_L(p_1, p_2)(Y) = wp_1(Y|n_1, \mu_1, \sigma_1^2) + (1 - w)p_2(Y|n_2, \mu_2, \sigma_2^2)$$

where $p_j(n_j, \mu_j, \sigma_j^2)$ $(j = 1, 2)$is the density of a $t$-distribution with $n_j$ degrees of freedom, mean $\mu_j$ and variance $\frac{n_j}{(n_j-2)}\sigma_j^2$ ($n_j > 2$ for finite variance).

Note that despite being very similar in shape to a normal distribution, the $t$-distribution has heavier tails. The smaller the number of degrees of freedom the heavier the tails. In practice, this means that samples from a $t$-distribution will have less observations away from the mean than samples from a normal distribution. Also, recall that for a normally distributed random variable $X \sim N(\mu, \sigma^2)$, and a chi-square distributed random variable $nU \sim \chi^2(n)$ with $n$ degrees of freedom, we have that $Y = \mu + \frac{X}{\sqrt{U/n}}$ will follow a $t$-distribution with $n$

degrees of freedom, mean $\mu$ and variance $\frac{n}{n-2}\sigma^2$, i.e. $Y \sim St_n(\mu, \sigma^2)$.

Now, if we introduce chi-square latent variables $n_j u_j \sim \chi^2(n_j)$ $(j = 1, 2)$, we can use them to re-scale the variances of normal distributions (so as to make them resemble $t$-distributions with heavier tails) such that the linear combination of $t$-densities above can be approximated by the linear combination of normal densities:

$$P_L(p_1, p_2)(Y) \approx w p_1^*(Y|\mu_1, \frac{\sigma_1^2}{u_1}) + (1 - w) p_2^*(Y|\mu_2, \frac{\sigma_2^2}{u_2})$$

where $p_j^*(Y|\mu_j, \frac{\sigma_j^2}{u_j})$ is the density of a normal distribution with mean $\mu_j$ and variance $\frac{\sigma_j^2}{u_j}$ $(j = 1, 2)$.

Therefore, the unimodality condition (6) for the linear combination of normal distributions can be rewritten as

$$(\mu_2 - \mu_1)^2 < \frac{27}{4} \frac{\sigma_1^2 \sigma_2^2}{(u_2 \sigma_1^2 + u_1 \sigma_2^2)}.$$

Note that, for heavier tails, $u_1, u_2 > 1$ imply that $\sigma_1^2 + \sigma_2^2 < u_2 \sigma_1^2 + u_1 \sigma_2^2$, and thus unimodality of the linear combination of the two $t$-densities would be achieved only when the means of the two $t$-distributed components are closer to each other comparatively with the required distance between the means of the Gaussian counterparts. That is, it is harder (in the sense that shorter distances between the means are required) to obtain unimodality in the linear combination of $t$-densities than it is in the linear combination of Gaussian densities.

## 3.3 The Geometric Combination

The geometric combination $P_G : \Delta^k \to \Delta$ of $k$ predictive densities for a time series $Y_t \in \Omega$, given the information set $D_t = \{y_t, D_{t-1}\}$, has the following general form (for $h = 1, 2, \ldots$):

$$P_G(p_1, ..., p_k)(Y_{t+h}|D_t) = c \prod_{j=1}^{k} p_j^{w_{jt}}(Y_{t+h}|D_t) \ , \tag{7}$$

where $c^{-1} = \int \prod_{j=1}^{k} p_j^{w_{jt}}(Y_{t+h}|D_t)dY_{t+h}$, $w_{jt} \in \mathbb{R}$, $j = 1, ..., k$, is the weight associated with the predictive density $p_j(Y_{t+h}|D_t)$ (obtained from model $\mathcal{M}_{jt}$) such that $\sum_{j=1}^{k} w_{jt} = 1$.

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 on the decision making in cases where predictive distributions are subjective opinions from experts (or interested parties). External Bayesianity is defined in the following section.

### 3.3.1 External Bayesianity

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

$$P\left(\frac{p_1}{\int \mathcal{L}p_1 d\mu}, \ldots, \frac{p_k}{\int \mathcal{L}p_k d\mu}\right) = \frac{\mathcal{L}P(p_1, \ldots, p_k)}{\int \mathcal{L}P(p_1, \ldots, p_k)d\mu} \ , \quad \mu \ a.e. \ , \tag{8}$$

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 (7) satifies (8) and therefore is EB. The reader can refer to Faria (1996) or Genest et. al. (1986) for more details.

### 3.3.2 The Geometric Combination of Exponential Family Densities

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

A density $p(y|\underline{\eta})$ where $\underline{\eta} = (\eta_1, \ldots, \eta_n) \in \Omega$ belongs to the $n$-parameter exponential family has the natural representation

$$p(y|\underline{\eta}) = h(y)c(\underline{\eta}) \exp[\underline{\eta}' \underline{d}(y)] \tag{9}$$

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 $\Omega$ 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) $\Omega$ is a $n$-dimensional open set. 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.

We can state the following result:

**Theorem 3.1** *Let $p_1(Y_{t+h}|D_t), \ldots, p_k(Y_{t+h}|D_t)$ be strongly unimodal densities from the regu-*

*lar exponential family. Then, the density of the geometric combination* $P_G(p_1, \ldots, p_k)(Y_{t+h}|D_t)$ *in (7) is also a strongly unimodal density from the regular exponential family.*

**Proof.** First, assume that a random variable $y$ whose density under a model $\mathcal{M}$, $p(y|\underline{\eta})$ belongs to the $n$-parameter strongly unimodal regular exponential family. Then, $p(y|\underline{\eta})$ 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 unimodal regular exponential family.

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

$$\begin{aligned} p^w(y|\underline{\eta}) &= [h(y)]^w [c(\underline{\eta})]^w \exp[w\underline{\eta}'\underline{d}(y)] \\ &= h^*(y)c^*(y)\exp[\underline{\eta}'\underline{d}^*(y)] \end{aligned} \tag{10}$$

where $h^*(y) = h^w(y)$, $c^*(\underline{\eta}) = c^w(\underline{\eta})$ and $\underline{d}^*(y) = (wd_1(y), \ldots, wd_n(y))$. 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 unimodal.

Now, the product of densities from the strongly unimodal 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 unimodal regular exponential family as above. Thus, given $\underline{w} = \{w_1, \ldots, w_k\}$ with $w_j \in (0,1) : \sum_{j=1}^k w_j = 1$, we can write that

$$\begin{aligned} P_G(y|\underline{\tilde{\eta}}) &= a(\underline{\tilde{\eta}}) \prod_{j=1}^k [h_j(y)]^{w_j} [c_j(\underline{\eta})]^{w_j} \exp[w_j \underline{\eta}_j' \underline{d}_j(y)] \\ &= \tilde{h}(y)\tilde{c}(\underline{\eta}) \exp[\sum_{j=1}^k \underline{\eta}_j' \underline{d}_j(y)] \ , \end{aligned} \tag{11}$$

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

$\tilde{h}(y) = \prod_{j=1}^{k}[h_j(y)]^{w_j}$ 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 $n = \sum_{j=1}^{k} n_j$.

Now,

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

is concave and thus $P_G$ is unimodal.

$\triangle$

Notice that this result is rather interesting from a decision analysis viewpoint. It may in many cases give a decision maker a very good reason to adopt the geometric rather than the linear combination of models as we shall see.

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

As mentioned before, there are situations in which the predictive densities belong to the family of Student t-distributions. The following section shows that (similarly to the linear combination of Gaussian distributions) the density of the geometric combination of t-distributions may be multimodal.

### 3.3.3 Unimodality in the geometric combination of t-distributions

Let $p_j(Y)$ $(j = 1, \ldots, k)$ be the density of a $t$-distribution with $\nu_j$ degrees of freedom, mean $\mu_j$ and variance $\sigma_j^2$, i.e. $Y|\mathcal{M}_j \sim St_{\nu_j}(\mu_j, \sigma_j^2)$. In this case, the geometric combination of the

form (7) will have a density of the form:

$$P_G(p_1,\ldots,p_k)(y) = c\ \prod_{j=1}^{k}\left[1+\tau_j^{-1}\,(y-\mu_j)^2\right]^{-\frac{w_j(\nu_j+1)}{2}}\ , \tag{12}$$

where $c$ is a normalizing constant, and $\tau_j = \nu_j\sigma_j^2$.

The (natural) logarithmic kernel of $P_G(y)$ has the from

$$L_G(y) = -\sum_{j=1}^{k}\frac{w_j}{2}(\nu_j+1)\ln[1+\tau_j^{-1}(y-\mu_j)^2]\}\ . \tag{13}$$

Now, if we represent by $D_y^i$ the $i$-th derivative w.r.t $y$, and equal the first derivative of $L_G(y)$ w.r.t. $y$ to zero, i.e. $D_yL_G(y) = \frac{d}{dy}L_G(y) = 0$, we have stationary points defined by the equation:

$$\sum_{j=1}^{k}\frac{w_j(\nu_j+1)\tau_j^{-1}(y-\mu_j)}{[\tau_j^{-1}(y-\mu_j)^2+\nu_j]} = 0\ , \tag{14}$$

and thus,

$$\sum_{j=1}^{k} w_j(\nu_j+1)\tau_j^{-1}(y-\mu_j)\prod_{i=1,i\neq j}^{k}[\tau_i^{-1}(y-\mu_i)^2=\nu_i] = 0\ . \tag{15}$$

Note that this equation is a cubic function of $y$.

Now, assuming for simplicity that $k=2$ and $\nu_j=\nu$ $(j=1,2)$, equation (15) can be rewritten in the form

$$x^3 - bx - a = 0 \tag{16}$$

where

$$\begin{aligned} x &= (y-\bar{\mu}) \ \text{ with} \\ &\quad \bar{\mu} = \frac{1}{3}[(2-w)\mu_1 + (1+w)\mu_2]\ ; \\ a &= \frac{1}{27}(M^3+\nu T) \ \text{ with} \\ &\quad M^3 = w_1^*\mu_1^3 - 3w_{12}^*\mu_1^2\mu_2 + 3w_{21}^*\mu_1\mu_2^2 - w_2^*\mu_2^3 \ \text{ where} \end{aligned}$$

$$w_1^* = (w-2)(4w^2 - 7w + 7) \ ,$$

$$w_{12}^* = 4w^3 - 9w^2 + 15w - 8 \ ,$$

$$w_{21}^* = 4w^3 - 3w^2 + 9w - 2 \ ,$$

$$w_2^* = (w+1)(4w^2 - w + 4) \ \text{ and}$$

$$T = 9\{(w-1)[(w-2)\mu_1 - (w+4)\mu_2]\tau_1 + w[(w-5)\mu_1 - (w+1)\mu_2]\tau_2\} \ ;$$

$$b = S^2 - \nu\bar{T} \ \text{ with}$$

$$\bar{T} = (1-w)\tau_1 + w\tau_2 \ \text{ and } \ S^2 = \frac{1}{3}(w^2 - w + 1)(\mu_1 - \mu_2)^2 \ .$$

Smith (1978) obtain a simpler result in determining the critical points for the likelihood function of t-distributions. Using Thom's *Classification Theorem* of catastrophe theory (see e.g. Poston and Stewart, 1978), this corresponds to the manifold of a cusp catastrophe for the potential function $P_G(Y|w, \nu, \mu_1, \mu_2, \tau_1, \tau_2)$ for two t-distributions with differing means and precisions but identical degrees of freedom. In this manifold, the axis $b$ in (16) lying along the cusp is the *splitting factor* (and the orthogonal axis $a$ is the *normal factor*). Note that, the splitting factor $b = S^2 - \bar{T}^*\nu^*$ above can be seen as representing the symmetrical "split" in $P_G$, while the normal factor $a = \frac{1}{27}(M^3 + \nu T)$ as incorporating all assymetrical components of $P_G$.

The projection of the cusp point (a point in the manifold formed by the set $Q = \{y, w, \nu, \mu_1, \mu_2, \tau_1, \tau_2\}$ of all points at which the derivative of $P_G$ with respect to $y$ vanish) onto the control space $C$ formed by $(w, \nu, \mu_1, \mu_2, \tau_1, \tau_2)$ divides it into regions which depict the stationary points of $P_G$. The cusp manifold and its projection on the plane (a, b) can be seen in the following Figure (1).

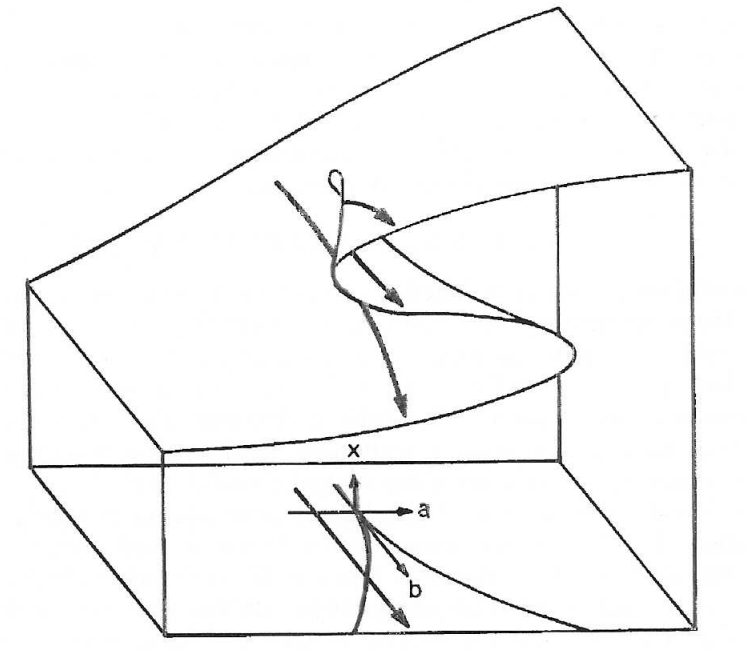


Figure 1: A cusp manifold with its projection over the $(a, b)$ plane formed by the normal and the splitting factors.

There will be be three roots as the solution of the cubic equation (16). Those roots for values of the control variables will determine whether the geometric combination is unimodal or not. A practical way of determining the regions of unimodality in the $(a, b)$ plane is by using Cardan's discriminant function $\delta(a, b)$ defined as:

$$\delta(a, b) = 27a^2 - 4b^3 \ . \tag{17}$$

When $\delta$ is positive ($\delta > 0$) there are three real roots for (16) with two possible values for the mode of $P_G$, that is we have bimodality. When $\delta < 0$ there is a single real root (and a conjugate pair of complex roots) and $P_G$ will be unimodal. In the case that $\delta = 0$ there are three real roots, of which two have the same value. This indicates that the a mode of $P_G$ is located on a catastrophe point. The cusp point occurs in the control space at $a = b = 0$, that

is $\delta(0,0) = 0$, and there are three real roots with the same value (all equal to 0). See Section 5.2 of Poston and Stewart (1978) for more details.

Here, we are interested on those points $(a, b)$ for which $\delta(a, b) < 0$ so that $P_G$ is unimodal. Those points satisfy $27a^2 < 4b^3$ or

$$(S^2 - \nu\bar{T})^3 < \frac{1}{108}(M^3 + \nu T)^2 \ . \tag{18}$$

Notice that while the geometric combination of natural exponential family densities is always unimodal, for $\tau_1 = \tau_2$ and $\nu > 1$ the weighted average mean $\bar{\mu}$ in the geometric combination of t-distributions goes from being a unique posterior mode in the region $\delta > 0$ to the unique anti-mode in the region $\delta < 0$ as the distance between the means $\mu_1$ and $\mu_2$ increases in the control space. Figure (2) shows a trajectory across the control space as $|\mu_2 - \mu_1|$ increases. Notice that for $\tau_2 = \tau_1$ and $\nu > 1$, the mean $\bar{\mu}$ changes from the unique mode to the unique anti-mode of $P_G$ as the distance between the individual means increases.

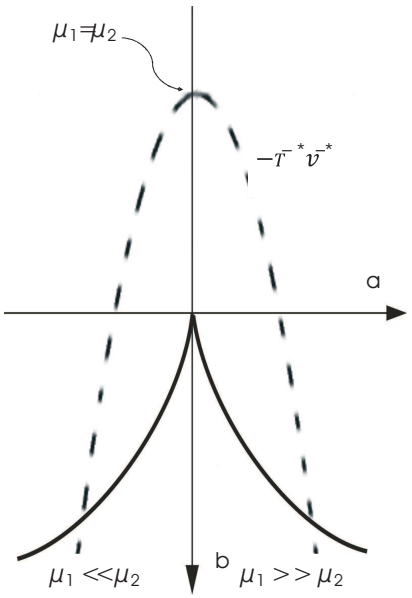


Figure 2: A trajectory across the control space as the absolute distance between the

means $|\mu_2 - \mu_1|$ increases.

As in our case the predictive densities are univariate, the first two moments of the geometric combination of t-distributions can easily be obtained computationally by numerical integration. However, while the mean and variance can characterise $P_G$ when its density is unimodal they are not appropriate parameters to characterise multimodal densities. In such cases, modes and spread around those would give a better description of the density's shape. In general, according to Cobb (1978), the first four moments would be appropriate for a proper characterisation. In this manuscript we look at the consequences on decisions of adopting the mean and the largest mode as point estimates in prediction as we shall see in the application section.

So far, we have not yet mentioned how the weights $w_1, \ldots, w_k$ ($\sum_j w_j = 1$) can be determined. In the following section we introduce some of the main methods for obtaining weights. In particular, for our application in this paper, we will adopt a Bayesian method called *outperformance* which allows an interpretation of a weight in terms of the probability of a scenario (which the model associated with that weight represents) occurring during the predictive horizon.

## 3.4 Weights for model combinations

One of the major difficulties associated with obtaining weights for the model combination approaches we have described in this paper, is that there is, up to date, no normative theory behind them to support their choice. However, there is a body of literature with several operational methods each giving a particular interpretation for the weights. Many of them were developed for linear combinations more due to the lack of alternative combining approaches

than to a methodological requirement.

The reader can refer to Winkler (1968), Bates and Granger (1969), DeGroot (1974), Bunn (1975), Smith and Makov (1978) and Cooke (1991) amongst others for different interpretations and methods for obtaining the weights.

In this paper, we adopt the *outperformance* approach proposed by Bunn (1975, 1978) which interprets a component $w_{jt}$ of the weights vector $\underline{w}_t = (w_1, \ldots, w_k)$ as the probability that model $\mathcal{M}_{jt}$ will produce the most appropriate forecast of $Y_t$.

The sufficient statistics with which to learn about $\underline{w}_t$ is assumed to be both $\underline{w}_{t-1}$ and the identity of the model that produced the closest forecast of $Y_{t+h}$. This identity, viewed as a random variable, is assumed to follow a Multinomial $(1, \underline{\theta}_t)$ distribution, $\underline{\theta}_t = (\theta_{1t}, \ldots, \theta_{k-1,t})$ where $\theta_{jt}$ is the parameter associated with the weight $w_{jt}$, while the $\underline{w}_t$ is interpreted as the prior mean of $\underline{\theta}_t$. The parameter vector $\underline{w}_t$ is then successively updated in the usual Bayesian framework in the light of forecasts. Using this method with the assumption that the $j$-th model relative performance about forecasting $Y_t$ is independent of every other model, it is easily checked that for $t \geq 1$

$$w_{j,t} = (1 - \rho_{t-1}) w_{j,t-1} + \rho_{t-1} (t-1)^{-1} r_{j,t-1}$$

for $j = 1, ..., k$, where $\rho_{t-1} = (t-1)/[\bar{\alpha}_{t-1} + (t-1)]$, with $\bar{\alpha}_{t-1} = \sum_{j=1}^{k} \alpha_{j,t-1}$, where $\alpha_{j,t-1}$ are the parameters of the conjugate Dirichlet prior distribution of $\underline{\theta}_t$, and $r_{j,t-1}$ is the number of successes of forecasting model $j$ up to time $t-1$.

A more attractive formulation of this approach (Bunn, 1978) allows the probability $\theta_{jt}$ that $\mathcal{M}_{jt}$ will be a more appropriate model than model $\mathcal{M}_{it}$ at time $t$, to be revised in the light of all the models relative performances. Pairwise comparisons between models are set up and a relative performance ranking is obtained. The weight $w_{jit}$ is assumed to be

the posterior mean of the $\theta_{ji,t-1}$ whose density function is now assumed to be a beta with parameters $(\alpha_{ji}, \alpha_{ij})_{t-1}$. These parameters are updated in the usual Bayesian way. Also assuming outperformance independence among estimators, the estimate of the probability of model $j$ outperform all other models, $w_{jt}$, can be obtained for $i \neq j$ as

$$w_{jt} \propto \prod_{u=1}^{k} w_{ujt} \quad .$$

Such a method of updating weights transfers directly onto both the linear and the geometric combinations.

Notice that the rule is fair in that it is symmetric in a model's success if a priori we set $w_{j,0} = w_{i,0}$ for $j, i = 1, ..., k$.

Now that we have introduced the geometric combination and some of its main characteristics as well as described a Bayesian method for obtaining the combining weights, we will show in the next section an application to the sales of beer in Zimbabwe.

# 4 Decision under multimodal predictive density

A Bayesian optimal decision $\hat{Y}_{t+h}$ is the one that minimises the expected value of a specified loss function $L(\hat{Y}_{t+h}, Y_{t+h})$ w.r.t. a density function $P(Y_{t+h}|D_t)$ for $Y_{t+h}$, i.e. the $\hat{Y}_{t+h}$ for which the infimum value of

$$EL_P(\hat{Y}) = \mathbb{E}_P[L(\hat{Y}_{t+h}, Y_{t+h})] = \int_{-\infty}^{\infty} L(\hat{Y}_{t+h}, Y_{t+h}) P(Y_{t+h}|D_t) dY_{t+h} \ , \tag{19}$$

is obtained. $\mathbb{E}_P$ denotes expectation w.r.t. the probability density function $P$.

In cases in which the expected loss has a single local minimum the optimal decision is that based on that choice. However, in cases in which the expected loss has more than one point of minimum, the selection of a particular point for decision must be made. However,

an infinitesimal change in the parameters of $L$ and/or $P$ can result in a major change in the decision.

It may be intuitive to think that when a decision maker's information is characterised by a continuous unimodal probability density function and the loss function is symmetric with just one minimum, the expected loss will have a single point of minimum. However, as shown by Smith et. al. (1981), this is not the case in general except for *bounded* monotonic increasing loss functions of $|\hat{Y}_{t+h} - Y_{t+h}|$ and strictly positive twice differentiable probability density functions. Note that the later condition include all symmetric unimodal densities as well as the Gamma and the Beta distributions. Nonetheless, there will exist a large set of loss functions which expected loss will have at least two local points of minima if the probability density is not strictly positive and the first derivative of its logarithm transformation tends to zero as $Y_{t+h}$ tends to infinity. This is the case for the log-normal, the inverted Gamma, Pareto and most distributions in the F family. Again, refer to Smith et. al. (1981) for a proof.

We have seen in Section 3 that the linear and the geometric combinations of predictive densities may not be unimodal. This can be the case for linear combinations of both normal and t-distributed component densities as well as for geometric combination of t-distributed components.

In the case of the linear combination of normal densities Smith et al. (1981) obtained the cusp catastrophe coordinates (i.e the normal and the splitting factors –$a$ and $b$ in an equation similar to (16) of section 3.3.3)– for the expected loss when Lindley's (1976) conjugate (to the Gaussian distribution) loss is adopted. In that case, the expected loss $\mathbb{E}_{P_L}[\hat{Y}_{t+h}] =$

$\sum_{j=1}^{k} w_j \mathbb{E}_{p_j}[\hat{Y}_{t+h} - \mu_j]$ where $\mathbb{E}_{p_j}[\hat{Y}_{t+h}] = \int_\Omega L(\hat{Y}_{t+h}, Y_{t+h}) p_j(y_{t+h}|D_t) dy_{t+h}$ and

$$L(\hat{Y}_{t+h}, Y_{t+h}) = L(\hat{Y}_{t+h} - Y_{t+h}) = 1 - \exp\{-\frac{1}{2}\lambda^{-1}(\hat{Y}_{t+h} - Y_{t+h})^2\} \tag{20}$$

with $\lambda > 0$. For $k = 2$, a common variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and a large $\lambda$, Smith et. al. (1981) showed that the expected loss above will have two points of minimum if

$$(\mu_1 - \mu_2)^2 > 4(\sigma^2 + \lambda) \ .$$

In this case, using the approximate symmetry of the expected loss, the lowest minimum will be the one nearest to $\mu_2$ if $w > \frac{1}{2}$, and the one nearest to $\mu_1$ if $w < \frac{1}{2}$.

A similar approach can be used for both the linear and the geometric combinations of t-densities.

Certainly that if $L(\hat{Y}_{t+h}, Y_{t+h}) = (\hat{Y}_{T+h} - Y_{T+h})^2$, then the expected loss function is always quadratic in $\hat{Y}_{t+h}$ and thus a minimum point can be obtained. The use of convex loss functions with unbounded decision parameter range, such as the quadratic loss above, has been criticised by Kadane and Chuang (1978). The use of bounded (conjugate) loss functions has been proposed by Lindley (1976).

# 5 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 non-similar RDLMs formulated for a series of beer sales from a brewery in Zimbabwe. We also show the consequences of decisions under the quadratic, the exponential and the logarithmic loss functions. For that we formulate three plausible RDLMs (each representing a past economic and weather scenario) to be combined

and tested in the last 12 month of data in the time series. The weight of each model in the combination methods could be chosen by the BDM as her subjective probability that the associated model represents the economic and weather scenario that she believes will prevail during the forecasting horizon. In this application the outperformance method described in Section 3.4 was adopted instead as a surrogate for the BDM assessment.

The underlying series comprises of 120 monthly deseasonalised log-transformed observations of total beer sales ($Y$) from April 1991 to December 2000. The series was deseasonalised as the original data presented a very strong and predictable seasonal behaviour with peaks at the end-of-the-year seasons (which coincide with the spring-summer seasons - from November to April) and troughs during autumn-winter seasons (from May to October). The deseasonalised sales data was then further transformed with the use of a logarithmic function in order to obtain a series with an approximately constant variability. The resulting series displayed in Figure 3(a) shows distinct trend patterns during three distinct periods of time. During the first period, which we call period $A$ (from April 1991 to December 1994), there was a linear but steep decline in sales. This was followed by a period B (from January 1995 to December 1997) of a positive linear sales trend, and a period C (from January 1998 to December 2000) of a negative linear trend. Period D in Figure 3 comprises of the last year of data and was used to test the models in this section.

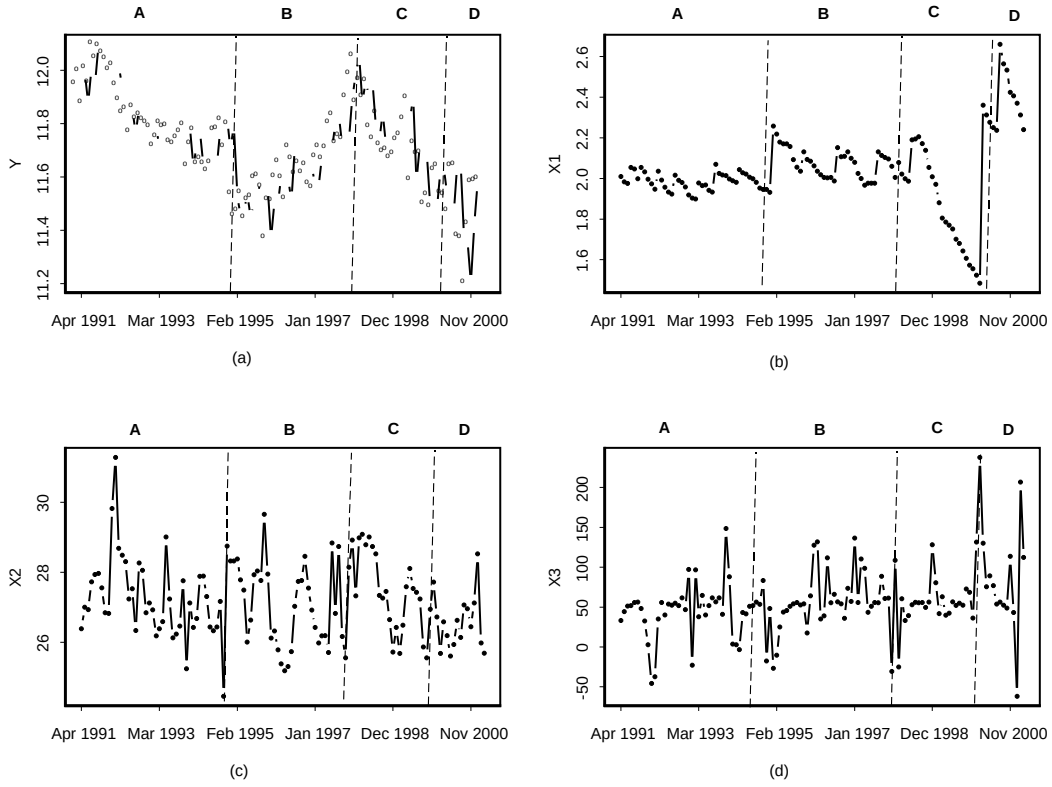


Figure 3: Deseasonalised series from April 1991 to March 2001 of (a) total beer sales ($Y$), (b) average deflated beer price per unit ($X_1$), (c) average maximum temperature ($X_2$) and (d) average rainfall ($X_3$).

Figure 3 (b), (c) and (d) shows the deseasonalised time series of beer price, maximum temperature and rainfall respectively.

## 5.1 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. Obviously the monthly (deflated)

average beer price ($X_1$) as well as the monthly average maximum temperature ($X_2$) were explanatory variables which were believed to influence the monthly average beer sales ($Y$) at any time period. However, each period A, B and C had other distinct explanatory factors believed influential only at those periods. Each formulated model is believed to 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 of a World Bank and International Monetary Fund (IMF) formulated economic structural adjustment programme (ESAP) implemented in Zimbabwe from 1990 to 1995 - see e.g. World Bank publications, 1996) which had a strong effect on increasing prices, but also by a drought which occurred between 1991 and 1992 and was characterised by a combination of low rainfall and high temperatures. The plausible model $\mathcal{M}_A$ formulated for period A's scenario included $X_1, X_2$ as well as monthly average rainfall ($X_3$) as explanatory variables. The regression vector was set as $\underline{F}'_A = (1, t, X_1, X_2, X_3)$ with the two initial terms (1 and $t$) used to model level and trend respectively. The initial mean and variance values were set as $(2, 0.2)$ for $X_1$, $(27, 2)$ for $X_2$ and $(50, 10)$ for $X_3$. The evolution matrix $\mathbf{G}_A$ was set as an $(5 \times 5)$ identity matrix. Discount factors were used to determine the observational ($V_a$) and the evolution ($\mathbf{W}_A$) variances. The observational variance was assumed unknown but constant was obtained by a discount factor of 0.99. The evolution variance was determined with the use of block component discount with the trend block of the variance having a factor of 0.99 and the regression component block having a 0.9 factor.

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 high temperatures, low rainfall and hence low crop sales following poor harvests. The model $\mathcal{M}_B$ for period B be included $X_1$, $X_2$, $X_3$ and total crop sales ($X_4$). The regression vector is $\underline{F}'_B = (1, t, X_1, X_2, X_3, X_5)$ with initial mean and variance values of $(2, 0.2)$, $(27, 2)$, $(50, 10)$ and $(6, 2)$ for $X_1, X_2, X_3, X_5$ respectively. The evolution matrix $\mathbf{G}_B$ was set as a $(6 \times 6)$ identity matrix. The observational variance, $V_B$, was set with the use of a discount factor of 0.99, while $\mathbf{W}_B$ had a trend block discount factor of 0.99 and a regression component discount factor of 0.9.

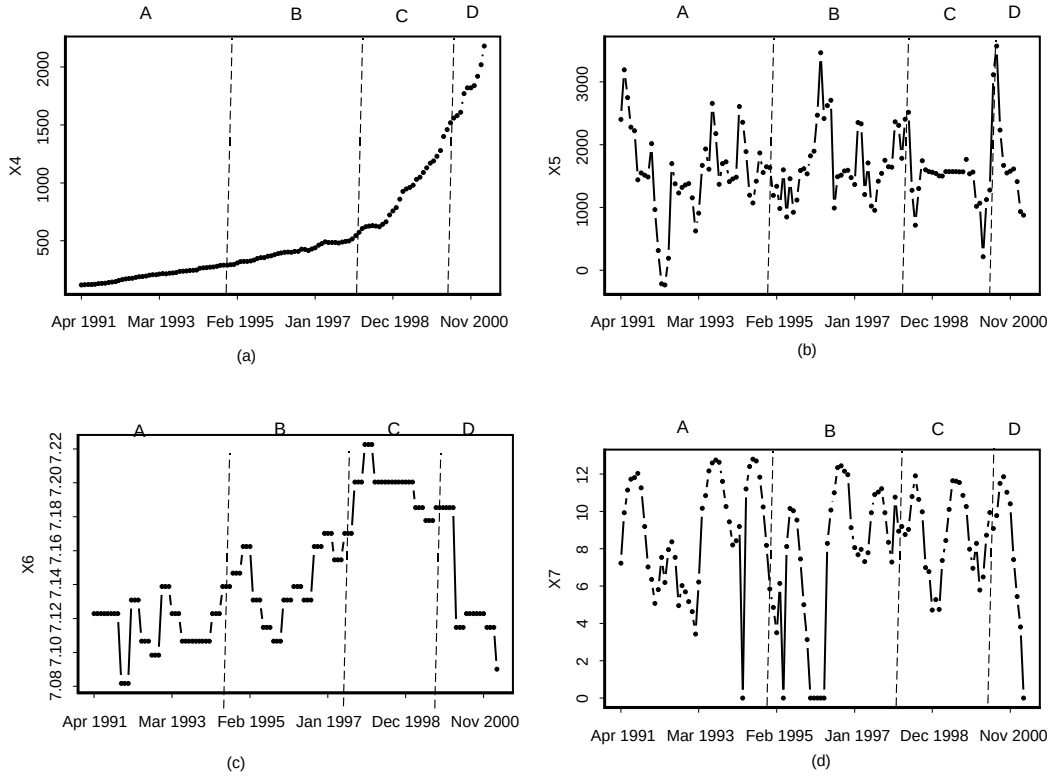


Figure 4: Transformed series from April 1991 to March 2001 of (a) CPI ($X_4$), (b) total crop sales ($X_5$), (c) employment ($X_6$) and (d) total maize production ($X_7$).

Period C had beer sales influenced by the well documented social, political and economical

events that occurred in Zimbabwe in the late nineties. In fact, two 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 began implementing the 1993 Land Designation Act which saw the redistribution of agricultural land with compensation covering buildings and infrastructure rather than land value. Those had an effect of significant output decrease on the commercial agricultural sector and related industries, Bond (1999) and Brett (2004). 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 $\mathcal{M}_C$ by: (i) the large increase in the (deseasonalised and log-transformed) consumer price index (CPI) ($X_4$); (ii) the decrease in (deseasonalised) total crop sales ($X_5$); (iii) the decrease in employment levels ($X_6$); as well as (iv) the decrease in (deseasonalised) total maize sales ($X_7$). Those explanatory variables can be seen in Figure 4(a), (b), (c) and (d) respectively.

The regression vector for model $\mathcal{M}_C$ was then set as $\underline{F}'_C = (1, t, X_1, X_2, X_4, X_5, X_6, X_7)$. The initial mean and variance values of $(2, 0.2)$, $(27, 2)$, $(100, 10)$, $(6, 2)$, $(7.1, 0.2)$ and $(8, 6)$ were chosen for $X_1, X_2, X_4, X_5, X_6$ and $X_7$ respectively. The evolution matrix $\mathbf{G}_C$ was set as a $(8 \times 8)$ identity matrix. A discount factor of 0.99 was chosen for the observational variance, $V_C$. The system variance, $\mathbf{W}_C$, had a trend block discount factor of 0.99 while the regression component factor was 0.9.

## 5.2 Predictive performances and losses

As mentioned above, each of the plausible RDLMs $\mathcal{M}_A$, $\mathcal{M}_B$ and $\mathcal{M}_C$ were formulated to represent a causal relationship (or scenario) prevalent in each period of time (A, B and

C) between major (structural) changes in the beer sales series behaviour. A fundamental condition in quantitative forecasting is that of continuity when (at least) some aspects of the (most recent) historical pattern (supposedly captured by the forecasting model) will continue in the forecasting horizon. Note that the approach we have adopted here (of giving weights for models $\mathcal{M}_A$, $\mathcal{M}_B$ and $\mathcal{M}_C$ and using their combination for forecasting) allows the BDM to explicitly model her beliefs about how past scenarios are likely to repeat in the forecasting period. This certainly can be extended for a scenario which did not happen yet but for which the BDM is able to formulate a Bayesian model. Predictive distributions from this model can also be included in the combination method.

Now, with the task of producing predictions for $h = 1, 2, \ldots, 12$ months ahead, in this section we will be interested in comparing the predictive performances of the linear and the geometric combinations of those models during period D (January to December 2001). 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 $\mathcal{M}_A$, $\mathcal{M}_B$ and $\mathcal{M}_C$. The initial required information for model estimation were automatically obtained by BATS with the use its reference initialisation. This feature uses some observed values to assign some 'reasonable' initial values to the model parameters.

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 its automatic 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 the forecasting period. The plot of Figure 5 shows the observed sales as well as the forecasts (means of predictive densities) for each of the component models

obtained by BATS.

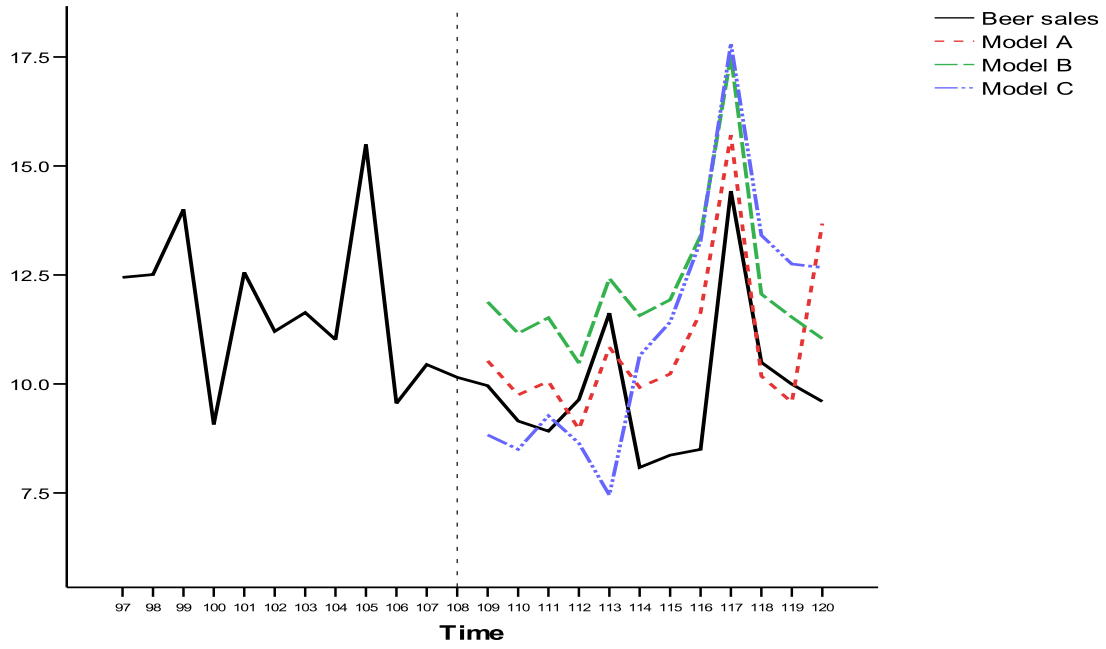


Figure 5: The observed sales (solid line) and the forecasts by $\mathcal{M}_A$ (dotted line), $\mathcal{M}_B$ (dash line) and $\mathcal{M}_C$ (dash-dotted line) made at $t = 108$ (March 2000) for period D (April 2000 to March 2001).

In cases where the combined density is multimodal (or skewed), the issue of what location and spread parameters to adopt by the BDM comes to light. In fact, the mean loses most of its usefulness as a descriptive statistics in such situations as it is expected to fall in the interval between the modes. 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 multimodal (or skewed) distributions.

As an example, the plot Figure (6) shows the one-step-ahead predictive densities, at time $t = 109$, for the linear and the geometric combination of the three formulated mod-

els ($\mathcal{M}_A, \mathcal{M}_B$ and $\mathcal{M}_C$) which had t-distributed predictive densities. All three component densities have 3 degrees of freedom, means $\mu_1 = 10.53$, $\mu_2 = 11.88$, $\mu_3 = 8.83$, and variances $\sigma_1^2 = 0.528$, $\sigma_2^2 = 0.424$ and $\sigma_3^2 = 0.815$. For the outperformance combination weights $w_1 = 0.348$, $w_2 = 0.275$ and $w_3 = 0.377$, the resulting LComb density has three modes at $y = 9.05$ , $y = 10.54$ and $y = 11.66$. The GComb in its turn has a slightly skewed but unimodal density with a single mode at $y = 10.51$. The observed beer sale value at $t = 109$ was 9.96.

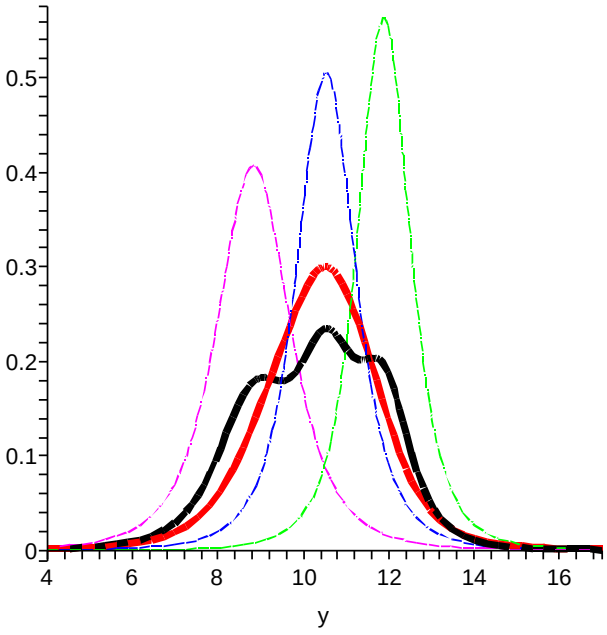


Figure 6: The one-step ahead predictive $t$-densities of $\mathcal{M}_A\mathcal{M}_B$ and $\mathcal{M}_C$ as well as the multimodal LComb and the skewed GComb models for April 2000.

The plot of Figure 7 shows the observed values of beer sales (dash-dotted line) from $t = 97$ (March 1999) to $t = 120$ (March 2001) as well as the means from the predictive densities of the best performing component model $\mathcal{M}_1$ (dashed line), as well as the (largest) modes from the predictive densities of both GComb (solid line) and LComb (dotted line). It can be

seen that in general all models overpredicted the sales most of the time with few expeptions. They all produced very similar short term forecasts during the initial four or five months of the forecasting period. After that, $\mathcal{M}_1$ produced more differentiated forecasts which were closest to the observed sales (except for $t = 120$) while both GComb and LComb continued to produce similar forecasts.

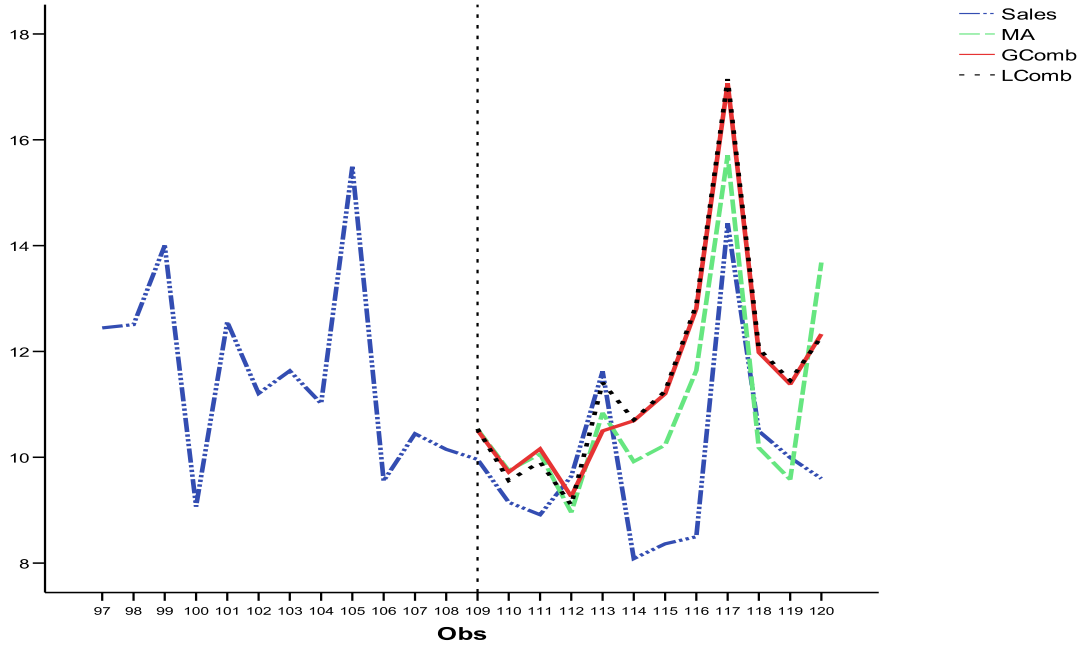


Figure 7: The observed sales (dash-dotted line) and the point forecasts from models $\mathcal{M}_A$ (dash line), LComb (solid line) and GComb (dotted line) in the forecasting period (from April 2000 to March 2001).

In cases where the combined density is multimodal (or skewed), the use of the density's mean as a location parameter in point forecasting loses most of its usefulness as a descriptive statistics. The modes themselves tend to coincide with means of the component distributions of the combination. Similarly, the density's variance fails to describe the peakedness (or

spread around the mean) of multimodal (or skewed) distributions. However, in a decision analysis context, where the BDM has a loss function associated with the consequences of her potential decisions, there may be a choice of location parameter which minimizes her expected loss.

Now, to verify the effects of different choices of location parameters for LComb and GComb we have calculated the cumulative losses associated with the quadratic, the logarithmic and the exponential loss functions under the mean and the largest mode. That is, for each forecasting period ($h = 1, \ldots, 12$), the losses incurred when the (i) mean and (ii) largest mode is adopted as point forecast under each loss were cumulatively computed over the 12 months period. The quadratic loss for each forecasting horizon $h$ was determined by the score function

$$L_Q(y_{T+h}, \hat{Y}_{t+h}) = (y_{T+h} - \hat{Y}_{t+h})^2$$

where $\hat{Y}_{T+h}$ is the chosen point forecast (mean or largest mode) for $y_{T+h}$ made at time $T = 108$ by each combination model. Similarly, the logarithmic loss was computed by

$$L_L(y_{T+h}, \hat{Y}_{t+h}) = \log(y_{T+h}/\hat{Y}_{T+h})$$

and the exponential loss by:

$$L_E(y_{T+h}, \hat{Y}_{T+h}) = 1 - \exp\{-\frac{1}{2}(y_{T+h} - \hat{Y}_{T+h})^2\} \quad .$$

The exponential loss function is a form of the conjugate (to Gaussian models) loss function advocated by Lindley (1976) and is related to choices of risk-averse utilities.

Table 1 shows the cumulative losses during the forecasting period for the best performing component model $\mathcal{M}_1$ as well as GComb and LComb when means, modes and largest modes were selected as point estimates (as indicated in the columns headings).

| Loss | $\mathcal{M}_1$ mean | $P_G$ mode | $P_L$ largest mode | $P_L$ mean |
|---|---|---|---|---|
| Quadratic | 38.35 | 55.59 | 55.40 | 56.23 |
| Logarithmic | -0.51 | -0.77 | -1.51 | -0.79 |
| Exponential | 5.59 | 7.55 | 11.54 | 7.04 |

Table 1: The cumulative quadratic, logarithmic and exponential losses by $\mathcal{M}_1$, LComb and GComb for different types of location parameters (means and largest modes) of their predictive densities from April 2000 to March 2001.

Note that $\mathcal{M}_1$ was the best model under any of the three types of loss. The remaining component models $\mathcal{M}_2$ and $\mathcal{M}_3$ had far worse performances overall and for simplicity were omitted from the table. This was in part surprising as $\mathcal{M}_3$ which represented the best scenario for period C was in principle expected to have performed better. However, this can be explained (at least in part) by the fact that it included explanatory variables such as the CPI which had a dramatic increase in period C, inducing $\mathcal{M}_3$ to predict far too low sales for period D. Also, by being the largest model of the three with six explanatory variables which values for period D were obtained from individual forecasting models, produced predictive densities with the largest variances of all component models.

LComb's largest mode produced the second best result under the quadratic loss (but only marginally better than GComb's mode). That changed under the logarithmic and the exponential losses with both LComb's mean and GComb's mode producing far better results. LComb's mode was the second best predictor under the logarithmic loss (but only marginally better than LComb's mean) and only third best under the exponential loss.

Those results show that while the largest mode seems to be a better choice of point pre-

dictor for LComb under quadratic loss, the mean seems a better choice under the logarithmic and exponential losses. That exemplifies how the choice of loss function can have an effect not only on what combination approach to adopt but also on the type of location parameters to choose as point forecasts in a decision making situation.

# 6 Conclusion

We have proposed the geometric combination of probability density functions as an alternative approach to both model selection and linear combination. Model combination has the advantage over model selection in that uncertainties about the individual models can be explicitly accounted for.

We have shown that in situations where unimodality of the resulting combined density is desired (e.g. in decision making situations when a single estimate of the location parameter is required), the geometric combination may be a preferable choice over its linear counterpart. In particular, this is the case when the models to be combined are from the regular exponential family. The linear combination model is typically multimodal in this case. For component models with densities from the Student $t$-distribution, conditions for unimodality of the density from the geometric combination have been established for the case of two component models. These conditions are less stringent than those for the linear approach. Therefore, for the same component models, unimodality under the geometric approach is obtained for larger distances between the means (relative to the variances) than under the linear approach.

Another advantage of geometric combinations are that they are externally Bayesian. So, whether to combine before or after new data become available is irrelevant. In fact, for such

combinations, Bayes theorem applied to combined prior densities will give the same result as combining the posterior densities themselves. When component densities 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 densities be combined first.

The application to the forecasting of beverage sales in Zimbabwe where three models with Student t predictive densities were combined, showed that the mode of the geometric combination performed well in general when chosen as point forecast. Under logarithmic loss, the choice of the geometric combination mode as point prediction was supported in this application. The largest mode of the linear combination density would be the preferred choice only marginally under the quadratic loss but not under the logarithmic and the exponential losses. In those cases the mean would be preferable to the largest mode.

The sales series presented three major (abrupt) changes in trend during its time span. Each component model was formulated to represent the causal relationship (or scenario) believed to have prevailed at each period of change in the trend. It is well known that a fundamental condition in quantitative forecasting is that of continuity when (at least) some aspects of the (most recent) historical pattern (supposedly captured by the forecasting model) will continue during the forecasting horizon. However, the combination approach we have adopted allows the BDM to explicitly model her beliefs about how past scenarios are likely to repeat in the forecasting period. Those beliefs when assessed as subjective probabilities can be used as the combining weights. Certainly, any other scenario which did not occur in the past but for which the BDM is able to formulate a statistical model can be included in the combination.

Although our application considered only Bayesian time series forecasting models the theoretical results here apply to any parametric statistical models. Further research directions include the investigation of other types of distributions including discrete ones, and an extension to multivariate response models. The combination of models with densities from different families is also worth investigating.

# References

Barndorff-Nielsen, O. E. (1973), Unimodality and exponential families. *Comm. Statist.*, **1**, 189–216)

Bates, J. M. & Granger, C. W. J. (1969), The Combination of Forecasts. *Operations Research Quartely*, **20**, 451–468.

Behboodian, J. (1970), On the Modes of a Mixture of Two Normal Distributions. *Technometrics*, **12**, 131–139.

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.

Central Statistical Office. (1991 to 2001), *Quarterly Digest of Statistics*, Government Printers.

Cobb, L. (1978), Stochastic catastrophe models and multimodal distributions. *Behavioural Science*, **23**, 360-374.

Cooke, R. M. (1991), *Experts in uncertainty : expert opinion and subjective probability in science*, Oxford University Press, New York.

DeGroot, M. H. (1974), Reaching a consensus. *jour J. Amer. Statist. Assoc.*, **69**, 118-121.

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, University of Warwick.

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

Lindley, D. (1976) A class of utility functions, *Annals of Statistics*, **4**, 1-10.

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

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

Poston, T. and Stewart, I. N. (1978), *Catastrophe Theory and its Applications.* Pitman, London.

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.

Smith, A. F. M. & Makov, U. E. (1978), A Quasi-Bayes procedure for Mixtures. *Journal of the Royal Statistical Society B*, **40**, 106-112.

Smith, J. Q. (1978), Problems in Bayesian statistics relating to discontinuous phenomena, catastrophe theory and forecasting. *PhD thesis*, University of Warwick.

Smith, J. Q. (1979), Mixture catastrophes and Bayes decision theory. *Mathematical Proceedings of the Cambridge Philosophical Society*, **86**, 91–101.

Smith, J. Q., Harrison, P. J. and Zeeman, E. C. (1981), The analysis of some discontinuous decision processes. *European Journal of Operational Research*, **7**, 30-43.

Tiao, G. C. and Zellner, A. (1964), Bayes Theorem and the use of Prior Knowledge in Regression Analysis. *Biometrika*, **51**, 219–230.

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.

Winkler, R. L. (1968), The Consensus of Subjective Probability Distributions. *Management Science*, **15**, 61–75.

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

Zimbabwe Department of Meteorological Services. (June 1989 to June 2002), *Zimbabwe Ministry of Transport and Communications*.