Network

Rimini Centre for Economic Analysis
Working Paper Series

wp 20-27

A Bayesian Dynamic Compositional Model for Large Density Combinations in Finance

Roberto Casarin
Stefano Grassi
Francesco Ravazzolo
Herman K. van Dijk

A Bayesian Dynamic Compositional Model for Large Density Combinations in Finance[*]

Roberto Casarin[†] Stefano Grassi[‡]
Francesco Ravazzolo[§] Herman K. van Dijk[¶]

[†]University Ca' Foscari of Venice
[‡]University of Rome Tor Vergata
[§]Free University of Bozen-Bolzano, BI Norwegian Business School and RCEA
[¶]Econometric Institute, Erasmus University Rotterdam, Norges Bank
and Tinbergen Institute

November 19, 2020

Abstract

A Bayesian dynamic compositional model is introduced that can deal with combining a large set of predictive densities. It extends the mixture of experts and the smoothly mixing regression models by allowing for combination weight dependence across models and time. A compositional model with Logistic-normal noise is specified for the latent weight dynamics and the class-preserving property of the logistic-normal is used to reduce the dimension of the latent space and to build a compositional factor model. The projection used in the dimensionality reduction is based on a dynamic clustering process which partitions the large set of predictive densities into a smaller number of subsets. We exploit the state space form of the model to provide an efficient inference procedure based on Particle MCMC. The approach is applied to track the Standard & Poor 500 index combining 3712 predictive densities, based on 1856 US individual stocks, clustered in relatively small number of model sets. For the period 2007-2009, which included the financial crisis, substantial predictive gains are obtained, in particular, in the tails using Value-at-Risk. Similar predictive gains are obtained for the US Treasury Bill yield using a large set of macroeconomic variables. Evidence obtained on model set incompleteness and dynamic patterns in the financial clusters provide valuable signals for improved modelling and more effective economic and financial decisions.

JEL codes: C11, C15, C53, E37.
Keywords: Density Combination, Large Set of Predictive Densities, Compositional Factor Models, Nonlinear State Space, Bayesian Inference.

[*]The present paper should not be reported as representing the views of Norges Bank. The views expressed are those of the authors and do not necessarily reflect those of Norges Bank. Roberto Casarin's research is supported by the Venice center of Economic and Risk Analytics. Stefano Grassi gratefully acknowledges financial support from the University of Rome 'Tor Vergata' under Grant "Beyond Borders" (CUP: E84I20000900005). This paper is part of the research activities at the Centre for Applied Macroeconomics and Commodity Prices (CAMP) at the BI Norwegian Business School.

1 Introduction

Predicting with large sets of data involving many model structures and explanatory variables is a topic of substantial interest to academic researchers as well as to professional and applied forecasters. It has been studied in several papers (e.g., see Stock and Watson, 1999, 2002, 2005, 2014, and Bańbura et al., 2010). The recent fast growth in (real-time) big data allows researchers to predict variables of interest more accurately (e.g., see Choi and Varian, 2012; Varian, 2014; Varian and Scott, 2014; Einav and Levin, 2014). Stock and Watson (2005, 2014), Bańbura et al. (2010) and Koop and Korobilis (2013) suggest that there are also potential gains from predicting using a large set of predictors.
However, predicting with large data sets, many predictors and high-dimensional models requires new modelling strategies, efficient inference methods and extra computing power possibly resulting from parallel computing. We refer to Granger (1998) for an early discussion of these issues.

We propose a Bayesian dynamic compositional model which deals with the combination of a large set of predictive densities using financial data. It extends Billio et al. (2013) and McAlinn and West (2019) in several directions.

In terms of methodology we introduce three innovations. First, we use the mixture o experts and/or smoothly mixing regression approaches (Jacobs et al., 1991, Jordan and Jacobs, 1994, Jordan and Xu, 1995, Peng et al., 1996, Wood et al., 2002, Geweke and Keane, 2007, Villani et al., 2009) and extend these by allowing the combination weights to be dependent between models as well as to learn over time. Learning about model set incompleteness is also specified. In this context a diagnostic analysis is presented to signal particular types of missing information.

Second, a dimension reduction of the latent weight variables is introduced by making use of the class-preserving property of the logistic-normal distribution. The dimension reduction involves modelling the combination weights of the large set of densities as a dynamic factor model with a small number of factors. The projection onto a low dimension latent space uses a dynamic clustering process that allocates the predictive densities into mutually exclusive groups. We contribute to the literature on modelling variables (Aitchinson and Shen, 1980; Aitchinson, 1982, e.g., see) and time-series on a bounded domain (e.g., see Wallis, 1987; Quintana and West, 1988; Grunwald et al., 1993; Cargnoni et al., 1997; Brunsdon and Smith, 1998; Dey et al., 2001; Kynclova et al., 2015; Snyder et al., 2017; Boonen et al., 2019).

Third, an efficient simulation-based Bayesian inferential procedure is derived. Given that the model can be represented as a nonlinear state space form where the measurement equation consists of a large finite mixture, Sequential Monte Carlo is used for efficient posterior approximation.[1] Also, we propose a Bayesian diagnostic analysis of the model set incompleteness and use De Finetti's diagrams (Ehm et al.,

[1] We implement parallel sequential filtering and clustering to exploit the computing power of graphics processing units (GPU).

2016) to study the evolution patterns in the model weights.

Using large financial data sets, the proposed approach is applied to two well-known problems in finance. In the first example we use 3712 predictive densities based on 1856 US individual stock return series and four clusters to construct a combined predictive density of a replication of S&P 500 returns over the sample 2007-2009, which includes the turbulence of the financial crisis. We estimate several features of this density, emphasizing that our method allows for a time-varying composition of the four clusters letting individual stocks to switch across them or eventually exit the model set, for example, after a default as in the Lehman Brothers case. Compared to the no-prediction ability benchmark and predictions from individual models estimated on the aggregate index, we find substantial accuracy gains in predicting means, volatilities and tail events, in particular, with respect to the economic value of such events like Value-at-Risk. Measures of model set incompleteness and dynamic patterns in the cluster-based weights provide valuable signals for improved economic and financial modelling and policy analysis.

In the second example, we consider for our predictive purposes the 3-month Treasury Bill yield series from the extended Stock and Watson (2005) dataset for the period 1959Q1 to 2011Q2. Assuming the existence of 5-7 clusters, we identify two clusters related to real activities; one cluster related to prices; and one cluster related to financial variables. The other clusters contain the remaining series. We find substantial gains in joint density predictions of 3-month Treasury Bill yield over the last 25 years for all horizons from one-quarter ahead to five-quarters ahead. The highest accuracy is achieved when the series is predicted using our combination schemes with cluster weights based on log score learning. A dominant cluster does not exist but we note that the cluster that includes Exports, Imports and GDP deflator receives a relatively large weight. Diagnostic analysis provides valuable signals that additional gains may be obtained with a better model set specification, more detailed cluster grouping and different learning rules for weights. This is left as a topic for further research.

The contents of this paper is structured as follows. Section 2 provides details of the methodological contributions of our approach. Section 3 describes the sequential inference. Section 4 contains novel empirical applications using a large set of US stocks and bond data. Section 5 presents conclusions and suggestions for further research.

2 A dynamic compositional combination model for large sets of predictive densities

In this section we present a new compositional combination model which makes use of dynamic mixture processes in order to deal with large sets of predictive densities. For a recent survey about the evolution of predictive density combinations, see Aastveit et al. (2019) and for background see Billio et al. (2013), McAlinn and West

(2019) and Bastuerk et al. (2019). When the number of models or experts is large the dimension of the latent space increases and overfitting issues can jeopardize the validity of the empirical analysis. We propose a dimensionality reduction strategy based on projections in the latent space and on dynamic clustering.

2.1 A dynamic mixture of density convolutions

A basic probabilistic approach to combine predictive information from different sources proceeds as follows. Let $\mathcal{I}_t$ be an information set at time t and $\mathcal{M} = \{M_1, \dots, M_n\}$ a set of models or experts. In the mixture of experts literature the conditional predictive probability density $f(y_t|\mathfrak{I}_{t-1}, \mathcal{M})$ of an economic variable of interest y_t is specified as a discrete mixture of conditional predictive probabilities of y_t coming from individual models, or experts $M_i \in \mathcal{M}$ with information sets $\mathfrak{I}_{i,t-1} \subset \mathfrak{I}_{t-1}$. The predictive distribution is the convex combination of predictive distributions with density given as:

$$f(y_t|\mathfrak{I}_{t-1}, \mathcal{M}) = \sum_{i=1}^{n} w_{it} f(y_t|\mathfrak{I}_{i,t-1}, M_i) \tag{1}$$

where $w_{it}, i = 1, \dots, n$ are the mixture weights such that $0 \leq w_{it} \leq 1$, $w_{1t} + \ldots + w_{nt} = 1$.

In this paper, we assume y_t follows a discrete random probability measure $G(\cdot)$ over the set of predictors from the models M_i, $i = 1, \dots, n$. The random measure is defined as:

$$G(y_t) = \sum_{i=1}^{n} \delta(a_{it} - y_t) w_{it} \tag{2}$$

with conditionally independent random atoms $a_{it} = \tilde{y}_{it} + \varepsilon_{it}$ $i = 1, \dots, n$ and possibly dependent random probability weights w_{it} $i = 1, \dots, n$, where $\delta(\cdot)$ denotes the Dirac delta.[2] We denote with $H_{0t} = H_{0t}^{W} \otimes H_{0t}^{A}$ the product distribution of the sequences of the atoms $a_{it}, i = 1, \dots, n$ and of the weights $w_{it}, i = 1, \dots, n$.

The first component of the atom $\tilde{y}_{it}$ follows the predictive density of the model M_i, i.e.

$$\tilde{y}_{it} \sim f(\tilde{y}_{it}|\mathfrak{I}_{i,t-1}, M_i) \tag{3}$$

as in standard mixture of expert models.

The random variable ε_{it}, being the difference between y_t and $\tilde{y}_{it}$, points towards two error sources. There may be predicting errors due to, for instance, sudden shocks in the series and there may be misspecification errors due to model set incompleteness. In this paper we focus on the latter, that is, a larger specification error in model M_i implies a larger error ε_{it}. Investigating the relative importance of a predicting error component is a topic for further research. In the following we assume the probability density functions of $\varepsilon_{it}, i = 1, \dots, n$

$$\varepsilon_{it} \sim g(\varepsilon_{it}|\sigma_{it}^2) \tag{4}$$

[2] We recall that $\delta(a - b) = 1$ if $a = b$ and zero otherwise.

are parametrized by the scaling process σ_{it}^2 which controls for the level of uncertainty due to misspecification. In summary the distribution H_{0t}^A has density given by

$$\prod_{i=1}^{n} \int_{\mathbb{R}} f(a_{it} - \varepsilon_{it} | \mathfrak{I}_{i,t-1}, M_i) g(\varepsilon_{it} | \sigma_{it}^2) d\varepsilon_{it} \tag{5}$$

As stated in the following, under these assumptions the predictive density combination model becomes a random finite mixture of convolutions of densities $f(y_t|\mathfrak{I}_{t-1}, \mathcal{M}, \sigma_t^2)$ where the process $\sigma_t^2 = \{\sigma_{1t}^2, \ldots, \sigma_{nt}^2\}$ controls for the overall uncertainty level about the predictive models used in the combination.

Proposition 2.1. *Let $f(y_t|\mathfrak{I}_{i,t-1}, M_i, \sigma_{it}^2)$ be a convolution of two densities:*

$$f(y_t|\mathfrak{I}_{i,t-1}, M_i, \sigma_{it}^2) = \int_{\mathbb{R}} f(y_t|\tilde{y}_{it}, \sigma_{it}^2) f(\tilde{y}_{it}|\mathfrak{I}_{i,t-1}, M_i) d\tilde{y}_{it} \tag{6}$$

where $f(y_t|\tilde{y}_{it}, \sigma_{it}^2) = g(y_t - \tilde{y}_{it}|\sigma_{it}^2)$, then integrating out the random atoms of $G(y_t)$ with respect to the measure H_{0t}^A one obtains a predictive density combination model

$$f(y_t|\mathfrak{I}_{t-1}, \mathcal{M}, \sigma_t^2) \;=\; \sum_{i=1}^{n} w_{it} f(y_t|\mathfrak{I}_{i,t-1}, M_i, \sigma_{it}^2) \tag{7}$$

When the uncertainty level tends to zero, $\max\{\sigma_{it}, i = 1, \ldots, n\} \to 0$ then $g(y_t - \tilde{y}_{it}|\sigma_{it}^2) \to \delta(y_t - \tilde{y}_{it})$ and one recovers the standard mixture of expert models in Eq. (1), i.e. $f(y_t|\mathfrak{I}_{t-1}, \mathcal{M}, \sigma_t^2) \to f(y_t|\mathfrak{I}_{t-1}, \mathcal{M})$.

In the next subsection we specify a stochastic process for the latent weights w_{it} with conditional distribution H_{0t}^W.

2.2 A compositional weight process

We represent the combination weights in our dynamic mixture model as a stochastic process on the simplex with logistic-normal noise. Exploiting the class-preserving property of the logistic-normal distribution, we project the weights onto a lower dimensional simplex while preserving their probabilistic properties.

In the following, we introduce some useful notation, definitions and results. For convenience, we omit the subindex t here. Let $\mathbb{R}_+^m$ be the positive orthant of $\mathbb{R}^m$ we introduce the m-dimensional standard simplex $\mathbb{S}^m = \{\mathbf{z} \in \mathbb{R}_+^m | z_1 + \ldots + z_m = 1\}$ and $\mathbb{V}^m = \{\mathbf{v} \in \mathbb{R}^m : v_1 + \ldots + v_m = 0\}$ subspaces of $\mathbb{R}^m$ of dimension $m-1$. We denote with $\boldsymbol{\iota}_m$ and I_m the m-dimensional unit vector and m-dimensional identity matrix, respectively and define the $((m-1) \times m$ matrix $D_m = (I_{m-1}, -\boldsymbol{\iota}_{m-1})$, and the $(m-1)$-dimensional square matrix $H_m = (I_{m-1} + \boldsymbol{\iota}_{m-1}\boldsymbol{\iota}_{m-1}')$.

Definition 2.1 (Composition function)**.** *The composition function is defined as $C_m(\mathbf{u}) : \mathbb{R}_+^m \to \mathbb{S}^m$, $\mathbf{u} \mapsto \mathbf{z} = C_m(\mathbf{u})$ where the i-th element of $\mathbf{z}$ is $z_i = u_i/z_m$, $i = 1, \ldots, m$, and $z_m = \mathbf{u}'\boldsymbol{\iota}_m$.*

In the following we denote with $\exp(\mathbf{x}) \in \mathbb{R}_+^m$ the component-wise exponential of $\mathbf{x} \in \mathbb{R}^m$, with $\log(\mathbf{v}) \in \mathbb{R}^m$ the component-wise logarithmic transforms of $\mathbf{v} \in \mathbb{R}_+^m$ and with $\mathbf{u} \circ \mathbf{v} \in \mathbb{R}^m$ the Hadamard product between vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$.

We assume the simplex space $\mathbb{S}^m$ is equipped with the following operations.

Definition 2.2 (Operations on the simplex)**.** *Let* $\mathbf{z}, \mathbf{v} \in \mathbb{S}^m$ *two elements of the simplex,* $a \in \mathbb{R}$ *a coefficient and* $\circ$ *the Hadamard element-wise product between vectors, then*

$$\mathbf{z} \oplus \mathbf{v} = C_m(\mathbf{u} \circ \mathbf{z}) \tag{8}$$

is a sum operation also called perturbation operation and

$$a \odot \mathbf{z} = C_m((z_1^a, \ldots, z_m^a)') \tag{9}$$

is a scalar product operation also called power transform.

For details and background, see Aitchinson (1986) and Aitchinson (1992). Billheimer et al. (2001) showed that $\mathbb{S}^m$ equipped with the perturbation and powering operations is a vector space. Moreover $\mathbb{S}^m$ is an Hilbert space, i.e. a complete, inner product vector space.

Random combination weights with values on the simplex can be defined by introducing a probability space with support on the simplex $\mathbb{S}^n$. A probability measure on the simplex can be obtained by introducing a system of coordinates and a probability distribution on the coordinates. The following transformations can be used to define different systems of coordinates (Egozcue et al., 2003).

Definition 2.3 (Logratio transformations)**.** *Given* $\mathbf{z} \in \mathbb{S}^m$*, we define the additive (alr), centered (clr) and isometric (ilr) logratio transformations:*

$$\begin{aligned} \mathrm{alr}(\mathbf{z}) &= \log(\mathbf{z}_{-m}/z_m) & (10)\\ \mathrm{clr}(\mathbf{z}) &= \log(\mathbf{z}/g(\mathbf{z})) & (11)\\ \mathrm{ilr}(\mathbf{z}) &= (D_m V)^{-1} D_m \log(\mathbf{z}) & (12) \end{aligned}$$

where $V = (\mathbf{v}_1, \ldots, \mathbf{v}_m)$ *is an orthonormal basis of* $\mathbb{V}^m$*,* $g(\mathbf{z}) = (z_1 z_2 \ldots z_m)^{1/m}$ *is the geometric mean,* $\mathbf{z}_{-m} = (z_1, \ldots, z_{m-1})$ *a subvector of* $\mathbf{z}$*.*

The clr$(\cdot)$ transformation is an one-to-one map from $\mathcal{S}^m$ to $\mathbb{V}_m$ subspace of $\mathbb{R}^m$, whereas the alr$(\cdot)$ and ilr$(\cdot)$ are one-to-one maps from $\mathbb{S}^m$ to $\mathbb{R}^{m-1}$. The inverse transformations are $\mathrm{clr}^{-1}(\mathbf{x}) = C_m(\exp(\mathbf{x}))$, $\mathrm{alr}^{-1}(\mathbf{x}) = C_m((\exp(\mathbf{x}_{-m}), 1))$ and $\mathrm{ilr}^{-1}(\mathbf{x}) = C_m(\exp(\mathbf{x}\Psi))$ where Ψ is a $((m-1) \times m)$-dimensional contrast matrix such that $\Psi\Psi' = I_m - m^{-1}\boldsymbol{\iota}_m\boldsymbol{\iota}_m'$. The following simple example illustrates the composition of alr$(\cdot)$ with alr$^{-1}(\cdot)$ which are the main transformations used in this paper.

Example 2.1. Let $\mathbf{z} = (z_1, z_2)$ be a vector in $\mathbb{S}^2$ then $\mathbf{x} = \mathrm{alr}(\mathbf{z})$ is in $\mathbb{R}$ with element $x_1 = \log(z_1/z_2)$. If we apply the inverse transformation we obtain a vector $\mathbf{v} = \mathrm{alr}^{-1}(\mathbf{x})$ in $\mathbb{S}^2$ with elements $v_1 = \exp(x_1)/(1 + \exp(x_1)) = z_1/(1 + z_2) = z_1$ since $z_2 = 1 - z_1$.

Note that it always possible to change the representation coordinates thanks to the following relationships: $\mathrm{clr}(\mathbf{z}) = \mathrm{D}_\mathrm{m}'\mathrm{H}_\mathrm{m}^{-1}\mathrm{alr}(\mathbf{z})$ and $\mathrm{clr}(\mathbf{z}) = (\mathrm{D_m V})^{-1}\mathrm{alr}(\mathbf{z})$ where $\mathbf{x} \in \mathbb{S}^m$, (see Pawlowsky-Glahn and Buccianti, 2011, pp. 102-103).

A probability distribution on $\mathbb{S}^m$ can be defined by assuming a normal distribution for the elements of the logratio transformations in Definition 2.3. All coordinate systems provide the same probability distribution, called logistic-normal, but its parametrization and reference measure will depend on the system chosen (see Pawlowsky-Glahn and Buccianti, 2011, ch. 7). Each coordinate system has its own drawback, the alr$(\cdot)$ depends on the choice of the coordinate used in the common denominator, ilr depends on the choice of the orthonormal basis V and the clr returns a singular distribution.

We choose the following definition of logistic-normal distribution (see Aitchinson and Shen, 1980; Aitchinson, 1982, 1986) induced by the alr$(\cdot)$ transformation because it exhibits some useful properties for modelling purposes. We exploit the fact that the logistic-normal family is closed under perturbation and power transformations to propose a random weight process. The resulting family of compositional processes is invariant under the subcomposition and amalgamation operations which will be used to provide a graphical representation of high dimensional weight vectors.

Definition 2.4 (Logistic-normal distribution)**.** *The random vector* $\mathbf{z} \in \mathcal{S}^m$ *follows a logistic-normal distribution* $\mathcal{L}_m(\boldsymbol{\mu}, \Upsilon)$ *if its probability density function is*

$$p(\mathbf{z}|\boldsymbol{\mu},\Upsilon) = |2\pi\Upsilon|^{-1/2}\left(\prod_{j=1}^{m} z_j\right)^{-1} \exp\Bigg(-\frac{1}{2}\left(\log(\mathbf{z}/z_m)-\boldsymbol{\mu}\right)' \\ (\Upsilon)^{-1}\left(\log(\mathbf{z}_{-m}/z_m)-\boldsymbol{\mu}\right)\Bigg) \tag{13}$$

with parameters $\boldsymbol{\mu} \in \mathbb{R}^{m-1}$ *and* Υ $(m-1)$*-dimensional positive symmetric matrix, where* $z_m = 1 - \mathbf{z}'_{-m}\boldsymbol{\iota}_{m-1}$.

As stated in the following the logistic-normal distribution is related to the normal and the log-normal distributions.

Proposition 2.2. *Let* $\mathbf{v} \sim \mathcal{N}_m(\boldsymbol{\mu},\Upsilon)$*, and define* $\mathbf{u} = \exp(\mathbf{v})$ *and* $\mathbf{z} = \mathrm{alr}^{-1}(\mathbf{v})$*. Then* $\mathbf{u}$ *follows the m-variate log-normal distribution,* $\Lambda_m(\boldsymbol{\mu},\Upsilon)$*, and* $\mathbf{z}$ *follows a m-variate logistic-normal distribution,* $\mathcal{L}_m(D_m\boldsymbol{\mu}, D_m\Upsilon D_m')$*.*

Proof See Appendix A.

The following example illustrates the inverse relationship between the logistic-normal and the normal distribution.

Example 2.2. Let $\mathbf{z} = (z_1, z_2)$ be a vector in $\mathbb{S}^2$ with distribution $\mathcal{L}_2(D_2\boldsymbol{\mu}, D_2\Upsilon D_2')$ and apply the transformations used used in Example 2.1. The (i,j)-th element of the Jacobian of $\mathbf{z} = \mathrm{alr}^{-1}(\mathbf{u})$ is $\partial_j z_i = u_i\delta(i-j) - u_i u_j$ and the Jacobian is $|J| = u_1 u_2$ and $\log(z_1/z_2) = \log(\exp(u_1)/(z_2(1+\exp(u_1)))$ with $z_2 = 1-\exp(u_1)/(1+\exp(u_1))$. It follows that by substituting $\log(z_1/z_2) = u_1$ in the density of $\mathbf{u}$ and by multiplying the by Jacobian, the distribution of $\mathbf{z}$ is $\mathcal{N}(D_2\boldsymbol{\mu}, D_2\Upsilon D_2)$. Top plots in Figure 1 show the density function of $\mathcal{L}_2(\mu, v^2)$ for different values of μ and v^2.

Distributions other than the logistic-normal can be used for weights such as the Dirichlet distribution, but as noted in Aitchinson and Shen (1980) this distribution may be too simple to be realistic in the analysis of compositional data since the components of a Dirichlet composition have a correlation structure determined solely by the normalization operation in the composition. Extensions of the logistic-normal distribution can be considered such as the logistic skew-normal (Mateu-Figueras et al., 2005) or the logistic Student-t (Aitchinson and Shen, 1980; Katz and King, 1999), but we leave this topic for future research.

Another relevant property of the logistic-normal distribution is the class-preserving property of the composition of the logistic-normal vectors (see Aitchinson and Shen, 1980). This property is used to build a stochastic process in the simplex with logistic-normal noise.

Proposition 2.3 (Class-preserving property)**.** *Let A a $(c \times d-1)$ matrix and $\mathbf{z} \sim \mathcal{L}_d(\boldsymbol{\mu}, \Upsilon)$ a logistic-normal vector. Define the following transform $\mathbf{w} = \phi_A(\mathbf{z})$ from $\mathbb{S}^d$ to $\mathbb{S}^c$, with $d < c$,*

$$w_i = \prod_{j=1}^{d-1} \left(\frac{z_j}{z_d}\right)^{a_{ij}} \left(1 + \sum_{i=1}^{c} \prod_{j=1}^{d-1} \left(\frac{z_j}{z_d}\right)^{a_{ij}}\right)^{-1} \tag{14}$$

$i = 1, \ldots, c$, then $\mathbf{w} = (w_1, \ldots, w_c)'$ follows the logistic-normal $\mathcal{L}_{c+1}(A\boldsymbol{\mu}, A\Upsilon A')$.

Proof See Appendix A.

We provides in the following the statistical interpretation of the coefficients A and illustrates how this property can be used to define dependent variables through transformation of a common latent factor.

Example 2.3. Let $\mathbf{z} = (z_1, z_2)$ be a vector in $\mathbb{S}^2$ with distribution $\mathcal{L}_2(0, \upsilon^2)$ and let $A = (1, a)'$ a (2×1) matrix then the random vector $\mathbf{w} = (w_1, w_2, w_3)$ with $w_1 = (z_1/z_2)/((z_1/z_2)+(z_1/z_2)^a+1)$, $w_2 = (z_1/z_2)^a/((z_1/z_2)+(z_1/z_2)^a+1)$ and $w_3 = 1-w_1-w_2$, follows a logistic-normal $\mathcal{L}_2(0, \upsilon^2(1,a)'(1,a))$. The covariance between w_1 and w_2 is defined as $\mathbb{C}ov(w_1, w_2) = \mathbb{C}ov(\log(w_1/w_3), \log(w_2/w_3)) = \upsilon^2 a$. This provides an interpretation of the coefficient matrix A appearing in Proposition 2.3.

Compositions are usually represented by means of the De Finetti's, or ternary, diagram. A point in the diagram has coordinates $(z, v)' \in \mathbb{R}^2$ given by the following map from $\mathbb{S}^3 \mapsto \mathbb{R}^2$

$$(z, v) = (Bz_1, Cz_2, Az_3) \tag{15}$$

with vertices $A = (z_0 + 0.5, v_0 + \sqrt{3}/2)'$, $B = (z_0, v_0)'$ and $C = (z_0 + 1, v_0)'$ where $(z_0, v_0)'$ are coordinates arbitrarily chosen. See Cannings and Edwards (1968) and Pawlowsky-Glahn et al. (2015) for further details. In this diagram the weights $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$, correspond to the vertices B, C and A, respectively. The black square corresponds to the barycentre of the triangle $(1/3, 1/3, 1/3)$. The last line in Fig. 1 provides the De Finetti's diagrams of the relationships given in the Example 2.3 and rappresented in the second line of the same plot.

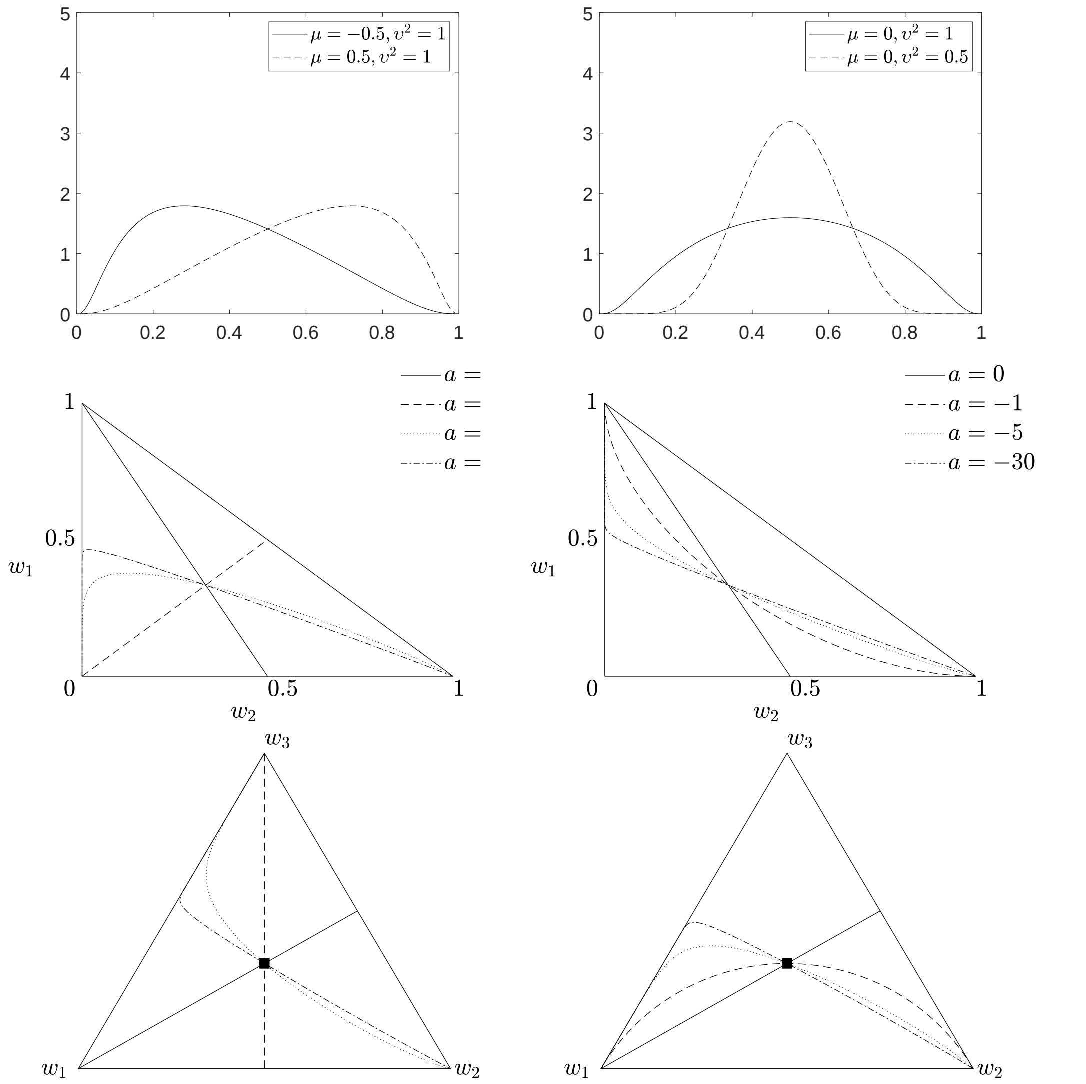

Figure 1: Top: distribution $\mathcal{L}_2(\mu, \upsilon^2)$ for different values of υ^2 (left) and μ (right). Middle: values of the random variables w_1 and w_2 as functions of the random variable z_1 for different level of positive (left) and negative (right) covariance $\upsilon^2 a$. Bottom: representation of the relationships in the De Finetti's diagrams.

In this paper, we apply these results as follows. First, we assume that the n-dimensional vector of weights $\mathbf{w}_t = (w_{1t}, \dots, w_{nt})'$ of our mixture model in Eq. (7) relates to a lower dimensional vector $\mathbf{z}_t \in \mathbb{S}^m$ as follows

$$\mathbf{w}_t \quad = \quad \phi_{A_t}(\mathbf{z}_t) \tag{16}$$

where the map $\phi_A(\cdot)$ from $\mathbb{S}^m$ to $\mathbb{S}^n$ is defined in Proposition 2.3, and A_t is a time-varying $((n-1) \times (m-1))$ projection matrix.

Second, we apply the operations in Definition 2.2 and assume that the latent factors follow a random walk process with normal-logistic-normal innovations

$$\mathbf{z}_t = \mathbf{z}_{t-1} \oplus \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathcal{L}_m(\mathbf{0}_m, \Upsilon) \tag{17}$$

which is a flexible model for the latent weights. Thanks to the class-preserving property of the logistic-normal the combination weights have conditional distribution H_{0t}^{W} which is a logistic-normal as stated in the following.

Corollary 2.1 (Weight dynamics)**.** *Let* $\mathbf{z}_t$ *be a process defined in Eq.* (17)*,* A_t *an* $((n-1)\times(m-1))$ *matrix, then* $\mathbf{w}_t$ *defined in Eq* (16) *belongs to the simplex* $\mathbb{S}^n$ *and follows the logistic-normal distribution* $\mathcal{L}_n(A_t\mathrm{alr}(\mathbf{z}_{\mathrm{t-1}}), \mathrm{A_t \Upsilon A_t'})$*.*
Proof See Appendix A.

The class-preserving property allows for a projection from the large dimensional simplex $\mathbb{S}^n$ onto the possibly lower space $\mathbb{S}^m$. Nevertheless, the effectiveness of the dimensionality reduction strategy and the probabilistic properties of the random vector $\mathbf{w}_t$ depend crucially on the choice of the sequence of matrices A_t, which will be discussed in the next section.

2.3 Dimensionality reduction and dynamic clustering

A contribution of this paper is to simplify the complexity of the combination exercise by reducing the dimension of the latent space of the weights while preserving crucial aspects such as their time variations and probabilistic properties.

Dimensionality reduction techniques are widely used in machine learning to reduce the dimension of high-dimensional datasets (e.g., see Casarin and Veggente, 2020, and references therein). The dependence structure in the data is used to reduce substantially the data dimension and to extract relevant information. In this paper we exploit the similarities between models or experts by specifying a dynamic clustering process in the space of predictive densities $f(\tilde{y}_t|\mathfrak{I}_{i,t-1}, M_i)$. The n predictors are clustered into m different groups with $m < n$, following some (time-varying) features $\boldsymbol{\psi}_{it} \in \mathbb{R}^d$, $i = 1, \ldots, n$, of the predictive densities. This allows to learn sequentially model dependence, a feature well documented on data but largely ignored in the predictive density combination literature.

Let $N_{1t}, \ldots, N_{mt}$ be the m groups of predictors such that each predictor belongs only to one group, i.e. $N_{it} \cap N_{jt} = \emptyset$ and $N_{1t} \cup N_{2t} \cup \ldots N_{mt} = \{1, 2, \ldots, n\}$. Given the clustering of the predictors, we specify the $(n \times m)$ allocation matrix Ξ_t with elements $\xi_{ij,t} = 1$ if $i \in N_{jt}$ and $\xi_{ij,t} = 0$ otherwise, and a $(n \times m)$ coefficient matrix B_t. The two matrices allow us to project the n-dimensional latent variable $\mathbf{w}_t$ onto a reduced dimension latent space by using the elements of $\tilde{A}_t = (\Xi_t \circ B_t)$ in (16). With this specification, if the i-th predictive density is allocated to the k-th cluster then its mixture weight w_{it} will be driven by the latent factor z_{kt} which is uniquely associated to the cluster k.

The dynamics of the allocation matrix Ξ_t is given in the following. Let $\mathbf{c}_{jt} \in \mathbb{R}^d$, $j = 1, \ldots, m$ be the centroids defined as

$$\mathbf{c}_{jt} = \frac{1}{n_{jt}} \sum_{i \in N_{jt}} \boldsymbol{\psi}_{it} \tag{18}$$

where $n_{jt} = \mathrm{Card}(N_{jt})$ is the number of predictive densities in the j-th cluster at time t. At time $t+1$ the allocation matrix updates as follows: $\xi_{ij,t+1} = 1$ if $j = j^*$

and $\xi_{ij,t+1} = 0$ otherwise where $j^* = \arg\min\{||\boldsymbol{\psi}_{it+1} - \mathbf{c}_{jt}||,\ j = 1, \ldots, m\}$ and $||\cdot||$ is the Euclidean norm. The centroids update as follows

$$\mathbf{c}_{jt+1} = \mathbf{c}_{jt} + \lambda_t(\mathbf{m}_{jt+1} - \mathbf{c}_{jt}) \tag{19}$$

where

$$\mathbf{m}_{jt+1} = \frac{1}{n_{jt+1}} \sum_{i \in N_{jt+1}} \boldsymbol{\psi}_{it} \tag{20}$$

and $\lambda_t \in [0, 1]$. Note that the choice $\lambda_t = n_{jt+1}/(n_{jt}^c + n_{jt+1})$, with $n_{jt}^c = \sum_{s=1}^{t} n_{js}$, implies a dynamic clustering with forgetting driven by the processing of the blocks of observations. In the application we fix $\lambda = 0.99$. The grouping of the predictors can change over time, following our dynamic clustering rule, but the number of clusters is assumed constant to preserve the interpretability of the factors.

For the elements $b_{ij,t}$ of the coefficient matrix B_t we consider two alternative specifications. In the first one all coefficients in the cluster have the same weights:

$$b_{ij,t} = \frac{1}{n_{jt}} \mathbb{I}(\xi_{j,it} = 1) \tag{21}$$

This specification may have the undesirable property that the projection coefficients $a_{ij,t} = \xi_{ij,t} b_{ij,t}$ are constant within a group. For this reason, we also propose a second specification where each model contributes to the combination following its predicting performance g_{it} (e.g. the log score in Eq. (C.14) in the Supplementary Material C), i.e.

$$b_{ij,t} = \mathbb{I}(\xi_{j,it} = 1) \sum_{s=1}^{t} \frac{\exp\{g_{is}\}}{\bar{g}_{it}} \tag{22}$$

where $\bar{g}_{it} = \sum_{l \in N_{it}} \sum_{s=1}^{t} \exp\{g_{ls}\}$.

In order to use the projection matrix $\tilde{A}_t$ in the Eq. (16) it is possible to choose A_t as the $((n-1) \times (m-1))$-dimensional matrix obtained by removing the last column and the last row of $\tilde{A}_t$.

Remark 1. *Given the results in the preceding subsection, the chosen specification for A_t returns a random weight vector $\mathbf{w}_t$ with degenerate distribution on $\mathbb{S}^n$.*

Proof See Appendix A.

The degeneracy is related to the rank deficiency of the matrix $A_t \Upsilon A_t'$. The first source of degeneracy is due the presence of zeros in the last n_{mt} rows of the matrix A_t and can be removed by apply the class-preserving transformation to obtain a subvector of $\mathbf{w}_t$. Let us denote with $n_t = (n_{1t} + \ldots + n_{m-1t})$ the number of models assigned to the first $m-1$ clusters and let us assume that they correspond to the first n_t elements of $\tilde{\mathbf{y}}_t$. Note that this simplifying assumption is not restrictive since it is possible to permute the rows of $\tilde{A}_t$ in order to have the coefficients for the last cluster in the last $n - n_t$ rows. Then one can obtain a random vector $(w_{1t}, \ldots, w_{n_t}, (1 - w_{1t} - \ldots - w_{n_t}))$ with logistic-normal distribution by applying the class-preserving transformation. The residual weight $1 - w_{1t} - \ldots - w_{n_t}$ is then assigned to the models in the last cluster. This can be obtained in two steps. First, we drop out the last $n - n_t$ rows of the projection matrix A_t and apply the class-preserving transformation by using the $(n_t \times (m-1))$-dimensional matrix $A_t^* = (I_{n_t}, O_{n-n_t})A_t$ as projection matrix. Second, we set $(w_{n_t+1}, \ldots, w_{nt}) = \kappa \mathbf{a}_t$ where $\mathbf{a}_t$ represents a $n - n_t$ dimensional vector containing the elements in the last $n - n_t$ rows of the last column of $\tilde{A}_t$. The following results establishes the relationship between our logistic-normal linear model with distribution on the simplex $\mathbb{S}^{n_t+1}$ and a Gaussian nonlinear model in $\mathbb{R}^{n_t}$.

Corollary 2.2. *Let us define the Gaussian random walk process $\mathbf{v}_t \in \mathbb{R}^{m-1}$ $t = 1, 2, \ldots$ with $\mathbf{v}_t \sim \mathcal{N}_{m-1}(\mathbf{v}_{t-1}, \Upsilon)$ and the following transformed processes $\mathbf{x}_t = A_t^* \mathbf{v}_t$, $\mathbf{w}_t^* = \mathrm{alr}^{-1}(\mathbf{x}_\mathrm{t})$, $\mathbf{z}_t = \mathrm{alr}^{-1}((\mathbf{v}_\mathrm{t}, 1))$ and $\mathbf{w}_t^* = \phi_A(\mathbf{z}_t)$. The transformed vectors have the following distributions*

$$\begin{aligned} \mathbf{z}_t &\sim \mathcal{L}_m(\mathbf{z}_{t-1}, \Upsilon) && (23) \\ \mathbf{w}_t^* &\sim \mathcal{L}_{n_t+1}(A_t^* \mathbf{v}_{t-1}, A_t^* \Upsilon A_t^{*\prime}) && (24) \end{aligned}$$

Proof See Appendix A.

The relationships with the Gaussian process given in the previous Corollary can be exploited in the inference procedures to easily generate samples for the latent weights or to derive filtering and smoothing recursions using the usual operations on $\mathbb{R}^{n_t}$. Finally note that the second source of degeneracy is intrinsic to our projection strategy based on the allocation matrix Ξ_t, which implies the rank deficiency of the matrix $A_t^* \Upsilon A_t^*$, and on the assumption that the relationship between $\mathbf{w}_t$ and $\mathbf{z}_t$ is not subject to errors. Our approach can be extended to account for contemporaneous uncertainty with the addition of logistic-normal noise in the equation for $\mathbf{w}_t$. This is left for further research.

2.4 A compositional state-space model

We summarize the model defined in the previous subsections and show that it is a conditionally linear state-space model on the simplex. Extending the $\odot$ product operation to the case of a matrix of real numbers allow us to write the transform ϕ_A, as a linear matrix operation between simplices of different dimensions, denoted with $\boxplus$.

Remark 2. *Let $\mathbf{z} \in \mathbb{S}^m$ be a composition, A a $(n \times m)$ real matrix and define the matrix multiplication $A \boxplus \mathbf{z} = C_n \left(\prod_{j=1}^m z_j^{a_{1j}}, \ldots, \prod_{j=1}^m z_j^{a_{n-1j}} \right)$. If A is such that $A\boldsymbol{\iota}_m = \mathbf{0}_n$ and $a_{im} = -1$, $i = 1, \ldots, n-1$ and $a_{n,j} = 0$ $j = 1, \ldots, m$, the transform defined in Proposition 2.3 can be written as $\phi_A(\mathbf{z}) = A \boxplus \mathbf{z}$.*

This result enables us to obtain the following state-space representation of our compositional combination model. Let $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{nt})'$ and $\tilde{\mathbf{y}}_t = (\tilde{y}_{1t}, \ldots, \tilde{y}_{nt})'$,

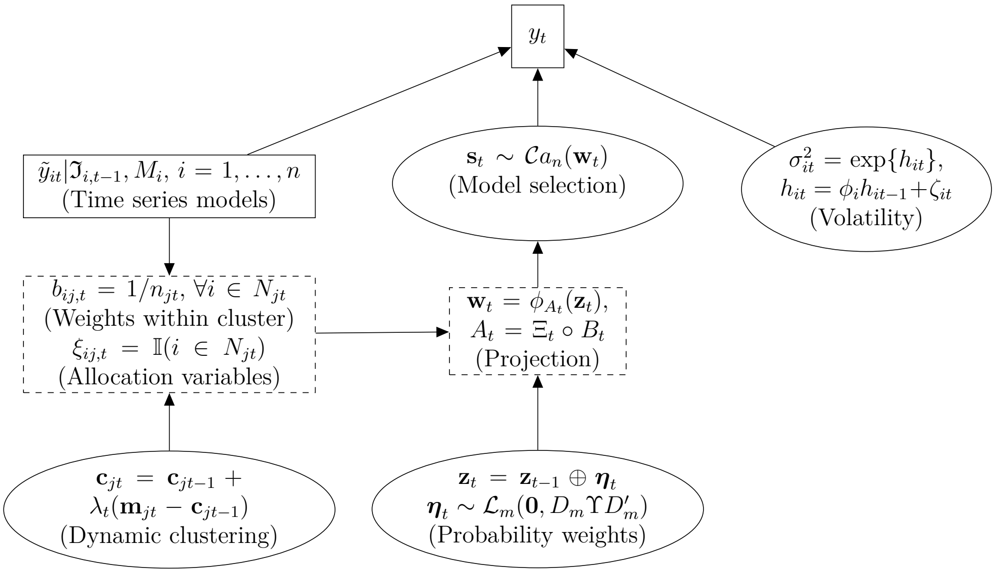
y_t
$\tilde{y}_{it}|\Im_{i,t-1}, M_i,\ i = 1, \dots, n$
(Time series models)
$\mathbf{s}_t \sim \mathcal{C}a_n(\mathbf{w}_t)$
(Model selection)
$\sigma^2_{it} = \exp\{h_{it}\}$,
$h_{it} = \phi_i h_{it-1} + \zeta_{it}$
(Volatility)
$b_{ij,t} = 1/n_{jt},\ \forall i \in N_{jt}$
(Weights within cluster)
$\xi_{ij,t} = \mathbb{I}(i \in N_{jt})$
(Allocation variables)
$\mathbf{w}_t = \phi_{A_t}(\mathbf{z}_t)$,
$A_t = \Xi_t \circ B_t$
(Projection)
$\mathbf{c}_{jt} = \mathbf{c}_{jt-1} +$
$\lambda_t(\mathbf{m}_{jt} - \mathbf{c}_{jt-1})$
(Dynamic clustering)
$\mathbf{z}_t = \mathbf{z}_{t-1} \oplus \boldsymbol{\eta}_t$
$\boldsymbol{\eta}_t \sim \mathcal{L}_m(\mathbf{0}, D_m \Upsilon D'_m)$
(Probability weights)

Figure 2: Directed acyclic graph of state space model. It exhibits the hierarchical structure of the observations on the endogenous variable y_t and the predicted variables $\tilde{y}_{it}$ (rectangles, solid line), the latent variables $\mathbf{s}_t$ and $\mathbf{z}_t$ (ellipses) and the link functions ϕ_{A_t}, Ξ_t and B_t (rectangles, dashed line). The directed arrows show the causal dependence structure of the model.

then the model in Equations (7), (16) and (17) writes as

$$\begin{array}{rcll} y_t &=& (\tilde{\mathbf{y}}_t + \boldsymbol{\varepsilon}_t)'\mathbf{s}_t, & \boldsymbol{\varepsilon}_t \sim \mathcal{N}_n(\mathbf{0}, \mathrm{diag}\{\sigma^2_{\mathrm{t}}\}) \qquad (25) \\ \mathbf{s}_t &\sim& \mathcal{C}a_n(\mathbf{w}_t), & \mathbf{w}_t = A_t \boxplus \mathbf{z}_t \qquad (26) \\ \mathbf{z}_t &=& \mathbf{z}_{t-1} \oplus \boldsymbol{\eta}_t, & \boldsymbol{\eta}_{kt} \sim \mathcal{L}_m\left(\mathbf{0}, D_n \Upsilon D'_m\right) \qquad (27) \end{array}$$

where $\mathbf{s}_t$ is a model-selection categorical process and $\mathcal{C}a_n(\mathbf{w_t})$ denotes the n-dimensional Multinoulli, or categorical, distribution with parameter $\mathbf{w_t} \in \mathbb{S}^n$.

A graphical representation of the dimensionality reduction strategy implemented in our model is given in Fig. 2. The observable variable y_t (top box) depends on three latent processes (ellipses). The process s_t allows for a dynamic selection of the models (see George and McCulloch, 1993, for model selection with latent variables). The stochastic volatility process σ^2_{it} accounts for model set incompleteness as proposed in Billio et al. (2013). The compositional model for the factors $\mathbf{z}_t$ allows for well-defined probability weights (see Pawlowsky-Glahn and Buccianti, 2011; Pawlowsky-Glahn et al., 2015, for an introduction to compositional models).

In Fig. 3-4 we provide simulated paths from our compositional model with five independent normal predictors $\tilde{y}_{it} \sim \mathcal{N}(-2+i, 0.1i)$ $i = 1, \dots, 5$ and observation noise $\varepsilon_t \sim \mathcal{N}(0, 0.2)$. The common latent factors z_{1t}, z_{2t}, z_{3t} (i.e. $m = 3$) are associated with the first, second and third clusters, respectively, and have noise

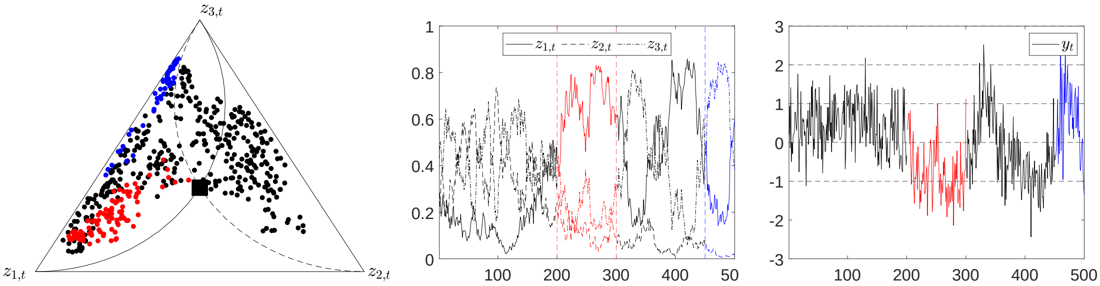

Figure 3: De Finetti's diagram (left) and the time series plot (middle) of the ternary $(z_{1,t}, z_{2,t}, z_{3,t})$ and observations y_t (right) with point predictions from the five predictors (horizontal dashed lines). In the panels, colors indicate different sub-samples.

distribution $\boldsymbol{\eta}_t \sim \mathcal{L}_2(\mathbf{0}_2, 0.2D_3D_3')$ i.i.d. $t = 1, \ldots, 500$. The projection matrix

$$A_t = \begin{pmatrix} b_{11,t} & 0 & 0 \\ b_{21,t} & 0 & 0 \\ 0 & b_{32,t} & 0 \\ 0 & b_{42,t} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad b_{ij,t} \sim \Lambda_1(0, 0.2), \forall i, j, t \tag{28}$$

allocates models M_1 and M_2 to the first cluster, models M_3 and M_4 to the second cluster and model M_5 to the third cluster.

Fig. 3 exhibits the trajectories of the common factors by using the De Finetti's, or ternary, diagram (left) and a time series plot (middle). In this diagram, weight vectors with probability 1 assigned to one cluster and 0 to the others, i.e. $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$, correspond to the vertices. Red and blue dots close to the vertices z_{1t} and z_{3t}, in the diagram, correspond to samples where cluster 1 and cluster 3, respectively receive weights close to one (blue and red subtrajectories in the time series plot). The barycentre of the triangle (black square) corresponds to the case of equally weighted clusters, i.e. $(1/3, 1/3, 1/3)$.

As reference lines, we report in the diagram two deterministic trajectories

$$\begin{aligned} \mathbf{r}_{1t} &= \mathrm{alr}^{-1}((-10 + 20\frac{\mathrm{t}}{\mathrm{T}}, -20 + 40\frac{\mathrm{t}}{\mathrm{T}}, -30 + 60\frac{\mathrm{t}}{\mathrm{T}})) \\ \mathbf{r}_{2t} &= \mathrm{alr}^{-1}((-20 + 40\frac{\mathrm{t}}{\mathrm{T}}, -10 + 20\frac{\mathrm{t}}{\mathrm{T}}, -30 + 60\frac{\mathrm{t}}{\mathrm{T}})) \end{aligned}$$

$t = 1, \ldots, T$ (dashed and solid lines, respectively). In the first trajectory, a weight close to 1 is assigned to cluster 1 at the beginning of the period, i.e. $t = 0$, and to cluster 3 at the end of the period, i.e. $t = T$. In the second trajectory unit weight is given to cluster 2 at $t = 0$ and to cluster 3 at $t = T$.

Realizations of y_t given random samples from the five predictors are in the right plot. Horizontal lines report the point predictions for the five predictors. For example, when the weight of the cluster 3, z_{3t} (middle plot, dotted-dashed, blue), increases, the weight of the fifth predictor w_{5t} increases resulting in values for y_t close to 2 (right, plot blue line).

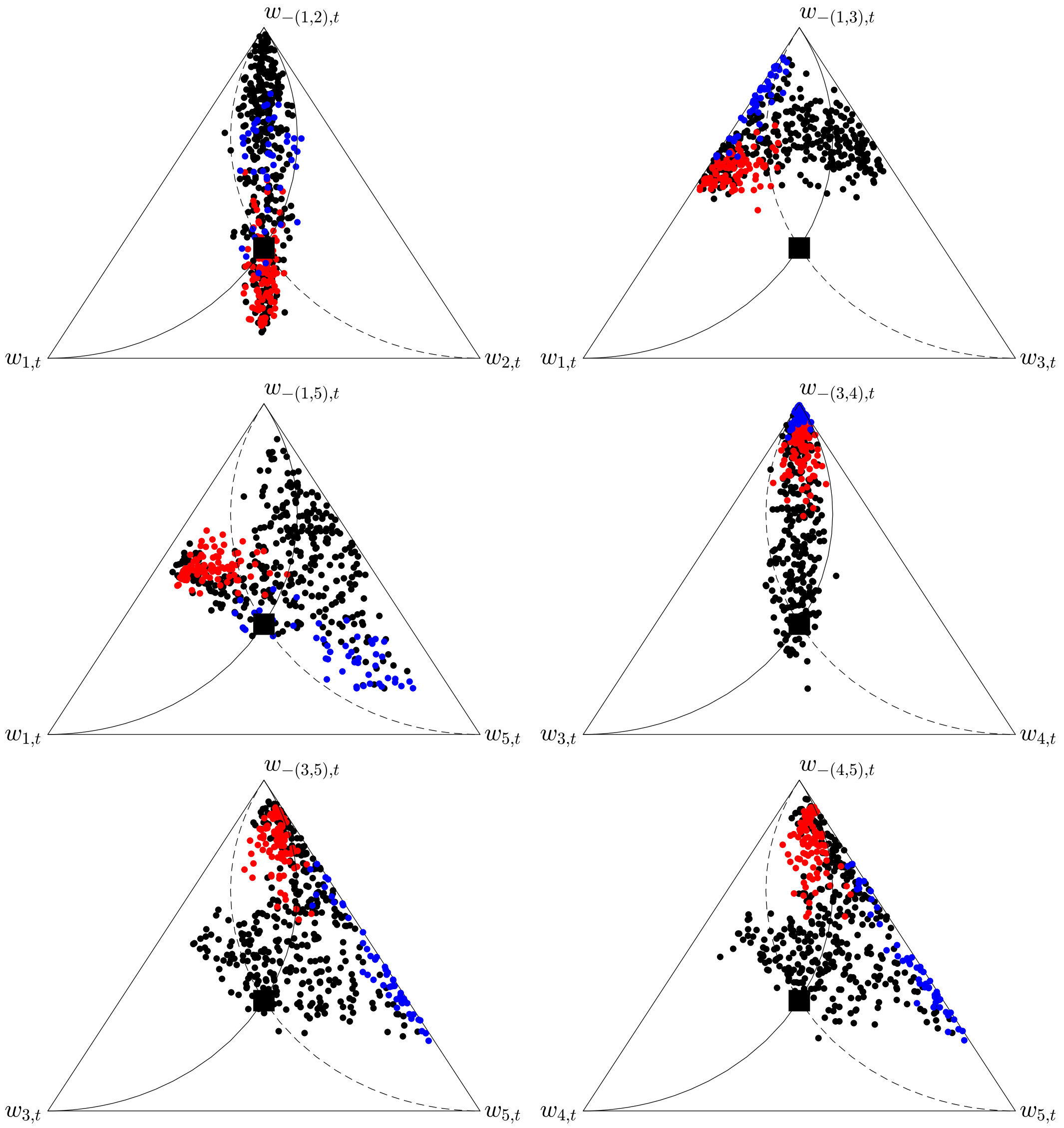

Figure 4: De Finetti's diagram of the ternaries $(w_{i,t}, w_{j,t}, w_{-(i,j)t})$, $j > i$ with $w_{-(i,j)t}$ the amalgamation of w_{lt}, $l \neq i, j$. In each plot the ternary samples (dots), the equal weight composition (square) and the reference lines (dashed and dotted). In the panels, colors indicate different sub-samples.

A way commonly found for reducing dimensionality of probabilistic weights is to sum some weights into a new weight which is called amalgamation.

Definition 2.5 (Amalgamation). *Given the composition $\mathbf{w} \in \mathbb{S}^{m-1}$, and a collection of indices $A = \{i_1, \ldots, i_d\} \subset \{1, \ldots, d\}$, $m - d > 0$, and the complement set $\bar{A} = \{1, \ldots, n\}/A$ the value*

$$w_A = \sum_{i \in A} w_i$$

is called amalgamated component. The vector $(\mathbf{w}_{\bar{A}}, w_A)'$ is called amalgamated composition which is in $\mathbb{S}^{m-d}$, where $\mathbf{w}_{\bar{A}}$ is the vector containing w_j with $j \in \bar{A}$.

See Egozcue et al. (2003), Egozcue and Pawlowskky-Glahn (2005) and Fiŝerová and Hron (2011) for further details on amalgamation and subcomposition operations. The amalgamation can be used in combination with the De Finetti's diagram for running weight comparisons and representation for high dimensional weight vectors. Figure 4 illustrates the pairwise comparison of the weight dynamics. In each diagram, the dots represent the ternary $(w_{i,t}, w_{j,t}, w_{-(i,j),t})$ with $i \neq j$ where $w_{-(i,j),t} = \sum_{l\neq i,j} w_{l,t}$ is the amalgamation of $w_{l,t}$ with $l \neq i, j$ into a new weight $w_{-(i,j),t}$. One can see that w_{1t} and w_{2t} move together (first plot) since they are driven by a common factor z_{1t}. Whereas w_{1t} does not depend on w_{3t} (second plot) and depends negatively on w_{5t} (third plot). Also w_{3t} and w_{4t} move together and have negative dependence with w_{5t} (last three plots).

2.5 Multiple-prediction extensions

In the case the same set of predictive densities is used for predicting a set of variables $y_{1t}, \ldots, y_{Kt}$, the model can be extended as follows:

$$\begin{aligned} y_{kt} &= (\tilde{\mathbf{y}}_t + \boldsymbol{\varepsilon}_t)'\mathbf{s}_{kt}, \quad && \boldsymbol{\varepsilon}_{kt} \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}^2_{\mathrm{kt}})) && (29)\\ \mathbf{s}_{kt} &\sim \mathcal{C}a_n(\mathbf{w}_{kt}), && \mathbf{w}_{kt} = A_{kt} \boxplus \mathbf{z}_{kt}, && (30)\\ \mathbf{z}_t &= \mathbf{z}_{t-1} \oplus \boldsymbol{\eta}_t, && \boldsymbol{\eta}_t \sim \mathcal{L}_{Km}(\mathbf{0}, I_m \otimes \Upsilon) && (31) \end{aligned}$$

where the projection matrices A_{kt} are driven by a common clustering process Ξ_t and a variable-specific learning coefficient B_{kt} which reflects the ability of each predictive density to predict the variable of interest y_{kt}.

3 Bayesian inference

The analytic solution of the optimal filtering problem is generally not known, also the clustered-based mapping of the predictor weights onto the subset of latent variables requires the solution of an optimization problem which is not available in closed form.

3.1 Prior and posterior distributions

Let $\boldsymbol{\theta} \in \Theta$ be the parameter vector of the combination model, that is $\boldsymbol{\theta} = (\boldsymbol{\phi}, \Upsilon)$ with $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$. We assume independent normal prior distributions $\mathcal{N}(0, s^2)$ for $\phi_1, \ldots, \phi_n$ and inverse Wishart distribution $\mathcal{IW}_m(a, S)$ for Υ and denote with $\pi(\boldsymbol{\theta})$ the joint distribution.

In the following, $\mathbf{u}_{1:t} = (\mathbf{u}_1, \ldots, \mathbf{u}_t)$ indicates the collection of vectors $\mathbf{u}_t$ from time 1 to time t and $\boldsymbol{\omega}_t = (\mathbf{s}_t, \mathbf{h}_t, \mathbf{z}_t)$ the collection of latent variables with values in the latent space $\mathcal{W} = (\{0,1\}^n \times \mathbb{R}^n \times \mathbb{S}^m)$. The joint posterior distribution is

$$\pi(\boldsymbol{\theta}|\mathbf{y}_{1:T}) = \frac{L(\mathbf{y}_{1:T}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(\mathbf{y}_{1:T})} \qquad (32)$$

where $m(\mathbf{y}_{1:T})$ is the marginal likelihood or model evidence and $L(\mathbf{y}_{1:T}|\boldsymbol{\theta})$ the likelihood function given in integral form

$$\int_{\mathcal{W}} \prod_{t=1}^{T} f(y_t|\mathbf{s}_t, \mathbf{h}_t) f(\mathbf{s}_t|\mathbf{z}_t, A_t) f(\mathbf{h}_t|\mathbf{h}_{t-1}, \boldsymbol{\phi}) f(\mathbf{z}_t|\mathbf{z}_{t-1}, \Upsilon)\pi(\boldsymbol{\theta})d\boldsymbol{\omega}_{1:T} \tag{33}$$

with $f(y_t|\mathbf{s}_t, \mathbf{h}_t)$, $f(s_t|\mathbf{z}_t, A_t)$ and $f(\mathbf{z}_t|\mathbf{z}_{t-1}, \Upsilon)$ the distributions in Eq. (25)-(27) and $f(\mathbf{h}_t|\mathbf{h}_{t-1}, \boldsymbol{\phi})$ the conditional distribution of the log-volatility process.

3.2 Posterior approximation

The joint posterior is not tractable thus we apply Markov chain Monte Carlo (MCMC) (Robert and Casella, 2004) combined with Sequential Monte Carlo (Kitagawa, 1998; Liu and West, 2001; Doucet et al., 2001). More specifically we consider the Particle Metropolis Hastings (PMH) algorithm discussed in (Andrieu et al., 2010; Andrieu and Roberts, 2009). At the k-th iteration of the PMH a candidate $\boldsymbol{\theta}^*$ is drawn from the proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(k-1)})$ where $\boldsymbol{\theta}^{(k-1)}$ is the previous iteration value of the MCMC chain, and it is accepted with probability

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(k-1)}) = \min\left\{1, \frac{\hat{L}_N(\mathbf{y}_{1:T}|\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(k-1)}|\boldsymbol{\theta}^*)}{\hat{L}_N(\mathbf{y}_{1:T}|\boldsymbol{\theta}^{(k-1)})\pi(\boldsymbol{\theta}^{(k-1)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(k-1)})}\right\} \tag{34}$$

with

$$\hat{L}_N(\mathbf{y}_{1:T}|\boldsymbol{\theta}^*) = \prod_{t=1}^{T} \hat{f}_N(y_t|y_{1:t-1}) \tag{35}$$

the product of approximated predictive likelihood functions

$$\hat{f}_N(y_t|y_{1:t}) = \int_{\mathcal{W}} \hat{f}_M(y_t|\mathbf{s}_t, \mathbf{h}_t) f(\mathbf{s}_t|\mathbf{z}_t, A_t) f(\mathbf{h}_t|\mathbf{h}_{t-1}, \boldsymbol{\phi}) f(\mathbf{z}_t|\mathbf{z}_{t-1}, \Upsilon)\pi_N(\boldsymbol{\omega}_t)d\boldsymbol{\omega}_t \tag{36}$$

In the approximated likelihood $\hat{f}_M(y_t|\mathbf{s}_t, \mathbf{h}_t)$ is an unbiased standard Monte Carlo estimator of $f(y_t|\Im_{i,t-1}, M_i, \sigma^2_{it})$ obtained with M independent draws from the predictive distributions, i.e.

$$\int_{\mathbb{R}} f(y_t|\tilde{y}_{it}, \sigma^2_{it}) f(\tilde{y}_{it}|\Im_{i,t-1}, M_i)d\tilde{y}_{it} \approx \frac{1}{M}\sum_{j=1}^{N} f(y_t|\tilde{y}^j_{it}, \sigma^2_{it}) \tag{37}$$

and $\pi_N(\boldsymbol{\omega}_t)$ is the approximated filtering distribution obtained by a Sequential Monte Carlo (SMC) algorithm, i.e.

$$\begin{aligned}
&\pi(\boldsymbol{\omega}_t|\mathbf{y}_{1:t}) = \\
= \; &\frac{L(y_t|\mathbf{s}_t, \mathbf{h}_t)}{L(y_t|\mathbf{y}_{1:t})}\int_{\mathcal{W}} f(\mathbf{s}_t|\mathbf{z}_t, A_t) f(\mathbf{h}_t|\mathbf{h}_{t-1}, \boldsymbol{\phi}) f(\mathbf{z}_t|\mathbf{z}_{t-1}, \Upsilon)p(\boldsymbol{\omega}_{t-1}|\mathbf{y}_{1:t-1})\boldsymbol{\omega}_{t-1} \\
&\approx \frac{1}{N}\sum_{j=1}^{N} \frac{L(y_t|\mathbf{s}^j_t, \mathbf{h}^j_t)}{L(y_t|\mathbf{y}_{1:t})} f(\mathbf{s}^j_t|\mathbf{z}^j_t, A_t) f(\mathbf{h}^j_t|\mathbf{h}^j_{t-1}, \boldsymbol{\phi}) f(\mathbf{z}^j_t|\mathbf{z}^j_{t-1}, \Upsilon)
\end{aligned} \tag{38}$$

In our SMC, given the initial set of particle $\Phi_t = \{\boldsymbol{\omega}_t^j, \tilde{\gamma}_t^j\}_{j=1}^N$ and the projection matrix $A_t = \Xi_t \circ B_t$ we iterate the clustering, prediction, updating and resampling steps detailed in the following.

First, we evaluate the dynamic clustering process in Eq. (19)-(22) by using the predictive distribution of the models or experts at time $t+1$ and obtain the updated Ξ_{t+1} and B_{t+1}.

Second, we approximate the state predictive density as follows:

$$\pi_N(\boldsymbol{\omega}_{t+1}|\mathbf{y}_{1:t}) = \sum_{j=1}^N p(\boldsymbol{\omega}_{t+1}|\boldsymbol{\omega}_t)\tilde{\gamma}_t^j \delta(\boldsymbol{\omega}_t^j - \boldsymbol{\omega}_t) \tag{39}$$

where $p(\boldsymbol{\omega}_{t+1}|\boldsymbol{\omega}_t) = f(\mathbf{s}_{t+1}|\mathbf{z}_{t+1}, A_{t+1})f(\mathbf{h}_{t+1}|\mathbf{h}_t, \boldsymbol{\phi})f(\mathbf{z}_{t+1}|\mathbf{z}_t, \Upsilon)$ The approximated state filtered density is easily obtained

$$\pi_N(\boldsymbol{\omega}_{t+1}|\mathbf{y}_{1:t+1}) = \sum_{j=1}^N \gamma_{t+1}^j \delta(\boldsymbol{\omega}_{t+1}^j - \boldsymbol{\omega}_{t+1}) \tag{40}$$

where $\gamma_{t+1}^j \propto \tilde{\gamma}_t^j \hat{f}_M(y_{t+1}|\mathbf{s}_{t+1}^j, \mathbf{h}_{t+1}^j)$ is a set of normalized weights

Since the systematic resampling of the particles introduces extra Monte Carlo variations and reduces the efficiency of the importance sampling algorithm, we do resampling only when the effective sample size (ESS) is below a given threshold. See Casarin and Marin (2009) for ESS calculation. At the $t+1$-th iteration if $\text{ESS}_{t+1}^j < \kappa$, simulate $\Phi_{t+1} = \{\boldsymbol{\omega}_{t+1}^{\theta\, k_j}, \tilde{\gamma}_{t+1}^j\}_{j=1}^N$ from $\{\boldsymbol{\omega}_{t+1}^{\theta\, j}, \gamma_{t+1}^j\}_{j=1}^N$ (e.g., multinomial resampling) and set $\tilde{\gamma}_{t+1}^j = 1/N$. We denote with k_j the index of the j-th re-sampled particle in the original set Φ_{t+1}. If $\text{ESS}_{t+1} \geq \kappa$ set $\Phi_{t+1} = \{\boldsymbol{\omega}_{t+1}^{\theta\, j}, \tilde{\gamma}_{t+1}^j\}_{j=1}^N$.

In the application with large number of predictors we apply the parallel evaluation of the dynamic clustering process and of the SMC algorithm as detailed in the Supplementary Material B.

4 Empirical applications

As a first application we focus on the daily stock market case, briefly mentioned in the previous section. We report results on several features of the combined predictive density of a replication of the daily Standard & Poor 500 (S&P500) index, including the economic value of tail events like Value-at-Risk.

The second application considers quarterly bond market data and using the extended Stock and Watson (2005) dataset, which includes 142 series sampled from 1959Q1 to 2011Q2, we predict the 3-month Treasury Bill interest rates.

In both exercises, we study incompleteness diagnostics and the weight patterns of the clusters over time which provide valuable signals that may lead to improved financial modelling and predicting.

4.1 Predicting and tracking the S&P500

Many investors of mutual funds, hedge funds and exchange-traded funds try to replicate the performance of the S&P500 index by holding a set of stocks, which are not necessarily the exact same stocks included in the index. We collected 1856 individual stock daily prices quoted in the NYSE and NASDAQ from Datastream over the sample March 18, 2002 to December 31, 2009, for a total of 2034 daily observations for each individual series. To control for liquidity we impose that each stock has been traded a number of days corresponding to at least 40% of the sample size. We compute log returns for all stocks. The cross-section average statistics of all series are reported in Table D.1 in Section D.1 of the Supplementary Material together with the results for S&P500.[3]

To ease on the computational workload, we apply an optimisation method to estimate the posterior modes of the parameters from a Normal GARCH(1,1) model and a Student-t GARCH(1,1) model[4] using rolling samples of 1250 trading days (about five years) for each stock return:

$$\begin{aligned} y_{it} &= c_i + \kappa_{it}\zeta_{it} && (41) \\ \kappa_{it}^2 &= \theta_{i0} + \theta_{i1}\zeta_{i,t-1}^2 + \theta_2\kappa_{i,t-1}^2, \qquad i = 1, 2, \ldots, n, && (42) \end{aligned}$$

where y_{it} is the log return of stock i at day t, $\zeta_{it} \sim \mathcal{N}(0,1)$ and $\zeta_{it} \sim \mathcal{T}(\nu_i)$ for the Normal and t-Student cases, respectively. The number of degrees of freedom ν_i is estimated in the latter model. We produce 784 one day ahead predictive densities from January 1, 2007 to December 31, 2009. Our out of sample period is associated with high volatility driven by the US financial crisis and includes, among others, events such as the acquisitions of Bear Stearns, the default of Lehman Brothers and all events of the following week.

In the dynamic clustering process we assume two clusters of predictive densities for the Normal GARCH(1,1) model and two clusters for the Student-t GARCH(1,1) model. The first two include Normal GARCH models with low (cluster one, labelled $n1$) and high (cluster two, labelled $n2$) volatility. The third cluster (labelled $t1$) includes Student-t GARCH models with low degrees of freedom and the fourth one (labelled $t2$) includes Student-t GARCH with high degrees of freedom. [5] The clustering of the densities is repeated every time a new prediction is produced and therefore the cluster composition varies over time.

Figure 5 presents results about these features. Normal models in cluster $n1$ differ substantially in terms of predicted variance (left plot, solid black line), having a rather low constant variance value over the entire period while cluster $n2$ has a variance more than double in size (left plot, dashed black line) including a shock in the latter part of 2008. Student-t models in cluster $t1$ have a relatively constant

[3] It has been suggested to make use of the information about shares outstanding in order to determine better the time behaviour of weights. We leave this as a topic for further research.

[4] Given our flat prior, these estimates are equal to maximum likelihood estimates and also are approximate Bayes mean estimates.

[5] Low degrees of freedom occur jointly with a large scale and high degrees of freedom occur jointly with a low scale.

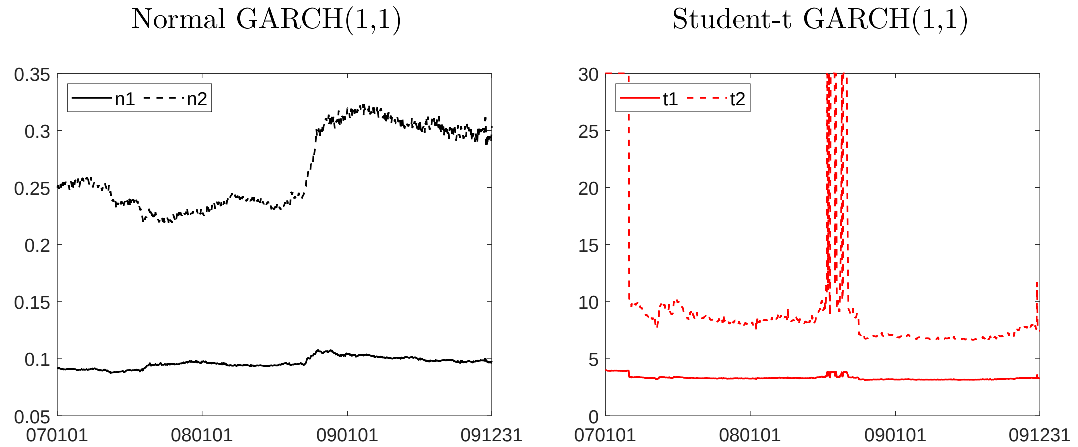

Figure 5: The figures present the average variance of the predictions from the two clusters for the Normal GARCH(1,1) models based on low (cluster 1, solid black line) and high (cluster 2, dashed black line) volatility in the left panel; and the average degree of freedom of the predictions from the two clusters for the Student-t GARCH(1,1) models based on low (cluster 3, solid red line) and high (cluster 4, dashed red line) degrees of freedom in the right panel. The degrees of freedom are bounded to 30.

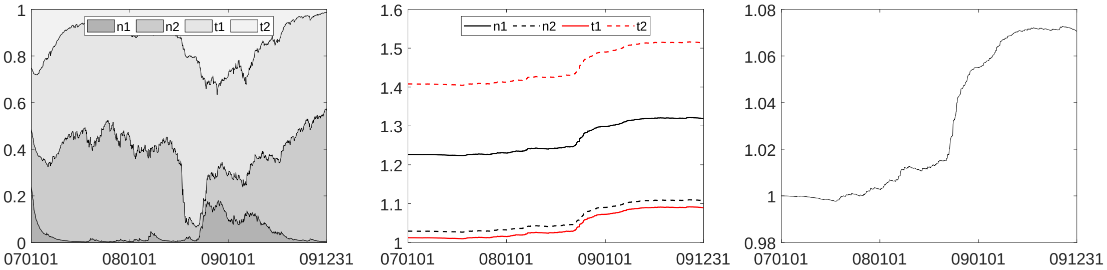

Figure 6: Left: the mean logistic-normal weights for the two Normal GARCH clusters, labeled in the graph "n1" and "n2", and for the two Student-t-GARCH clusters, labeled in the graph "t1" and "t2". Middle: posterior mean estimates of incompleteness measures in the four clusters in the scheme DCEW-SV. Right: average of the posterior mean estimates of the model set incompleteness measure.

thick tail over the entire period (right plot, solid red line) while cluster $t2$ has values around 10 for the degrees of freedom (right plot, dashed red line) except during the crisis period, where the density collapses toward a normal density with degrees of freedom larger than 30. The Lehman Brother effect is visible in the figure, with an increase of volatility in the normal cluster $n2$ and a decrease in the degrees of freedom in the Student-t cluster $t2$.

Diagnostic for the combination model with four clusters is shown in Figure 6. The average weights per cluster, that is the average of w_{it} over $i \in N_{jt}$, are in the left plot. De Finetti's diagrams in Figure 7 exhibit the pairwise comparison of the weight dynamics. In the diagrams the blue dots represent the ternary $(z_{i,t}, z_{j,t}, z_{-(i,j),t})$ where $z_{-(i,j),t} = \sum_{l \neq i,j} z_{l,t}$ is the other model's total weight.

There is evidence of time variations in the weights, with three distinct subperiods, and of fat tails playing an important role. Before the crisis, clusters $n2$ and $t1$ have almost equal high weights (blue dots in the fourth diagram,

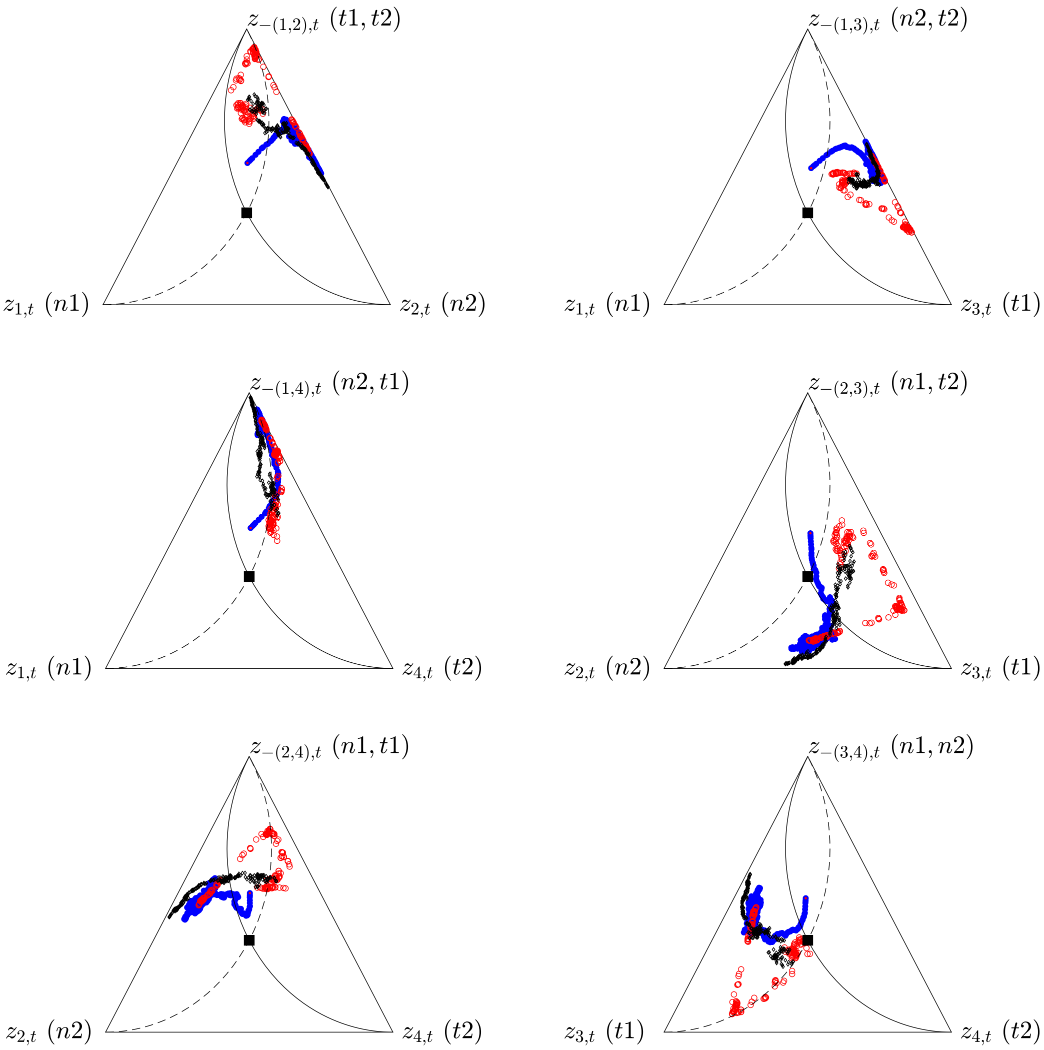

Figure 7: De Finetti's diagram for the pairwise subcomposition comparison between model weights. In each plot the trajectory of the ternary $(z_{it}, z_{jt}, z_{-(ij)t})$, $j > i$ (blue line), the starting point (red dot), the ending point (black dot) and the equal weight composition (square).

Figure 6), while clusters $n1$ and $t2$, both play a much less important role (third diagram). In the crisis period of 2008, cluster $t1$ receives almost all the weight with clusters $n1$ and $n2$ almost none (red dots on the dashed reference line in the sixth diagram). Some of the assets led the market experiencing large losses in that period. This results in very fat tailed densities and our combination scheme captures these features and assigns to cluster $t1$ more weight. In the period after the Lehman Brothers collapse cluster $t1$ receives again a substantial weight while the normal cluster with large variance $n2$ is getting gradually more weight (black dots, diagrams four and five).

We measure incompleteness for the model set Density Combination with Equal Weights and Stochastic Volatility, (DCEW-SV). Estimates of model set incompleteness are shown in Figure 6. We compute the incompleteness contribution of each cluster as the average value of the squared posterior residuals.

	RMSPE	LS	CRPS	avQS-T	avQS-L	Violation
WN	1.852	-9.045	1.017	0.429	0.425	3.57%
Normal GARCH	1.852	-4.164**	0.956**	0.139**	0.195**	2.93%
Student-t GARCH	1.852	-2.738**	0.937**	0.118**	0.154**	2.55%
GJR GARCH	1.852	-4.068**	0.955**	0.125**	0.158**	2.75%
EW-GARCH	1.853	-3.145**	1.018	0.144**	0.171**	2.80%
DCEW	**1.812****	**2.249****	**0.911****	**0.114****	**0.149****	0.90%
DCEW-SV	1.816**	2.206**	0.913**	**0.114****	**0.149****	**1.02%**

Table 1: Predicting results for next day S&P500 log returns. Bold numbers indicate the best statistic for each loss function. One or two asterisks indicate that differences in accuracy from the white noise (WN) benchmark are credibly different from zero at 5%, and 1%, respectively, using the Diebold-Mariano t-statistic for equal loss. The underlying p-values are based on t-statistics computed with a serial correlation-robust variance, using the pre-whitened quadratic spectral estimator of Andrews and Monahan (1992). The column "Violation" shows the number of times the realised value exceeds the 1% Value-at-Risk (VaR) predicted by the different models over the sample.

It is seen that $n1$ and $t2$ have the higher average incompleteness and $n2$ and $t1$ have lower average incompleteness. This diagnostic information confirms that clusters $n1$ and $t2$ give lower predictive accuracy.[6]

We also plot the average estimate of the overall model incompleteness. This estimate has a 7% increase in September 2008, which is due to the default of Lehman Brothers and related following events. Interestingly, the volatility does not reduce in 2009, a year with large positive returns opposite the large negative returns in 2008.

We compare the performance of our approach with results from five different basic models applied to the S&P500 log returns: a white noise model (or a random walk for prices), often used as a main benchmark in equity premium predictibility; the Normal GARCH(1,1) and the Student-t GARCH(1,1) models described above. In order to explore the sensitivity of our results for model set incompleteness in more detail, we include the Normal GJR GARCH(1,1) model in Glosten et al. (1993) that includes leverage effects in the model set. This model is a richer model than the standard GARCH and should fit the data better. In fact, leverage effect is considered among the stylised facts of financial returns and the added feature may become relevant in our analysis. Finally, since it might difficult to know which of the GARCH models perform better *ex-ante*, we apply also an equal weight combination of the three GARCH models, labeled EW-GARCH.

Out-of-sample predicting result are presented in Table 1. The first three columns deal with location and shape features of the predictive densities. It is seen that our combination schemes produce the lowest Root Mean Squared Prediction Error (RMSPE) and Cumulative Rank Probability Score (CRPS) and the highest Log Score (LS), see also Section C in the Supplementary Material for more details.

[6] We note that one may experiment with a larger set of individual models, see for example Geweke and Durham (2012).

The results indicate that the combination schemes are statistically superior to the no-predictability WN benchmark. The Normal GARCH(1,1) model, the Student-t GARCH(1,1) model and the Normal GJR GARCH(1,1) model fitted on the index also provide more accurate density predictions than the WN, but not on point predicting. For all three score criteria, the statistics given by the three individual models are inferior to our combination schemes.

We consider two statistics that refer to left and right tails of the predictive densities. These refer to weighted averages of the Gneiting and Raftery (2007) quantile scores that are based on quantile predictions that correspond to the predictive densities from the different models. In the Supplementary Material it is shown that avQS-T emphasizes both tails and avQS-L the left tail of the predictive density relative to the realization 1-step ahead. To study how the models perform in the left tail predictions over time, we consider the cumulative sum of avQS-L and the most accurate model at observation t produces the lowest cumavQS-L$_{i,h,t}$. The fourth and fifth columns of Table 1 show results for tail evaluation. Our schemes provides the lowest avQS-T and avQS-L statistics, confirming the accuracy of the method in the tails of the distribution. See Figure D.1 in the Supplementary Material for a comparison of performance over time.

As economic measure, we apply a Value-at-Risk (VaR) based measure, see Jorion (2006). We compare the accuracy of our models in terms of violations, that is the number of times that negative returns exceed the VaR predictions at time t, with the implication that actual losses on a portfolio are worse than had been predicted. Higher accuracy results in numbers of violation close to nominal value of 1%. Moreover, to have a gauge of the severity of the violations we compute the total losses by summing the returns over the days of violation for each model. When looking to VaR violations, reported in the final column of Table 1, the number for all individual models is high and above 1%, with the WN higher than 3%. The dramatic events in our sample, including the Lehman default and all the other features of the US financial crisis provide an explanation for the result. It is important to note that the two combination schemes provide the best statistics, with violations very close to the 1% theoretical value. The property of our combination schemes to assign higher weights to the fat tail cluster $t1$ helps to model more accurately the lower tail of the index returns and covers more adequately risks.

4.2 Treasury Bill predicting

We consider the extended Stock and Watson (2005) dataset, which includes 142 series sampled at a quarterly frequency from 1959Q1 to 2011Q2. A graphical description of the data is given in Figure D.2 in the Supplementary Material. The dataset includes only revised series and not vintages of real-time data.[7] In order to deal with stationary series, we apply the series-specific transformation suggested

[7] See Aastveit et al. (2018) for a real-time application, with fewer series, of combined density nowcasting and the role of model set incompleteness over vintages and time.

in Stock and Watson (2005). We also re-scale the series to have zero mean.

We split the sample size 1959Q3-2011Q2 in two periods. The initial 102 observations from 1959Q3-1984Q1 are used as initial in-sample period; the remaining 106 observations from 1985Q1-2011Q2 are used as an out-of-sample period. We evaluate combined predictive densities of the 3-month Treasury Bill rate for $h = 1, \dots, 5$ step-ahead horizons using the large database.[8] For all variables we apply a Gaussian autoregressive model of the first order AR(1) and a Dynamic Factor Model (DFM) with 5 factors described in Stock and Watson (2012).[9]

Let y_t be the variable of interest to be predicted (i.e., Treasury Bill rates), the AR(1) model

$$y_{it} = \alpha_i + \beta_i y_{it-1} + \zeta_{it}, \quad \zeta_{it} \sim \mathcal{N}(0, \sigma_i^2) \tag{43}$$

is estimated following a Bayesian inference approach and assuming a diffuse informative Normal-Inverse-Gamma prior with null means and variances equal to 100 for the independent prior distributions of α_i and β. For the variance σ_i^2 we use an Inverse-Gamma with degrees of freedom equal to the number of lags (one) and intercept, that is two. The AR(1) models are estimated recursively and $h-$step ahead (Bayesian) Student-t predictive densities are constructed using a direct approach extending each vintage with the new available observation; see for example Koop (2003) for the exact formula of the mean, standard deviation and degrees of freedom.

We also consider the DFM with 5 factors described in Stock and Watson (2012) as another benchmark. More precisely:

$$\tilde{y}_t = \Lambda \boldsymbol{f}_t + \boldsymbol{\varepsilon}_{ft}, \quad \Phi(\mathrm{L}) \boldsymbol{f}_t = \boldsymbol{\eta}_{ft} \tag{44}$$

where the y_t is the variable of interest, $\boldsymbol{f}_t = (f_{1,t}, \dots, f_{r,t})'$ is an r vector of latent factors (in our case $r = 5$ or 7), Λ is a $1 \times r$ matrix of factors loadings, $\Phi(\mathrm{L})$ is an $(r \times r)$ matrix lag polynomial of order 2 , $\boldsymbol{\varepsilon}_{ft}$ is a (1×1) vector of idiosyncratic components and $\boldsymbol{\eta}_{rt}$ is an r vector of innovations. In this formulation the term $\Lambda \boldsymbol{f}_t$ is the common component of y_t. Bayesian estimation of the model described in equation (44) is carried out using Gibbs Sampling given in Koop and Korobilis (2009).

The clusters of predictive densities are identified by the dynamic clustering process, where our predictive densities are grouped in clusters depending on the persistence in the series. We are interested in the interpretation and behaviour of the clusters over the full sample and, for convenience, we impose that the cluster allocation of each model is fixed over the predicting vintages.[10] Note that in the finance exercise this assumption is relaxed. We assume alternatively 5 and 7

[8] We restrict the presentation to results for $h = 1$, 3, 5 horizons.

[9] We note that one more experiment would be to make use of a DFM structure for the models in our combination approach and compare results with our choice of the AR model structure. Given the extensive work that we did empirically, we prefer to leave this as a topic for future research.

[10] We also experimented just using the initial sample and the results were similar.

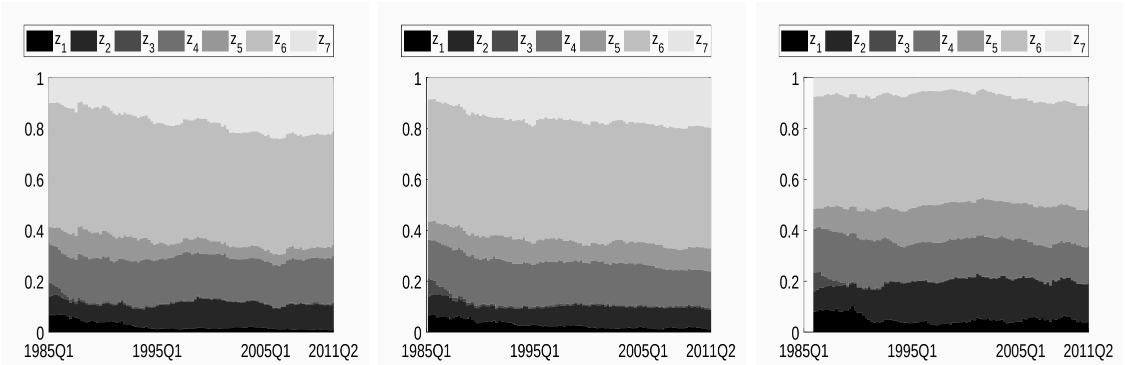

Figure 8: In each plot the logistic-normal weights (different lines) based on the UDCLS7 scheme for the horizons $h = 1$, 3, 5 steps ahead predictive horizons (different columns).

clusters.[11] In the grouping, we identify two clusters related to real activities; one cluster related to prices; and one cluster related to financial variables. The other clusters contains the remaining series. A detailed description of the 5 and 7 clusters is provided in Tables D.2-D.3 in the Supplementary Material.

As described in Section 2, we consider two alternative strategies for the specification of the weights b_{ijt}: equal weights and score recursive weights, where in the second case we fix the log scores for the various horizons h. We note that we keep the volatility of the incompleteness term constant, for convenience. In the present analysis, the number of components matters more.

We construct dynamic combination models (D) with equal (EW) and log-score driven (LS) cluster weights, and with 5 and 7 clusters. We obtain four models: DCEW5, DCLS5, DCEW7 and DCLS7. For the variance-covariance matrix of the combination weights we use a very informative prior.

In Figure 8 shows the time patterns of the weights based on the DCLS7 scheme. The 6-th cluster has a large weight, but several other clusters have also large positive weights, namely, clusters 2, 4, and 5 while clusters 1 and 7 do not receive much weight. Apparently, variables such as Exports, Imports and GDP deflator included in the 6-th cluster play an important role in predicting interest rate. Also note that cluster 3, which includes the 3-month Treasury Bills, has the lowest weight in Figure 8. This confirms evidence in Ludvigson and Ng (2009) that relying only on the term structure information for predicting yields gives less accurate results than applying a large database including real and inflation factors.

Figure 9 shows the De Finetti's diagram of the weights within each cluster, which are driven by the common factors z_{jt}, $j = 1, \ldots, 7$ and the coefficients b_{ijt}. The ternary diagrams indicate there are large differences across clusters: for clusters 2, 4, 5 and 6, only a few models have most of the weights, which we are able to identify as models 20, 1, 22 and 4. Also, within these clusters there are more heterogeneity over time and models, indicating that individual model performances change over time even if overall information given by each cluster is stable. As

[11] Interestingly, Stock and Watson (2012) find that a factor model with 5 factors provides superior predictions to factor models with less factors. We also investigate combinations with a lower number of clusters, precisely 2 and 3 clusters, but predictions are less accurate.

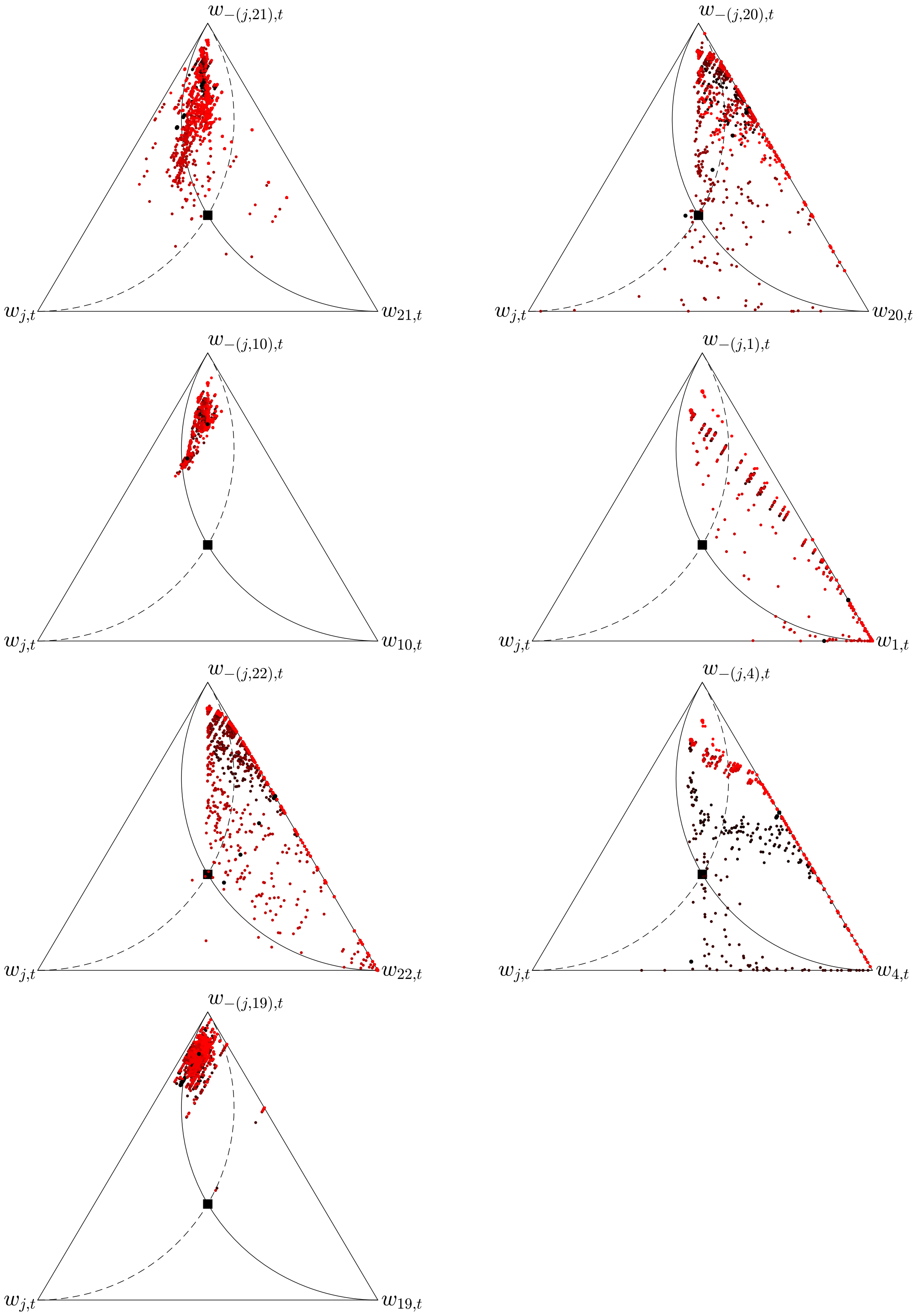

Figure 9: De Finetti's diagrams for the pairwise comparison between model weights at horizon $h = 1$*. In each plot the estimated ternaries* $(z_{it}, z_{jt}, z_{-(ij)t})$*,* $j > i$ *and* $t = 1, \ldots, 106$ *(red dots) for each cluster (different plots) where we choose the weight of the following reference models* $21, 20, 10, 1, 22, 4, 19$ *for the clusters* $1, \ldots, 7$*, respectively.*

		h=1			h=3			h=5	
	PE	LS	CRPS	PE	LS	CRPS	PE	LS	CRPS
					3-month Treasury Bills				
AR	0.569	-1.058	0.363	0.518	-1.038	0.343	0.545	-1.041	0.358
BDFM	0.553*	-1.190	0.359	0.516	-1.092	0.392	0.517	-1.089	0.401
DCEW5	0.519	-0.778**	0.288**	**0.509**	-0.772**	0.283	0.525	-0.791**	0.292*
DCLS5	0.740	-1.254	0.448	0.532	-1.210	0.381	0.584	-1.286	0.424
DCEW7	0.525	-0.783**	0.289*	0.514	-0.768**	0.284*	0.522	-0.786**	0.289*
DCLS7	**0.512****	**-0.773****	**0.284***	0.514	**-0.770****	**0.284***	**0.511***	**-0.793****	**0.289***

Table 2: Predicting results for $h = 1, 3, 5$ steps ahead 3-month Treasury Bill yields. Root mean square Prediction error (PE), Logarithmic Score (LS) and the Continuous Rank Probability Score (CRPS) are reported. Bold numbers indicate the best statistic for each horizon and loss function. One or two asterisks indicate that differences in accuracy versus the AR benchmark are credibly different from zero at 5%, and 1%, respectively, using the Diebold-Mariano t-statistic for equal loss. The underlying p-values are based on t-statistics computed with a serial correlation-robust variance, using the pre-whitened quadratic spectral estimator of Andrews and Monahan (1992).

regards to the other clusters (1, 3 and 7) similar weights occur across models since the dots in the diagrams concentrate around the point $(1/n_j, 1/n_j, (n_j - 2)/n_j)$.

Table 2 reports the results to predict 3-month Treasury Bills for three different horizons and using three different scoring measures. For all variables, horizons and scoring measures our methodology provides more accurate predictions than the AR(1) benchmark and the DFM benchmark. The DFM model provides more accurate predictions than the AR(1) when focusing on the mean square prediction error and CRPS metrics, but not for LS evaluation. On the contrary, several of our combination schemes outperform this benchmark for all three metrics and horizons. The predictive gains are similar across different horizons, that is up to 10% relative to the AR benchmark in terms of RMSPE metrics and even larger for the log score and CRPS measures.[12] The combination that provides the largest gain is the one based on seven clusters and log score weights within clusters (DCLS7), resulting in the best statistics 8 times out of 9 cases in the Table. In most of the cases, the difference is statistically credible at the 1% level. Fan charts in Figure S.4 of the Supplementary Material show that the predictions are accurate even at our longest horizon, $h = 5$. Note that the combination based on 5 clusters and equal weights yields also accurate predictions.

We conclude that combining joint model predictions using multiple clusters with cluster-based weights provides substantial predictive gains, confirming recent evidence that machine learning type of tools are useful for predicting financial markets, see for example Gu et al. (2020) and Bianchi et al. (2020). Of course, additional gains may be obtained by playing with a more detailed cluster grouping and different performance scoring rules for weights associated with models inside

[12] One would expect that RMSPE's are monotonic decreasing over longer horizons. This is not everywhere observed and is due to the fact of model misspecification.

a cluster. This is left as a topic for further research.

5 Conclusions

We proposed a Bayesian semi-parametric model to construct a time-varying weighted combination of a large set of predictive densities. The dimensionality reduction is based on clustering the set of predictive densities in mutually exclusive subsets. This modelling strategy reduces the dimension of the latent spaces and leads to a more parsimonious combination model. We provide several theoretical properties of the weights and propose an efficient procedure for posterior approximation of the latent weights.

We applied the methodology to large financial data sets and find substantial gains in point and density predicting for stock returns and Treasury yields. In a stock market application, we show, using our methodology, how 3712 predictive densities based on 1856 US individual stocks replicate the daily S&P500 index return and predict accurately the economic value of tail events like Value-at-Risk. In a bond market exercise, we show that combining models with cluster-based weights increases predictive accuracy substantially; weights across clusters are very stable over time and horizons. Furthermore, weights within clusters are very volatile, indicating that individual model performances are very unstable, strengthening the use of density combinations.

The line of research presented in this paper can be extended in several directions. For example, the importance of the model clusters change following the variable to predict. This calls for the use of multivariate combination models. Some clusters have a substantial weight while others have only little weight and such a pattern may vary over time. This may lead to the construction of alternative models with asymmetric distributions in the combination weights with the aim to get more accurate out-of-sample predicting and to improve policy analysis. We notice also potential fruitful applications of our approach to dynamic portfolio allocation and to study the effects of the COVID-19 pandemic on financial predictions.

References

Aastveit, K. A., Mitchell, J., Ravazzolo, F., and van Dijk, H. K. (2019). The evolution of forecast density combinations in economics. In *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press.

Aastveit, K. A., Ravazzolo, F., and van Dijk, H. K. (2018). Combined density nowcasting in an uncertain economic environment. *Journal of Business Economics & Statistics*, 36(1):131–145.

Aitchinson, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society Series, Series B*, 44:139–177.

Aitchinson, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.

Aitchinson, J. (1992). On criteria for measures of compositional difference. *Mathematical Geology*, 24:365–379.

Aitchinson, J. and Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, 67:261–272.

Aldrich, E. M. (2014). Gpu computing in economics. In L., K. J. and Schmedders, K., editors, *Handbook of Computational Economics, Vol. 3*. Elsevier.

Aldrich, E. M., Fernández-Villaverde, J., Gallant, A. R., and Rubio Ramırez, J. F. (2011). Tapping the supercomputer under your desk: Solving dynamic equilibrium models with graphics processors. *Journal of Economic Dynamics and Control*, 35:386–393.

Andrews, D. W. K. and Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, 60:953–966.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society Series, Series B*, 72:269–342.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient monte carlo computations. *Annals of Statistics*, 37:697–725.

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25:71–92.

Bastuerk, N., Borowska, A., Grassi, S., Hoogerheide, L., and van Dijk, H. K. (2019). Forecast density combinations of dynamic models and data driven portfolio strategies. *Journal of Econometrics*, 210:170–186.

Bianchi, D., Buechner, M., and Tamoni, A. (2020). Bond risk premiums with machine learning. *The Review of Financial Studies*.

Billheimer, D., Guttorp, P., and Fagan, W. F. (2001). Statistical interpretation of species composition. *Journal of the America Statistical Association*, 96:1205–1214.

Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. K. (2013). Time-varying combinations of predictive densities usif nonlinear filtering. *Journal of Econometrics*, 177:213–232.

Boonen, T. J., Guillen, M., and Santolino, M. (2019). Forecasting compositional risk allocations. *Insurance: Mathematics and Economics*, 84:79–86.

Brunsdon, T. and Smith, T. (1998). The time series analysis of compositional data. *Journal of Official Statistics*, 14:237 – 253.

Cannings, C. and Edwards, A. W. F. (1968). Natural selection and the De Finetti diagram. *Annals of Human Genetics*, 31:421–428.

Cargnoni, C., Müller, P., and West, M. (1997). Bayesian forecasting of multinomial time series through conditionally gaussian dynamic models. *Journal of the American Statistical Association*, 92(438):640–647.

Casarin, R., Grassi, S., Ravazzolo, F., and van Dijk, H. K. (2015). Parallel sequential Monte Carlo for efficient density combination: the DeCo Matlab toolbox. *Jounal of Statistical Software*, 68(3):1–30.

Casarin, R. and Marin, J. M. (2009). Online data processing: Comparison of Bayesian regularized particle filters. *Electronic Journal of Statistics*, 3:239–258.

Casarin, R. and Veggente, V. (2020). Random projection methods in economics and finance. In Petr, H., Uddin, M., and Abedin, M. Z., editors, *The Essentials of Machine Learning in Finance and Accounting*, pages 1–20. Routledge Taylor & Francis.

Choi, H. and Varian, H. (2012). Predicting the present with Google trends. *Economic Record*, 88:2–9.

Clark, T. E. and McCracken, M. W. (2011). Testing for unconditional predictive ability. In Clements, M. and Hendry, D., editors, *Oxford Handbook of Economic Forecasting*. Oxford University Press, Oxford.

Clark, T. E. and McCracken, M. W. (2015). Nested forecast model comparisons: a new approach to testing equal accuracy. *Journal of Econometrics*, 186:160–177.

Clark, T. E. and Ravazzolo, F. (2015). The macroeconomic forecasting performance of autoregressive models with alternative specifications of time-varying volatility. *Journal of Applied Econometrics*, 30:551–575.

Dey, D., Iyengar, M., and Ravishanker, N. (2001). Compositional time series analysis of mortality proportions. *Communications in Statistics - Theory and Methods*, 30(11):2281–2291.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Stastistics*, 13:253–263.

Doucet, A., Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag.

Dziubinski, M. P. and Grassi, S. (2013). Heterogeneous computing in economics: A simplified approach. *Computational Economics*, 43:485–495.

Egozcue, J. J. and Pawlowskky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37:795–828.

Egozcue, J. J., Pawlowskky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35:279–300.

Ehm, W., Gneiting, T., Jordan, A., and Kruger, F. (2016). Of quantiles and expectiles: consistent scoring functions, choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B*, (78):505–562.

Einav, L. and Levin, J. (2014). Economics in the age of big data. *Science*, 346(6210):715–718.

Favirar, R., Rebolledo, D., Chan, E., and Campbell, R. (2008). A parallel implementation of K-Means clustering on GPUs. *Proceedings of 2008 International Conference on Parallel and Distributed Processing Techniques and Applications*, 2:14–17.

Fiŝerová, E. and Hron, K. (2011). On the interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, 43:455–468.

Gentle, J. E. (2007). *Matrix Algebra: Theory, computations, and applications in statistics*. Springer Texts in Statistics. Springer, 1 edition.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88:881–889.

Geweke, J. and Durham, G. (2012). Massively parallel sequential Monte Carlo for Bayesian inference. Working papers, National Bureau of Economic Research, Inc.

Geweke, J. and Keane, M. (2007). Smoothly mixing regressions. *Journal of Econometrics*, 138:252–290.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48:1779–1801.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378.

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold and quantile weighted scoring rules. *Journal of Business and Economic Statistics*, 29:411–422.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782.

Granger, C. W. J. (1998). Extracting information from mega-panels and high-frequency data. *Statistica Neerlandica*, 52:258–272.

Groen, J. J. J., Paap, R., and Ravazzolo, F. (2013). Real-time inflation forecasting in a changing world. *Journal of Business & Economic Stastistics*, 31:29–44.

Grunwald, G. K., Raftery, A. E., and Guttorp, P. (1993). Time series of continuous proportions. *Journal of the Royal Statistical Society B*, 55:103–116.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5):2223–2273.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Journal of Neural Computation*, 3:79–87.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Journal Neural Computation*, 6:181–214.

Jordan, M. I. and Xu, L. (1995). Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 8:1409–1431.

Jorion, P. (2006). *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill, New York.

Katz, J. N. and King, G. (1999). A statistical model for multiparty electoral data. *American Political Science Review*, 93:15–32.

Kitagawa, G. (1998). Self-organizing state space model. *Journal of the American Statistical Association*, 93:1203–1215.

Koop, G. (2003). *Bayesian Econometrics*. John Wiley and Sons.

Koop, G. and Korobilis, D. (2009). Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends in Econometrics*, 3:267–358.

Koop, G. and Korobilis, D. (2013). Large time-varying parameter VARs. *Journal of Econometrics*, 177:185–198.

Kynclova, P., Filzmoser, P., and Hron, K. (2015). Modeling compositional time series with vector autoregressive models. *Journal of Forecasting*, 34(4):303–314.

Lee, A., Yau, C., Giles, M. B., Doucet, A., and Holmes, C. C. (2010). On the utility of graphic cards to perform massively parallel simulation with advanced Monte Carlo methods. *Journal of Computational and Graphical Statistics*, 19:769–789.

Lerch, S., Thorarinsdottir, T., Ravazzolo, R., and Gneiting, T. (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, 32(1):106–127.

Liu, J. and West, M. (2001). Combined parameter and state estimation in simulation based filtering. In Doucet, A., de Freitas, N., and Gordon, N., editors, *Sequential Monte Carlo Methods in Practice*. Springer-Verlag.

Ludvigson, S. C. and Ng, S. (2009). Macro factors in bond risk premia. *The Review of Financial Studies*, 22(12):5027–5067.

Marcellino, M., Stock, J. H., and Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135:499–526.

Mateu-Figueras, G., Pawlowsky-Glahn, V., and BarceloÂ´-Vidal, C. (2005). The additive logistic skew-normal distribution on the simplex. *Stochastic Environmental Research and Risk Assessment*, 19:205–214.

McAlinn, K. and West, M. (2019). Dynamic bayesian predictive synthesis in time series forecasting. *Journal of Econometrics*, 210(1):155 – 169.

Morozov, S. and Mathur, S. (2012). Massively parallel computation using graphics processors with application to optimal experimentation in dynamic control. *Computational Economics*, 40:151–182.

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional Data Analysis: Theory and Applications*. Wiley.

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. Wiley.

Peng, F., Jacobs, R. A., and Tanner, M. A. (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, 91:953–960.

Quintana, J. and West, M. (1988). Time series analysis of compositional data. In Bernardo, J., DeGroot, M., Lindley, D., and Smith, A., editors, *Bayesian Statistics 3*, pages 747–756. Elsevier.

Ravazzolo, F. and Vahey, S. V. (2014). Forecast densities for economic aggregates from disaggregate ensembles. *Studies in Nonlinear Dynamics and Econometrics*, 18:367–381.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Verlag.

Snyder, R. D., Ord, J. K., Koehler, A. B., McLaren, K. R., and Beaumont, A. N. (2017). Forecasting compositional time series: A state space approach. *International Journal of Forecasting*, 33(2):502 – 512.

Stock, J. H. and Watson, W. M. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44:293–335.

Stock, J. H. and Watson, W. M. (2002). Forecasting using principal components from a large number of predictors. *Journal of American Statistical Association*, 97:1167–1179.

Stock, J. H. and Watson, W. M. (2005). Implications of dynamic factor models for VAR analysis. Technical report, NBER Working Paper No. 11467.

Stock, J. H. and Watson, W. M. (2012). Disentangling the channels of the 2007-09 recession. *Brookings Papers on Economic Activity*, pages 81–156, Spring.

Stock, J. H. and Watson, W. M. (2014). Estimating turning points using large data sets. *Journal of Econometris*, 178:368–381.

Varian, H. (2014). Machine learning: New tricks for econometrics. *Journal of Economics Perspectives*, 28:3–28.

Varian, H. and Scott, S. (2014). Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5:4–23.

Villani, M., Kohn, R., and Giordani, P. (2009). Regression density estimation using smooth adaptive Gaussian mixtures. *Journal of Econometrics*, 153:155–173.

Wallis, F. K. (1987). Time series analysis of bounded economic variables. *Journal of Time Series Analysis*, 8:115–123.

Wood, S. A., Jiang, W., and Tanner, M. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika*, 89:513–528.

A Proofs of the results in the paper

Proof of Proposition 2.1 Let $a = (a_{1t}, \ldots, a_{nt})$ then our probability density combination model writes as

$$f(y_t|\Im_{t-1}, \mathcal{M}, \sigma_t^2) = \int_{\mathbb{R}^n \times \mathbb{R}^n} G(y_t) H_{0t}^A(da) \tag{A.1}$$

$$= \int_{\mathbb{R}^n \times \mathbb{R}^n} \sum_{i=1}^n w_{it} \delta(\tilde{y}_{it} + \varepsilon_{it} - y_t) \prod_{i=1}^n f(\tilde{y}_{it}|\Im_{i,t-1}, M_i) g(\varepsilon_{it}|\sigma_{it}^2) d\tilde{y}_{it} d\varepsilon_{it} \tag{A.2}$$

$$\sum_{i=1}^n w_{it} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \delta(\tilde{y}_{it} + \varepsilon_{it} - y_t) g(\varepsilon_{it}|\sigma_{it}^2) d\varepsilon_{it} \right) f(\tilde{y}_{it}|\Im_{i,t-1}, M_i) d\tilde{y}_{it} \tag{A.3}$$

where last line follows from Fubini's theorem and the conditional independence assumption for the atoms. Solving the integral with respect to ε_{it} one obtains

$$f(y_t|\Im_{t-1}, \mathcal{M}, \sigma_t^2) = \sum_{i=1}^n w_{it} \int_{\mathbb{R}} g(y_t - \tilde{y}_{it}|\sigma_{it}^2) f(\tilde{y}_{it}|\Im_{i,t-1}, M_i) d\tilde{y}_{it} \tag{A.4}$$

which concludes the proof.

Proof of Proposition 2.2 The Jacobian of the inverse transformation $\mathbf{v} = \log(\mathbf{u})$ is the diagonal matrix $\mathrm{J}(\mathbf{v}) = \mathrm{diag}((\mathrm{u}_1^{-1}, \ldots, \mathrm{u}_\mathrm{m}^{-1})')$ where $\mathrm{diag}(\mathbf{a})$ is the diagonal matrix with the elements of the vector $\mathbf{a}$ on the main diagonal. The probability density function of $\mathbf{u}$ is

$$\begin{aligned} p(\mathbf{u}|\boldsymbol{\mu}, \Upsilon) &= |(2\pi)\Upsilon|^{-1/2}|\mathrm{J}(\mathbf{v})| \exp\Big((\log(\mathbf{u}) - \boldsymbol{\mu})'\Upsilon^{-1}(\log(\mathbf{u}) - \boldsymbol{\mu})\Big) \\ &= |(2\pi)\Upsilon|^{-1/2}\left(\prod_{j=1}^{m} u_j\right)^{-1} \exp\left(-\frac{1}{2}(\log(\mathbf{u}) - \boldsymbol{\mu})'\Upsilon^{-1}(\log(\mathbf{u}) - \boldsymbol{\mu})\right) \quad \text{(A.5)} \end{aligned}$$

which is the pdf of the m-variate log-normal distribution $\Lambda_m(\boldsymbol{\mu}, \Upsilon)$.

The transformation $\mathbf{z} = \mathrm{alr}^{-1}(\mathbf{v}) = C_m((\exp(\mathbf{v}_{-m}), 1)')$ implies $z_m = 1 - \boldsymbol{\iota}'_{m-1}\mathbf{z}_{-m}$ and

$$\mathbf{z}_{-m} = \mathbf{u}_{-m}/(1 + \boldsymbol{\iota}'_{m-1}\mathbf{u}_{-m}) \quad \text{(A.6)}$$

where $\mathbf{u}_{-m} = \exp(\mathbf{v}_{-m} - v_m\boldsymbol{\iota}_{m-1}) = \exp(D_m\mathbf{v})$. By the properties of the log-normal distribution it follows that $\mathbf{u}_{-m} \sim \Lambda_{m-1}(D_m\boldsymbol{\mu}, D_m\Upsilon D'_m)$. From Eq. A.6 one obtains $\mathbf{z}_{-m} = (I_{m-1} - \mathbf{z}_{-m}\boldsymbol{\iota}_{m-1})\mathbf{u}_{m-1}$ and the inverse transformation

$$\mathbf{u}_{-m} = (I_m - \mathbf{z}_{-m}\boldsymbol{\iota}'_{m-1})^{-1}\mathbf{z}_{-m} = \frac{1}{1 + \boldsymbol{\iota}'_{m-1}\mathbf{z}_{-m}}\mathbf{z}_{-m} \quad \text{(A.7)}$$

where the last equality follows from the Sherman-Morrison-Woodbury's formula $(A + \mathbf{x}\mathbf{y}')^{-1} = A^{-1} - A^{-1}\mathbf{x}\mathbf{y}'A^{-1}(1 + \mathbf{y}'A^{-1}\mathbf{x})^{-1}$ (e.g., see Gentle, 2007, p. 221). The Jacobian of this transformation is

$$\mathrm{J}\mathbf{u}_{-\mathrm{m}} = \frac{1}{\mathrm{d}}\mathrm{I}_{\mathrm{m}-1} - \frac{1}{\mathrm{d}^2}\boldsymbol{\iota}_{\mathrm{m}-1}\mathbf{z}'_{-\mathrm{m}} \quad \text{(A.8)}$$

where $d = 1 + \boldsymbol{\iota}'_{m-1}\mathbf{z}_{-m}$. By applying the determinant rule $|A + \mathbf{x}\mathbf{y}'| = |A|(1 + \mathbf{y}'A^{-1}\mathbf{x})$ one obtains

$$|\mathrm{J}\mathbf{u}_{-\mathrm{m}}| = \frac{1}{\mathrm{d}^{\mathrm{m}-1}}(1 - \frac{1}{\mathrm{d}}\mathbf{z}'_{-\mathrm{m}}\boldsymbol{\iota}_{\mathrm{m}-1}) = \frac{1}{\mathrm{d}^{\mathrm{m}-1}}\frac{1 + \boldsymbol{\iota}'_{\mathrm{m}-1}\mathbf{z}_{-\mathrm{m}} - \boldsymbol{\iota}'_{\mathrm{m}-1}\mathbf{z}_{-\mathrm{m}}}{\mathrm{d}} = \mathrm{d}^{-\mathrm{m}} \quad \text{(A.9)}$$

and the density function

$$\begin{aligned} p(\mathbf{z}|D_m\boldsymbol{\mu}, D_m\Upsilon D'_m) &= |(2\pi)D_m\Upsilon D_m|^{-1/2}|\mathrm{J}\mathbf{u}_{-\mathrm{m}}|\left(\prod_{\mathrm{j}=1}^{\mathrm{m}-1}\frac{\mathrm{z_j}}{\mathrm{d}}\right)^{-1} \\ &\quad \exp\Big(\log(\mathbf{z}_{-m}c) - D_m\boldsymbol{\mu})'(D_m\Upsilon D'_m)^{-1}(\log(\mathbf{z}_{-m}c) - D_m\boldsymbol{\mu})\Big) \\ &= |(2\pi)D_m\Upsilon D_m|^{-1/2}\left(\prod_{j=1}^{m} z_j\right)^{-1} \\ &\quad \exp\Big(\log(\mathbf{z}_{-m}/z_m) - D_m\boldsymbol{\mu})'(D_m\Upsilon D'_m)^{-1}(\log(\mathbf{z}_{-m}/z_m) - D_m\boldsymbol{\mu})\Big)\text{(A.10)} \end{aligned}$$

$$\begin{tikzcd}
(\mathbb{R}^{m-1},\mathcal{N}_{m-1}) \arrow[r, "\exp"] & (\mathbb{R}_{+}^{m-1},\Lambda_{m-1}) \arrow[d, "C_m"] \\
 & (\mathbb{S}^{m},\mathcal{L}_{m}) \arrow[ul, "\text{alr}"]
\end{tikzcd}$$

Figure A.1: Relationship between the space of real vector $\mathbb{R}^{m-1}$ endowed with the normal distribution, the positive real $\mathbb{R}_{+}^{m-1}$ endowed with the log-normal distribution and the simplex $\mathbb{S}^{m}$ endowed with the logistic-normal distribution $\mathcal{L}_{m}$.

where $c = (1+\boldsymbol{\iota}_{m-1}'\mathbf{u}_{-m})$ and we used $z_m = 1-\boldsymbol{\iota}_{m-1}'\mathbf{z}_{-m} = 1-\boldsymbol{\iota}_{m-1}'\mathbf{u}_{-m}/(1+\boldsymbol{\iota}_{m-1}'\mathbf{u}_{-m})$ which gives $(1+\boldsymbol{\iota}_{m-1}'\mathbf{u}_{-m})^{-1} = z_m$. In this proof we show that the logistic-normal can be obtained by transforming log-normal variables, an alternative proof can be obtained by considering the Jacobian of the inverse $\text{alr}(\cdot)$ transformation (see Fig. A.1).

Proof of Proposition 2.3 The weights w_i can be written as

$$\begin{aligned}
w_i &= \exp\Big(\sum_{j=1}^{d-1} a_{ij}\log(z_j/z_d)\Big)\left(1+\sum_{i=1}^{c}\exp\Big(\sum_{j=1}^{d-1} a_{ij}\log(z_j/z_d)\Big)\right)^{-1} \\
&= \exp(u_i)\left(1+\sum_{i=1}^{c}\exp(u_i)\right)^{-1} \qquad (A.11)
\end{aligned}$$

where $\mathbf{a}_i = (a_{i1},\ldots,a_{id-1})$ and $u_i = \mathbf{a}_i\,\text{alr}(\mathbf{z})$, $i=1,\ldots,c$. By applying the definition of $\text{alr}^{-1}(\cdot)$ one obtains $\mathbf{w} = (w_1,\ldots,w_c)' = \text{arl}^{-1}(\mathbf{u}^*)$ and $w_{c+1} = 1-w_1-\ldots-w_c$, where $\mathbf{u}^* = (\mathbf{u}',\kappa)'$ with $\kappa\in\mathbb{R}$, $\mathbf{u} = (u_1,\ldots,u_c)' = A\,\text{alr}(\mathbf{z})$. From the properties of logistic-normal it follows $\text{alr}(\mathbf{z}) \sim \mathcal{N}_{\mathrm{d-1}}(\boldsymbol{\mu},\Upsilon)$ and $A\,\text{alr}(\mathbf{z}) \sim \mathcal{N}_{\mathrm{c}}(\mathrm{A}\boldsymbol{\mu},\mathrm{A}\Upsilon\mathrm{A}')$. In conclusion $\mathbf{w}\sim\mathcal{L}_{c+1}(A\boldsymbol{\mu},A\Upsilon A')$

Proof of Corollary 2.1 From the definition of perturbation operator $\oplus$ and the properties of the log-normal distribution given in Proposition 2.2 it follows that the log-ratio process $\text{alr}(\mathbf{z}_{\mathrm{t}}) = \text{alr}(\mathbf{z}_{\mathrm{t-1}})+\text{alr}(\boldsymbol{\eta}_{\mathrm{t}})$ follows $\mathcal{N}_{m-1}(\text{alr}(\mathbf{z}_{\mathrm{t-1}}),\Upsilon)$ and the random composition process $\mathbf{z}_t = \text{alr}^{-1}(\text{alr}(\mathbf{z}_{\mathrm{t}})) \sim \mathcal{L}_{\mathrm{m}}(\text{alr}(\mathbf{z}_{\mathrm{t-1}}),\mathrm{D}_{\mathrm{m}}\Upsilon\mathrm{D}_{\mathrm{m}}')$. Then by setting $\mathbf{w}=\mathbf{w}_t$, $A=A_t$ and $\mathbf{z}=\mathbf{z}_t$ in Proposition 2.3 one obtains the result.

Proof of Remark 1 Without loss of generality, we assume that the $n-n_t$ elements in the cluster m correspond to the last $n-n_t$ elements of $\tilde{\mathbf{y}}_t$. Under this assumption the projection matrix can be partitioned as follows

$$\tilde{A}_t = \left(\begin{array}{c|c} A_t^* & \mathbf{0}_{n_t} \\ \hline O_{(n-n_t)\times(m-1)} & \mathbf{a}_t \end{array}\right)$$

where $\mathbf{0}_n$ and $O_{n\times m}$ denote the null vector and matrix and $\mathbf{a}_t$ is a $n-n_t$ dimensional vector such that $\mathbf{a}_t'\boldsymbol{\iota}_{n-n_t}=1$. Matrix A_t is thus equal to $(A_t^{*\prime},O_{n-n_t}')'$ and the scale

$$\begin{tikzcd}
(\mathbb{R}^{d-1}, \mathcal{N}_{d-1}) \arrow[d, dashed, "\mathrm{alr}^{-1}"'] & (\mathbb{R}^{m-1}, \mathcal{N}_{m-1}) \arrow[l, dashed, "A"'] \arrow[d, "\mathrm{alr}^{-1}"] \\
(\mathbb{S}^{d}, \mathcal{L}_{d}) & (\mathbb{S}^{m}, \mathcal{L}_{m}) \arrow[l, "\phi_A"]
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathbf{x}_t \arrow[d, dashed, "\mathrm{alr}"'] & \mathbf{v}_t \arrow[l, dashed, "A"'] \arrow[d, "\mathrm{alr}^{-1}"] \\
\mathbf{w}_t^* & \mathbf{z}_t \arrow[l, "\phi_A"]
\end{tikzcd}$$

Figure A.2: Relationships between Gaussian and logistic-normal representation of the latent probability space (left) and of the latent process (left) involved in our compositional factor model. In the directed edges the arrow indicates the results of the transformation, and the edge label defines the transformation applied. Note in that in this diagram we allow for mapping onto subspaces of $\mathbb{R}^n$.

matrix of the logistic-normal distribution is thus

$$A_t \Upsilon A_t' = \left(\begin{array}{c|c} A_t^* \Upsilon^* A_t^{*\prime} & O_{n_t \times (n-n_t)} \\ \hline O_{(n-n_t)\times(n_t)} & O_{(n-n_t)\times(n-n_t)} \end{array} \right)$$

where Υ^* is the matrix given by the first n_t rows and columns of Υ.

Proof of Corollary 2.2 For the easy of notation, in the following we assume $d = n_t + 1$ and $A = A_t^*$ where A_t^* is the projection matrix defined in the proof of Remark 1. In the first part of the proof we show that by applying a chain of transformations, a Gaussian process in $\mathbb{R}^m$ has a logistic-normal process representation on the simplex $\mathbb{S}^d$ (dashed lines in Figure A.2). The process $\mathbf{x}_t = A\mathbf{v}_t$ has a Gaussian distribution, $\mathbf{x}_t \sim \mathcal{N}_{d-1}(A\mathbf{v}_{t-1}, A\Upsilon A')$, and the transformed process $\mathbf{w}_t^* = \mathrm{alr}^{-1}((\mathbf{x}_{\mathrm{t}}, 1))$ is in $\mathbb{S}^d$ and follows $\mathcal{L}_d(A\mathbf{v}_{t-1}, A\Upsilon A')$. In the second part we show that our compositional model in $\mathcal{S}^d$ has an equivalent representation in a subspace of $\mathbb{R}^{d-1}$ (solid lines in Figure A.2). By Propositions 2.2 $\mathbf{z}_t$ is in $\mathbb{S}^m$ and follows $\mathcal{L}_m(\mathbf{v}_{t-1}, \Upsilon)$ with $\mathbf{v}_{t-1} = \mathrm{alr}(\mathbf{z}_{\mathrm{t-1}})$. By Corollary 2.1 the process $\mathbf{w}_t^* = \phi_A(\mathbf{z}_t)$ is in $\mathbb{S}^d$ and follows $\mathcal{L}_d(A\mathbf{v}_{t-1}, A\Upsilon\tilde{A}')$.

Supplementary Material

B Parallel implementation

B.1 Parallel sequential filtering

With regard to the filtering part, we use M parallel conditional SMC filters, where each filter is conditioned on the predictor vector sequence $\tilde{\mathbf{y}}_s$, $s = 1, \ldots, t$. We initialise independently the M particle sets: $\Phi_0^j = \{\boldsymbol{\omega}_0^{ij}, \tilde{\gamma}_0^{ij}\}_{i=1}^N$, $j = 1, \ldots, M$. Each particle set Φ_0^j contains N i.i.d. random variables $\boldsymbol{\omega}_0^{ij}$ with random weights $\tilde{\gamma}_0^{ij}$. We initialise the set of predictors, by generating i.i.d. samples $\tilde{\mathbf{y}}_1^j$, $j = 1, \ldots, M$, from $p(\tilde{\mathbf{y}}_1|\mathbf{y}_0)$ where $\mathbf{y}_0$ is an initial set of observations for the variable of interest.

Then, at the iteration $t+1$ of the combination algorithm, we approximate the predictive density $p(\tilde{\mathbf{y}}_{t+1}|\mathbf{y}_{1:t})$ as follows

$$p_M(\tilde{\mathbf{y}}_{t+1}|\mathbf{y}_{1:t}) = \frac{1}{M}\sum_{j=1}^{M} \delta(\tilde{\mathbf{y}}_{t+1}^j - \tilde{\mathbf{y}}_{t+1})$$

where $\tilde{\mathbf{y}}_{t+1}^j$, $j = 1, \ldots, M$, are i.i.d. samples from the predictive densities and $\delta_x(y)$ denotes the Dirac mass at x.

We assume an independent sequence of particle sets $\Phi_t^j = \{\boldsymbol{\omega}_{1:t}^{ij}, \tilde{\gamma}_t^{ij}\}_{i=1}^N$, $j = 1, \ldots, M$, is available at time $t+1$ and that each particle set provides the approximation

$$p_{N,j}(\boldsymbol{\omega}_t|\mathbf{y}_{1:t}, \tilde{\mathbf{y}}_{1:t}^j) = \sum_{i=1}^{N} \tilde{\gamma}_t^{ij} \delta(\boldsymbol{\omega}^{ij} - \boldsymbol{\omega}_t) \tag{B.12}$$

of the filtering density, $p(\boldsymbol{\omega}_t|\mathbf{y}_{1:t}, \tilde{\mathbf{y}}_{1:t}^j)$, conditional on the j-th predictor realisation, $\tilde{\mathbf{y}}_{1:t}^j$. Then M independent conditional SMC algorithms are used to find a new sequence of M particle sets, which include the information available from the new observation and the new predictors. Each SMC algorithm iterates, for $j = 1, \ldots, M$, the steps given in section 3.

After collecting the results from the different particle sets, it is possible to obtain the following empirical predictive density

$$p_{M,N}(\mathbf{y}_{t+1}|\mathbf{y}_{1:t}) = \frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N} \delta(\mathbf{y}_{t+1}^{ij} - \mathbf{y}_{t+1}) \tag{B.13}$$

For horizons $h > 1$, we apply a direct predicting approach (see Marcellino et al., 2006) and compute predictive densities $p_{M,N}(\mathbf{y}_{t+h}|\mathbf{y}_{1:t})$ following the steps previously described.

B.2 Parallel dynamic clustering

The parallel evaluation of the dynamic clustering process can be described as follows, see also Favirar et al. (2008) and the reference therein. Assume, for simplicity, the n data points can be split in P subsets, $N_p = \{(p-1)n_p+1, \dots, pn_p\}$, $p = 1\dots, P$, with the equal number of elements n_P. P is chosen according to the number of available cores.

1. Assign P sets of n_P data points to different cores.
2. For each core p, $p = 1, \dots, P$
   - 2a. find $j_i = \arg\min\{j = 1, \dots, m| \, ||\boldsymbol{\psi}_{it} - \mathbf{c}_{jt}||\}$, for each observation $i \in N_p$ assigned to the core p.
   - 2.b find the local centroid updates $\mathbf{m}_{p,jt+1}$, $j = 1, \dots, m$
3. Find the global centroid updates $\mathbf{m}_{jt+1} = 1/P \sum_{p=1}^{P} \mathbf{m}_{p,jt+1}$, $j = 1, \dots, m$
4. Update the centroids as in Eq. (19).

The dynamic clustering is parallel in point 2) and 3) and this can be used in the GPU context as we do in this paper.

C Predictive evaluation

To measure the predictive ability of our methodology, we consider several statistics for point and density predctions previously proposed in the literature. Assume we have n different approaches to predict the variable y.

Point predictions. We compare point predictions in terms of Root Mean Square Prediction Errors (RMSPE)

$$RMSPE_{i,h} = \sqrt{\frac{1}{t^*} \sum_{t=\underline{t}}^{\overline{t}} e_{i,t+h}},$$

where $t^* = \overline{t} - \underline{t} + h$, $\overline{t}$ and $\underline{t}$ denote the beginning and end of the evaluation period, and $e_{i,t+h}$ is the h-step ahead square prediction error of model i.

Density predictions. The complete predictive densities are evaluated as follows. Let $f(y_{t+h}|\mathfrak{I}_{it})$ be a candidate density obtained from the approach i. The Logarithmic Score (LS) is then given as:

$$LS_{i,h} = -\frac{1}{t^*} \sum_{t=\underline{t}}^{\overline{t}} \ln f(y_{t+h}|\mathfrak{I}_{it}) \tag{C.14}$$

for all i and choose the model for which this score is minimal, or, as we report in our tables and use in the learning strategies, its opposite is maximal.

We also evaluate density predictions based on the continuous rank probability score (CRPS); see, for example, Gneiting and Raftery (2007), Gneiting and Ranjan (2013), Groen et al. (2013) and Ravazzolo and Vahey (2014). The CRPS for the model i measures the average absolute distance between the empirical cumulative distribution function (CDF) of y_{t+h}, which is simply a step function in y_{t+h}, and the empirical CDF that is associated with model i's predictive density:

$$\begin{aligned} \mathrm{CRPS}_{i,t+h} &= \int_{-\infty}^{+\infty} \Big(F(z|\mathfrak{I}_{it}) - \mathbb{I}_{[y_{t+h},+\infty)}(z)\Big)^2 \mathrm{d}z \\ &= \mathbb{E}_t|\tilde{y}_{i,t+h} - y_{t+h}| - \frac{1}{2}\mathbb{E}_t|\tilde{y}^*_{i,t+h} - \tilde{y}'_{i,t+h}|, \end{aligned} \tag{C.15}$$

where $F(\cdot|\mathfrak{I}_{it})$ is the CDF from the predictive density $f(y_{t+h}|\mathfrak{I}_{it})$ of model i and $\tilde{y}^*_{i,t+h}$ and $\tilde{y}'_{i,t+h}$ are independent random variables with common sampling density equal to the posterior predictive density $f(y_{t+h}|\mathfrak{I}_{it})$. We report the sample average CRPS:

$$\mathrm{CRPS}_{i,h} = -\frac{1}{t^*}\sum_{t=\underline{t}}^{\bar{t}} \mathrm{CRPS}_{i,t+h}. \tag{C.16}$$

Smaller CRPS values imply higher precisions and, as for the log score, we report the average $\mathrm{CRPS}_{i,h}$ for each model i in all tables.

Tail predictions. Given that our approach produces complete predictive densities for the variable of interest, it is particularly suitable to compute tail events. We consider two statistics and an economic measure for tail events. We compute weighted averages of Gneiting and Raftery (2007) quantile scores that are based on quantile predictions that correspond to the predictive densities from the different models, i.e.,

$$\mathrm{QS}(\alpha, i, t) = \Big(\mathrm{I}\{y_{t+1} \leqq F^{-1}(\alpha, i)\} - \alpha\Big)\Big(F^{-1}(\alpha|\mathfrak{I}_{it}) - y_{t+1}\Big), \tag{C.17}$$

with $F^{-1}(\alpha|\mathfrak{I}_{it})$ is the 1-step ahead quantile prediction using prediction i for level $\alpha \in (0,1)$. It can be shown that integrating (C.17) over $\alpha \in (0,1)$ will result in the CRPS measure (C.15), see Gneiting and Ranjan (2011). Gneiting and Ranjan (2011), Groen et al. (2013) and Lerch et al. (2017) propose to integrate weighted versions of (C.17) over α, with these weights being fixed functions of α chosen such to emphasize in the predictive evaluation a certain area of the underlying predictive density. We use a discrete approximation to this integration and use weights that emphasize both tail and the left tail of the predictive density:

$$\begin{aligned}
\text{avQS-T}_i &= \frac{1}{T-t_0-1}\sum_{s=t_0-1}^{T-1}\left(\frac{1}{99}\sum_{j=1}^{99}(2\alpha_j-1)^2\text{QS}(\alpha_j,i,s+1)\right)\\
\text{avQS-L}_{i,h} &= \frac{1}{T-t_0-1}\sum_{s=t_0-1}^{T-1}\left(\frac{1}{99}\sum_{j=1}^{99}(1-\alpha_j)^2\text{QS}(\alpha_j,i,s+1)\right)
\end{aligned} \tag{C.18}$$

where $\alpha_j = j/100$ and $\text{QS}(\alpha_j, i, s+1)$ is defined in (C.17) for a quantile j. In (C.18), avQS-T emphasizes both tails and avQS-L the left tail of the predictive density relative to the realization 1-step ahead. To study how the models perform in the left tail prediction over time, we consider the cumulative sum of avQS-L:

$$\text{cumavQS-L}_{i,h,t} = \sum_{s=t_0-1}^{t} \text{avQS-L}_{i,h,s} \tag{C.19}$$

The most accurate model at observation t produces the lowest $\text{cumavQS-L}_{i,h,t}$.

Finally, following Clark and Ravazzolo (2015), we apply the Diebold and Mariano (1995) t-tests for equality of the average loss (with loss defined as squared error, log score, or CRPS). In our tables presented below, differences in accuracy that are statistically different from zero are denoted by one, two, or three asterisks, corresponding to significance levels of 10%, 5%, and 1%, respectively. The underlying p-values are based on t-statistics computed with a serial correlation-robust variance, using the pre-whitened quadratic spectral estimator of Andrews and Monahan (1992). Monte Carlo evidence in Clark and McCracken (2015) and Clark and McCracken (2011) indicates that, with nested models, the Diebold-Mariano test compared against normal critical values can be viewed as a somewhat conservative (conservative in the sense of tending to have size modestly below nominal size) test for equal accuracy in the finite sample. Since the AR benchmark is always one of the model in the combination schemes, we treat each combination as nesting the baseline, and we report p-values based on one-sided tests, taking the AR as the null and the combination scheme in question as the alternative.

D Additional details on empirical results

D.1 Additional details on the S&P500 application

Table D.1 reports the cross-section average statistics, together with statistics for the S&P500. Some series have much lower average returns than the index and volatility higher than the index up to 400 times. Heterogeneity in skewness is also very evident with the series with lowest skewness equal to -42.5 and the one with highest skewness equal to 27.3 compared to a value equal to -0.18 for the index. Finally, maximum kurtosis is 200 times higher than the index value. The inclusion in our sample of the crisis period explains such differences, with some stocks that realized enormously negative returns in 2008 and impressive positive returns in

	Subcomponents			S&P500
	Lower	Median	Upper	
Average	-0.002	0.000	0.001	0.000
St dev	0.016	0.035	0.139	0.019
Skewness	-1.185	0.033	1.060	-0.175
Kurtosis	8.558	16.327	65.380	9.410
Min	-1.322	-0.286	-0.121	-0.095
Max	0.122	0.264	1.386	0.110

Table D.1: Average cross-section statistics for the 1856 individual stock daily log returns in our dataset for the sample 18 March 2002 to 31 December 2009. The columns "Lower", "Median" and "Upper" refer to the cross-section 10% lower quantile, median and 90% upper quantile of the 3712 statistics in rows, respectively. The rows "Average", "St dev", "Skewness", "Kurtosis", "Min" and "Max" refers to sample average, sample standard deviation, sample skewness, sample kurtosis, sample minimum and sample maximum statistics, respectively. The column "S&P500" reports the sample statistics for the aggregate S&P500 log returns.

2009.

Figure D.1 shows for the time series of the full sample the cumulative avQS-L for the Student-t GARCH(1,1) model, the best ex-post GARCH model, the combination of GARCH models and DCEW model set. We note that our method requires some observations in the beginning to catch up with the other models. However, from August 2007 when stock markets start to experience large stress, it provides the most accurate tail predictions. The gap between the three models increases steadily over time and it becomes substantially larger after the collapse of Bear Stearns. With the default of the Lehman brothers, the accuracy of all three schemes reduces sharply until November/December 2008 when central banks and governments from around the World started to take actions which reduced the volatility in financial markets. Our DCEW, however, provides the lowest statistic until the end of the sample.

D.2 Additional details on the Treasury Bill predicting

We consider the extended Stock and Watson (2005) dataset, which includes 142 series sampled at a quarterly frequency from 1959Q1 to 2011Q2. A graphical description of the data is given in Figure D.2.

For each variable we estimate a Gaussian autoregressive model of the first order, AR(1),

$$y_{it} = \alpha_i + \beta_i y_{it-1} + \zeta_{it}, \quad \zeta_{it} \sim \mathcal{N}(0, \sigma_i^2) \tag{D.20}$$

using the first 60 observations from each series. Then we identify the clusters of parameters by applying our clustering algorithm on the vectors, $\hat{\boldsymbol{\theta}}_i = (\hat{\alpha}_i, \hat{\beta}_i, \hat{\sigma}_i^2)'$, of least square estimates of the AR(1) parameters. A detailed description of the 5 and 7 clusters is provided in Tables D.2-D.3.

Table D.2: Predictors classification in 5 clusters (columns).

1	2	3	4	5
NAPMprodn	Exports	RGDP	Cons-Dur	Cons-Serv
CapacityUtil	PGDP	Cons	Imports	FixedInv
Emptotal	PCED	Cons-NonDur	GovFed	NonResInv
Empgdsprod	CPI-ALL	GPDInv	IPfuels	NonResInv-Struct
Empdblegds	PCED-Core	Gov	Ul5wks	NonResInv-Bequip
Empservices	CPI-Core	GovStateLoc	U5-14wks	Res.Inv
EmpTTU	PCED-DUR-HHEQ	IPconsgds	Orders(NDCapGoods)	IPtotal
Empwholesale	PCED-DUR-OTH	IPconsdble	PCED-DUR	IPproducts
EmpFIRE	PCED-NDUR	IP:consnondble	PCED-DUR-MOTOR	IPfinalprod
Avghrs	PCED-NDUR-FOOD	Empmining	PCED-NDUR-OTH	IP:buseqpt
HStartsTotal	PCED-NDUR-CLTH	EmpCPStotal	PFI-NRES	IPmatls
BuildPermits	PCED-NDUR-ENERGY	Overtimemfg	PFI-NRES-EQP	IPdblemats
HStartsNE	PCED-SERV	Umeanduration	Pimp	IP:nondblemats
HStartsMW	PCED-SERV-HOUS	U15-26wks	LaborProd	IPmfg
HStartsSouth	PCED-SERV-HOUSOP	Orders(ConsGoods)	RealCompHour	Empconst
HStartsWest	PCED-SERV-H0-ELGAS	Comspotprice(real)	3moT-bill	Empmfg
PMI	PCED-SERV-HO-OTH	OilPrice(Real)	6moT-bill	Empnondbles
NAPMnewordrs	PCED-SERV-TRAN	RealAHEgoods	5yrT-bond	Empretail
NAPMvendordel	PCED-SERV-MED	RealAHEmfg	10yrT-bond	EmpGovt
NAPMInvent	PCED-SERV-REC	UnitLaborCost	Reservesnonbor	Helpwantedindx
NAPMcomprice	PCED-SERV-OTH	Aaabond	ExrateSwitz	Helpwantedemp
Consumerexpect	PGPDI	Baabond	ExrateJapan	EmpCPSnonag
fygm10-fygm3	PFI	Exrateavg	DJIA	EmpHours
Fyaaac-fygt10	PFI-NRES-STRPrInd	ExrateUK		Uall
Fyaaac-fygt10	PFI-RES	EXrateCanada		U15pwks
	Pexp	S&P500		U27pwks
	Pgov	S&Pindust		RealAHEconst
	PgovFed	S&Pdivyield		Conscredit
	Pgovstatloc	S&PPEratio		fygm1-fygm3
	FedFunds	fygm6-fygm3		
	1yrT-bond			
	M1			
	MZM			
	M2			
	MB			
	Reservestot			
	BUSLOANS			

Table D.3: Predictors classification in 7 clusters (columns).

1	2	3	4	5	6	7
FixedInv	Cons-Serv	Empmining	IPfuels	RGDP	Exports	NAPMprodn
NonResInv	NonResInv-Bequip	CPI-ALL	PCED	Cons	Imports	CapacityUtil
NonResInv-Struct	Res.Inv	PCED-NDUR	CPI-Core	Cons-Dur	Ul5wks	Empwholesale
IPproducts	GovStateLoc	PCED-NDUR-CLTH	PCED-DUR-OTH	Cons-NonDur	Orders(NDCapGoods)	Helpwantedindx
IP:buseqpt	IPtotal	PCED-NDUR-ENERGY	PCED-SERV	GPDInv	PGDP	Avghrs
IP:nondblemats	IPfinalprod	PCED-SERV-H0-ELGAS	PCED-SERV-HOUS	Gov	PCED-NDUR-FOOD	HStartsTotal
Emptotal	IP:consnondble	FedFunds	PCED-SERV-HO-OTH	GovFed	PCED-SERV-HOUSOP	BuildPermits
Empgdsprod	IPmfg	3moT-bill	PCED-SERV-TRAN	IPconsgds	PCED-SERV-MED	HStartsNE
Empmfg	Empdblegds	6moT-bill	PCED-SERV-REC	IPconsdble	PGPDI	HStartsMW
Empnondbles	Helpwantedemp	1yrT-bond	PCED-SERV-OTH	IPmatls	PFI	HStartsSouth
Empservices	Overtimemfg	5yrT-bond	PFI-NRES-STRPrInd	IPdblemats	PFI-NRES	HStartsWest
EmpTTU	Orders(ConsGoods)	10yrT-bond	Pimp	Empconst	PFI-RES	PMI
Empretail	PCED-Core	M1	PgovFed	EmpCPStotal	Pexp	NAPMnewordrs
EmpFIRE	PFI-NRES-EQP	MZM	Pgovstatloc	U5-14wks	Pgov	NAPMvendordel
EmpGovt	Comspotprice(real)	MB	M2	U15-26wks	BUSLOANS	OilPrice(Real)
EmpCPSnonag	RealAHEconst	Reservestot		U27pwks		NAPMcomprice
EmpHours	RealCompHour	Reservesnonbor		PCED-DUR		Conscredit
Uall	UnitLaborCost	ExrateUK		PCED-DUR-MOTOR		Consumerexpect
Umeanduration	S&P500	EXrateCanada		RealAHEgoods		fygm10-fygm3
U15pwks	fygm6-fygm3	S&Pindust		RealAHEmfg		Fyaaac-fygt10
NAPMInvent		DJIA		LaborProd		Fyaaac-fygt10
PCED-DUR-HHEQ				S&Pdivyield		
PCED-NDUR-OTH						
Aaabond						
Baabond						
Exrateavg						
ExrateSwitz						
ExrateJapan						
S&PPEratio						
fygm1-fygm3						

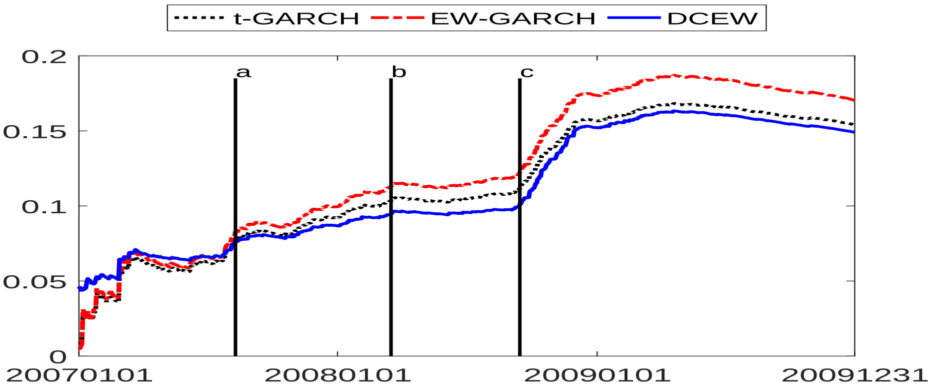

Figure D.1: Cumulative left quantile scores described in formula (C.19) *of the Student-t GARCH model, EW-GARCH model and DCEW. Timeline legend: a - 8/9/2007, BNP Paribas redemptions on three investment funds; b - 3/17/2008, collapse of Bear Stearns; c - 9/15/2008, Lehman bankruptcy.*

The left and right columns in Fig.D.3 show the clusters of series in the parameter space. The results show substantial evidence of different time series characteristics in several groups of series. The groups are not well separated when looking at the intercept values (see Fig. D.3, first and second row). However, the groups are well separated along two directions of the parameter space, which are the one associated with the variance and the one associated with persistence parameters (Fig.D.3, last row). The differences in terms of persistence, in the different groups, is also evident from the heat maps given in Fig.D.4. Different gray levels in the two graphs show the value of the variables (horizontal axis) over time (vertical axis). The vertical red lines indicate the different clusters. One can see for example that the series in the 2nd and 4th cluster (of 5) are more persistent then the series in the clusters 1, 3 and 5 (see also Fig. D.3, bottom left). Series in cluster 1, 2 and 4 are less volatile than series in the cluster 3 and 5. This information is also summarised by the mean value of the parameter estimates for the series that belong to the same cluster. See the values in Table D.4. Looking at the composition of the predictor groups (see also Tables D.2-D.3), we find for the five clusters that:

1. The first cluster comprises capacity utilisation, employment variables, housing (building permits and new ownership started) and manufacturing variables (new orders, supplier deliveries index, inventories).

2. The second cluster contains exports, a large numbers of price indexes (e.g. prices indexes for personal consumption expenditures, and for gross domestic product) some money market variables (e.g. M1 and M2).

3. The third cluster includes real gross domestic product, consumption and consumption of non-durables, some industrial production indexes, and some financial market variables (e.g., S&P industrial, corporate bonds and USD - GBP exchange rate).

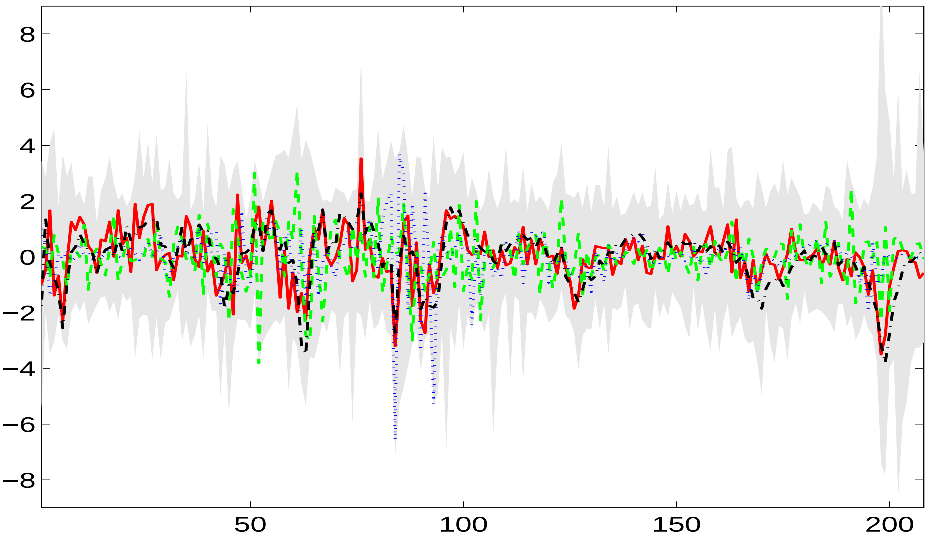

Figure D.2: Gray area: the set of series (standardised for a better graphical representation), at the monthly frequency, of the Stock and Watson dataset. Solid line: growth rate of real GDP (seasonally adjusted) for the US. Dashed line: inflation measured as the change in the GDP deflator index (seasonally adjusted). Dotted line: yields on US government 90-day T-Bills (secondary market). Dashed-dotted: total employment growth rate for private industries (seasonally adjusted).

4. The fourth cluster includes imports, some price indexes and financials such as government debt (3- and 6-months T-bills and 5- and 10-years T-bonds), stocks and exchange rates.

5. The fifth cluster mainly includes investments, industrial production indexes (total and many sector indexes), and employment.

Evidence is similar for the seven clusters.

D.3 Computing time

In this section we compare the computational speed of CPU with GPU in the implementation of our combination algorithm for both the financial and macro application. Whether CPU computing is standard in econometrics, GPU approach to computing has been received large attention in economics only recently. See, for example, Aldrich (2014) for a review, Geweke and Durham (2012) and Lee et al. (2010) for applications to Bayesian inference and Aldrich et al. (2011), Morozov and Mathur (2012) and Dziubinski and Grassi (2013) for solving DSGE models.

The CPU and the GPU versions of the computer program are written in MATLAB, as described in Casarin et al. (2015). In the CPU setting, our test machine is a server with two Intel Xeon CPU E5-2667 v2 processors and a total of 32 core. In the first GPU setting, our test machine is a NVIDIA Tesla K40c GPU. The Tesla K40c card is with 12GB memory and 2880 cores and it is installed in the

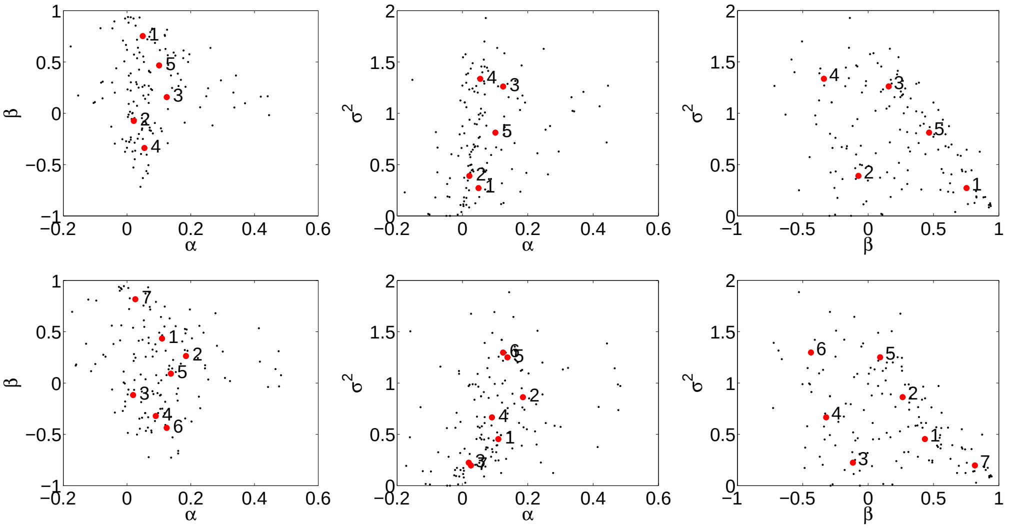

Figure D.3: Pairwise scatter plots of the series features: α_i and β_i (first column), α_i and σ_i^2 (second column) and β_i and σ_i^2 (last column). In each plot the red dots represent the cluster means. We assume alternatively 5 (left) and 7 (right) clusters.

CPU server. In the second GPU setting, our test machine is a NVIDIA GeForce GTX 660 GPU card, which is a middle-level video card, with a total of 960 cores. The test machine is a desktop Windows 8 machine, has 16 GB of Ram and only requires a MATLAB parallel toolbox license.

We compare two sets of combination experiments, the density combination based on 4 clusters with equal weights within clusters and time-varying volatility, DCEW-SV, see Section 4.1, and the density combination based on 7 clusters with recursive log score weights within clusters, DCLS7, see Section D.2, for an increasing number of particles N. In both sets of experiments we calculated, in seconds, the overall average execution time reported in Table D.5.

As the table shows, the CPU implementation is slower then the first GPU set-up in all cases. The NVIDIA Tesla K40c GPU provides gains in the order of magnitude from 2 to 4 times than the CPU. Very interestingly, even the second GPU set-up, which can be installed in a desktop machine, provides execution times comparable to the CPU in the financial applications and large gains in the macro applications. Therefore, the GPU environment seems the preferred one for our density combination problems and when the number of predictive density becomes very large a GPU server card gives the highest gains.

5 clusters			
k	α	β	σ^2
1	0.049	0.752	0.270
2	0.021	-0.074	0.390
3	0.124	0.157	1.260
4	0.054	-0.338	1.335
5	0.100	0.466	0.811

7 clusters			
k	α	β	σ^2
1	0.109	0.434	0.454
2	0.185	0.263	0.862
3	0.019	-0.116	0.224
4	0.090	-0.321	0.665
5	0.137	0.091	1.250
6	0.124	-0.437	1.297
7	0.026	0.817	0.197

Table D.4: Cluster means for the 5 (top table) and 7 (bottom table) cluster analysis. The first column, k, indicates the cluster number given in Fig. D.3 and the remaining three columns the cluster mean along the different directions of the parameter space.

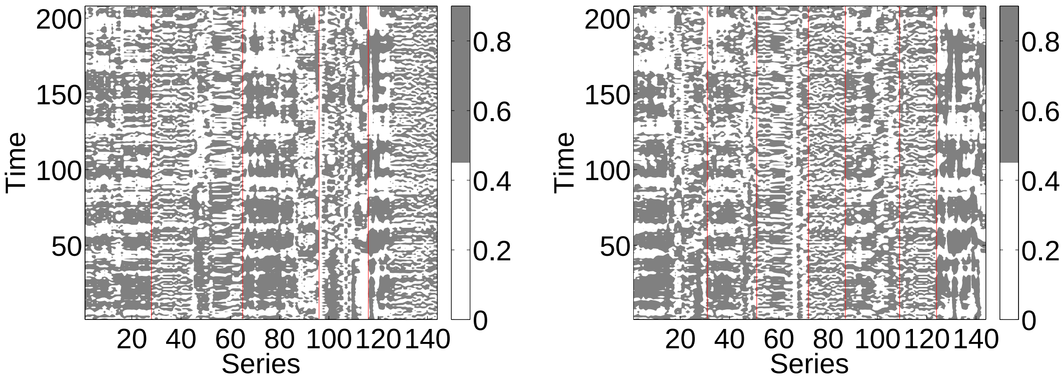

Figure D.4: Normal cumulative density function for the standardised series. The series are ordered by cluster label. We assume alternatively 5 (left) and 7 (right) clusters.

	DCEW-SV			DCLS7		
Draws	100	500	1000	100	500	1000
CPU	1032	5047	10192	5124	25683	51108
GPU 1	521	2107	4397	1613	6307	14017
GPU 2	1077	5577	13541	2789	13895	27691
Ratio 1	1.98	2.39	2.32	3.18	4.07	3.65
Ratio 2	0.96	0.90	0.75	1.84	1.85	1.85

Table D.5: Observed total time (in seconds) and CPU/GPU ratios for the algorithm on CPU and GPU on different machines and with different numbers of particles. The CPU is a 32 core Intel Xeon CPU E5-2667 v2 two processors and the GPU1 is a NVIDIA Tesla K40c GPU and the GPU2 is a NVIDIA GeForce GTX 660. "Ratio 1" refers to the CPU/GPU 1 ratio and "ratio 2" refers to the CPU/GPU 2 ratios. Number below 1 indicates the CPU is faster, number above one indicates that the GPU is faster.