Predictive Density Combinations with Dynamic Learning for Large Data Sets in Economics and Finance

Roberto Casarin  
University of Venice

Stefano Grassi  
University of Rome “Tor Vergata”

Francesco Ravazzolo  
Free University of Bozen-Bolzano  
BI Norwegian Business School

Herman K. van Dijk  
Erasmus University Rotterdam  
Tinbergen Institute  
Norges Bank

10th ECB Workshop on Forecasting Techniques: Economic Forecasting with Large Datasets, June 18-19, 2018.  
Disclaimer: The views expressed herein are solely those of the authors and do not necessarily reflect the views of Norges Bank.
Motivation: Average 7424 Density Forecasts of 1856 Stock Returns
Motivation

- **Problem and Practice** Many agencies handle this averaging informally. Our aim: Give this a **Bayesian probabilistic foundation** in order to evaluate practical issues like: Probabilities of (extreme) events: Recession probability; Turning point probability; Probabilistic warnings about defaults; Value-at-Risk etc etc.

- **Fast growth in Big Data** give more accurate measures. Analogy with weather forecasting using many satellite pictures. But multimodality, skewness etc in economics.

- **Parallel Computing: New Hardware and Software** give openings to solve complex problems. Machine learning with several hidden layers using neural networks have a direct connection with filtering methods in nonlinear time series models.
Three Theoretical Contributions

- **Flexible Bayesian Combination Model** is an extension of the Mixture of Experts Model, (Jordan and Jacobs, 2010 and many others) allowing for cross-section and time dependent **Bayesian weight learning** and **diagnostic learning** about model incompleteness.

- **Dimension reduction in three steps**
  - **Sequential clustering** of large set of predictions with learning weights.
  - **Dynamic learning** of weights on a **simplex of reduced dimension**. (Time series behaviour on a bounded domain).
  - **From small simplex to large simplex** using **class-preserving property** of the logistic-normal distribution.

- **Model Representation and Efficient Computation**
  - Model is a **nonlinear State Space Model**
  - **K-means clustering algorithm** and **Nonlinear Sequential Filtering**. Filtering is parallelised and directly connected to hidden layers in machine learning with neural network weights.
Review of the literature

- **Averaging** for forecast accuracy (Barnard (1963), Bates and Granger (1969)).
- **Parameter and model uncertainties** importance (BMA, Roberts (1965)).
- **Probabilistic predictions** provide larger set of information (Wallis).
- **Correlations** between forecasts and weights (Garratt, Mitchell and Vahey (2012)).
- Model performance **varies over time** possibly with persistence (Diebold and Pauly (1987), Guidolini and Timmermann (2009), Hoogerheide et al. (2010), Gneiting and Raftery (2007); Billio et al. (2013); Del Negro, Hasegawa and Schorfheide, 2015).
- Models might perform differently for multiple variables of interest (**specific weight** for each series, univariate models).
- Model set is possible **incomplete** (Geweke (2009), Geweke and Amisano (2010), Waggoner and Zha (2010)).
- **GPU computing improve speed**, see Casarin, et al. (2014), Dziubinski and Grassi(2013), Geweke and Durham (2012) and many others.
Review earlier density combination

Billio, Casarin, Ravazzolo and Van Dijk (2013, JE) propose a distributional state-space representation of the predictive densities and the combination scheme.


Extensions to Nowcasting in Aastveit, Ravazzolo and Van Dijk (2018, JBES); Forecasting Combinations and Portfolio Combinations in Basturk, Borowska, Grassi, Hoogerheide and Van Dijk (2018). In this paper: Large Data

Bayesian Foundational paper McAllinn and West (2018)

Background: Evolution of Density Combinations in Economics, Aastveit, Mitchell, Ravazzolo and Van Dijk (2018)

software on http://www.francescoravazzolo.com/pages/DeCo.html
Mixture representation extends mixture of experts model

Basic Idea for one economic variable $y_t$

- **Basic practice** $\sum_{i=1}^{n} w_{it} \tilde{y}_{it}$.

- **Formally**: conditional predictive probability of $y_t$, given $\tilde{y}_t = (\tilde{y}_{1t}, \ldots, \tilde{y}_{nt})'$, is discrete mixture of conditional predictive probabilities of $y_t$ given $\tilde{y}_{it}$, $i = 1, \ldots, n$ from $n$ individual models with weights, $w_{it}$, $i = 1, \ldots, n$. Fundamental density combination

  $$f(y_t | \tilde{y}_t) = \sum_{i=1}^{n} w_{it} f(y_t | \tilde{y}_{it})$$

- **Discrete/continuous mixture representation**
  Under standard regularity conditions the marginal predictive density of $y_t$ has the following discrete/continuous representation:

  $$f(y_t | I) = \sum_{i=1}^{n} w_{it} \int_{\mathbb{R}} f(y_t | \tilde{y}_{it}) f(\tilde{y}_{it} | I_i) d\tilde{y}_{it}$$
We specify a Gaussian combination density
\[ f(y_t|\tilde{y}_{it}) = \mathcal{N}(y_t|\tilde{y}_{it}, \sigma^2_t) \]
with stochastic volatility, \( \sigma^2_t \) that determines overall uncertainty and indicates model incompleteness. If \( \sigma^2_t \) tends to zero, one obtains the static mixture of experts model from Jacobs and Jordan and Geweke and Keane. Also choice of the model set is important. This leads to diagnostic learning.

Dynamic weights are periodically updated with Bayesian learning. (Here simple random walk).
Number of predictions is large, say $n$, reduce this to a small set, say $m$. Dimension reduction in three steps

- **Step 1:** Sequential clustering of predictions to $m$ clusters with learning following some features of the predictive densities. Grouping can change over time. Weights driven by a model-specific predictive performance measure like log score (or equal weights).

- **Step 2:** Unrestricted weights $v_t$ of clusters follow a $m$-variate normal random walk process or more involved Bayesian learning and are mapped to low dimensional simplex to have probabilistic interpretation of model weights, $z_t$.

- **Step 3:** Map weights $z_t$ from small $m-1$ simplex back to large $n-1$ simplex to $w_t$ using the class-preserving property of the logistic-normal distribution from Aitchinson’s geometry of the simplex so that typical feature of model probability weight is preserved.
Summary of Data Driven Density Combinations for Large Datasets and connection with Machine Learning

<table>
<thead>
<tr>
<th>Densities for Large Datasets (7400 densities)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential clustering step</td>
</tr>
<tr>
<td>Cluster 1</td>
</tr>
<tr>
<td>$z_1 = g(v_1)$</td>
</tr>
<tr>
<td>Cluster 2</td>
</tr>
<tr>
<td>$z_2 = g(v_2)$</td>
</tr>
<tr>
<td>Cluster 3</td>
</tr>
<tr>
<td>$z_3 = g(v_3)$</td>
</tr>
<tr>
<td>Cluster 4</td>
</tr>
<tr>
<td>$z_4 = g(v_4)$</td>
</tr>
<tr>
<td>Updating Clustered Density Combinations</td>
</tr>
<tr>
<td>Original series with weights $w$</td>
</tr>
<tr>
<td>Densities with weights for all series</td>
</tr>
</tbody>
</table>

Results

- **Hidden layer weights integrated using filtering. Also neural networks from machine learning.**
- **Mixtures on large space have same properties as reduced space.**
Nonlinear State Space Model

\[ y_t \sim \sum_{i=1}^{n} w_{it} \mathcal{N} \left( \tilde{y}_{it}, \sigma_t^2 \right) \]

\[ \tilde{w}_t \sim \mathcal{L}_{n-1} \left( \tilde{B}_t D_m v_{t-1}, \tilde{B}_t D_m \Sigma D_m' \tilde{B}_t' \right), \]

\[ \tilde{w}_t = (w_{1,t}, \ldots, w_{n-1,t})' \text{ and } w_{n,t} = 1 - \tilde{w}_t' \iota_{n-1} \text{ and } \tilde{B}_t \text{ contains weights.} \]
Representation result: (II) Under regular conditions the Model is Generalized Linear Model with Nonlinear Local Level Model

Corollary

Let $s_t$ be an allocation vector, with $s_t \sim M_n(1, w_t)$, where $M_n(1, w_t)$ denotes the multinomial distribution, and let $\sigma_t$ be a time-varying variance. Then, the state space model given in the Proposition can be written as

\begin{align*}
    y_t &= \tilde{y}_t' s_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_t) \\
    s_{i,t} &= \begin{cases} 
        1 & \text{with probability } w_{i,t} \\
        0 & \text{otherwise} 
    \end{cases} \\
    w_t &= \phi_B(z_t) \\
    z_t &= z_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{L}_{m-1}(\mathbf{0}, D_m \Sigma D_m')
\end{align*}

where $\phi_B(z_t)$ is a nonlinear logistic transformation
Two Numerical approximations (prior is diffuse but proper)

- (1) Clustering-based mapping of the predictors requires the solution of an optimization problem which is not available in analytic form;
- (2) Analytic solution of the optimal filtering problem is generally not known;
- Apply sequential numerical approximation algorithms which, at time $t$ iterate over the following steps:

1. **Parallel k-means clustering on GPU**, based on features $\psi_{it}$ where the centroids are updated dynamically as follows:

   $$c_{jt+1} = c_{jt} + \lambda_t (m_{jt+1} - c_{jt})$$

   where

   $$m_{jt+1} = \frac{1}{n_{jt+1}} \sum_{i \in N_{jt+1}} \psi_{it}$$

   and $\lambda_t \in [0, 1]$. This implies a sequential clustering with forgetting driven by the processing of the blocks of observations.

2. **Sequential (over time dimension) Monte Carlo approximation** of the nonlinear SSM using GPU (more later)
Empirical 1: Financial Time Series: Data are 1856 individual stock prices, quoted in NYSE and NASDAQ. Predict Features of a Replication of S&P500 index.

- Better accuracy than standard models about many density features: means, volatilities and, in particular, tails.
- Prediction of the statistical and economic accuracy of a tail estimates: event like Value-at-Risk.
Empirical 1: Financial time series: Models and Estimation

- Data: 1856 individual stock daily prices quoted in the NYSE and NASDAQ from Datastream over the sample March, 18, 2002 to December, 31, 2009.

- Models and Estimation: Estimate a Normal GARCH(1,1) model and a $t$-GARCH(1,1) model via posterior mode using rolling samples of 1250 trading days (about five years) for each stock return:

\[
\begin{align*}
    y_{it} &= c_i + \kappa_{it} \varepsilon_{it} \\
    \kappa_{it}^2 &= \theta_{i0} + \theta_{i1} \varepsilon_{i,t-1}^2 + \theta_2 \kappa_{i,t-1}^2
\end{align*}
\]

where $y_{i,t}$ is the log return of stock $i$ at day $t$, $\varepsilon_{it} \sim N(0, 1)$ and $\varepsilon_{it} \sim T(\nu_i)$ for the Normal and Student-$t$ cases, respectively.

- Produce 784 one day ahead density forecasts for the period January 1, 2007 to December 31, 2009.
Empirical 1: Dynamic behaviour of weights of 4 clusters

Normal GARCH(1,1) with **low (cluster 1) vol.** and **high (cluster 2) vol.**; t-GARCH(1,1) with **low (cluster 3)** and **high (cluster 4) d.o.f.**. Three subperiods: Weights adjust to time instability: Lehman Brothers Default
Empirical 1: Time movement of joint dependence of weights using Canonical Correlations

Each cluster versus all others. Change in cluster 2 and 3 after Lehman
### Empirical 1: Forecasting Features of Densities: Means; Log Scores: Tail probabilities and Risk

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

**Table:** Forecasting results for next day S&P500 log returns. avQS-T and avQS-L: average tails (T) and average left tail (L) scores described in Gneiting and Raftery (2007). Violation: percentage of times the realization exceeds the 1% Value at Risk (VaR).
Empirical 1: Cumulative left quantile score for 3 cases

Empirical 1: Model incompleteness measured by $\sigma_t$

Variance increases in DCEW-SV scheme in September 8, 2008. Signal for including possibly models with jumps in volatility.
Empirical 2: Macroeconomic Time Series. Extend Stock and Watson (2002,2005) as follows:

- **Joint prediction model** for the group of variables Real GDP, GDP deflator, Treasury Bill rate, Employment instead of a single variable 'beats' all alternatives considered.

- **Learning about dynamic behaviour of 5-7 sectors.** Factor-model combination shows rise and decline of sectors. Dominant factor?

- **Accurate probability of turning point and of recession measured over time**

- Bayesian estimation of AR(1) process for all series and group them using residual variance and the persistence parameters. Identify $m = 5, 7$ clusters.
- Combination: univariate and multivariate combinations, 5 and 7 clusters, equal and log score weights.
- Produce from 1 to 5-step ahead recursive AR(1) forecasts for the out-of-sample: 1985Q1-2011Q2.
- Evaluation criteria: MSPE, CRPS and Log Score.
Empirical 2: Dynamic behaviour of cluster weights of GDP growth: dominant is 6 (Exports, Imports, GDP deflator)
Empirical 2: Dynamic behaviour of model weights $w$ of GDP growth, 1-step ahead

Cluster 1

Cluster 2

Cluster 3
Empirical 2: 8 Competitive models for Forecasting

- UDC versus MDC (univariate versus multivariate combination)
- EW versus LS (equal weights versus recursive log score weights);
- 5 versus 7 (5 cluster versus 7 clusters);
- 8 models: UDCEW5, MDCEW5, UDCLS5, UDCLS7, MDCLS5, UDCEW7, MDCEW7, UDCLS7, MDCLS7
Empirical 2: GDP Forecasting Results

<table>
<thead>
<tr>
<th></th>
<th>h=1</th>
<th></th>
<th>h=5</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSPE</td>
<td>LS</td>
<td>CRPS</td>
<td>MSPE</td>
</tr>
<tr>
<td>AR</td>
<td>0.647</td>
<td>-1.002</td>
<td>0.492</td>
<td>0.682</td>
</tr>
<tr>
<td>BDFM</td>
<td>0.649</td>
<td>-1.091</td>
<td>0.382**</td>
<td>0.655</td>
</tr>
<tr>
<td>UDCEW5</td>
<td>0.644</td>
<td>-0.869</td>
<td>0.333**</td>
<td>0.658*</td>
</tr>
<tr>
<td>MDCEW5</td>
<td>0.63</td>
<td>-0.928</td>
<td>0.326**</td>
<td>0.636*</td>
</tr>
<tr>
<td>UDCLS5</td>
<td>0.773</td>
<td>-1.306</td>
<td>0.464</td>
<td>0.715</td>
</tr>
<tr>
<td>MDCLS5</td>
<td>0.725</td>
<td>-1.145</td>
<td>0.505</td>
<td>0.557*</td>
</tr>
<tr>
<td>UDCEW7</td>
<td>0.649</td>
<td>-0.875</td>
<td>0.334**</td>
<td>0.657*</td>
</tr>
<tr>
<td>MDCEW7</td>
<td>0.642</td>
<td>-0.979</td>
<td>0.334**</td>
<td>0.654*</td>
</tr>
<tr>
<td>UDCLS7</td>
<td>0.646</td>
<td>-0.868*</td>
<td>0.332**</td>
<td>0.657*</td>
</tr>
<tr>
<td>MDCLS7</td>
<td><strong>0.596</strong></td>
<td><strong>-0.586</strong></td>
<td><strong>0.275</strong></td>
<td><strong>0.610</strong></td>
</tr>
</tbody>
</table>
Empirical 2: Probabilities of negative quarterly growth

One quarter ahead, three quarters ahead and five quarters ahead probabilities over time of negative quarterly growth given by the combination approach and ex-post NBER.
Conclusions

- **Time-varying combinations** of large sets of predictive densities based on sequential clustering analysis can deal with **big data**.

- Combination weights are driven by **cluster-specific latent processes much smaller** than the number of available predictors.

- **The proposed model is a nonlinear SSM with finite mixtures and dynamic logistic-normal weights.** Density combination evaluated by nonlinear sequential filtering. Interesting connections with machine learning with neural nets.

- Forecasting financial time series: Three improvements: **Forecast accuracy of moments; dynamic learning about clusters; efficient estimates of risk**

- Substantial gains in point and density forecasting of **joint** set of US real GDP, GDP deflator, Treasury Bill returns and employment growth based on log score learning. **Rise and decline of sectors; probabilities of recession**.

- Parallelisation: Handling big data sets with GPU can be easy and fast.
Topics for Further Research

- **Modelling and estimation.** Use more information on clustering and explore diagnostic analysis about model incompleteness and richer model set.


- **More efficient parallel computing.**