# 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"

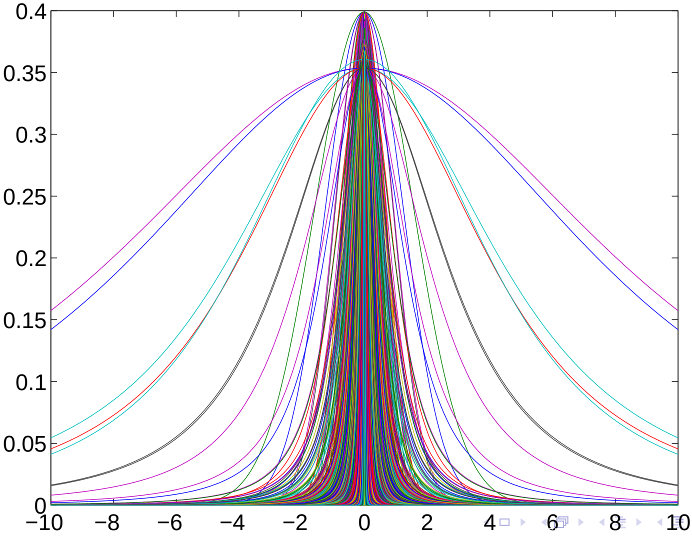
Francesco Ravazzolo
Free University of Bozen-Bolzano
BI Norwegian Business School

**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 extension of 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 performances differs over regions of interest/quantiles (mixture of predictives; generalized LOP: Fawcett, Kapetanios, Mitchell and Price, 2014; subsets of interest: Pelenis, 2014).
- Model performance varies over time possibly with persistence (Diebold and Pauly (1987), Guidolin 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
- Casarin, Grassi, Ravazzolo and Van Dijk, (2015, JSS) created **a MATLAB toolbox DECO** for estimating the density combination scheme of BCRVD (2013).
- 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 $\tilde{\mathbf{y}}_t = (\tilde{y}_{1t}, \ldots, \tilde{y}_{nt})'$, 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**

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

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

$$f(y_t|I) = \sum_{i=1}^{n} w_{it} \int_{\mathbb{R}} f(y_t|\tilde{y}_{it}) f(\tilde{y}_{it}|I_i) d\tilde{y}_{it}$$

# Mixture representation with learning about model incompleteness and dynamic weight behaviour

- We specify a Gaussian combination density $f(y_t|\tilde{y}_{it}) = \mathcal{N}\left(y_t|\tilde{y}_{it}, \sigma_t^2\right)$ with stochastic volatility, $\sigma_t^2$ **that determines overall uncertainty and indicates model incompleteness.** If $\sigma_t^2$ 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).
- **Step2: 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

| | Densities for Large Datasets (7400 densities) | | | |
|---|---|---|---|---|
| | ↘ | ↓ | ↓ | ↙ |
| | Sequential clustering step | | | |
| | ↓ | ↓ | ↓ | ↓ |
| Hidden layers with unrestricted cluster weights $v$ and scaled probabilistic weights $z$ | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
| | ↓ | ↓ | ↓ | ↓ |
| | $z_1 = g(v_1)$ | $z_2 = g(v_2)$ | $z_3 = g(v_3)$ | $z_4 = g(v_4)$ |
| | ↘ | ↓ | ↓ | ↙ |
| | Updating Clustered Density Combinations | | | |
| | | ↓ | ↓ | |
| | Original series with weights $w$ | | | |
| | ↙ | ↓ | ↓ | ↘ |
| | Densities with weights for all series | | | |

**Results**

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

# Representation result: (I) Nonlinear SSM as finite mixture with large sets of members and transition function as logistic-normal weight

Model: **marginal** predictive density of $y_t$ with normal combination density, time-varying volatility for model incompleteness; dynamic learning of clustered weights

$$f(y_t|I) = \sum_{i=1}^{n} w_{it} \int_{\mathbb{R}} \mathcal{N}\left(y_t|\tilde{y}_{it}, \sigma_t^2\right) f(\tilde{y}_{it}|I_i) d\tilde{y}_{it}$$

## Nonlinear State Space Model

$$\mathbf{y}_t \sim \sum_{i=1}^{n} w_{it} \mathcal{N}\left(\tilde{y}_{it}, \sigma_t^2\right) \quad (1)$$

$$\tilde{\mathbf{w}}_t \sim \mathcal{L}_{n-1}\left(\tilde{B}_t D_m \mathbf{v}_{t-1}, \tilde{B}_t D_m \Sigma D_m' \tilde{B}_t'\right), \quad (2)$$

$\tilde{\mathbf{w}}_t = (w_{1,t}, \ldots, w_{n-1,t})'$ and $w_{n,t} = 1 - \tilde{\mathbf{w}}_t' \boldsymbol{\iota}_{n-1}$ and $\tilde{B}_t$ contains weights.

# Representation result: (II) Under regular conditions the Model is Generalized Linear Model with Nonlinear Local Level Model

## Corollary

*Let* $\mathbf{s}_t$ *be an allocation vector, with* $\mathbf{s}_t \sim \mathcal{M}_n(1, \mathbf{w}_t)$*, where* $\mathcal{M}_n(1, \mathbf{w}_t)$ *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*

$$\mathbf{y}_t = \tilde{\mathbf{y}}_t' \mathbf{s}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0}, \sigma_t) \tag{3}$$

$$\mathbf{s}_{i,t} = \begin{cases} 1 & \textit{with probability } w_{i,t} \\ 0 & \textit{otherwise} \end{cases} \tag{4}$$

$$\mathbf{w}_t = \boldsymbol{\phi}_{\boldsymbol{B}}(\mathbf{z}_t) \tag{5}$$

$$\mathbf{z}_t = \mathbf{z}_{t-1} + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \mathcal{L}_{m-1}\left(\mathbf{0}, D_m \Sigma D_m'\right) \tag{6}$$

*where* $\boldsymbol{\phi}_{\boldsymbol{B}}(\mathbf{z}_t)$ *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 $\boldsymbol{\psi}_{it}$ where the centroids are updated dynamically as follows:

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

where

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

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)

# Two Groups of Empirical Contributions

- **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.**
  - **Learning about** time behaviour of clusters of stocks. **Joint** dependence over time among weights. Signal of model incompleteness from diagnostic learning.
  - Prediction of the **statistical and economic accuracy of a tail estimates: event like Value-at-Risk**.

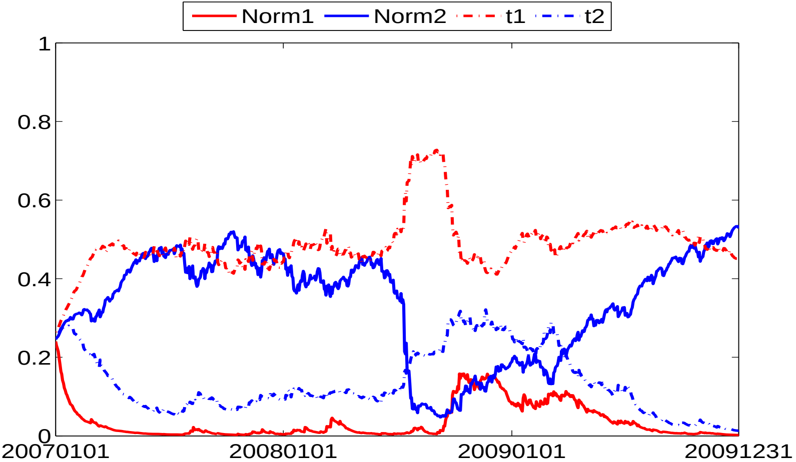
# Empirical 1: Financial time series: Models and Estimation

- 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:

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

  where $y_{i,t}$ is the log return of stock $i$ at day $t$, $\varepsilon_{it} \sim \mathcal{N}(0,1)$ and $\varepsilon_{it} \sim \mathcal{T}(\nu_i)$ 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

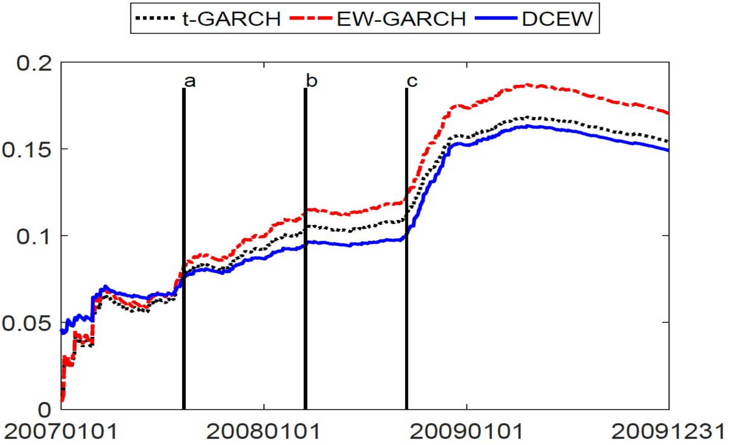

Each cluster versus all others. Change in cluster 2 and 3 after Lehman

# Empirical 1: Forecasting Features of Densities: Means; Log Scores: Tail probabilities and Risk

| | 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% |
| *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%** |

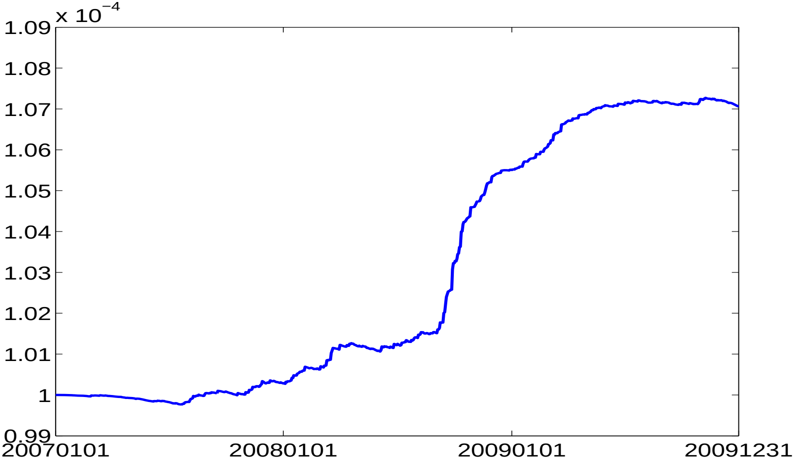
Table: Forecasting results for next day S&P500 log returns.
avQS-T and avQS-L: average tails (T) and average left tail (L) scores described in Gneiting and Raftery (2007).
Violation: percentage of times the realization exceeds the 1% Value at Risk (VaR).

# Empirical 1: Cumulative left quantile score for 3 cases




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. Stock market stress increases gap between models. Accuracy drops after Lehman default

# 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

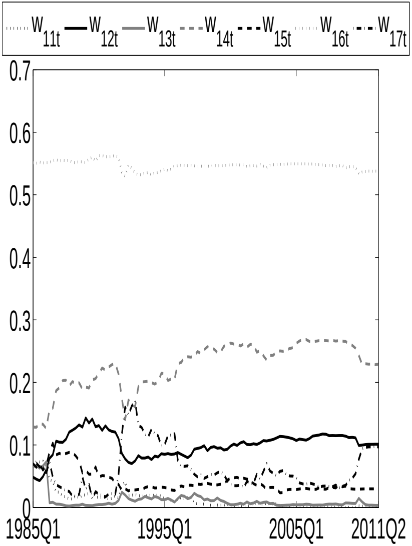
# Empirical 2: Stock and Watson (2005) dataset. Models and Estimation GDP, GDP deflator, 3-month Interest rates, Employment

- **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**

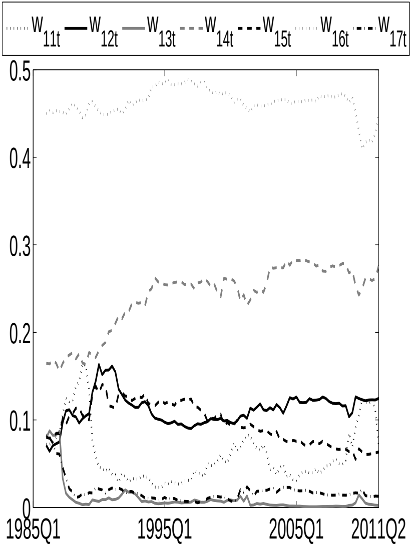
# Empirical 2: Stock and Watson (2005) dataset. Models and Estimation GDP, GDP deflator, 3-month Interest rates, Employment.

- Stock and Watson (2005) dataset: 142 series, standardized and sampled at a quarterly frequency from 1959Q1 to 2011Q2.
- 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)

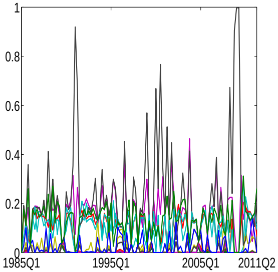



1-step ahed

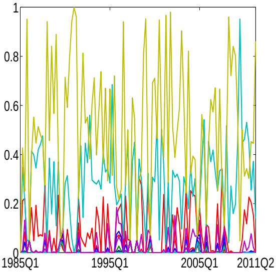



5-step ahead

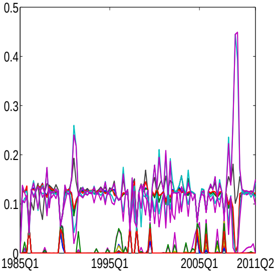
# Empirical 2: Dynamic behaviour of model weights $w$ of GDP growth, 1-step ahead




cluster 1




cluster 2




cluster 3

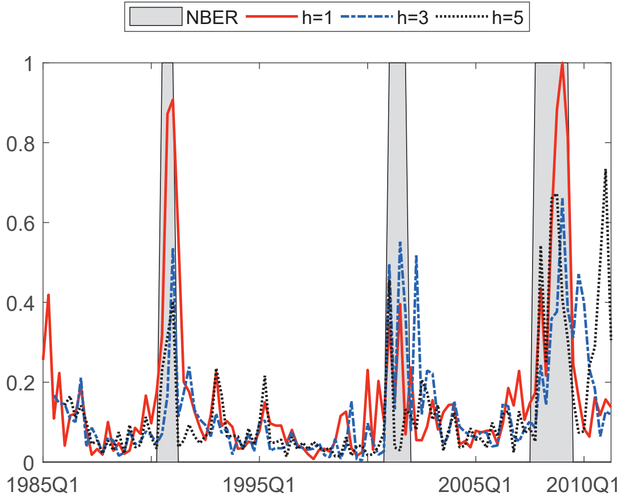
# Empirical 2: 8 Competitive models for Forecasting

- 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

| | h=1 | | | h=5 | | |
|---|---|---|---|---|---|---|
| | MSPE | LS | CRPS | MSPE | LS | CRPS |
| AR | 0.647 | -1.002 | 0.492 | 0.682 | -1.009 | 0.506 |
| BDFM | 0.649 | -1.091 | 0.382** | 0.655 | -1.099 | 0.388** |
| UDCEW5 | 0.644 | -0.869 | 0.333** | 0.658* | -0.912 | 0.343** |
| MDCEW5 | 0.63 | -0.928 | 0.326** | 0.636* | -0.844 | 0.324** |
| UDCLS5 | 0.773 | -1.306 | 0.464 | 0.715 | -1.38 | 0.481 |
| MDCLS5 | 0.725 | -1.145 | 0.505 | 0.557* | -1.005 | 0.358** |
| UDCEW7 | 0.649 | -0.875 | 0.334** | 0.657* | -0.891 | 0.338** |
| MDCEW7 | 0.642 | -0.979 | 0.334** | 0.654* | -1.009 | 0.342** |
| UDCLS7 | 0.646 | -0.868* | 0.332** | 0.657* | -0.914 | 0.342** |
| MDCLS7 | **0.596*** | **-0.586**** | **0.275**** | **0.610**** | **-0.634**** | **0.286**** |

# Empirical 2: Probabilities of negative quarterly growth




One quarter ahead, three quarters ahead and five quarter ahead probabilities over time of negative quarterly growth given by the 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 on **efficiency of filter** methods: New Filter obtained using the mixture of student density approximation constructed using an EM weighted importance sampling approach, HOVD(2012, JE), labeled M-Filter. Forthcoming Journal of Econometrics.
- **Forecasting and Policy**. More on Nowcasting and Multiperiod out-of-sample forecasting. Applications to **Policy Issues**: In the field of Finance using Decision Models. Current research on Bayesian Dynamic Modelling and Time-varying combinations of Equity Momentum Strategies using US industrial portfolios 1929-2015. **Time-varying combinations jointly for models and policies: mixture of mixtures**. Challenge to do this for macro-models.
- **More efficient parallel computing.**