Answering the Queen: Machine Learning and Financial Crises

Jérémy Fouliard, Michael Howell, Hélène Rey

European Central Bank

October 7, 2019

Forecasting the financial crisis of 2008

Visiting the LSE and being shown how terrible the situation was and had been, the Queen asked: "Why did nobody notice it?"

Motivation

- The Queen got a terse answer. The economic and finance profession have a bad record at predicting crises.

Motivation

- The Queen got a terse answer. The economic and finance profession have a bad record at predicting crises.
- After several months, the Economic Section of the British Academy wrote a three-page missive to Her Majesty blaming the lack of foresight of the crisis on the "failure of the collective imagination of many bright people".

Motivation

- The Queen got a terse answer. The economic and finance profession have a bad record at predicting crises.
- After several months, the Economic Section of the British Academy wrote a three-page missive to Her Majesty blaming the lack of foresight of the crisis on the "failure of the collective imagination of many bright people".
- This paper aims at predicting systemic crises well in advance (12 quarters ahead) using cutting-edge machine learning tools.

Short Literature Review

- **Early Warning Indicators of financial crises** : Kaminsky [1998], Kaminsky and Reinhart [1999], Borio and Lowe [2002], Gourinchas and Obstfeld (2012), Drehmann and Juselius [2013]; Demirgüç-Kunt and Detragiache [1998], Eichengreen and Rose [1997], Bongini et al. [2000], Frankel and Saravelos [2012], Schularick and Taylor [2012], Coudert and Idier [2017], Mian and Sufi [2018], Krishnamurthy and Muir [2018].

- **Machine Learning** : Davis and Karim [2008], Duttagupta and Cashin [2012], Ward [2014], Joy et al. [2017], Alessi and Detken [2018].

Short Literature Review

- **Theory of Financial Crises**: Booms go bust (credit growth) [Kindleberger (1978), Shularick and Taylor (2012), Reinhart and Rogoff (2008), Coimbra and Rey (2018); Household debt, Mian and Sufi (2018); behavioural explanations, Bordalo Gennaioli Shleifer (2019) (bubbles in asset prices); excessive risk taking (leverage, moral hazard) Allen and Gale, Rajan; search for yield (Stein); real shock amplified by a capital constraint (real economy; macro finance Gertler Kiyotaki, Brunnnermeier Sannikov, Krishnamurthy He); balance of payment (real exchange rate, capital flows); concentrated exposures of banking system (real estate, oil), ...

Many different variables, non linear interactions, time varying effects.

New Framework: Online learning: NOT BIG DATA but MODEL AGGREGATION

This framework is very suitable for crisis prediction in real time:

- **Multivariate** : Which variables cause a financial crisis?
- **Time-varying weights** : Causes of financial crises may be different over time.
- **Statistically robust** : overfitting is a problem in the literature.
- **Not "black-box"** : assess the role each model plays to predict the pre-crisis.
- **Theoretically grounded** : asymptotic properties of our aggregation rules ensure convergence.

- **More general than Bayesian Model Averaging**

New Framework: Online learning: NOT BIG DATA but MODEL AGGREGATION

This framework is very suitable for crisis prediction in real time:

- **Multivariate** : Which variables cause a financial crisis?
- **Time-varying weights** : Causes of financial crises may be different over time.
- **Statistically robust** : overfitting is a problem in the literature.
- **Not "black-box"** : assess the role each model plays to predict the pre-crisis.
- **Theoretically grounded** : asymptotic properties of our aggregation rules ensure convergence.

- **More general than Bayesian Model Averaging**
- This framework has been used to predict French electricity load (EDF); the tracking of climate models; the network traffic demand.

Sequential prediction with expert advice

Sequential predictions

Online learning is performed in a sequence of consecutive rounds where at time instance t the forecaster:

1. Receives a question.
2. Uses expert advice $\{f_{j,t} \in \mathcal{D} : j \in \mathcal{E}\}$
3. Predicts $\hat{y}_t \in \mathcal{Y}$
4. Receives true answer $y_t \in \mathcal{Y}$
5. Suffers a loss $\ell(\hat{y}_t, y_t)$.

Sequential prediction with expert advice

To combine experts' advice, the forecaster chooses a sequential aggregation rule $\mathcal{S}$ which consists in setting a time-varying weight vector $(p_{1,t}, ..., p_{N,t}) \in \mathcal{P}$:

$$\hat{y}_t = \sum_{j=0}^{N} p_{j,t} f_{j,t}$$

The forecaster and each expert incur a cumulative loss defined by :

$$L_T(\mathcal{S}) = \sum_{t=1}^{T} \ell(\sum_{j=0}^{N} p_{j,t} f_{j,t}) = \sum_{t=1}^{T} (\hat{y}_t - y_t)^2$$

Sequential prediction with expert advice

- How can we measure the performance of a sequential aggregation rule ?

Sequential prediction with expert advice

- How can we measure the performance of a sequential aggregation rule ?
- We do not have any ideas about the generating process of the observations.
- Forecaster's performance is relative. We define the regret :

$$R_{j,T} = \sum_{t=1}^{T} (\ell(\hat{y}_t, y_t) - \ell(f_{j,t}, y_t)) = \hat{L_T} - L_{j,T}$$

where $\hat{L_T} = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t)$ denotes the forecaster's cumulative loss and $L_{j,T} = \sum_{t=1}^{T} \ell(f_{j,t}, y_t)$ is the cumulative loss of expert j.

Sequential prediction with expert advice

We **minimize the regret** with respect to the best combination of experts:

$$R(\mathcal{S}) = \hat{L}_T(\mathcal{S}) - \inf_{q \in \mathcal{P}} L_T(q)$$

We only select aggregation rules with a "vanishing per-round regret" (regret goes to zero asymptotically).

The Regret can be bounded (bound depends on T, on the learning rate and on log(number of experts)).

Sequential prediction with expert advice

This approach is a **meta-statistic approach**: the aim is to find the best sequential combination of experts (who can be any economic models or judgement).

$$\hat{L}_T(\mathcal{S}) = \inf_{q \in \mathcal{P}} L_T(q) + R(\mathcal{S})$$

Forecaster's cumulative loss is the sum of :

- **An approximation error** : given by the cumulative loss of the best combination of experts.
- **An estimation error** : given by the regret. It measures the difficulty to approach the best combination of experts.

  The rule of the game is to minimize the regret and to find the best experts.

Sequential prediction with expert advice and delayed feedback

- We adapt the approach to incorporate delayed feedback.
- We take into account the fact that we only know whether we are in a pre-crisis period after 12 quarters.
- **Experts** at t are estimated on the batch sample using information available at **t-1**.
- **Aggregation rules** at t use only **t-12** information.

Aggregation rules

We used four aggregation rules :

1. The Exponentially weighted average aggregation rule (EWA) [Littlestone and Warmuth, 1994 ; Vovk,1990].
2. The Multiple learning rates aggregation rule (ML) [Gaillard, Erven and Stoltz, 2014]
3. The Online Gradient Descent aggregation rule (OGD) [Zinkevich, 2003].
4. The Ridge Regression aggregation rule (Ridge) [Azoury and Warmuth, 2001, Vovk, 2001].

Exponentially weighted average aggregation rule (EWA)

Convex aggregation rules combining experts' predictions with a time-varying vector $p_t = (p_{1,t}, ..., p_{N,t})$ in a simplex $\mathcal{P}$ of $\mathbb{R}^N$:

$$\forall j \in \{1, ..., N\}, p_{j,t} \geq 0 \text{ et } \sum_{k=1}^{N} p_{k,t} = 1$$

- We use the gradient-based version of the EWA aggregation rule.
- The weights are computable in a simple incremental way.
- Easy to interpret.

Gradient-based version of the EWA

- Weights are defined by :

$$p_{j,t} = \frac{exp(-\eta_t \sum_{s=1}^{t-1} \tilde{L_{j,s}})}{\sum_{k=1}^{N} exp(-\eta_t \sum_{s=1}^{t-1} \tilde{L_{k,s}})}$$

  where $\tilde{L_{j,s}} = \nabla \ell(\sum_{k=1}^{N} p_{k,s} f_{k,s}, y_s) \cdot f_{j,s}$ and where η_t is the learning rate.

- If j's advice $f_{j,s}$ points in the direction of the largest increase of the loss function (large inner products $\nabla \ell(\sum_{k=1}^{N} p_{k,s} f_{k,s}, y_s) \cdot f_{j,s}$ in the past) the weight assigned to expert j will be small.

The EWA Aggregation rule

Theorem

Theorem 1 [Cesa-Bianchi and Lugosi, 2003]

We assume that

- Functions $L(\cdot, y)$ are differentiable.

Theorem 1 [Cesa-Bianchi and Lugosi, 2003]

We assume that

- Functions $L(\cdot, y)$ are differentiable.
- The losses $L_{j,t}$ are bounded.

Therefore, for all learning rate $\eta_t > 0$,

$$\sup\{R_T(\mathcal{E}_\eta^{grad})\} \leq \frac{ln(N)}{\eta_t} + \eta_t \frac{T}{2}$$

EWA

The bound depends on three parameters :

- The number of experts N.
- The number of time instances T.
- The learning rate η_t.

EWA

For the gradient-based EWA aggregation rule, the forecaster chooses the parameter η_t with the best past performance :

$$\eta_t \in \arg\min_{\eta>0} \hat{L}_{t-1}(\mathcal{E}_\eta)$$

To find the value of η_t which minimizes the cumulative loss, at each time instance, we minimize the cumulative loss on a grid.

Data of systemic crisis episodes: off-the-shelf

- The ECB provides an official database of systemic crisis episodes [Lo Duca et al., 2017q] and additional smaller crises. Judgement of national authorities is involved.

- Our sample starts in 1985q1 (depending on data availability) and ends in 2018q1.
- Our sample includes 7 countries: France, Germany, Italy, Spain, Sweden, UK, US.
- Today, we focus on France, the UK, Germany and Spain. We predict systemic pre-crises.

Data of crisis episodes: off-the shelf

The ECB data uses a characteristic function $C_{n,t}$:

$$C_{n,t} = \begin{cases} 1 & \text{If there is a systemic crisis in country } n \text{ at time } t \\ 0 & \text{Otherwise} \end{cases}$$

Let's define the pre-crisis indicator $I_{n,t}$:

$$I_{n,t} = \begin{cases} 1 & \text{if } \exists h \in H = [0, 12] \text{ such that } C_{n,t+h} = 1 \\ 0 & \text{otherwise} \end{cases}$$

Variables

Our database contains commonly used Early Warning Indicators with transformations (1-y, 2-y, 3-y change and gap-to-trend). It could be enriched. They are not vintage data.

- **Real economy indicators** : Current account, Consumer Price Index, GDP, productivity, Unemployment rate, real echange rate.
- **Credit indicators** : Bank (or Total) credit to financial and non financial sector, M3, Household debt, Debt Service Ratios, bank equity.
- **Interest rates indicators** : 3-month rate, 10 years rate, slope of the yield curve (10y-3m), gap of 10-y rate to real GDP.
- **Housing indicators** : Residential real estate prices, Price-to-income ratio, Price-to-rent ratio.
- **Market indicators: Risk taking** : Real effective exchange rate, Stock prices, Financial Conditions Index, Risk Appetite Index, Cross-border flows, Total Liquidity Index.

Oecumenical Choice of Experts

Experts using country specific and all country panel information

We take as given the models used by some European Central Banks to predict pre-crisis periods and summarized by the Macro-prudential Research Network :

- These models are Dynamic Probit Models, logit models at the country or at the whole panel level, bayesian model averaging models (country specific or on the whole sample)

We add different classes of experts:

- Logit with elastic-net penalty with 5 different sets of variables
- Classifiation tree models at the country and at the panel level
- Random forest models at the country and at the panel level

We end up with 22 experts combining country specific and panel information.

Experts 1

- **Expert P1**: Dynamic Probit Model, Panel 1.
  Variables selected with a country-specific AUROC in- sample
- **Expert P2**: Panel logit fixed effect, Panel 2.
  Variables selected with a country-specific PCA analysis in- sample
- **Expert P3**: Panel logit fixed effect, Panel 3.
  Many credit variables and transforms of credit variables.
- **Expert P4**: Random Coefficient Logit, Panel 4. Share Price 1y change; Real GDP 1y change; Banking credit to private sector 1y change; Rent Price Index 1y change; Total Liquidity Index 1y change; Financial Condition Index 1y change.
- **Expert B1**: Bayesian Model Averaging, BMA 1 (panel)
  We pre-select indicators with a Panel Auroc analysis on the batch sample
- **Expert B2**: Bayesian Model Averaging, BMA 2 (country-specific)
  Country-specific Auroc analysis on the batch sample.

Experts 2

- **Expert T1**: Binary Classification Tree, (panel)
- **Expert T2**: Binary Classification Tree, (country-specific)
- **Expert F1**: Random Forests, (panel)
- **Expert F2**: Random Forests, (country-specific)
- **Expert B**: Logit regression, Bashful
  Best variables selected with a country-specific AUROC Analysis
- **Expert G**: Logit regression, Grumpy
  All variables available in 1987q3. We also introduce some logit with elastic-net penalty which are country-specific.
- **Expert Lh**: Logit with elastic-net penalty, Logit housing.
  Price-to-rent; Price to income; Price to rent 1y change; Price to income 1y change; Real estate price 1y change; Real estate price 2y change.

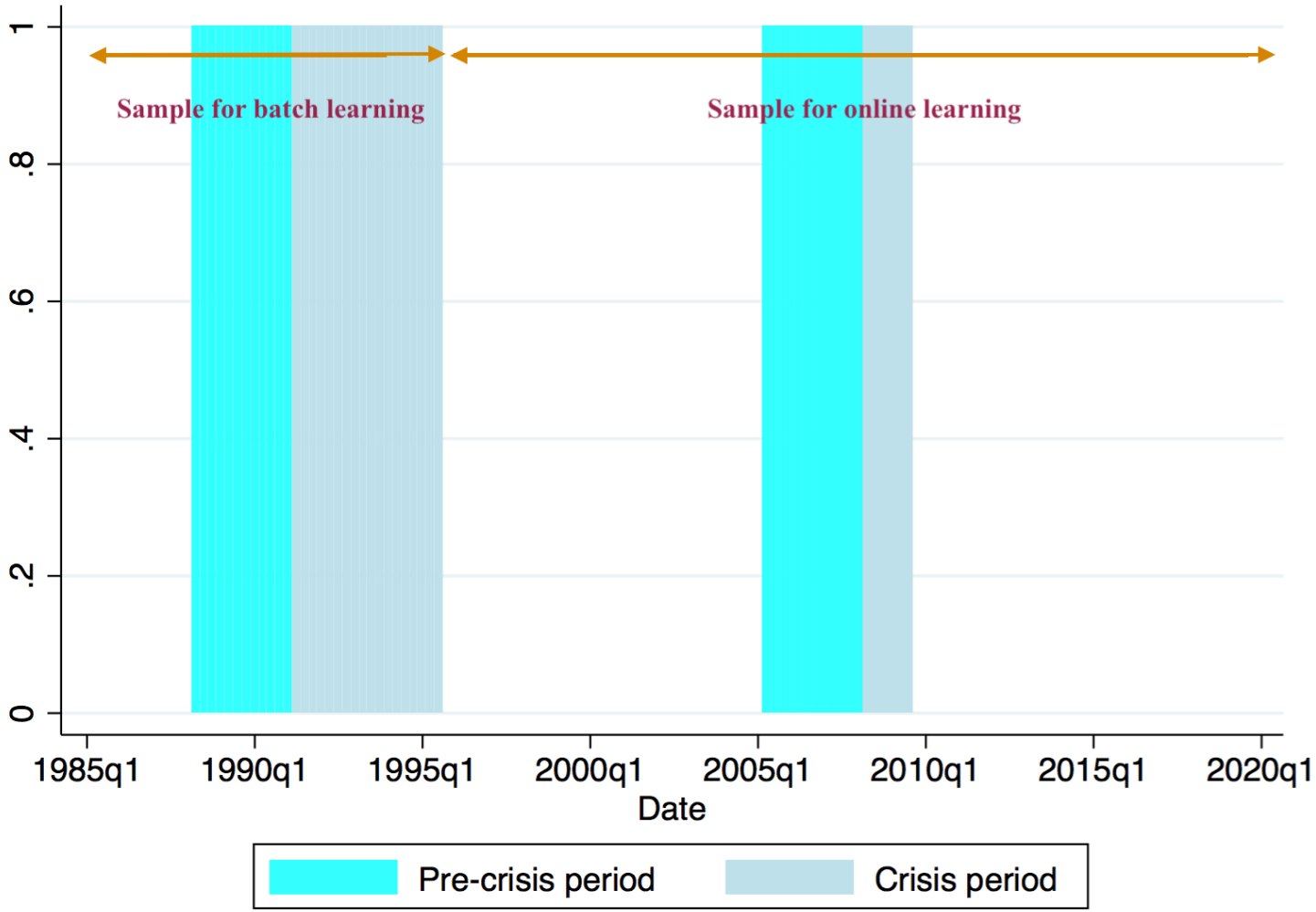
Experts 3

- **Expert Lr**: Logit with elastic-net penalty, Logit real economy. GDP (1y 2 y change); GDP (Per capita per person per hour); Multifactor productivity; Oil price 1y 2 y change.
- **Expert Lcr**: Logit with elastic-net penalty, Logit credit Every credit variable with every transformation (1y, 2y, 3y change and gap to trend)
- **Expert Lba**: Logit with elastic-net penalty, Logit risk taking Risk Appetite; Financial Condition Index; Share price Index; Equity holding; Liquid Assets with two transformations: 1y change and 2y change.
- **Expert Lm**: Logit with elastic-net penalty, Logit monetary. M3; Short term interest rate (nominal); Short term interest rate (real); Consumer prices.
- **Expert Lbu**: Logit with elastic-net penalty, Logit bubble.

Experts 4

- **Expert Lc**: Logit with elastic-net penalty, Logit CrossBorder Capital (Risk appetite; financial condition index; total liquidity index (including capital flows); share price index.).
- **Expert Lc1**: Logit with elastic-net penalty, Logit Combination 1 Housing + Real Economy
- **Expert Lc2**: Logit with elastic-net penalty, Logit Combination 2 Credit + Risk taking
- **Expert Lc3**: Logit with elastic-net penalty, Logit Combination 3 Monetary + risk taking

France


Sample for batch learning
Sample for online learning
0
.2
.4
.6
.8
1
1985q1
1990q1
1995q1
2000q1
2005q1
2010q1
2015q1
2020q1
Date
Pre-crisis period
Crisis period

Forecasting the pre-crisis period out-of-sample

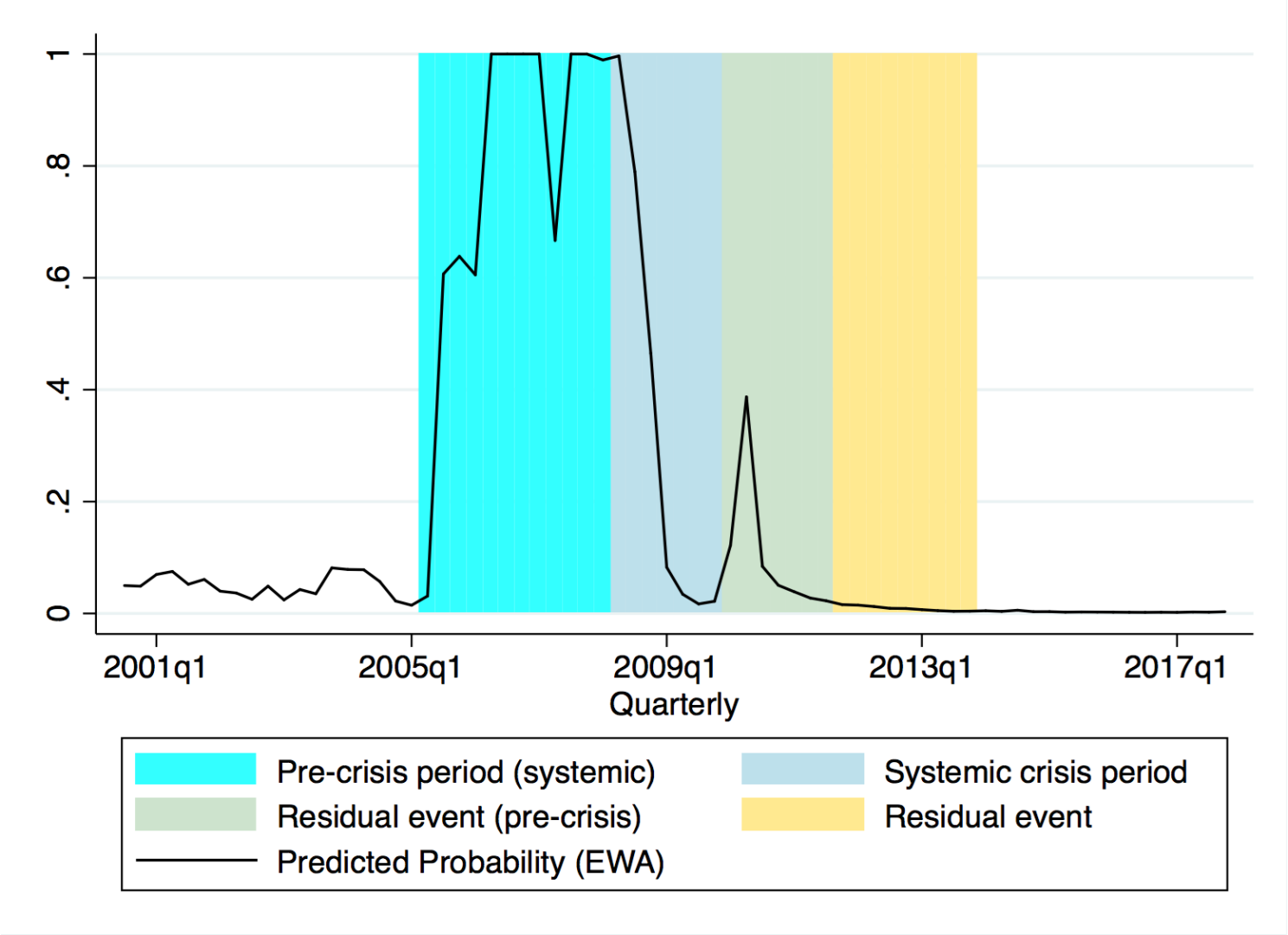


Figure 1: Predicted probability - EWA

Forecasting the pre-crisis period out-of-sample

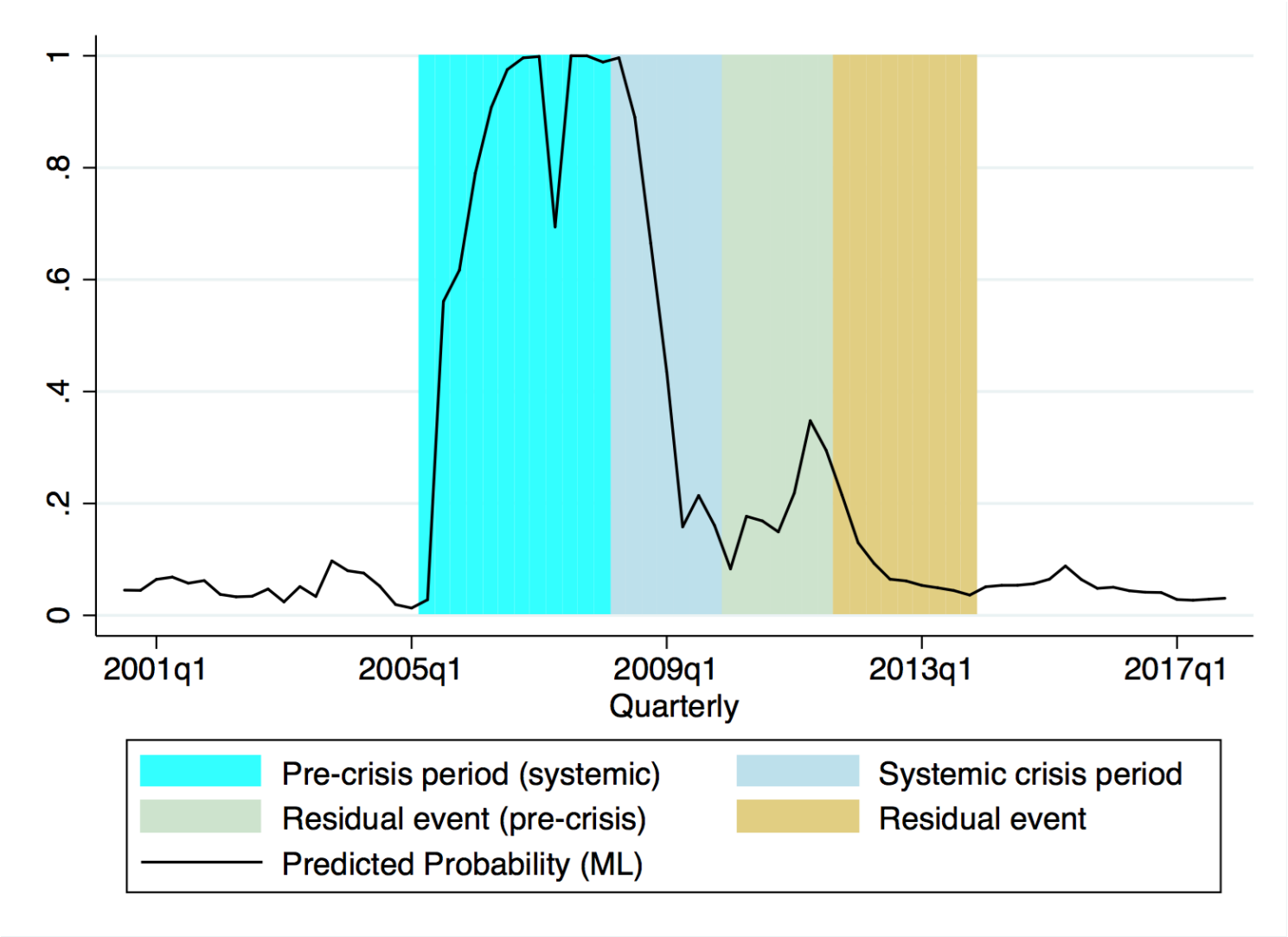


Figure 2: Predicted probability - ML

Forecasting the pre-crisis period out-of-sample

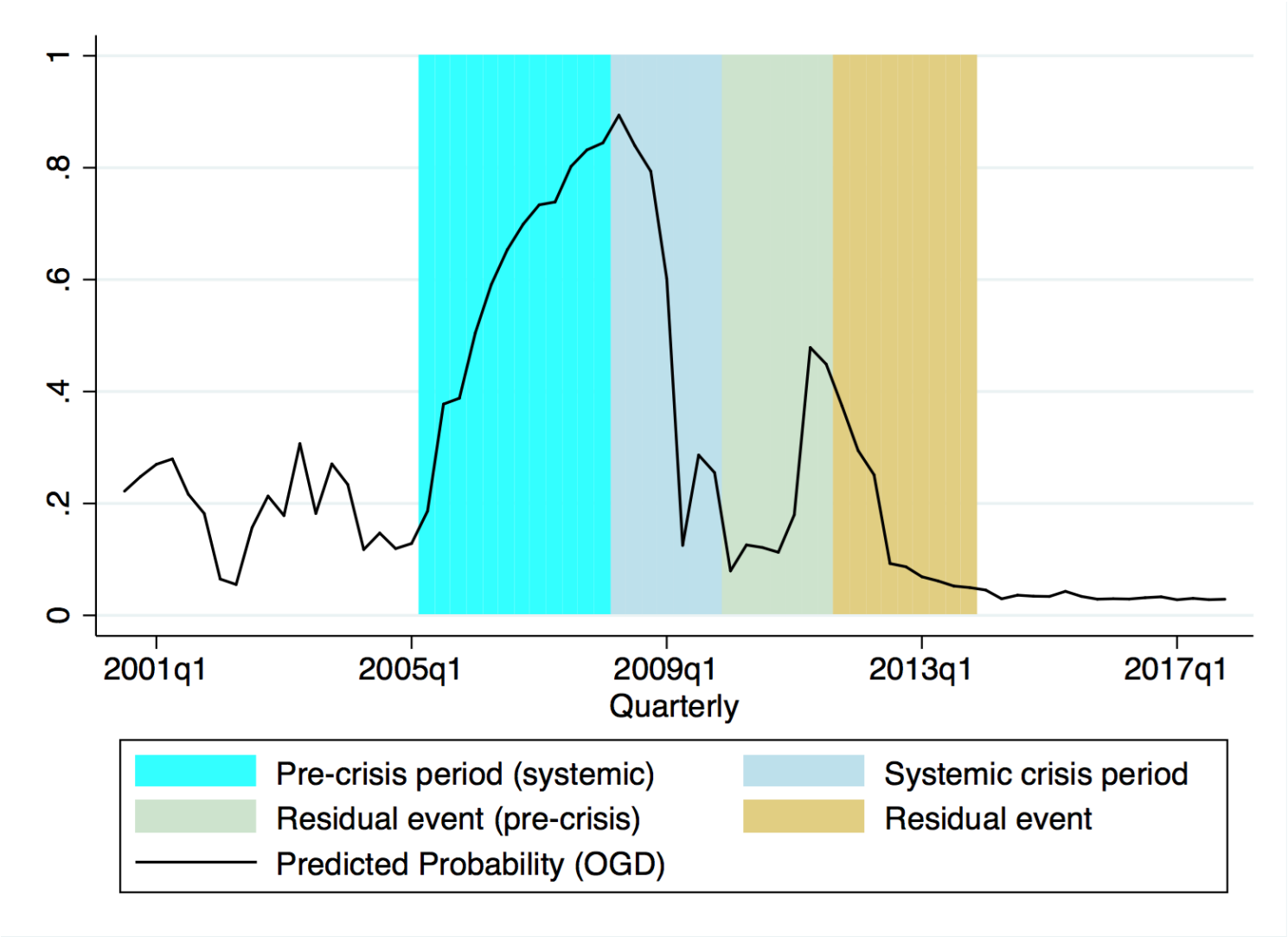


Figure 3: Predicted probability - OGD

Forecasting the pre-crisis period out-of-sample

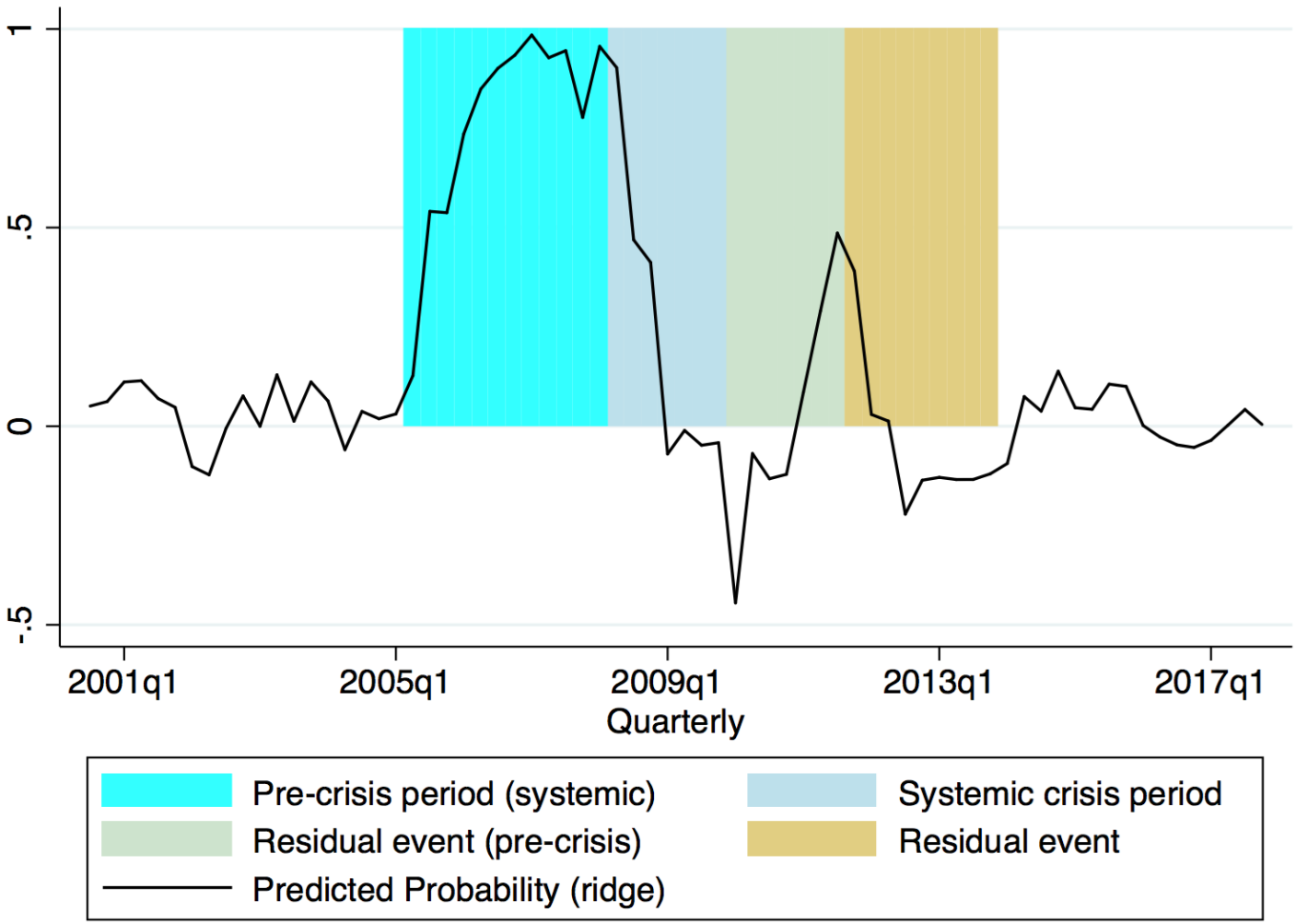


Figure 4: Predicted probability - Ridge

France

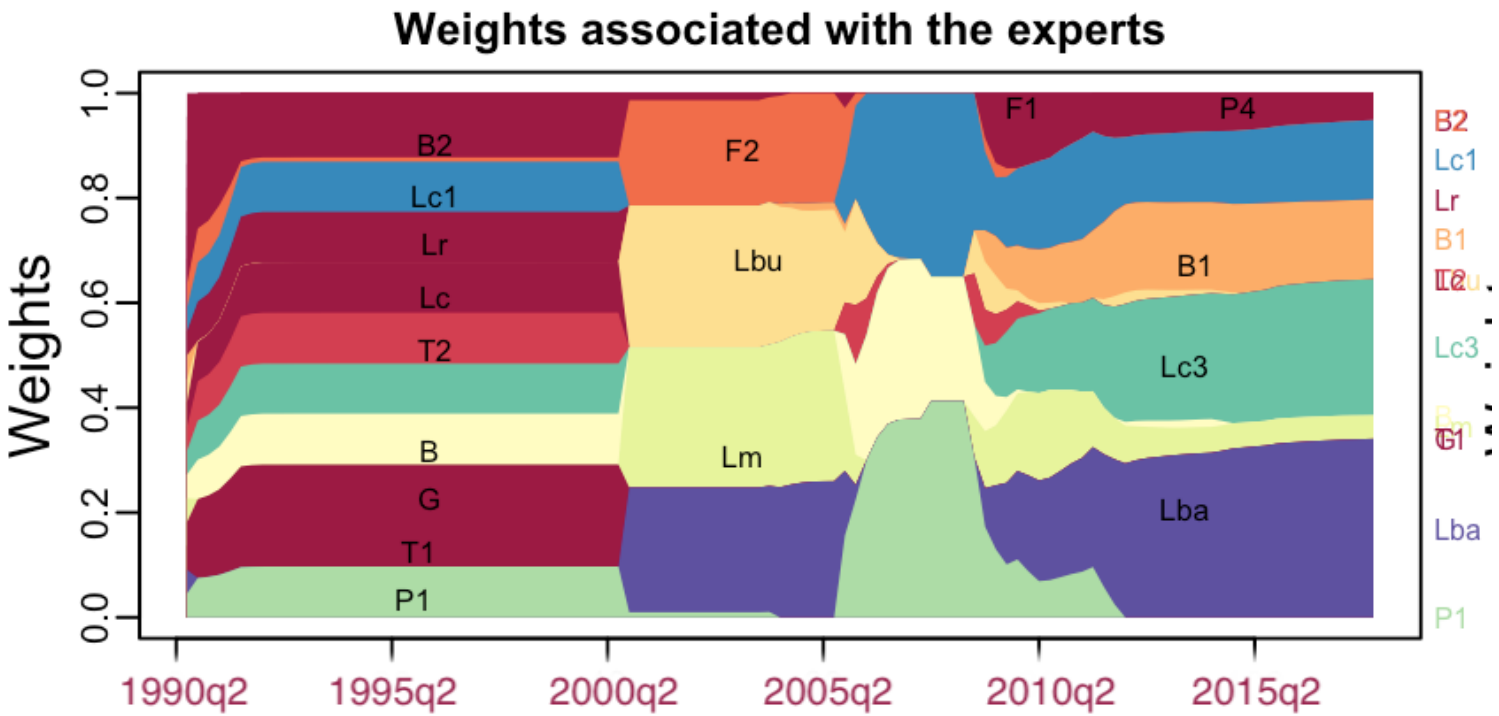


Figure 5: Weights - EWA

France

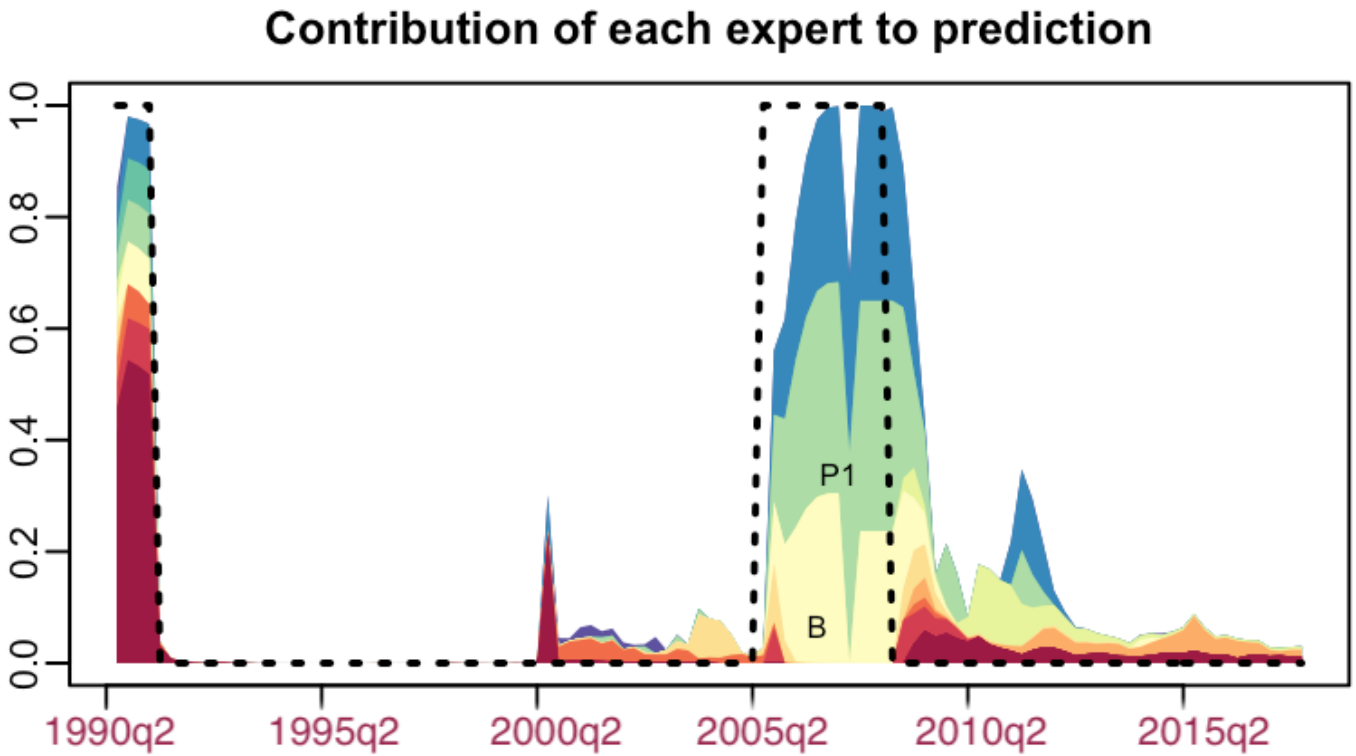


Figure 6: Weights - EWA

France: Sequence of Experts

- Period [1987Q3-2000Q1]: estimation in-sample. Winners are models with many variables. Probably overfitting
- Period [2000Q2-2003Q2]: out-of-sample with frozen weights (delayed feedback).
- Period [2003Q3-2017Q4]: out-of-sample with sequential weights adjustment.

France: Out-of-sample: Endogenous reweighting

- Period [2000Q2-2005Q1]: Quiet period. Models dominating are the random forest **F2** and selastic net models models with monetary **Lm**, bubble variables **Lbu**, risk taking **Lba**.
- Period [2005q2-2008q2]: pre systemic crisis. **B** (Price-to-rent ratio; Price-to-income ratio; Unemployment ratio; Total Credit to non-financial corporations, % GDP; gap to trend; 1y change; Total Credit to private non-financial sector 1y change; Banking credit to private non-financial sector 1y change; GDP1y change; M3 1y change; Financial Conditions Index 2y change; 3y change; Total Liquidity Index 3y change; Bank equity.); **P1** (Total Credit to private non-financial sector 1y change, Total credit to non-financial firms 2y change, Real GDP 2y change, Rent Price Index 2y change); **Lc1** (housing; real economy).
- Period [2008q3-2009q4]: systemic crisis: **B**; **P1**; **Lc1** and **Lba**, **Lm**, **Lc3** (risk taking and monetary).

France: Out-of-sample: Endogenous reweighting

- Period [2010q1-2010q4]: pre euro-crisis. Models dominating are **P1**, **Lbu**, **Lm** (monetary), **Lba** (risk appetite, financial conditions, share price), **Lc1** (housing; real economy, **Lc3** (risk taking and monetary), **B1** (Consumer Prices, Short-term interest rates (nominal), Banking credit to private sector 1y change, Banking credit to private sector 2y change, Rent Price Index 2y change.)
- Period [2010q4-2017q4]: Models dominating are **Lba** (risk appetite, financial conditions, share price), **Lc1** (housing; real economy), **Lc3** (risk taking and monetary), **B1** (Consumer Prices, Short-term interest rates (nominal), Banking credit to private sector 1y change, Banking credit to private sector 2y change, Rent Price Index 2y change.), **Lm** (monetary), **P4** (Share Price 1y change; Real GDP 1y change; Banking credit to private sector 1y change; Rent Price Index 1y change; Total Liquidity Index 1y change; Financial Condition Index 1y change).

Contributions to crisis prediction

- Spike in probability due to **P1** (Total Credit to private non-financial sector 1y change, Total credit to non-financial firms 2y change, Real GDP 2y change, Rent Price Index 2y change.); **Lc1** (housing; real economy) and **B** (Price-to-rent ratio; Price-to-income ratio; Unemployment ratio; Total Credit to non-financial corporations, % GDP; gap to trend; 1y change; Total Credit to private non-financial sector 1y change; Banking credit to private non-financial sector 1y change; GDP1y change; M3 1y change; Financial Conditions Index 2y change; 3y change; Total Liquidity Index 3y change; Bank equity.)
- Summary: For France, total and bank Credit to private non-financial sector, price to rent, financial conditions, unemployment, GDP are important variables to signal a crisis.

Root Mean Square Errors

Online aggregation Rule	RMSE
EWA	0.23
ML	0.25
OGD	0.30
Ridge	0.23
Uniform	0.41
Best convex combination (ex post)	0.24

Table 1: RMSE France

AUROC

- The ROC curve represents the ability of a binary classifier by plotting the true positive rate against the false positive rate for all thresholds.
- The AUROC is the area under the ROC curve :

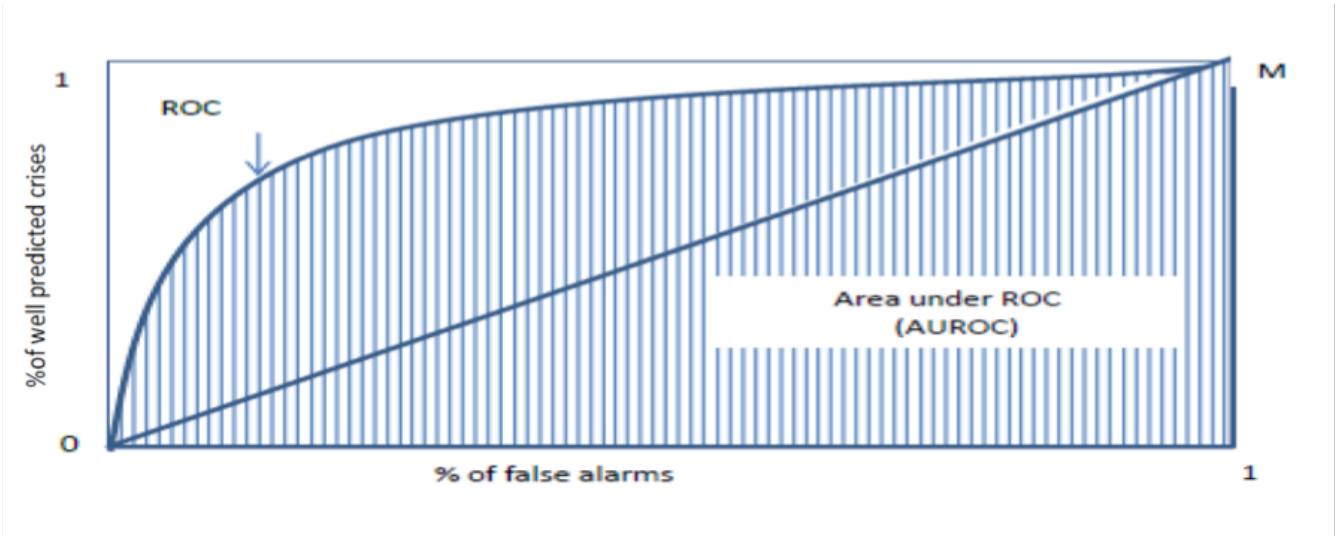


Figure 7: ROC curve

AUROC: France, out-of-sample

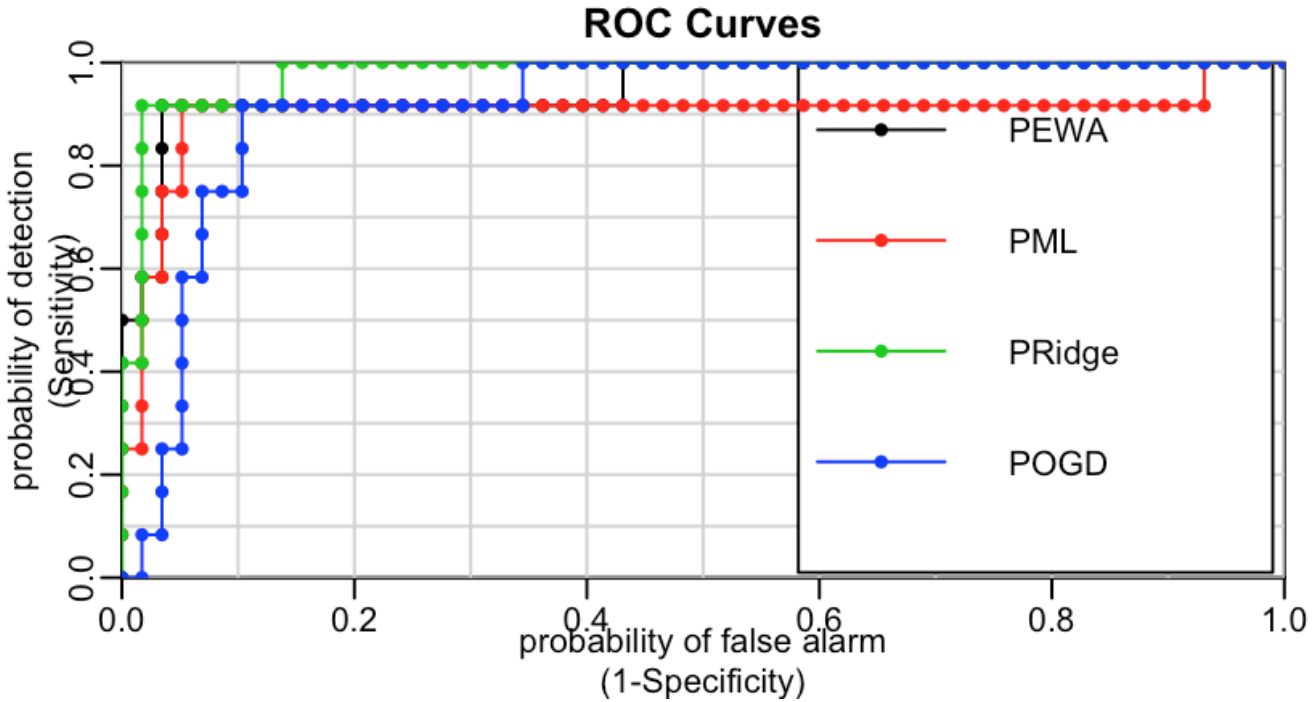


Figure 8: AUROC-EWA=0.95; ML=0.90; OGD=0.92; Ridge=0.98

AUROC: France, out-of-sample

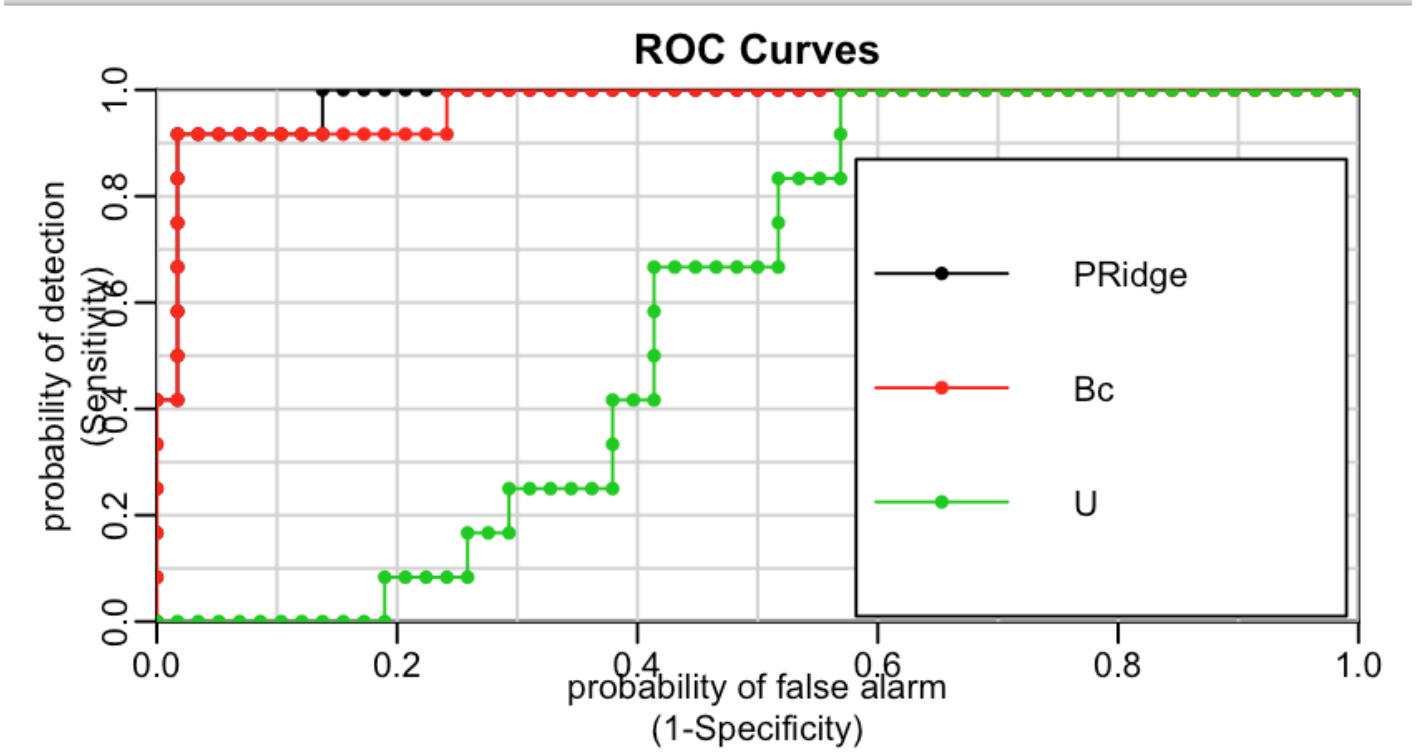
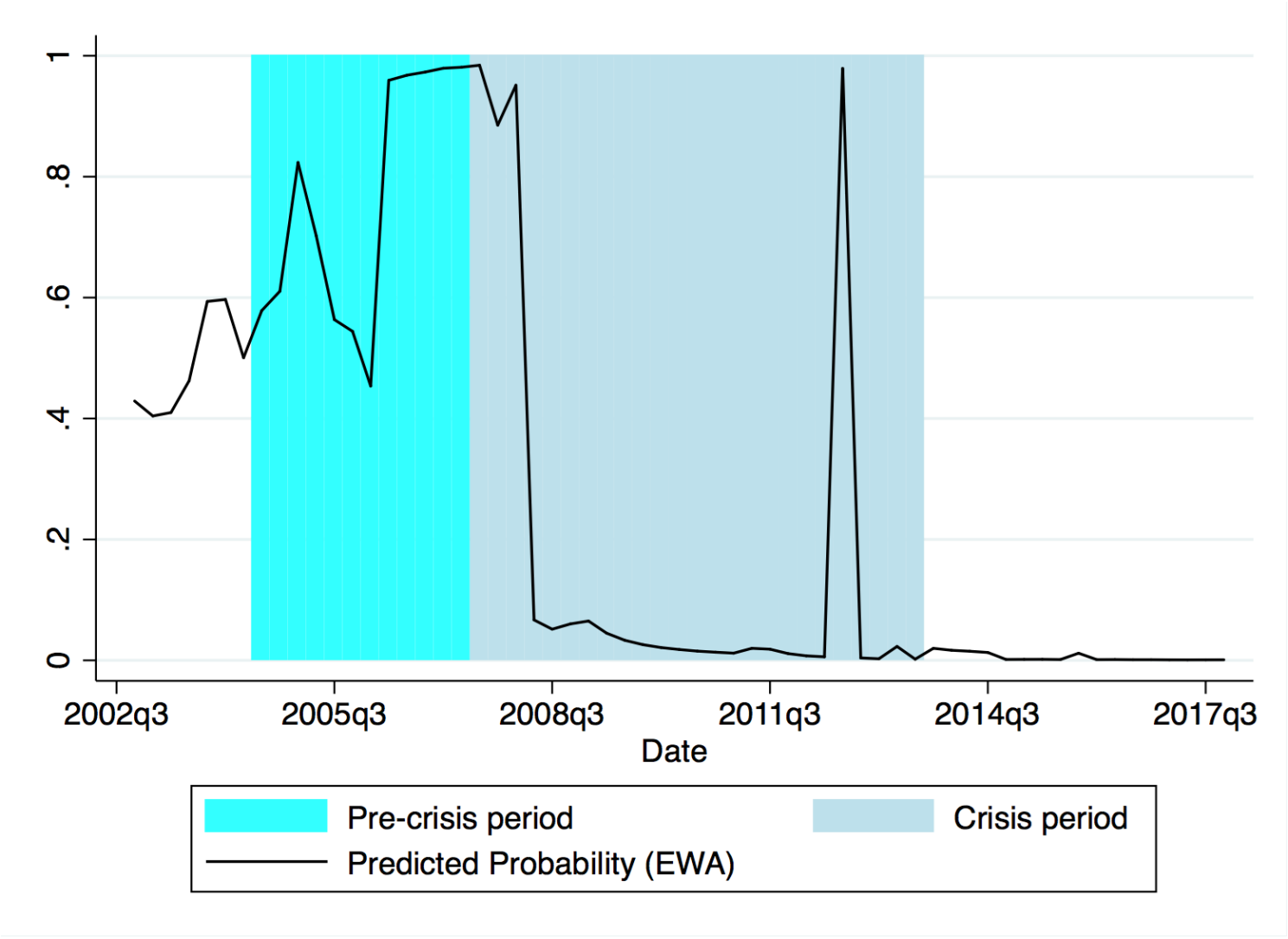


Figure 9: AUROCs Ridge=0.98; Best fixed convex combination (ex post)=0.95; Uniform=0.59

Germany: pre-crisis period out-of-sample (different dates from France)



Germany: pre-crisis period out-of-sample

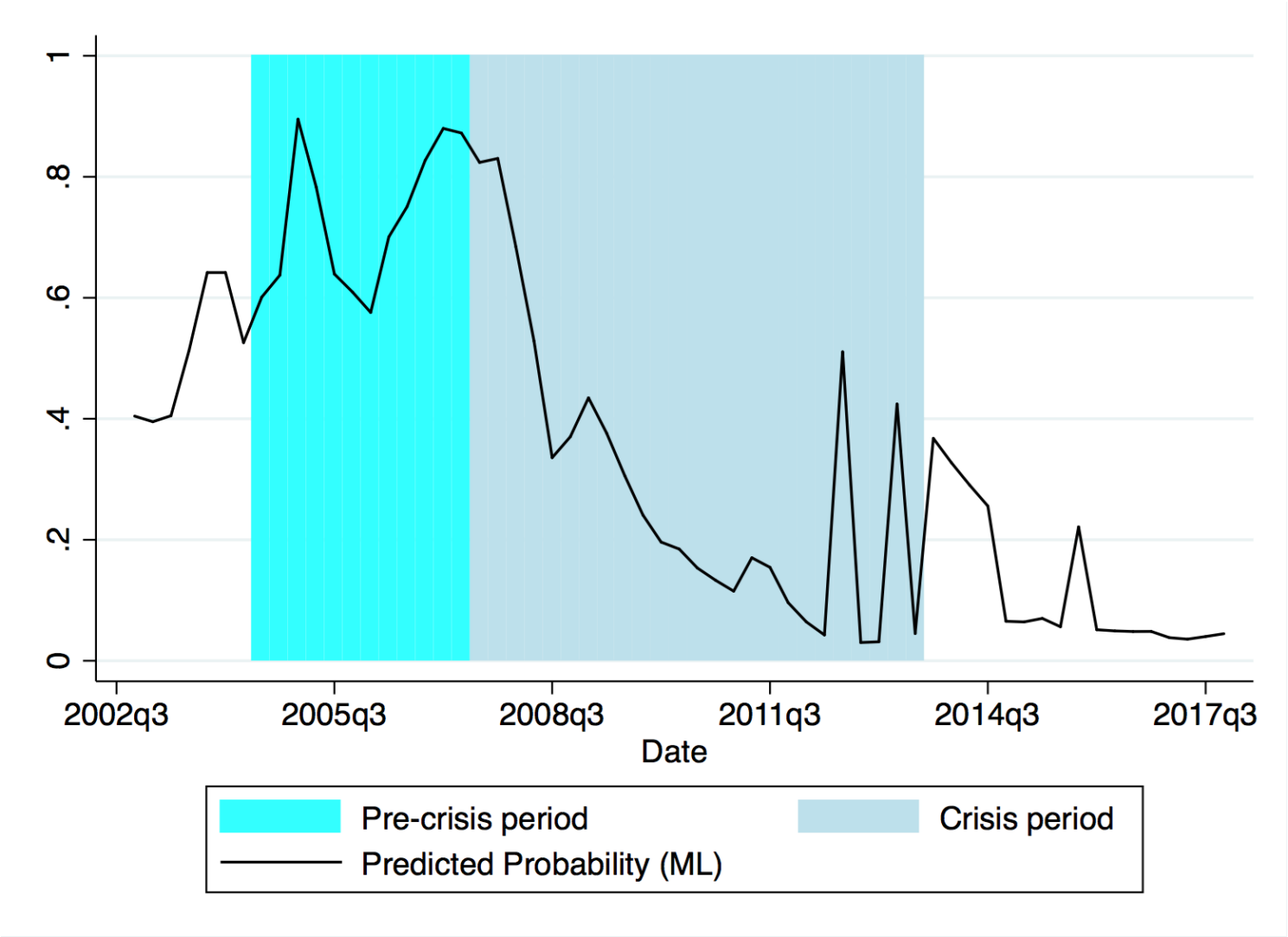


Figure 11: Predicted probability - ML

Germany: pre-crisis period out-of-sample

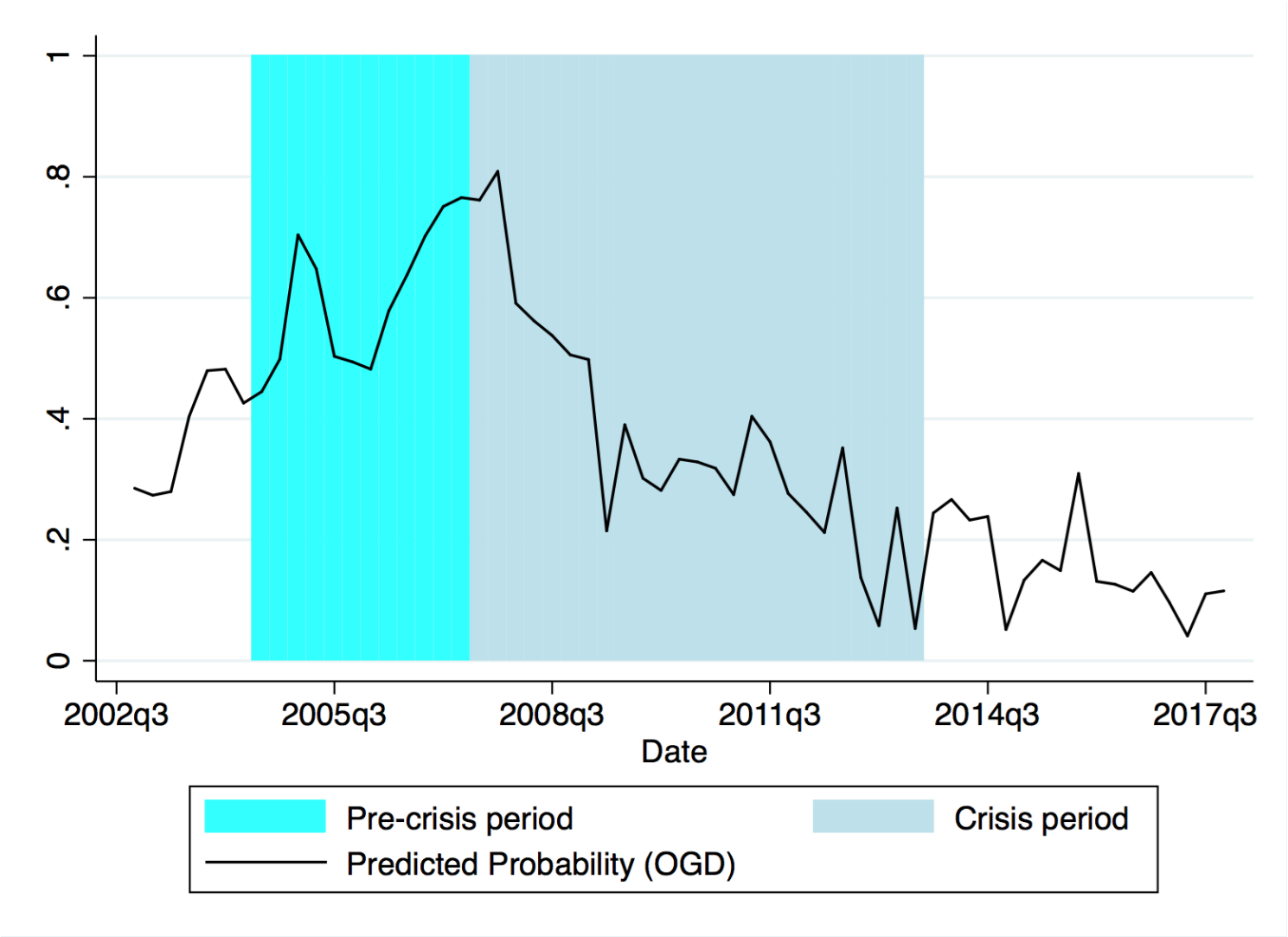


Figure 12: Predicted probability - OGD

Germany: pre-crisis period out-of-sample

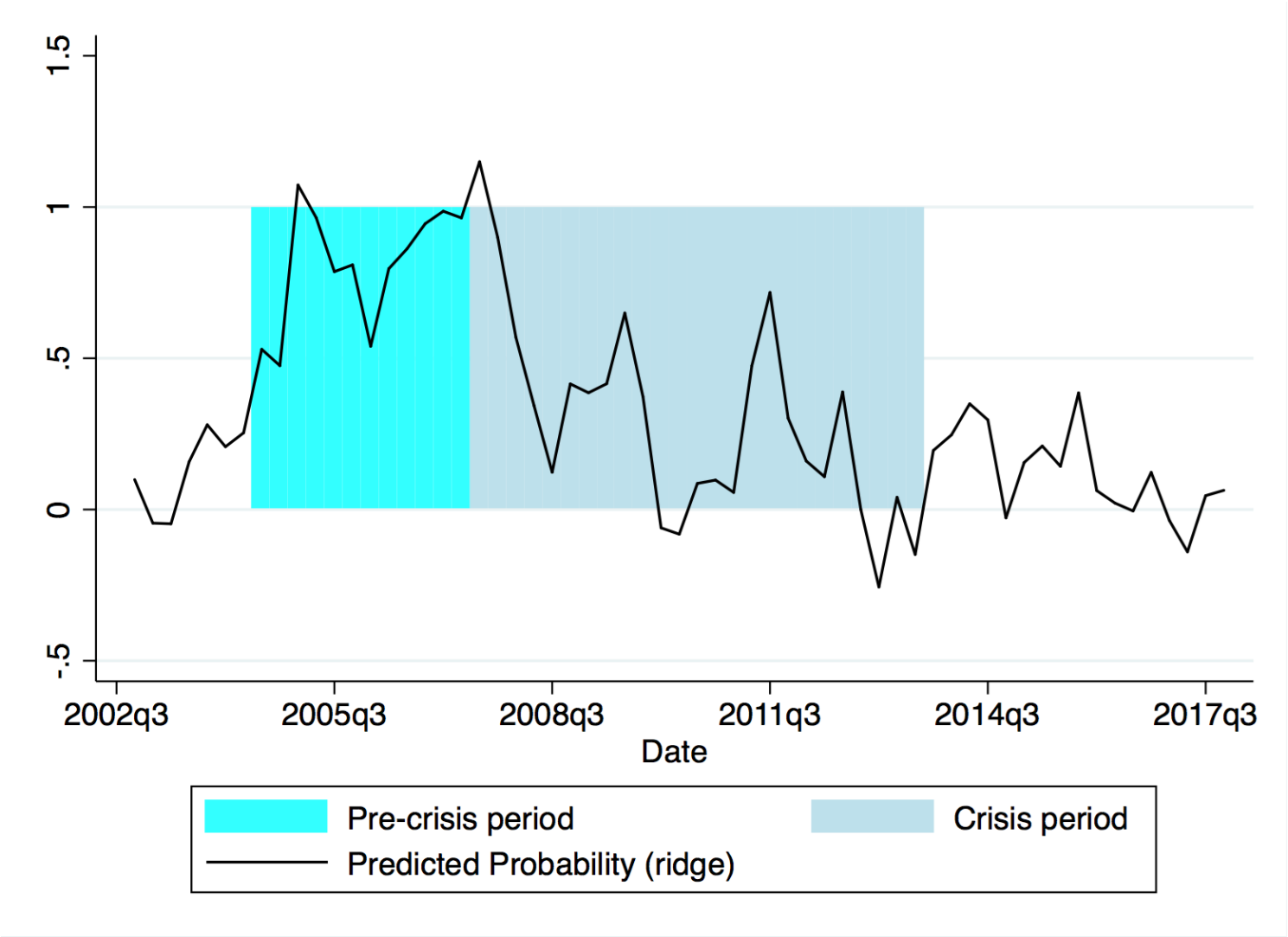


Figure 13: Predicted probability - Ridge

Germany

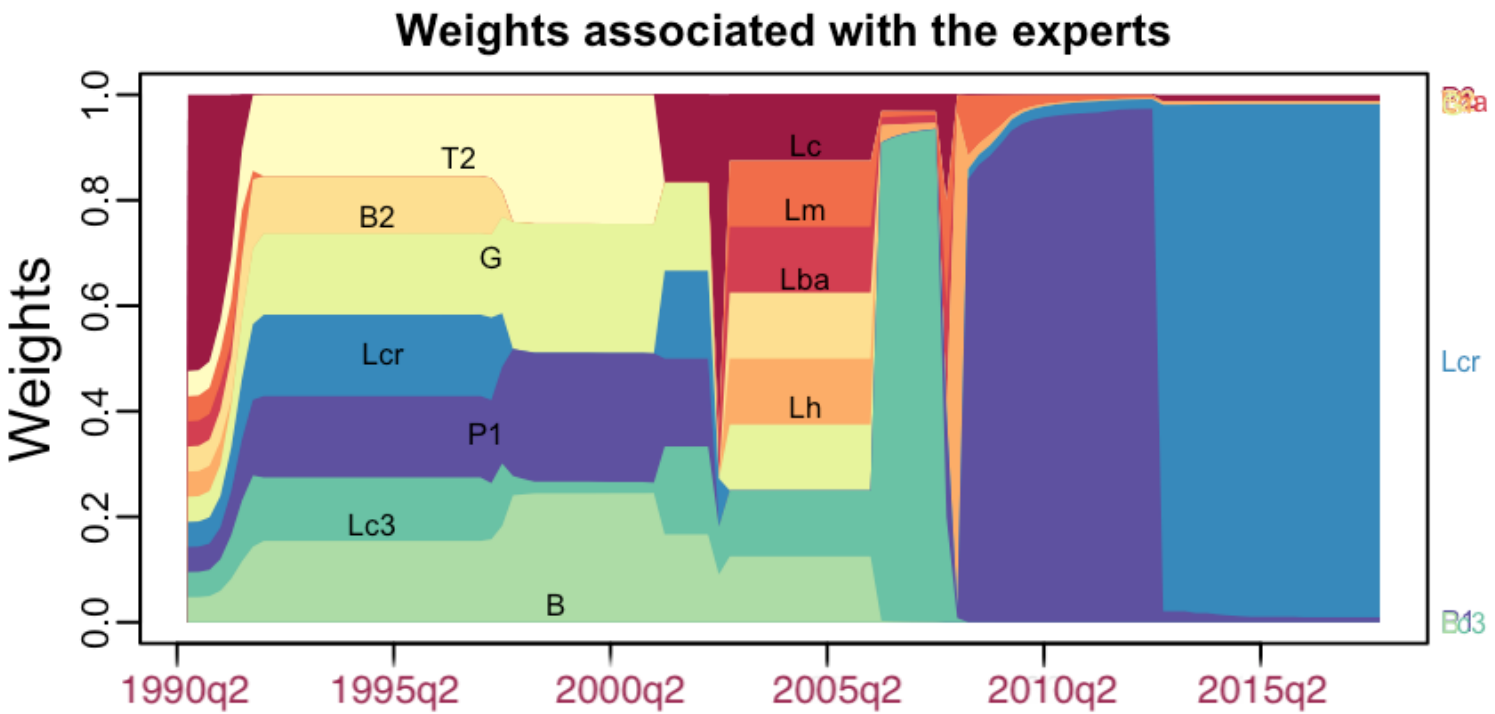


Figure 14: Weights - EWA

Germany

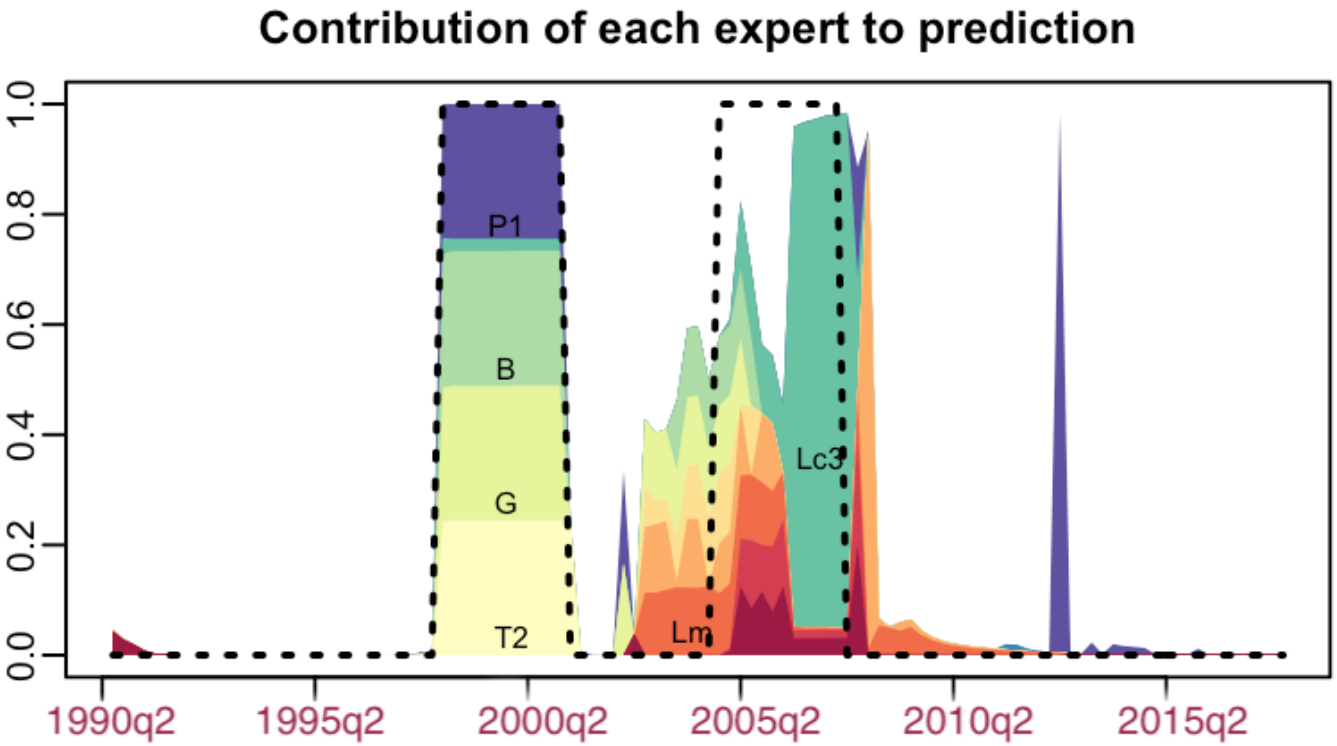


Figure 15: Weights - EWA

Contributions to crisis prediction: Germany

- Spike in probability due to **Lc3** (monetary and risk taking) and earlier **Lm Lc Lba Lh**. Strange spike in euro area crisis due to **P1** (Share Price Indices, Rent Price Index 1y change, Total Credit to non-financial corporations 2y change, Equity Holdings.)

- Summary: For Germany, monetary and risk taking variables seem to be more important than for France to signal a crisis.

AUROC: Germany, out-of-sample

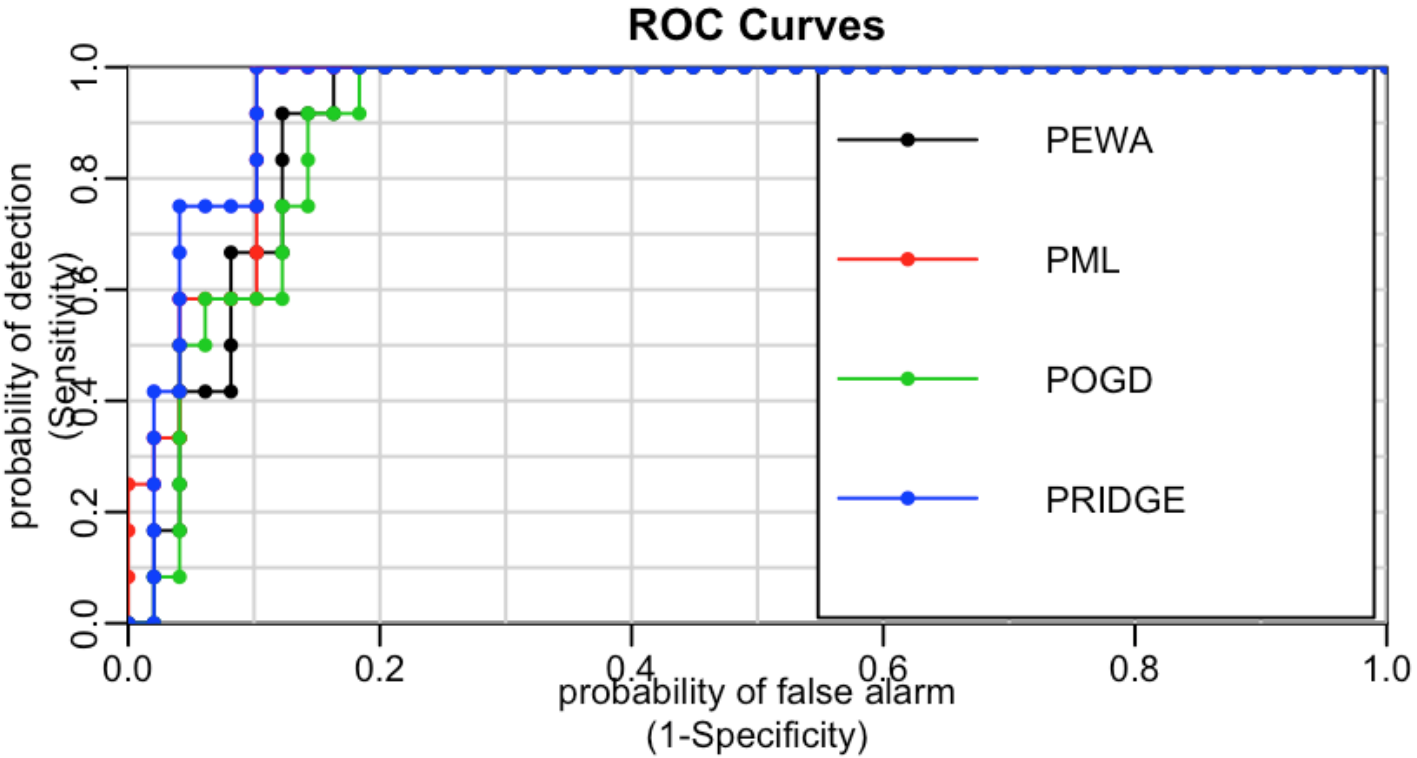


Figure 16: AUROCs EWA= 0.92; ML=0.95; OGD=0.92; Ridge=0.98. Quasi-real time.

AUROC: Germany, out-of-sample

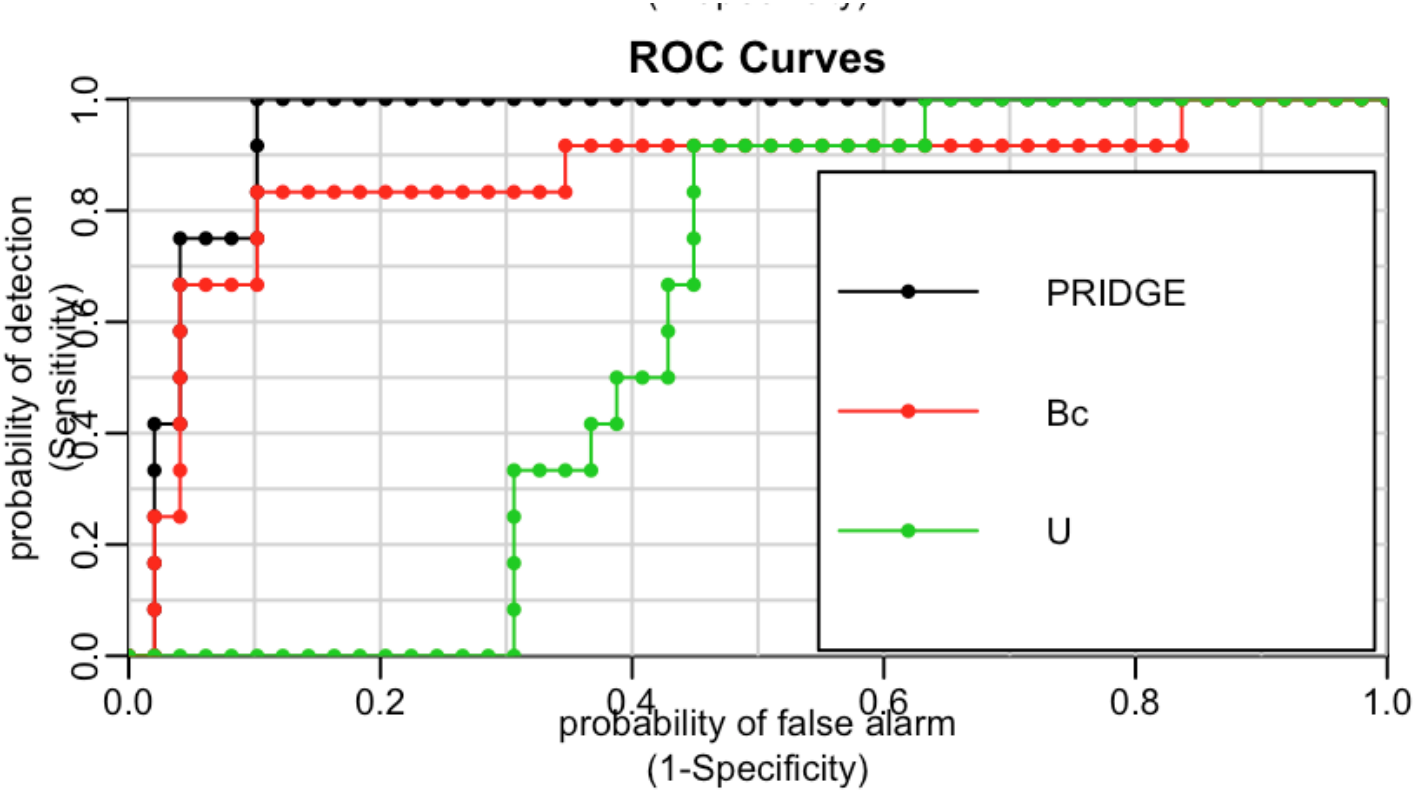


Figure 17: AUROCs Ridge=0.98; Best fixed convex combination (ex post)=0.86; Uniform=0.6

Spain: pre-crisis period out-of-sample

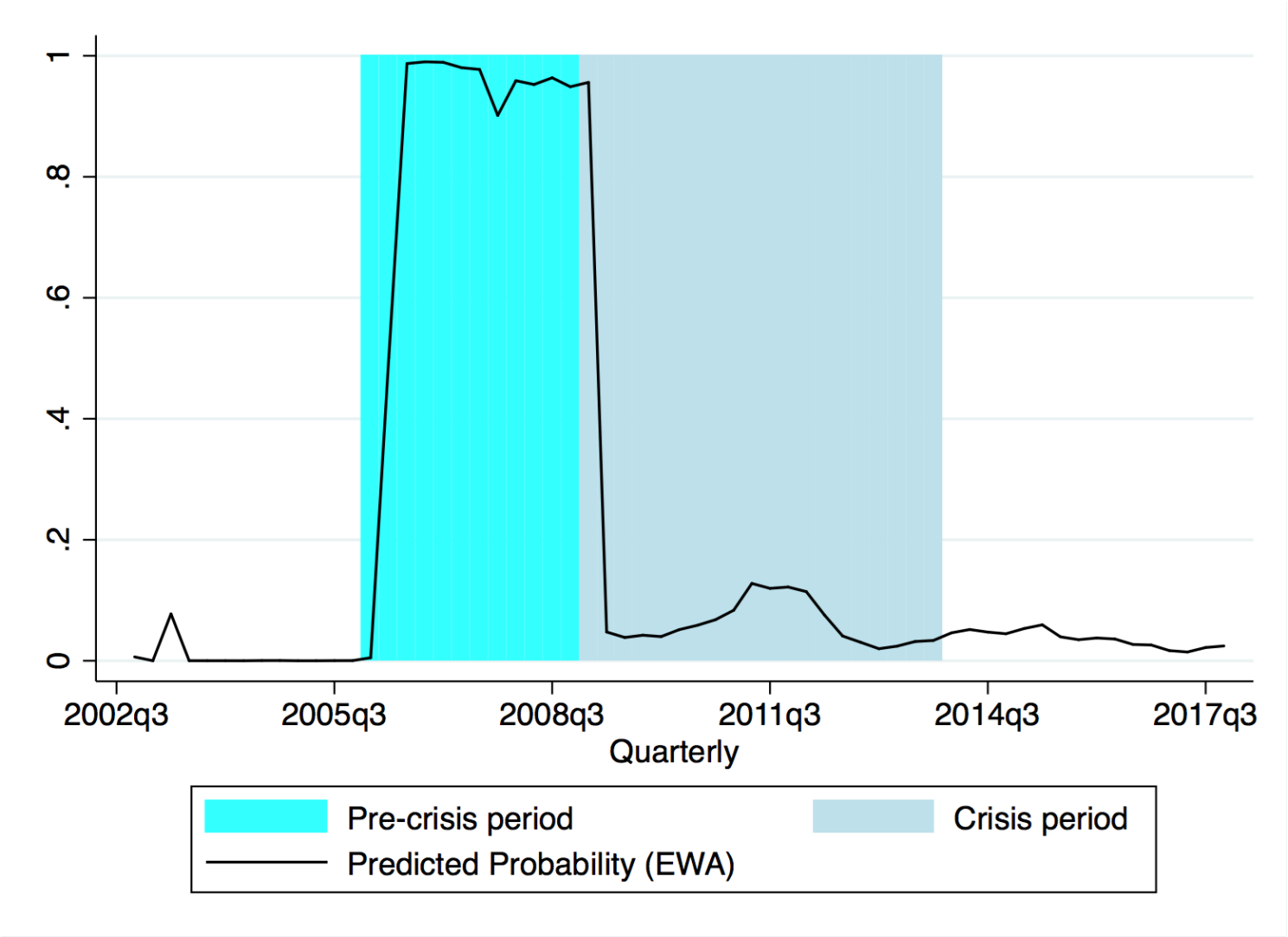


Figure 18: Predicted probability - EWA

Spain: pre-crisis period out-of-sample

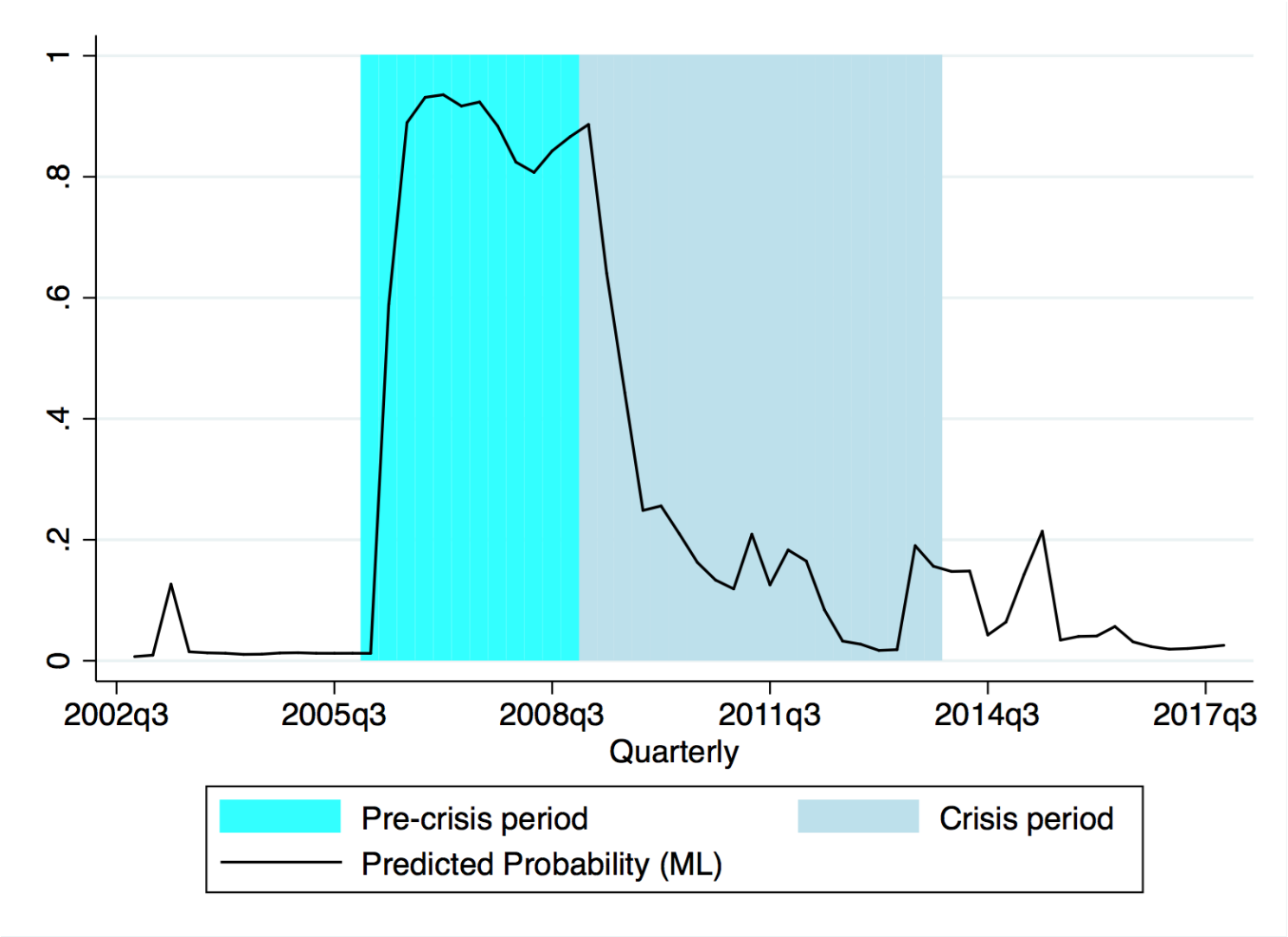


Figure 19: Predicted probability - ML

Spain: pre-crisis period out-of-sample

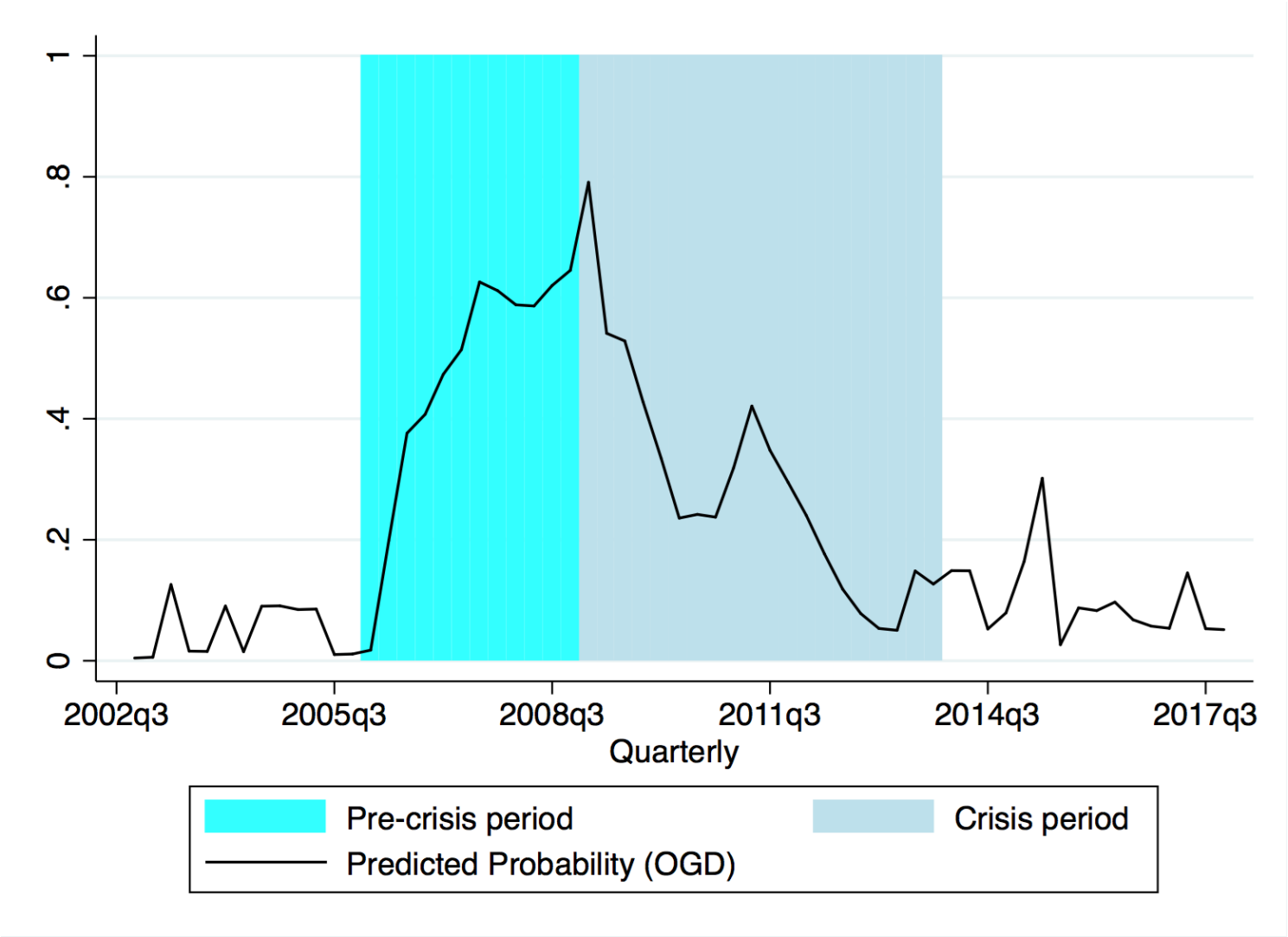


Figure 20: Predicted probability - OGD

Spain: pre-crisis period out-of-sample

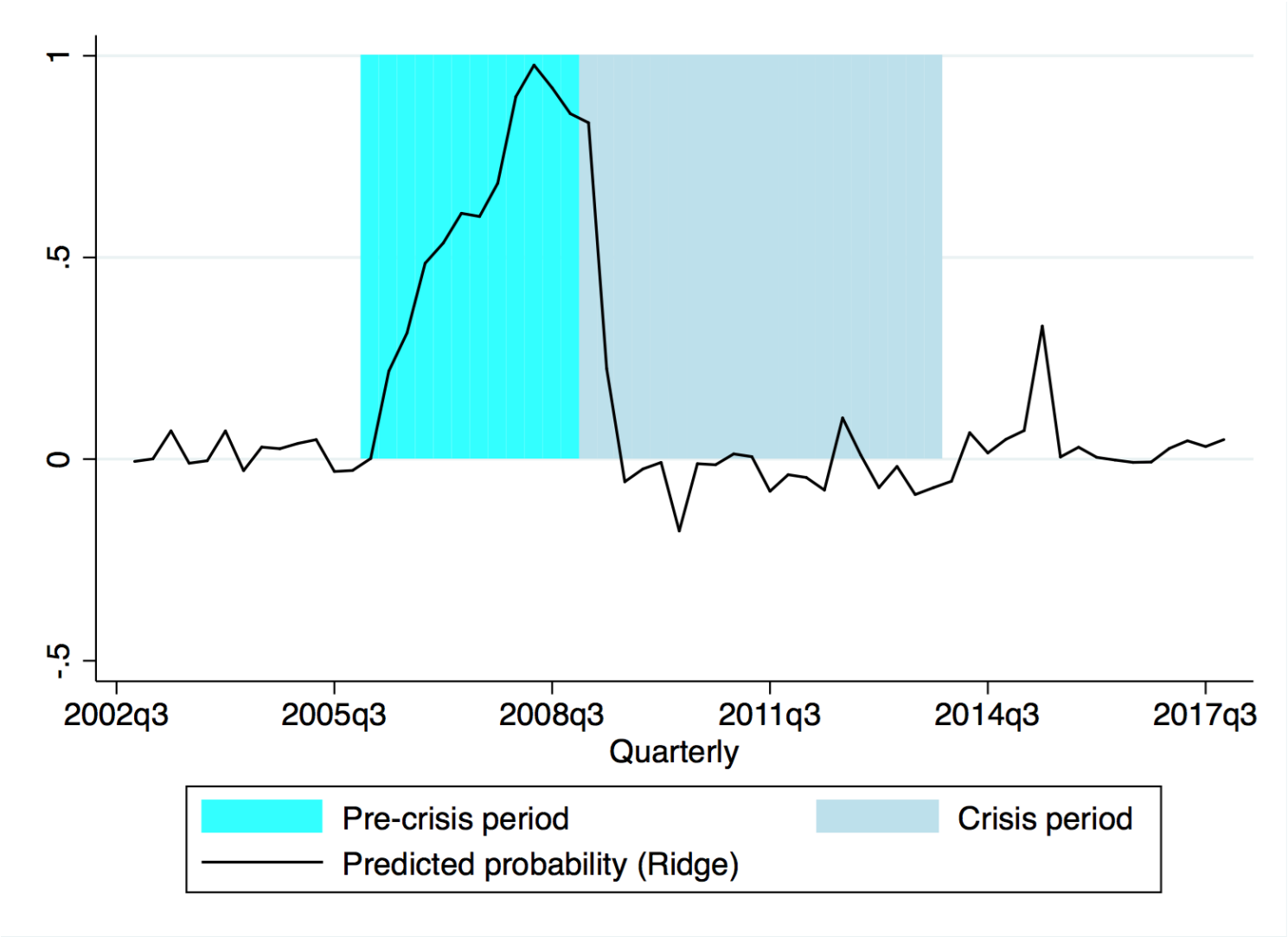


Figure 21: Predicted probability - Ridge

Spain

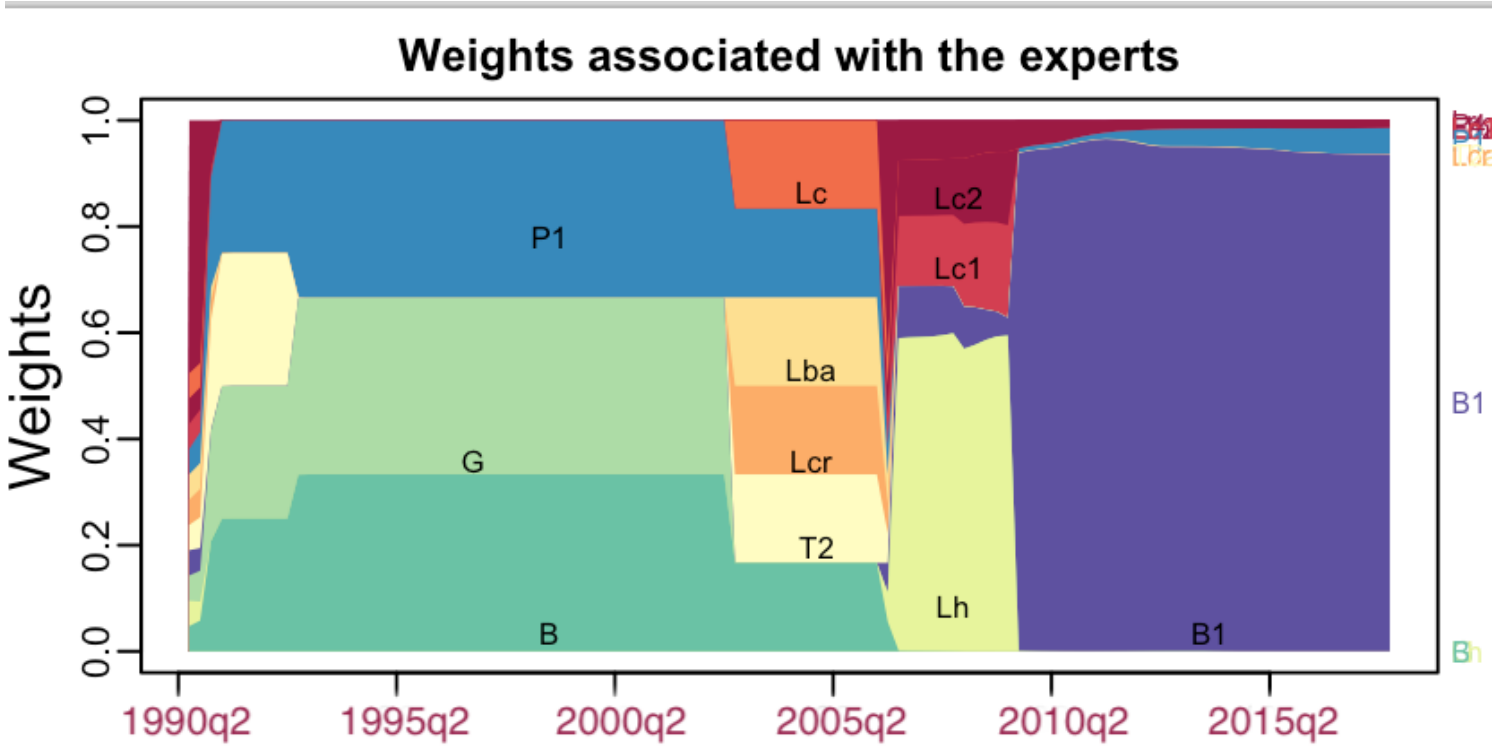


Figure 22: Weights - EWA

Spain

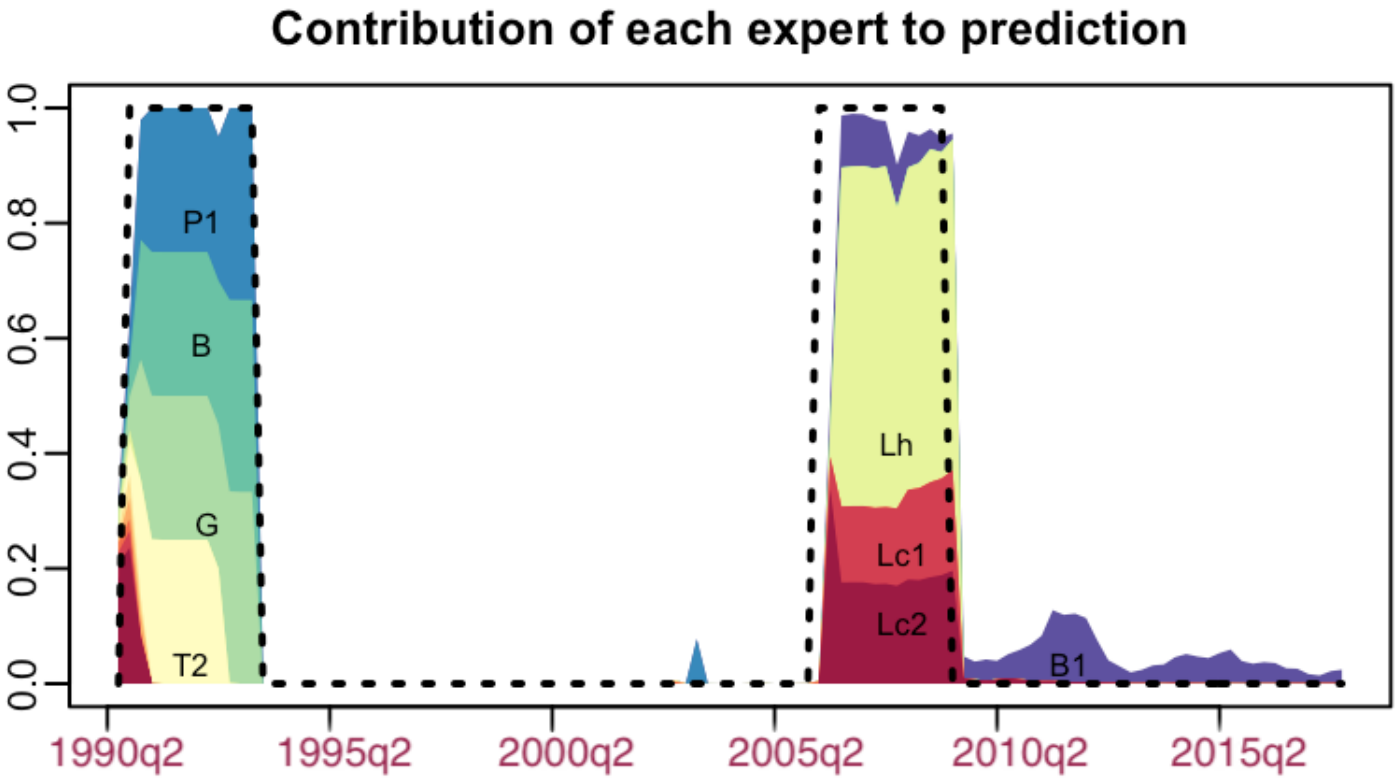


Figure 23: Weights - EWA

Contributions to crisis prediction: Spain

- Spike in probability due to **Lc1** (housing; real economy) and **Lc2** (credit and risk taking) and **Lh** (housing); a bit of **B1** Consumer Prices; Short-term interest rates (nominal); Banking credit to private sector 1y change; Total Credit to non-financial corporations 2y change; Total credit to non-financial sector – gap to trend; Consumer prices : 1y change and 3y change; M3 : 1y change and 2y change.

- Summary: Housing variables are really important to signal a crisis in Spain.

AUROC: Spain, out-of-sample

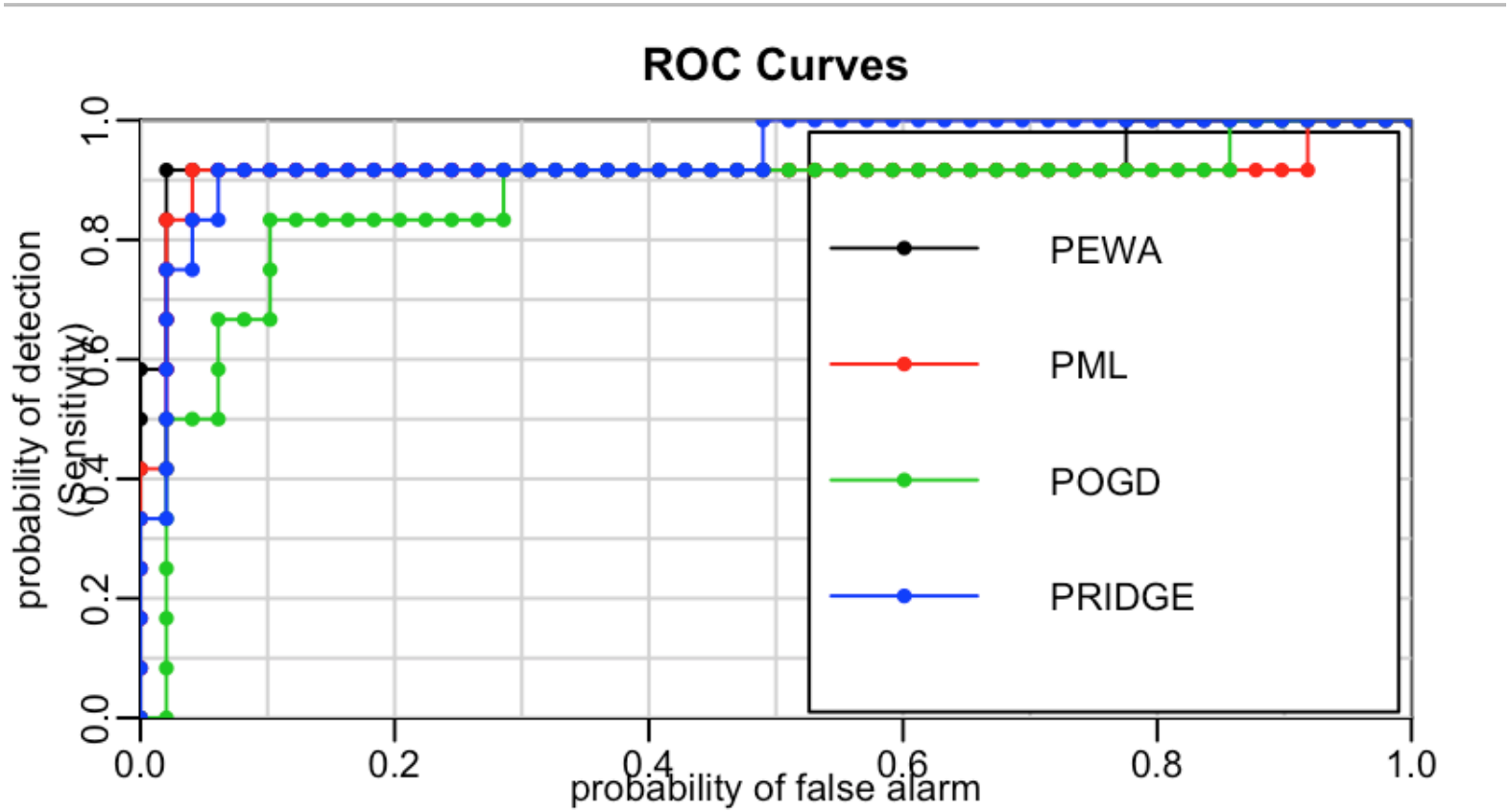


Figure 24: AUROC. EWA=0.93; ML=0.91; OGD= 0.86; Ridge=0.94. quasi-real time.

AUROC: Spain, out-of-sample

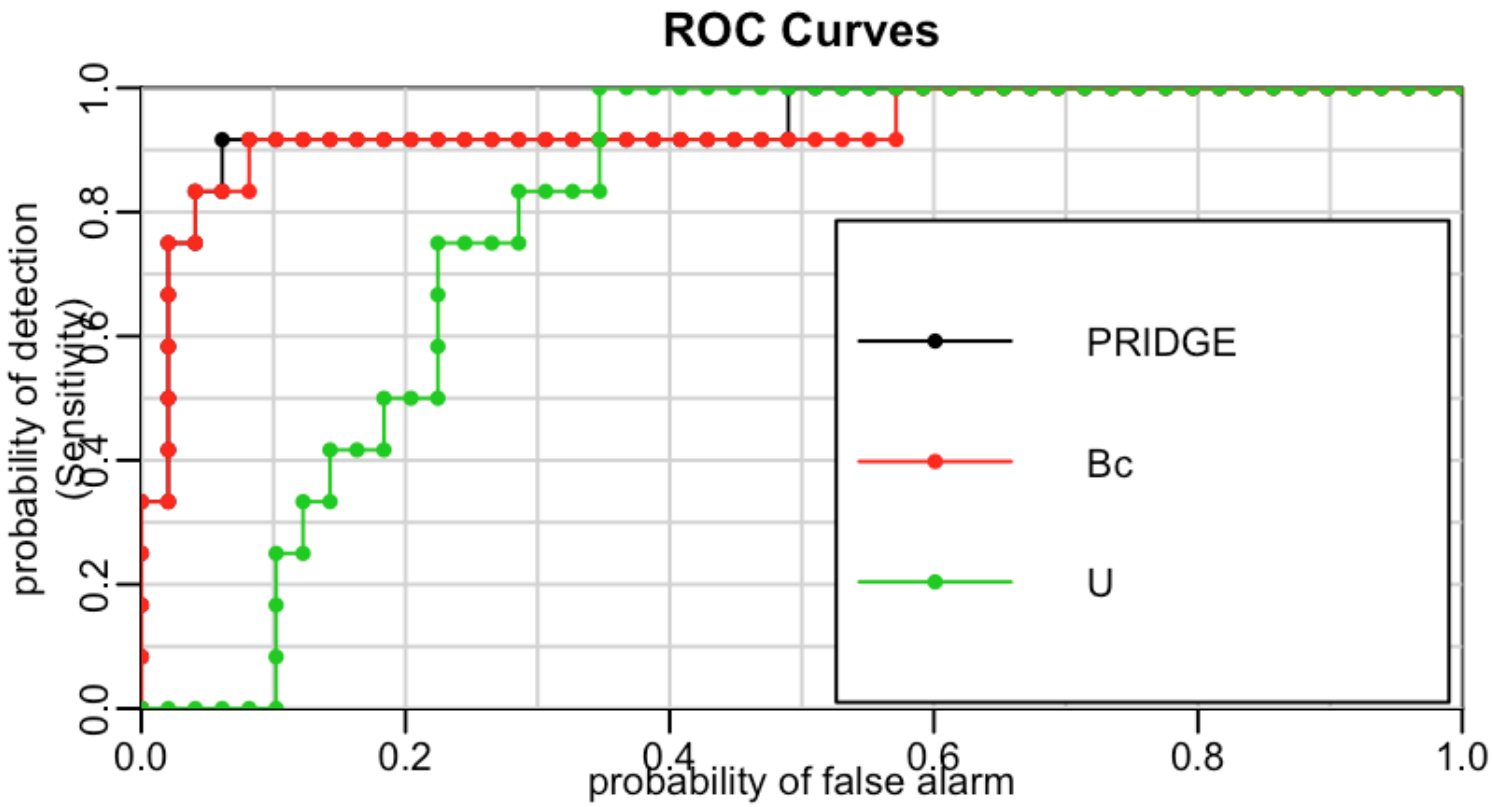


Figure 25: AUROC. Best convex (ex post)= 0.93; Uniform= 0.38; Ridge=0.94. quasi-real time.

UK: pre-crisis period out-of-sample

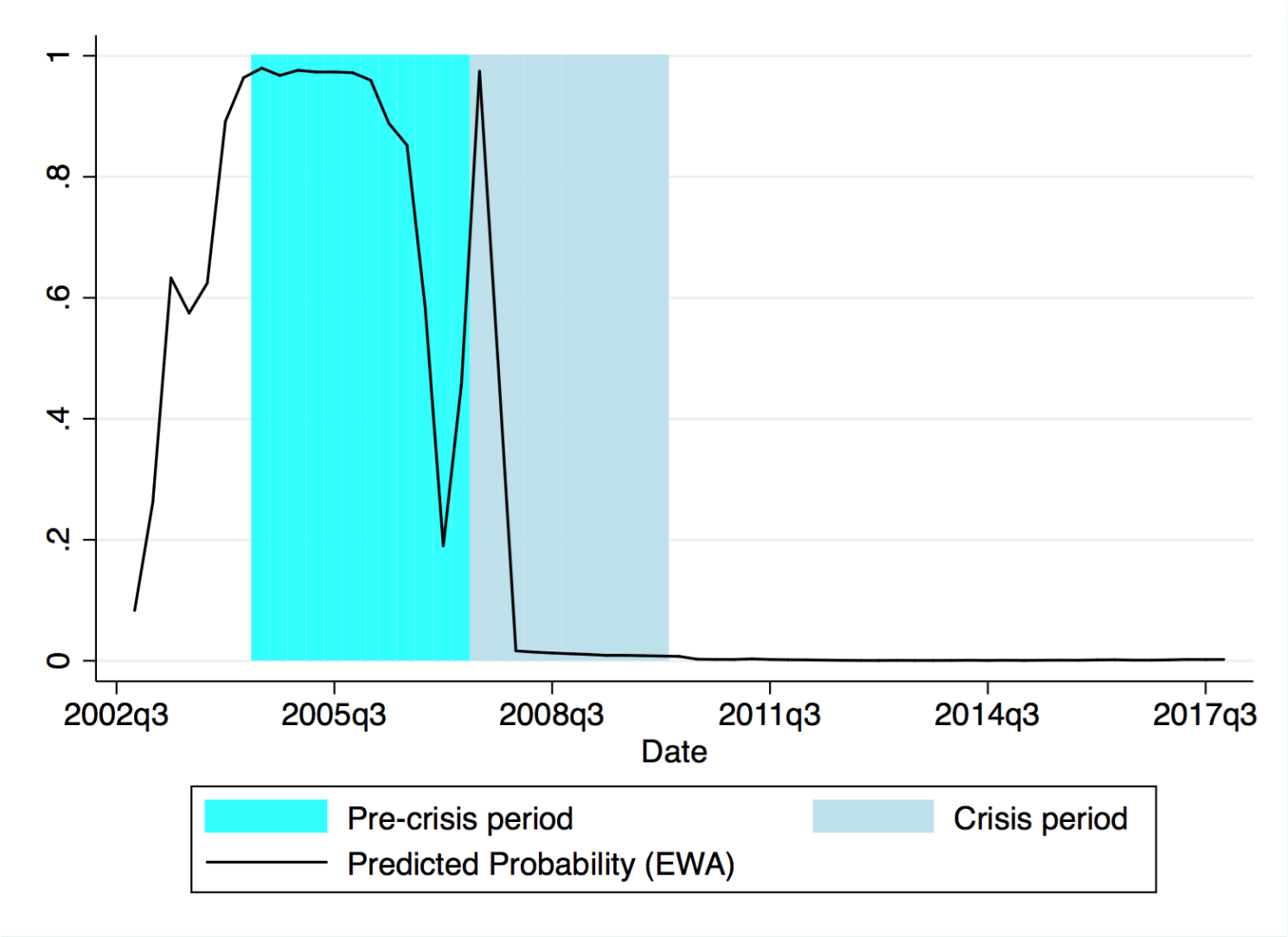


Figure 26: Predicted probability - EWA

UK: pre-crisis period out-of-sample

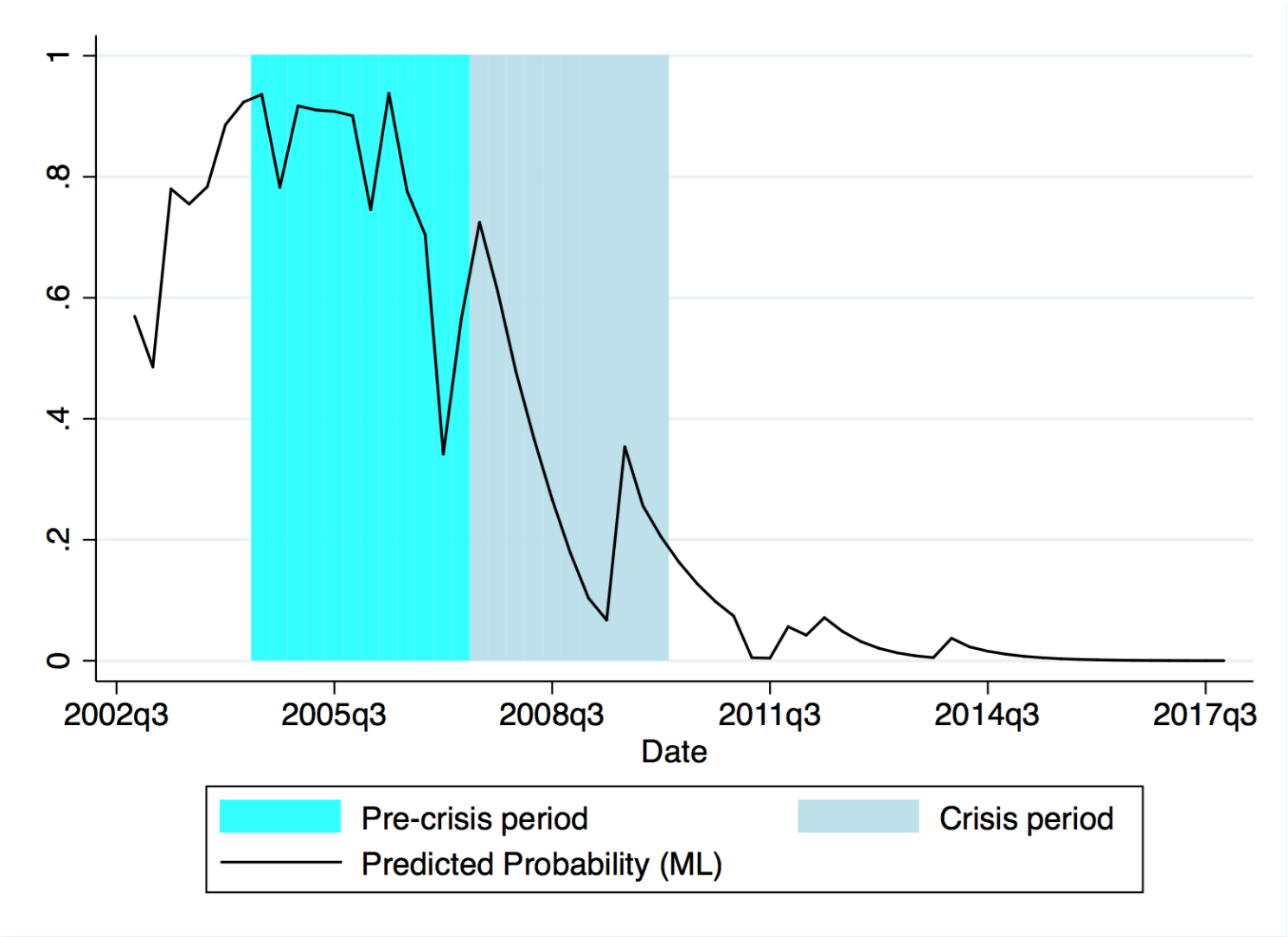


Figure 27: Predicted probability - ML

UK: pre-crisis period out-of-sample

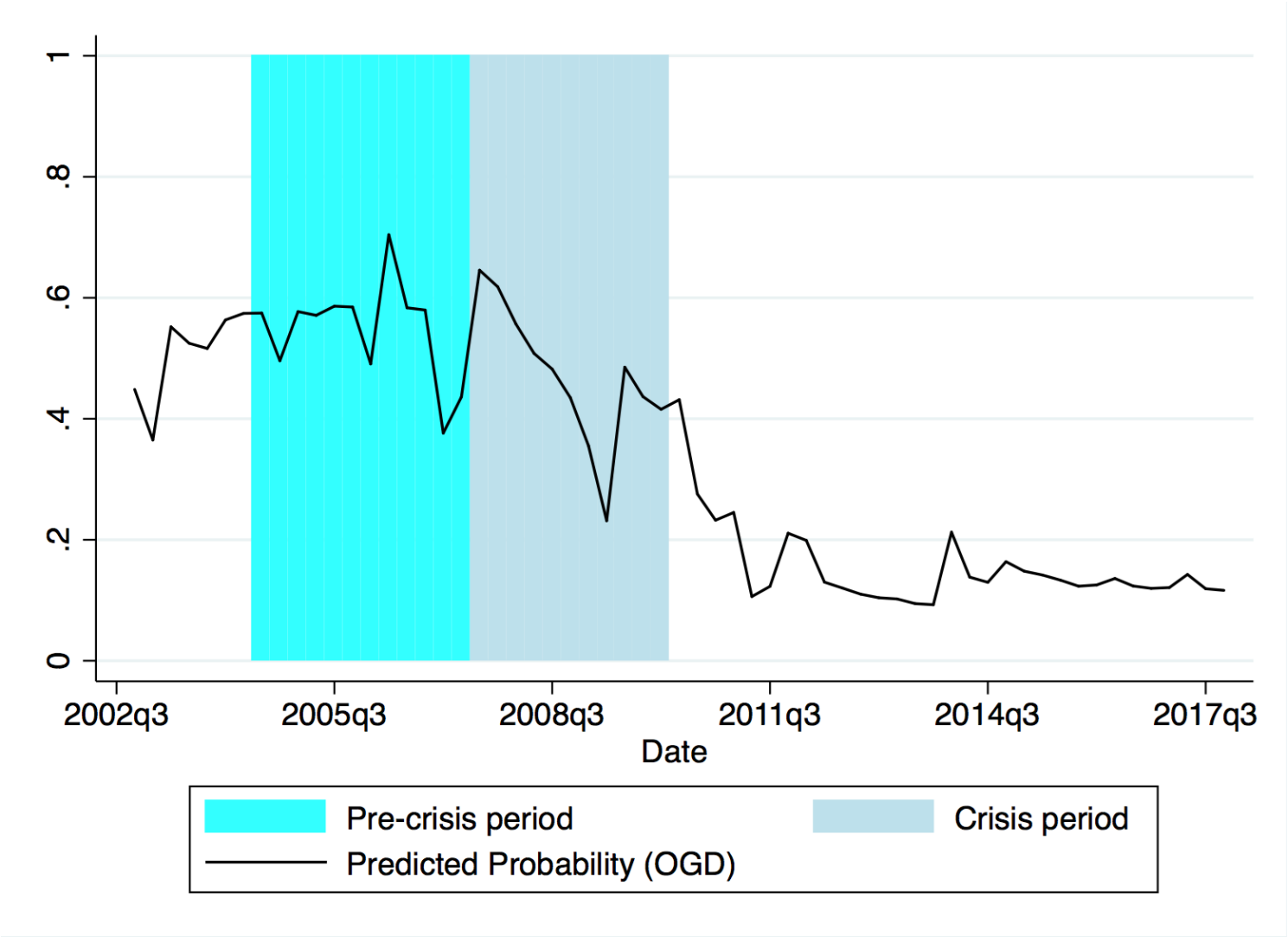


Figure 28: Predicted probability - OGD

UK: pre-crisis period out-of-sample

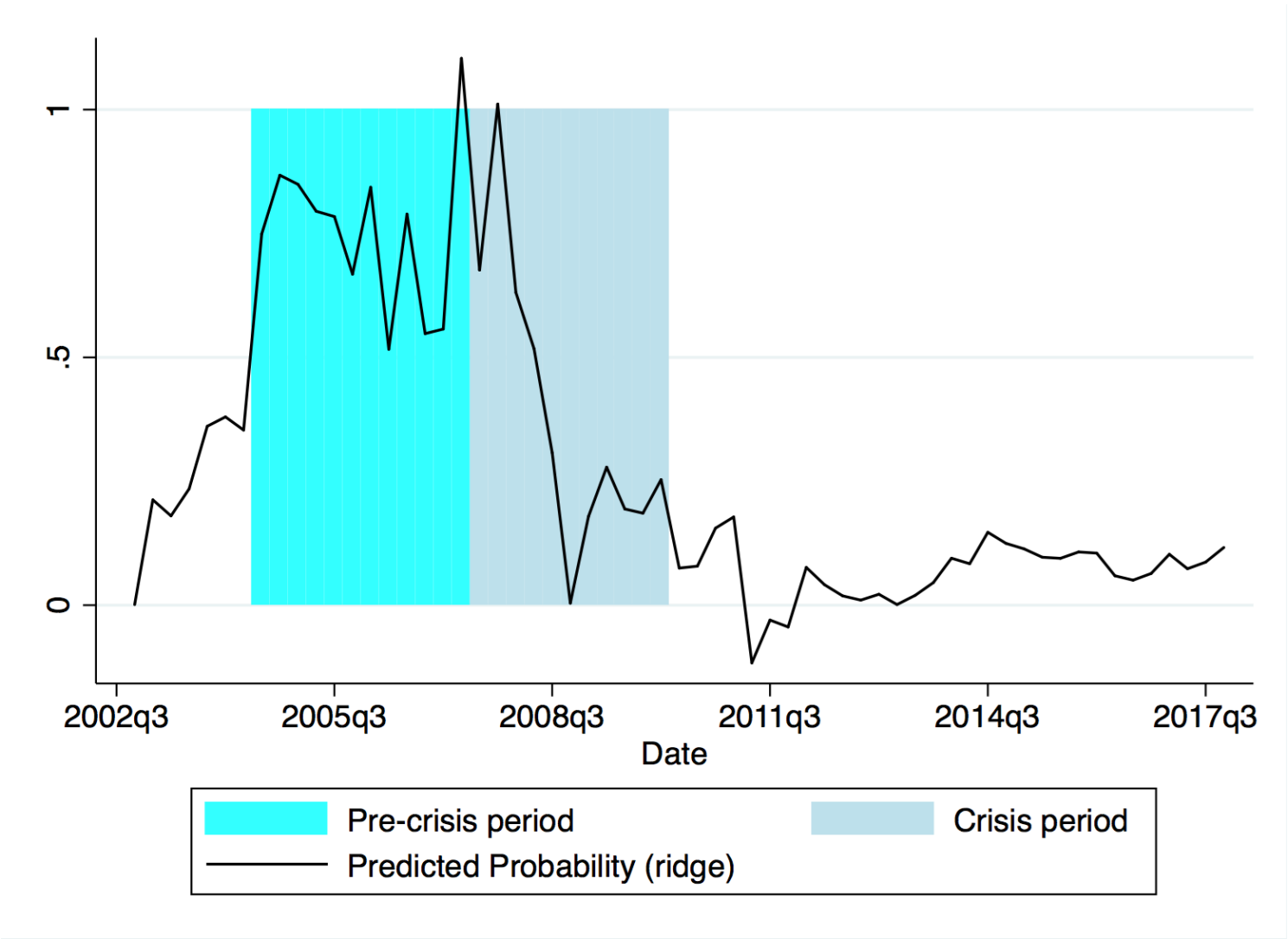


Figure 29: Predicted probability - Ridge

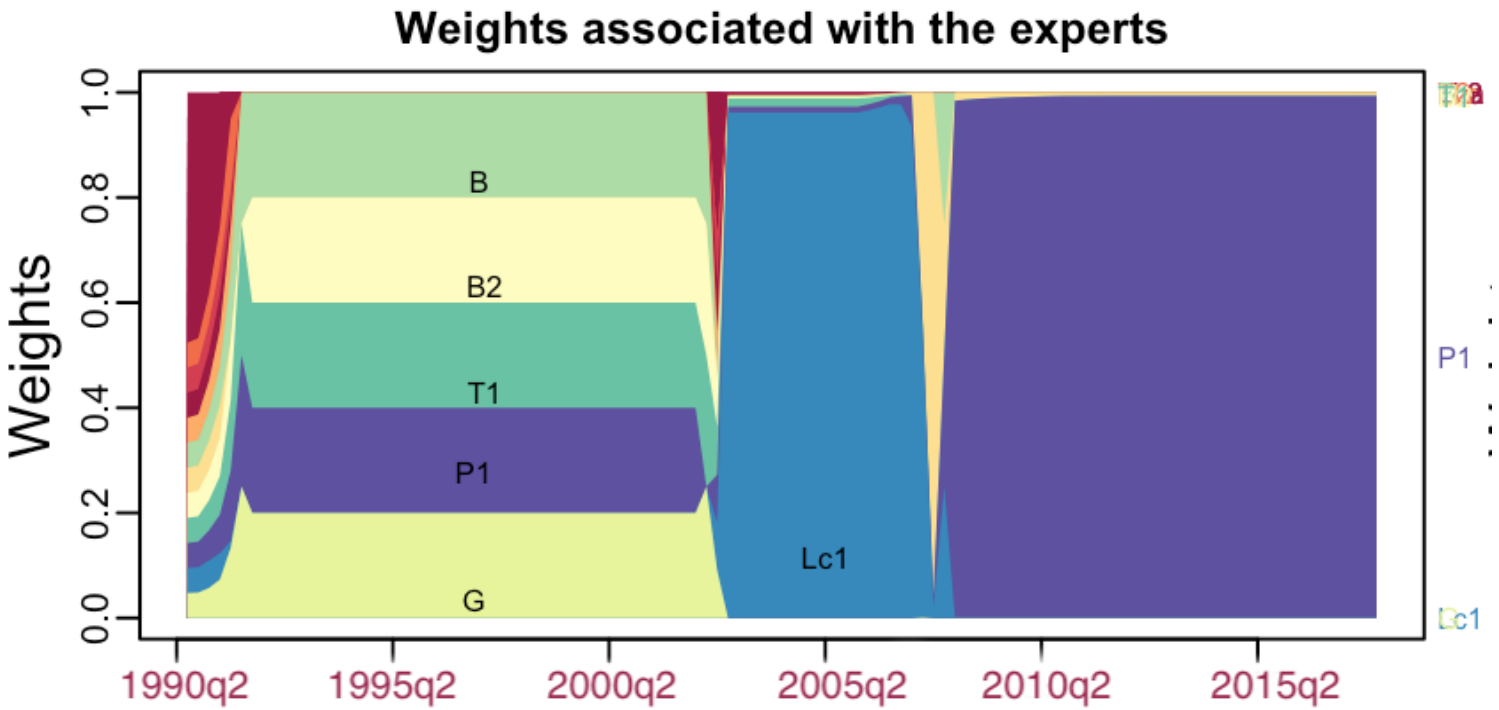


Figure 30: Weights - EWA

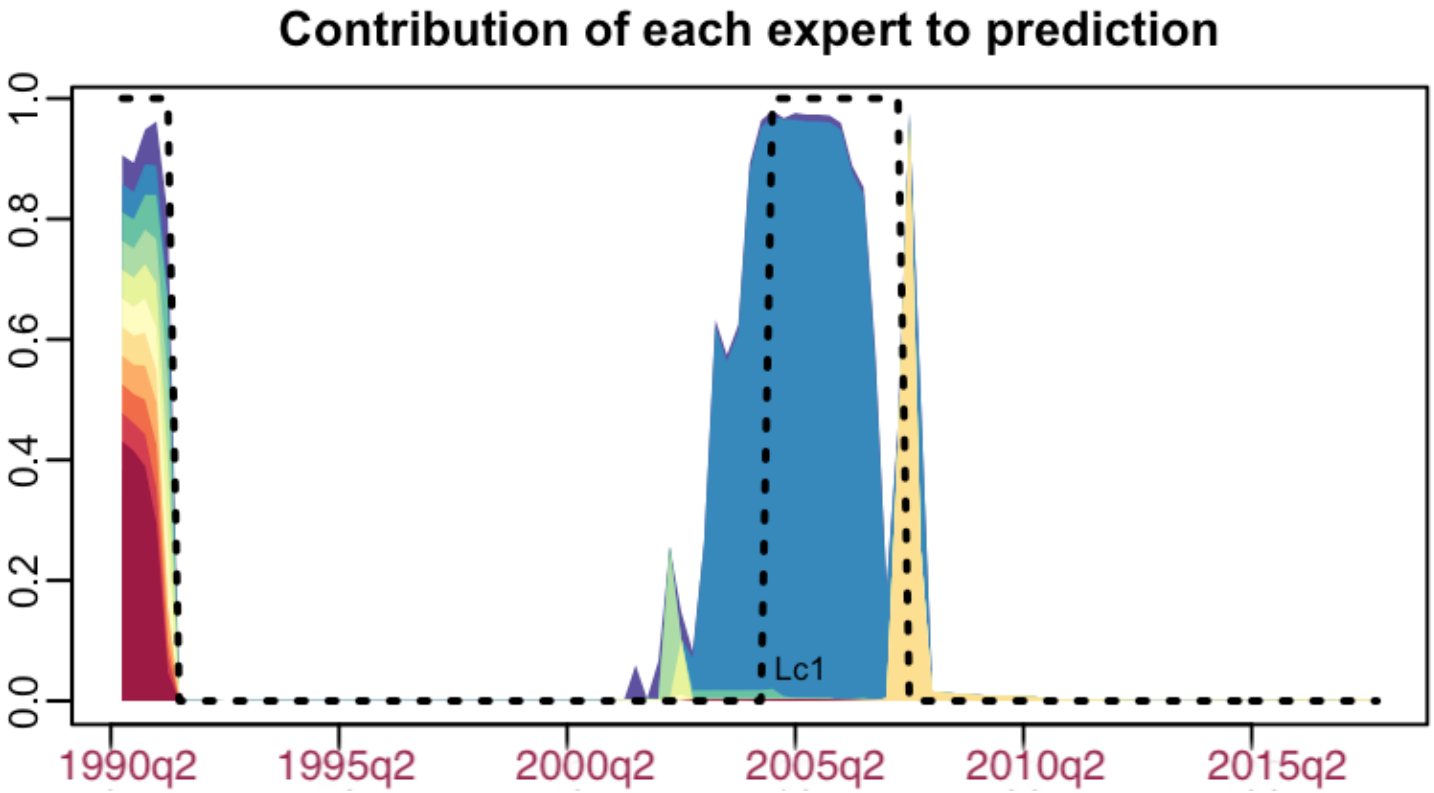


Figure 31: Weights - EWA

Contributions to crisis prediction: UK

- Spike in probability due to **Lc1** (housing; real economy). Thin spike at the end of the pre crisis period is due to **Lh**.

AUROC: UK, out-of-sample

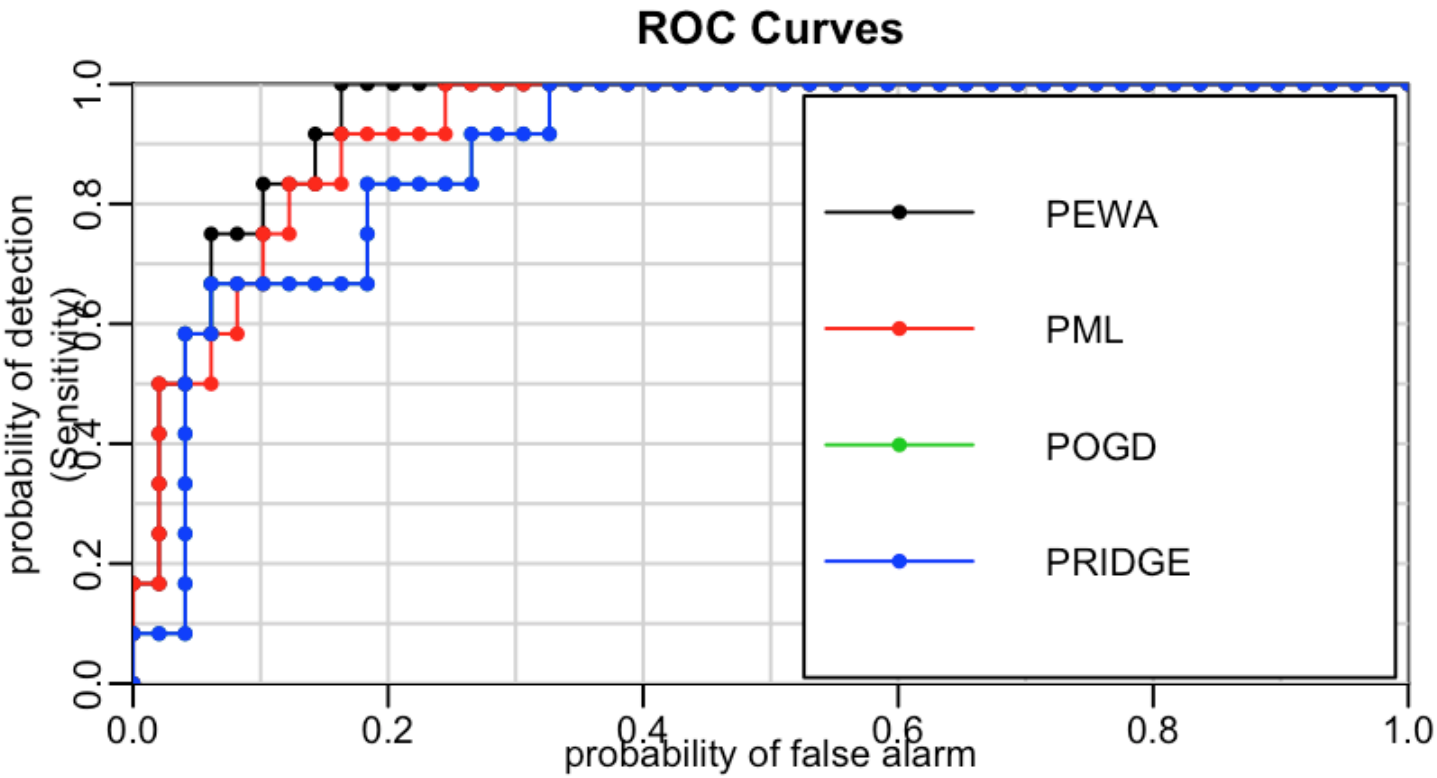


Figure 32: AUROC. EWA=0.95; ML=0.93; OGD= 0.89; Ridge=0.89. quasi-real time.

AUROC: UK, out-of-sample

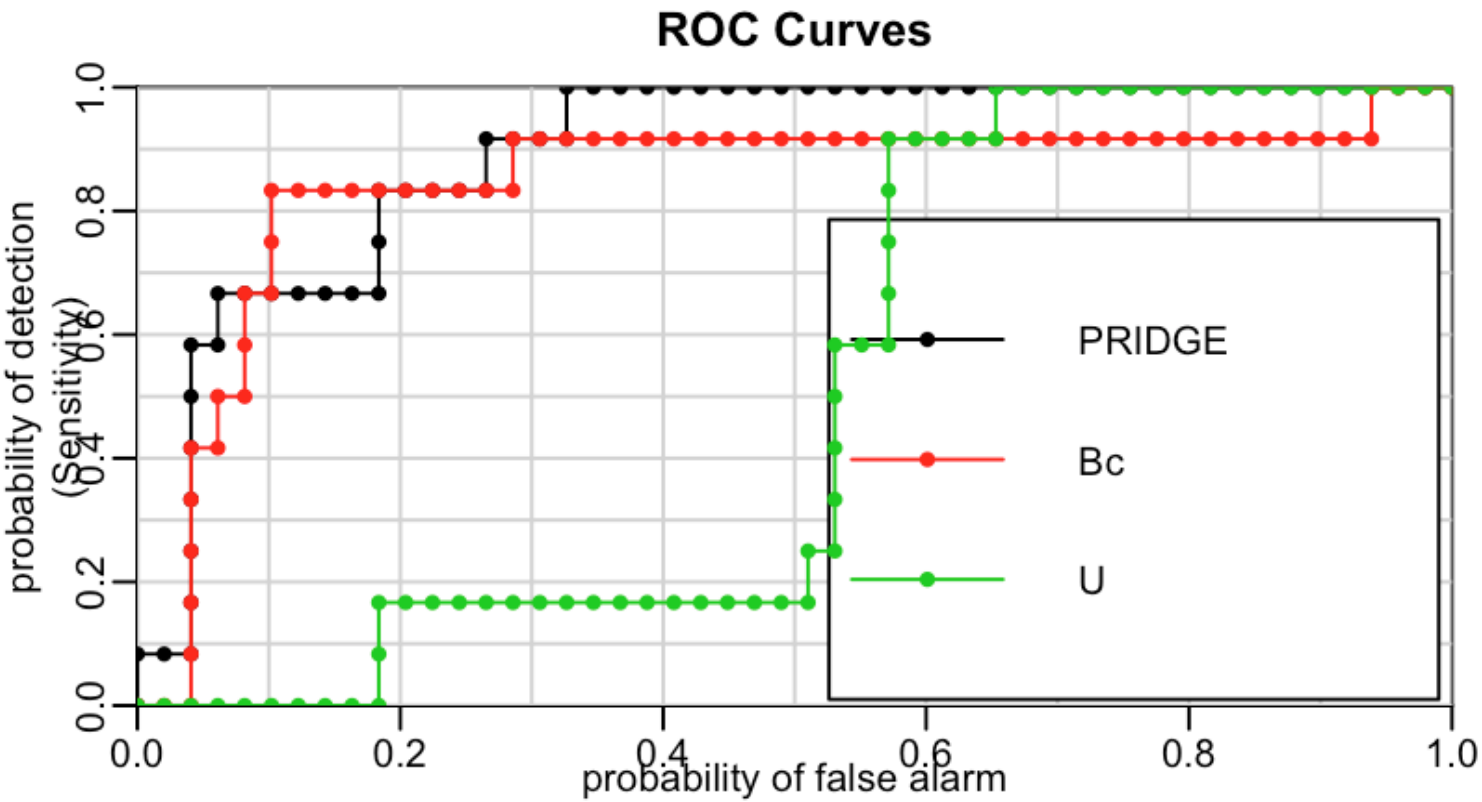


Figure 33: AUROC. Best convex (ex post)= 0.85; Uniform= 0.5; Ridge=0.89. quasi-real time.

Conclusions

- This approach gives strong out-of-sample forecasting results to predict financial crises. Also gives interesting information on relevant variables.
- We now work on the historical database as well (we predict the Great Depression out-of-sample). We can investigate which models are better for Great Depression versus Lehman Brothers.
- We can apply same framework to predicting recessions.
- We could improve data and improve experts.
- Open questions :
  - Using more microeconomic data from bank databases?
  - Revisions and lags in the data? Approach is now quasi real time.
  - Causality?
  - How to test for the effect of macroprudential policies on crisis probabilities?

The ML aggregation rule

In the Multiple learning rates aggregation rule :

The ML aggregation rule

In the Multiple learning rates aggregation rule :

- Each expert has a specific online learning rate.

The ML aggregation rule

In the Multiple learning rates aggregation rule :

- Each expert has a specific online learning rate.
- The rest of the framework does not change

The ML aggregation rule

The ML aggregation rule

Let's consider that every expert j incurs a loss $\ell_{j,t} \in [0,1]$. As for the rest of the framework, the loss function is given by :

The ML aggregation rule

Let's consider that every expert j incurs a loss $\ell_{j,t} \in [0,1]$. As for the rest of the framework, the loss function is given by :

$$\ell(\hat{y}_t) = \ell(\sum_{j=0}^{N} p_{j,t} f_{j,t}) = \mathbf{p_t}^T \tilde{\ell}_t = \sum_{j=1}^{N} p_{k,t} \ell_{k,t}$$

The ML aggregation rule

Let's consider that every expert j incurs a loss $\ell_{j,t} \in [0,1]$. As for the rest of the framework, the loss function is given by :

$$\ell(\hat{y}_t) = \ell(\sum_{j=0}^{N} p_{j,t} f_{j,t}) = \mathbf{p_t}^T \tilde{\ell}_t = \sum_{j=1}^{N} p_{k,t} \ell_{k,t}$$

where $\mathbf{p_t} = (p_{1,t}, ..., p_{N,t})$ and $\tilde{\ell}_t = (\ell_{1,t}, ..., \ell_{N,t})$

The ML aggregation rule

We define the $\mathcal{M}_{\eta}^{grad}$ multiple learning rates aggregation rule strategy :

The ML aggregation rule

We define the $\mathcal{M}_{\eta}^{grad}$ multiple learning rates aggregation rule strategy :

- *Parameter :* a vector $\eta = (\eta_1, ..., \eta_N)$ of learning rates.

The ML aggregation rule

We define the $\mathcal{M}_{\eta}^{grad}$ multiple learning rates aggregation rule strategy :

- *Parameter :* a vector $\eta = (\eta_1, ..., \eta_N)$ of learning rates.
- *Initialization :* a uniform vector $\mathbf{w_0}$

The ML aggregation rule

We define the $\mathcal{M}_{\eta}^{grad}$ multiple learning rates aggregation rule strategy :

- *Parameter :* a vector $\eta = (\eta_1, ..., \eta_N)$ of learning rates.
- *Initialization :* a uniform vector $\mathbf{w_0}$

For each round $t = 1, 2, ..., T$

The ML aggregation rule

We define the $\mathcal{M}_{\eta}^{grad}$ multiple learning rates aggregation rule strategy :

- *Parameter :* a vector $\eta = (\eta_1, ..., \eta_N)$ of learning rates.
- *Initialization :* a uniform vector $\mathbf{w_0}$

For each round $t = 1, 2, ..., T$

1. form the mixture p_t defined component-wise by $p_{j,t} = \frac{\eta_j w_{j,t}}{\eta' \mathbf{w_{t-1}}}$

The ML aggregation rule

We define the $\mathcal{M}_{\eta}^{grad}$ multiple learning rates aggregation rule strategy :

- *Parameter :* a vector $\eta = (\eta_1, ..., \eta_N)$ of learning rates.
- *Initialization :* a uniform vector $\mathbf{w_0}$

For each round $t = 1, 2, ..., T$

1. form the mixture p_t defined component-wise by $p_{j,t} = \frac{\eta_j w_{j,t}}{\eta' \mathbf{w_{t-1}}}$
2. observe the loss vector ℓ_t and incur loss $\ell_t = \mathbf{p_t}' \tilde{\ell}_t$.

The ML aggregation rule

We define the $\mathcal{M}_{\eta}^{grad}$ multiple learning rates aggregation rule strategy :

- *Parameter :* a vector $\eta = (\eta_1, ..., \eta_N)$ of learning rates.
- *Initialization :* a uniform vector $\mathbf{w_0}$

For each round $t = 1, 2, ..., T$

1. form the mixture p_t defined component-wise by $p_{j,t} = \frac{\eta_j w_{j,t}}{\eta' \mathbf{w_{t-1}}}$
2. observe the loss vector ℓ_t and incur loss $\ell_t = \mathbf{p_t}' \tilde{\ell}_t$.
3. for each expert j perform the update $w_{j,t} = w_{j,t-1}(1 + \eta_j(\ell_t - \ell_{j,t}))$

Multiple learning rates

Theorem

Theorem 3 [Gaillard, Stoltz Erven, 2014]

Multiple learning rates

Theorem

Theorem 3 [Gaillard, Stoltz Erven, 2014]

For all sequences of loss vectors $\ell_t \in [0,1]^N$, the cumulative loss run with learning rates $\eta_k \leq \frac{1}{2}$ is bounded by

Multiple learning rates

Theorem

Theorem 3 [Gaillard, Stoltz Erven, 2014]

For all sequences of loss vectors $\ell_t \in [0,1]^N$, the cumulative loss run with learning rates $\eta_k \leq \frac{1}{2}$ is bounded by

$$L_T(\mathcal{M}_\eta^{grad}) = \sum_{t=1}^{T} \ell_t \leq \min_{1\leq j\leq N} \left\{ \sum_{t=1}^{T} \ell_{j,t} + \frac{1}{\eta_j} ln(\frac{1}{w_{j,0}}) + \eta_k \sum_{t=1}^{T} (\ell_t - \ell_{j,t}) \right\}$$

Multiple learning rates

- The vector η is theoretically calibrated. :

$$\eta_{j,t-1} = \frac{1}{1 + \sum_{s=1}^{t-1} (\hat{\ell}_s - \ell_{j,s})^2}$$

Multiple learning rates

- The vector η is theoretically calibrated. :

$$\eta_{j,t-1} = \frac{1}{1 + \sum_{s=1}^{t-1} (\hat{\ell}_s - \ell_{j,s})^2}$$

- The loss function of the ML aggregation rule is bounded. Nevertheless, for the moment, the ML aggregation rule does not theoretically compete with the best convex combination of predictors.

Multiple learning rates

- The vector η is theoretically calibrated. :

$$\eta_{j,t-1} = \frac{1}{1 + \sum_{s=1}^{t-1} (\hat{\ell}_s - \ell_{j,s})^2}$$

- The loss function of the ML aggregation rule is bounded. Nevertheless, for the moment, the ML aggregation rule does not theoretically compete with the best convex combination of predictors.

Now we can relax the convex weights vector assumption. $\Rightarrow$ *Next*
$\Leftarrow$ *Contents*

Aggregation rules

2 kinds of aggregation rule :

Aggregation rules

2 kinds of aggregation rule :

- **Follow-The-Leader** : the forecaster chooses the convex weight vector p_t to minimize the cumulative loss :

$$p_t = \arg\min_{p \in \mathcal{P}} \widehat{L_{t-1}}(\mathcal{S}) = \arg\min_{p \in \mathcal{P}} \sum_{i=1}^{t-1} \widehat{\ell_i}(\sum_{j=0}^{N} p_{j,t} f_{j,t})$$

Aggregation rules

2 kinds of aggregation rule :

- **Follow-The-Leader** : the forecaster chooses the convex weight vector p_t to minimize the cumulative loss :

$$p_t = \arg\min_{p \in \mathcal{P}} \widehat{L_{t-1}}(\mathcal{S}) = \arg\min_{p \in \mathcal{P}} \sum_{i=1}^{t-1} \widehat{\ell_i}(\sum_{j=0}^{N} p_{j,t} f_{j,t})$$

- **Follow-The-Regularized leader** : the forecaster chooses the weight vector p_t to minimize the cumulative loss plus a regularization term :

Aggregation rules

2 kinds of aggregation rule :

- **Follow-The-Leader** : the forecaster chooses the convex weight vector p_t to minimize the cumulative loss :

$$p_t = \arg\min_{p \in \mathcal{P}} \widehat{L_{t-1}}(\mathcal{S}) = \arg\min_{p \in \mathcal{P}} \sum_{i=1}^{t-1} \widehat{\ell}_i(\sum_{j=0}^{N} p_{j,t} f_{j,t})$$

- **Follow-The-Regularized leader** : the forecaster chooses the weight vector p_t to minimize the cumulative loss plus a regularization term :

$$p_t = \arg\min_{p \in \mathcal{P}} \widehat{L_{t-1}}(\mathcal{S}) + R(p_t) = \arg\min_{p \in \mathcal{P}} \sum_{i=1}^{t-1} \widehat{\ell}_i(\sum_{j=0}^{N} p_{j,t} f_{j,t}) + R(p_t)$$

Aggregation rules

2 kinds of aggregation rule :

- **Follow-The-Leader** : the forecaster chooses the convex weight vector p_t to minimize the cumulative loss :

$$p_t = \arg\min_{p\in\mathcal{P}} \widehat{L_{t-1}}(\mathcal{S}) = \arg\min_{p\in\mathcal{P}} \sum_{i=1}^{t-1} \widehat{\ell}_i(\sum_{j=0}^{N} p_{j,t} f_{j,t})$$

- **Follow-The-Regularized leader** : the forecaster chooses the weight vector p_t to minimize the cumulative loss plus a regularization term :

$$p_t = \arg\min_{p\in\mathcal{P}} \widehat{L_{t-1}}(\mathcal{S}) + R(p_t) = \arg\min_{p\in\mathcal{P}} \sum_{i=1}^{t-1} \widehat{\ell}_i(\sum_{j=0}^{N} p_{j,t} f_{j,t}) + R(p_t)$$

In the last case, $R(p_t)$ can be :

- a linear/convex function : **Online Gradient Descent.**

Aggregation rules

2 kinds of aggregation rule :

- **Follow-The-Leader** : the forecaster chooses the convex weight vector p_t to minimize the cumulative loss :

$$p_t = \arg\min_{p\in\mathcal{P}} \widehat{L_{t-1}}(\mathcal{S}) = \arg\min_{p\in\mathcal{P}} \sum_{i=1}^{t-1} \widehat{\ell}_i(\sum_{j=0}^{N} p_{j,t} f_{j,t})$$

- **Follow-The-Regularized leader** : the forecaster chooses the weight vector p_t to minimize the cumulative loss plus a regularization term :

$$p_t = \arg\min_{p\in\mathcal{P}} \widehat{L_{t-1}}(\mathcal{S}) + R(p_t) = \arg\min_{p\in\mathcal{P}} \sum_{i=1}^{t-1} \widehat{\ell}_i(\sum_{j=0}^{N} p_{j,t} f_{j,t}) + R(p_t)$$

In the last case, $R(p_t)$ can be :

- a linear/convex function : **Online Gradient Descent.**

Aggregation rules

2 kinds of aggregation rule :

- **Follow-The-Leader** : the forecaster chooses the convex weight vector p_t to minimize the cumulative loss :

$$p_t = \arg\min_{p \in \mathcal{P}} \widehat{L_{t-1}}(\mathcal{S}) = \arg\min_{p \in \mathcal{P}} \sum_{i=1}^{t-1} \widehat{\ell}_i(\sum_{j=0}^{N} p_{j,t} f_{j,t})$$

- **Follow-The-Regularized leader** : the forecaster chooses the weight vector p_t to minimize the cumulative loss plus a regularization term :

$$p_t = \arg\min_{p \in \mathcal{P}} \widehat{L_{t-1}}(\mathcal{S}) + R(p_t) = \arg\min_{p \in \mathcal{P}} \sum_{i=1}^{t-1} \widehat{\ell}_i(\sum_{j=0}^{N} p_{j,t} f_{j,t}) + R(p_t)$$

In the last case, $R(p_t)$ can be :

- a linear/convex function : **Online Gradient Descent**.
- a square-ℓ_2-norm regularization : **Ridge regression**.

Online Gradient Descent aggregation rule - General Framework

We define the $\mathcal{O}_\eta$ OGD aggregation rule :

Online Gradient Descent aggregation rule - General Framework

We define the $\mathcal{O}_\eta$ OGD aggregation rule :

Online Gradient Descent aggregation rule - General Framework

We define the $\mathcal{O}_\eta$ OGD aggregation rule :

- *Parameter :* a learning rate η_t.

Online Gradient Descent aggregation rule - General Framework

We define the $\mathcal{O}_\eta$ OGD aggregation rule :

- *Parameter :* a learning rate η_t.
- *Initialization :* an arbitrary vector p_1.

Online Gradient Descent aggregation rule - General Framework

We define the $\mathcal{O}_\eta$ OGD aggregation rule :

- *Parameter :* a learning rate η_t.
- *Initialization :* an arbitrary vector p_1.

For each round $t = 1, 2, ..., T$, the vector p_{t+1} is selected according to :

Online Gradient Descent aggregation rule - General Framework

We define the $\mathcal{O}_\eta$ OGD aggregation rule :

- *Parameter :* a learning rate η_t.
- *Initialization :* an arbitrary vector p_1.

For each round $t = 1, 2, ..., T$, the vector p_{t+1} is selected according to :

$$p_{j,t+1} = P_j(p_{j,t} - \eta_t \partial \ell(\sum_{j=1}^{N} p_{j,t} f_{j,t}, y_t))$$

Online Gradient Descent aggregation rule - General Framework

We define the $\mathcal{O}_\eta$ OGD aggregation rule :

- *Parameter :* a learning rate η_t.
- *Initialization :* an arbitrary vector p_1.

For each round $t = 1, 2, ..., T$, the vector p_{t+1} is selected according to :

$$p_{j,t+1} = P_j(p_{j,t} - \eta_t \partial \ell(\sum_{j=1}^{N} p_{j,t} f_{j,t}, y_t))$$

where $P_j = \arg\min_{p_j} d(p, y) = \arg\min_{p_j} || \sum_{j=1}^{N} p_{j,t} f_{j,t} - y_t||$

The Online Gradient Descent aggregation rule

Theorem

Theorem 2 [Zinkevich, 2003]

The Online Gradient Descent aggregation rule

Theorem

Theorem 2 [Zinkevich, 2003]

If $\eta_t = t^{\frac{1}{2}}$, the regret is :

The Online Gradient Descent aggregation rule

Theorem

Theorem 2 [Zinkevich, 2003]

If $\eta_t = t^{\frac{1}{2}}$, the regret is :

$$R(\mathcal{O}_\eta^{grad}) \leq \frac{\sqrt{T}}{2} + 4(\sqrt{T} - \frac{1}{2})$$

The Online Gradient Descent aggregation rule

The Online Gradient Descent aggregation rule

The bound of the regret is obtained using two facts :

The Online Gradient Descent aggregation rule

The bound of the regret is obtained using two facts :

- $\max |\hat{y}_t - y_t| = 1$

The Online Gradient Descent aggregation rule

The bound of the regret is obtained using two facts :

- $\max |\hat{y}_t - y_t| = 1$
- $\max \frac{\partial \ell(\hat{y}_t)}{\partial \hat{y}_t} = \max 2(\hat{y}_t - y_t) = 2$

$\Rightarrow Next$
$\Leftarrow Contents$

The ridge regression aggregation rule

In this framework, weights $u_t = (u_{1,t}, ..., u_{N,t}) \in \mathbb{R}$ are not choosen in a simplex $\mathcal{P}$ anymore. The forecaster prediction is defined by :

The ridge regression aggregation rule

In this framework, weights $u_t = (u_{1,t}, ..., u_{N,t}) \in \mathbb{R}$ are not choosen in a simplex $\mathcal{P}$ anymore. The forecaster prediction is defined by :

$$\tilde{y}_t = \sum_{j=1}^{N} u_{j,t} f_{j,t}$$

The ridge regression aggregation rule

In this framework, weights $u_t = (u_{1,t}, ..., u_{N,t}) \in \mathbb{R}$ are not choosen in a simplex $\mathcal{P}$ anymore. The forecaster prediction is defined by :

$$\tilde{y}_t = \sum_{j=1}^{N} u_{j,t} f_{j,t}$$

The ridge regression aggregation rule

We define by :

The ridge regression aggregation rule

We define by :

- $f_t = (f_{1,t}, ..., f_{N,t})$ the vector of forecaster predictions.

The ridge regression aggregation rule

We define by :

- $f_t = (f_{1,t}, ..., f_{N,t})$ the vector of forecaster predictions.
- the euclidian norm of a vector $u \in \mathbb{R}$:

The ridge regression aggregation rule

We define by :

- $f_t = (f_{1,t}, ..., f_{N,t})$ the vector of forecaster predictions.
- the euclidian norm of a vector $u \in \mathbb{R}$:

$$||u_2|| = \sqrt{\sum_{j=1} u_j^2}$$

The rest of the framework does not change.

The ridge regression aggregation rule

The ridge regression aggregation rule

We choose u_t as follows :

The ridge regression aggregation rule

We choose u_t as follows :

$$u_t \in \arg\min_{v \in \mathbb{R}^N} \left\{ \lambda ||v||_2^2 + \sum_{s=1}^{t=1} (y_s - \sum_{j=1}^{N} v_j f_{j,s})^2 \right\}$$

The ridge regression aggregation rule

We choose u_t as follows :

$$u_t \in \arg\min_{v \in \mathbb{R}^N} \left\{ \lambda ||v||_2^2 + \sum_{s=1}^{t=1} (y_s - \sum_{j=1}^{N} v_j f_{j,s})^2 \right\}$$

An explicit solution is given by :

$$u_t = (\lambda I_N + M_{t-1})^{-1} \sum_{s=1}^{t-1} y_s f_s$$

Theorem 4 [Vovk, 2001]

Ridge Regression

Theorem

Theorem 4 [Vovk, 2001]

Since $\hat{y}_t \in [0, 1]$:

Ridge Regression

Theorem

Theorem 4 [Vovk, 2001]

Since $\hat{y}_t \in [0, 1]$:

$$R(\mathcal{R}_\eta) \leq \inf_{v \in \mathbb{R}^N} \left\{ \lambda ||v_2^2|| \right\} + N \times ln(1 + \frac{T}{\lambda N})$$

Theorem 4 [Vovk, 2001]

Since $\hat{y}_t \in [0, 1]$:

$$R(\mathcal{R}_\eta) \leq \inf_{v \in \mathbb{R}^N} \left\{ \lambda ||v_2^2|| \right\} + N \times ln(1 + \frac{T}{\lambda N})$$

Since N is very large in our case, this online aggregation rule should have a great performance.

Tuning

Tuning

As for η, we calibrate λ with a grid such as :

$$\lambda \in \arg\min_{\eta>0} \widehat{L_{t-1}}(\mathcal{R}_\lambda)$$

Tuning

As for η, we calibrate λ with a grid such as :

$$\lambda \in \arg\min_{\eta>0} \widehat{L_{t-1}}(\mathcal{R}_\lambda)$$

$\Rightarrow Next$
$\Leftarrow Contents$