THE MEMORY OF EXTRAPOLATING P-SPLINES

ALBA CARBALLO[1], MARÍA DURBÁN[1], DAE-JIN LEE[2] AND PAUL EILERS[3]

[1] Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain
[2] BCAM - Basque Center for Applied Mathematics, Bilbao, Spain
[3] Department of Biostatistics, Erasmus University Medical Center, Rotterdam, The Netherlands

INTRODUCTION: FORECASTING WITH P-SPLINES

Smooth models in which forecasting with smoothing models is needed:

- Hourly temperatures at a weather station.
- Yearly number of deaths.

It may be important to know how much of the past information we are using to forecast. We introduce a concept as a tool to provide that information: **memory of a P-Spline**.

Consider the case of a univariate Gaussian data, with ordered regressor $\boldsymbol{x}$ and response variable $\boldsymbol{y}$.

Smooth model:

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon},$$

$f(\cdot)$ unknown smooth function.

Given n observations $\boldsymbol{y}$ of the response variable, we can predict new values $\boldsymbol{y}_p$ by fitting and forecasting simultaneously:

$$\hat{\boldsymbol{y}}_+ = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{M}\boldsymbol{B} + \lambda\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\boldsymbol{M}\boldsymbol{y}_+ = \boldsymbol{H}_+\boldsymbol{y}_+,$$

where:

- $\boldsymbol{B}$: B-spline basis built from a set of knots which range covers all values of the extended explicative variable.
- $\boldsymbol{M}$: diagonal weight matrix with diagonal elements equal to 1 if the data is observed and extra zeros if the data is forecasted.
- $\boldsymbol{y}_+$: extended response variable.
- $\lambda\boldsymbol{D}'\boldsymbol{D}$: penalty matrix, with λ the smoothing parameter and $\boldsymbol{D}$ a difference matrix of order q .

The last columns of $\boldsymbol{H}_+$ are all zeros (as the corresponding diagonal elements of $\boldsymbol{M}$ are zero):

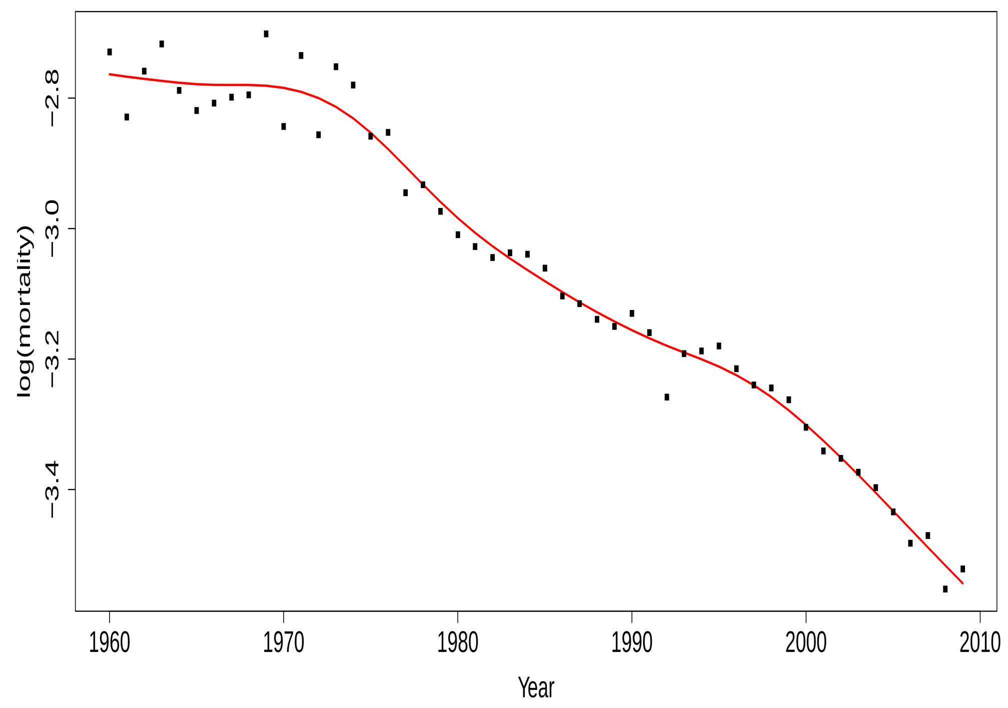
$$\hat{\boldsymbol{y}}_+ = \begin{bmatrix} \boldsymbol{H} & \boldsymbol{O} \\ \boldsymbol{H}_p & \boldsymbol{O} \end{bmatrix} \boldsymbol{y}_+,$$

Therefore, $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_p = \boldsymbol{H}_p\boldsymbol{y}$.

The data

Log mortality rates of Spanish men aged 73 between 1960 and 2009 (50 observations).

Fitted curve with P-splines method Lambda=23.63



DEFINITION: MEMORY OF A P-SPLINE

The predicted values are:

$$\hat{\boldsymbol{y}}_p = \boldsymbol{H}_p\boldsymbol{y},$$

summarizing the rows and columns of $\boldsymbol{H}_p$ we find how the past is affecting the forecast. We have noticed:

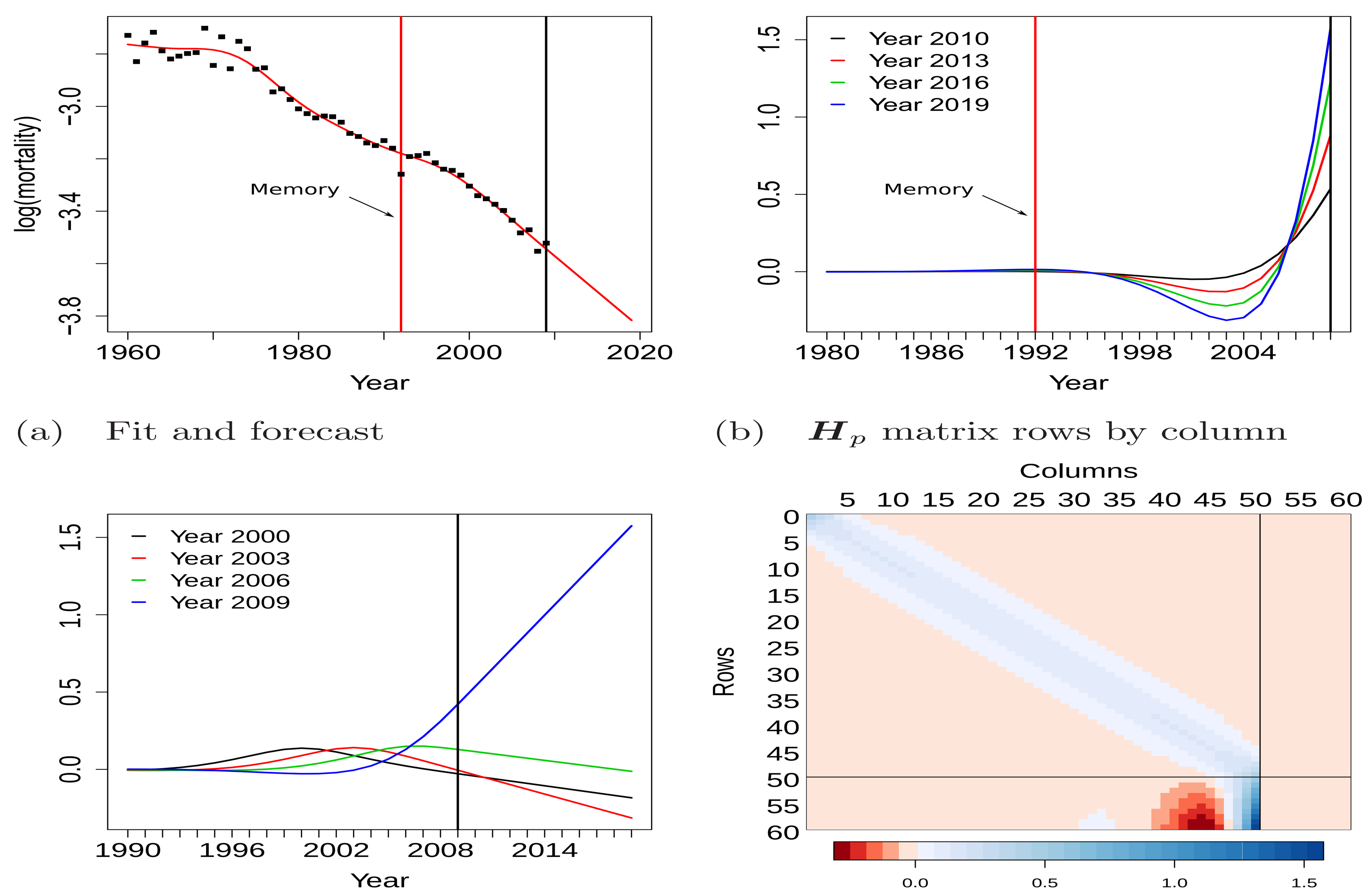
- All rows of $\boldsymbol{H}_p$ follow a similar pattern, see panel (b).
- Each row of $\boldsymbol{H}_p$ gives the contributions of past years in each future value.
- The contribution of each observation in the past decays gradually as we move away from the present.
- Each column of $\boldsymbol{H}_p$ gives the contribution of each observation of the past in the future values. These contributions are a polynomial function of time (of order $d-1$, where d is the order of the penalty), see panel (c) (where $d = 2$).

The sum of the absolute value of the columns of $\boldsymbol{H}_p$, standardized by their sum, can be considered as a distribution $\mathcal{W}$.

The **memory of the P-spline** is the 99th percentile of the $\mathcal{W}$ distribution.

ILLUSTRATION

Forecast up to 2019, 10 new observations. Memory = 18. Lambda = 23.63



(a) Fit and forecast

(b) $\boldsymbol{H}_p$ matrix rows by column

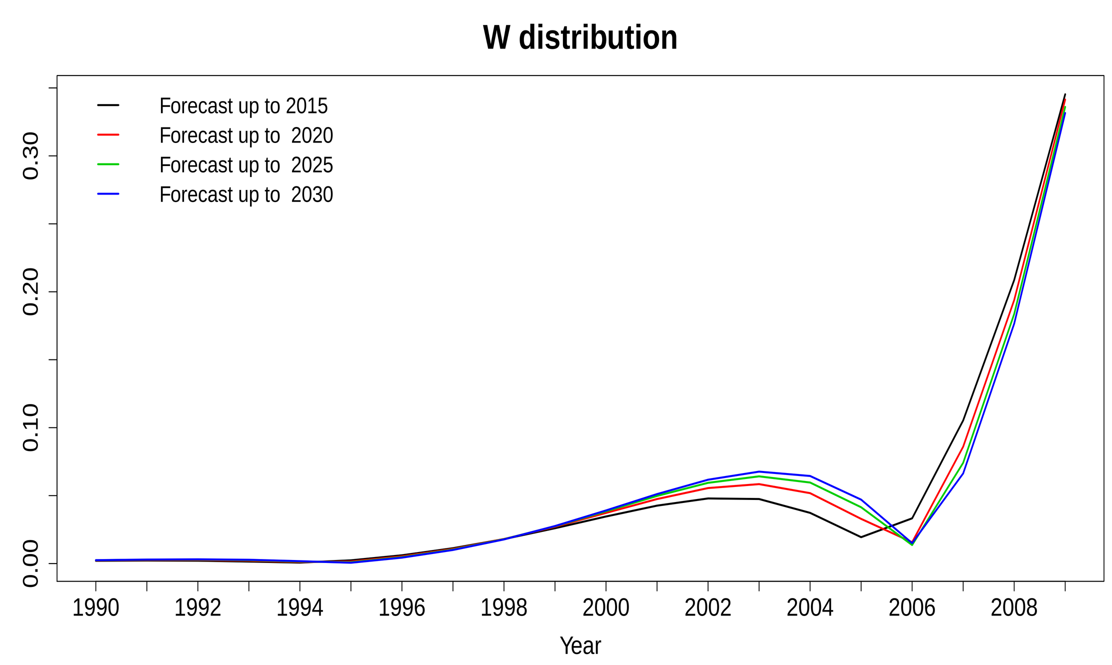
(c) $\boldsymbol{H}_+$ matrix columns by row

(d) $\boldsymbol{H}_+$ matrix

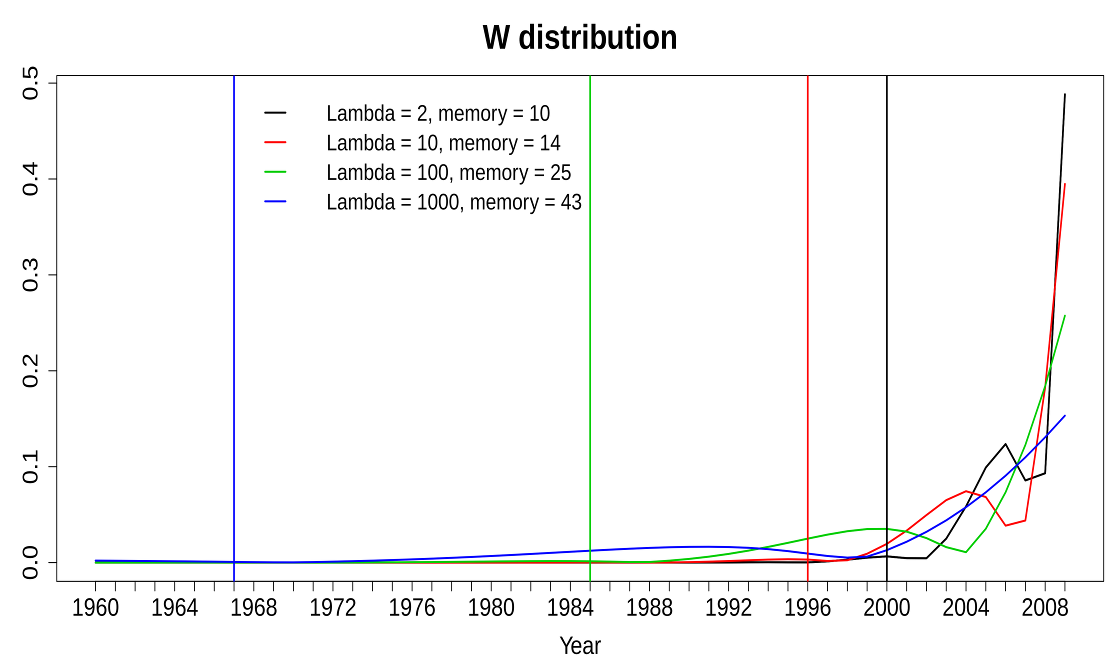
The **memory of the P-spline** is 18. What has happened more than 18 years backward, before 1992, has no influence on the future.

PROPERTIES OF THE MEMORY OF A P-SPLINE

1.- The memory does not depend on the prediction horizon.



2.- For a given basis, the memory depends on the smoothing parameter. The smaller (larger) the smoothing parameter is, the smaller (greater) the influence of the past on the predicted values.



REFERENCES

[1] CURRIE, I. D., DURBÁN, M., AND EILERS, P. H. C. (2004). *Smoothing and forecasting mortality rates.* Statistical Modelling 4(4):279–298.

[2] EILERS, P. H. C. AND MARX, B. D. (1996). *Flexible Smoothing with B-Splines and Penalties.* Statistical Science, 11:89–121.

ACKNOWLEDGEMENTS

Research supported by the Spanish Ministry of Economy and Competitiveness grants MTM2014-52184 and SEV-2013-0323. The research of Dae-Jin Lee was also supported by the Basque Government through the BERC 2014-2017 program.