Revisiting Whittaker-Henderson Smoothing

Guillaume Biessy

Revisiting Whittaker-Henderson Smoothing

Guillaume Biessy*, PhD, LinkPact† and Sorbonne Université‡

September 16, 2023

Introduced nearly a century ago, Whittaker-Henderson smoothing remains one of the most commonly used methods by actuaries for constructing one-dimensional and two-dimensional experience tables for mortality and other Life Insurance risks. This paper proposes to reframe this smoothing technique within a modern statistical framework and addresses six questions of practical interest regarding its use.

Firstly, we adopt a Bayesian view of this smoothing method to build credible intervals. Next, we shed light on the choice of observation vectors and weights to which the smoothing should be applied by linking it to a maximum likelihood estimator introduced in the context of duration models. We then enhance the precision of the smoothing by relaxing an implicit asymptotic approximation on which it relies. Afterward, we select the smoothing parameters based on maximizing a marginal likelihood. We later improve numerical performance in the presence of a large number of observation points and, consequently, parameters. Finally, we extrapolate the results of the smoothing while preserving consistency between estimated and predicted values through the use of constraints.

Keywords: smoothing methods, duration models, experience tables, maximum likelihood, generalized additive models, empirical Bayesian approach, marginal likelihood, extrapolation.

1 Introduction

1.1 Notations

In this paper, vector names are written in bold characters and matrix names in uppercase letters. If $\mathbf{y}$ is a vector and A is a matrix, $\mathrm{Var}(\mathbf{y})$ denotes the variance-covariance matrix

*guillaume.biessy78@gmail.com
†LinkPact, 80 boulevard Pasteur, 75015 Paris, France
‡Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, 75005 Paris, France

associated with $\mathbf{y}$, $\mathbf{diag}(A)$ represents the diagonal of matrix A, and $\text{Diag}(\mathbf{y})$ is the diagonal matrix such that $\mathbf{diag}(\text{Diag}(\mathbf{y})) = \mathbf{y}$. The sum of the diagonal elements of A is denoted as $\text{tr}(A)$ and its transpose as A^T. In the case where A is invertible, A^{-1} denotes its inverse, A^{-T} is the inverse of its transpose, and $|A|$ is the product of the eigenvalues of A. For a non-invertible matrix A, A^- refers to the Moore-Penrose pseudoinverse of A, and $|A|_+$ denotes the product of the non-zero eigenvalues of A. By writing the eigendecomposition as $A = U\Sigma V^T$, where U and V are two orthogonal matrices and Σ is a diagonal matrix containing the eigenvalues of A, and by denoting Σ^- as the matrix obtained by replacing the non-zero eigenvalues in Σ with their inverses while keeping the zero eigenvalues unchanged, the pseudoinverse is given by $A^- = V\Sigma^- U^T$. The Kronecker product of two matrices A and B is denoted as $A \otimes B$, and their Hadamard product, or element-wise product, is denoted as $A \odot B$. $\mathbf{vec}(A)$ represents the vector obtained by stacking the columns of matrix A together. Finally, the symbol $\propto$ signifies a proportional relationship between the expressions on both sides of it.

1.2 Origin

Whittaker-Henderson smoothing is a gradation method aimed at correcting the effect of sampling fluctuations on an observation vector. It is applied to evenly-spaced discrete observations. Initially proposed by Whittaker (1922) for constructing mortality tables and further developed by the works of Henderson (1924), it remains one of the most popular methods among actuaries for constructing experience tables in life insurance. Extending to two-dimensional tables, it can be used for studying various risks, including but not limited to mortality, disability, long-term care, lapse, mortgage default, and unemployment.

1.3 The one-dimensional case

Let $\mathbf{y}$ be a vector of observations and $\mathbf{w}$ a vector of positive weights, both of size n. The estimator associated with Whittaker-Henderson smoothing is given by:

$$\hat{\mathbf{y}} = \underset{\boldsymbol{\theta}}{\text{argmin}}\{F(\mathbf{y}, \mathbf{w}, \boldsymbol{\theta}) + R_{\lambda,q}(\boldsymbol{\theta})\} \tag{1}$$

where:

- $F(\mathbf{y}, \mathbf{w}, \boldsymbol{\theta}) = \sum_{i=1}^{n} w_i(y_i - \theta_i)^2$ represents a fidelity criterion to the observations,
- $R_{\lambda,q}(\boldsymbol{\theta}) = \lambda \sum_{i=1}^{n-q} (\Delta^q \boldsymbol{\theta})_i^2$ represents a smoothness criterion.

In the latter expression, Δ^q denotes the forward difference operator of order q, such that for any $i \in [1, n-q]$:

$$(\Delta^q \boldsymbol{\theta})_i = \sum_{k=0}^{q} \binom{q}{k} (-1)^{q-k} \theta_{i+k}.$$

Let us define $W = \text{Diag}(\mathbf{w})$, the diagonal matrix of weights, and $D_{n,q}$ as the order q difference matrix of dimensions $(n-q) \times n$, such that $(D_{n,q}\boldsymbol{\theta})_i = (\Delta^q \boldsymbol{\theta})_i$ for all $i \in [1, n-q]$. The most commonly used difference matrices of order 1 and 2 have the following forms:

$$D_{n,1} = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix} \quad \text{and} \quad D_{n,2} = \begin{bmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & -2 & 1 \end{bmatrix}.$$

The fidelity and smoothness criteria can be rewritten in matrix form as:

$$F(\mathbf{y}, \mathbf{w}, \boldsymbol{\theta}) = (\mathbf{y} - \boldsymbol{\theta})^T W (\mathbf{y} - \boldsymbol{\theta}) \quad \text{and} \quad R_{\lambda,q}(\boldsymbol{\theta}) = \lambda \boldsymbol{\theta}^T D_{n,q}^T D_{n,q} \boldsymbol{\theta}.$$

The associated estimator for smoothing becomes:

$$\hat{\mathbf{y}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \left\{ (\mathbf{y} - \boldsymbol{\theta})^T W (\mathbf{y} - \boldsymbol{\theta}) + \boldsymbol{\theta}^T P_\lambda \boldsymbol{\theta} \right\} \tag{2}$$

where $P_\lambda = \lambda D_{n,q}^T D_{n,q}$.

1.4 The two-dimensional case

In the two-dimensional case, let us consider a matrix Y of observations and a matrix Ω of non-negative weights, both of dimensions $n_x \times n_z$. The estimator associated with the Whittaker-Henderson smoothing can be written as:

$$\widehat{Y} = \underset{\Theta}{\text{argmin}} \{ F(Y, \Omega, \Theta) + R_{\lambda,q}(\Theta) \}$$

where:

- $F(Y, \Omega, \Theta) = \sum_{i=1}^{n_x} \sum_{j=1}^{n_z} \Omega_{i,j} (Y_{i,j} - \Theta_{i,j})^2$ represents a fidelity criterion to the observations,
- $R_{\lambda,q}(\Theta) = \lambda_x \sum_{j=1}^{n_z} \sum_{i=1}^{n_x - q_x} (\Delta^{q_x} \Theta_{\bullet,j})_i^2 + \lambda_z \sum_{i=1}^{n_x} \sum_{j=1}^{n_z - q_z} (\Delta^{q_z} \Theta_{i,\bullet})_j^2$ is a smoothness criterion.

This latter criterion can be written as the sum of two one-dimensional regularization criteria, with orders q_x and q_z, applied respectively to all rows and all columns of Θ. It is also possible to adopt matrix notations by defining $\mathbf{y} = \mathbf{vec}(Y)$, $\mathbf{w} = \mathbf{vec}(\Omega)$, and $\boldsymbol{\theta} = \mathbf{vec}(\Theta)$ as the vectors obtained by stacking the columns of the matrices Y, Ω, and Θ, respectively. Additionally, let us denote $W = \mathrm{Diag}(\mathbf{w})$ and $n = n_x \times n_z$. The fidelity and smoothness criteria can then be rewritten as linear combinations of the vectors $\mathbf{y}$, $\mathbf{w}$, and $\boldsymbol{\theta}$:

$$\begin{aligned} F(\mathbf{y}, \mathbf{w}, \boldsymbol{\theta}) &= (\mathbf{y} - \boldsymbol{\theta})^T W (\mathbf{y} - \boldsymbol{\theta}) \\ R_{\lambda,q}(\boldsymbol{\theta}) &= \boldsymbol{\theta}^T (\lambda_x I_{n_z} \otimes D_{n_x,q_x}^T D_{n_x,q_x} + \lambda_z D_{n_z,q_z}^T D_{n_z,q_z} \otimes I_{n_x}) \boldsymbol{\theta}. \end{aligned}$$

and the estimator associated with the smoothing takes the form:

$$\hat{\mathbf{y}} = \underset{\boldsymbol{\theta}}{\mathrm{argmin}} \left\{ (\mathbf{y} - \boldsymbol{\theta})^T W (\mathbf{y} - \boldsymbol{\theta}) + \boldsymbol{\theta}^T P_\lambda \boldsymbol{\theta} \right\}$$

of Equation 2 except in this case $P_\lambda = \lambda_x I_{n_z} \otimes D_{n_x,q_x}^T D_{n_x,q_x} + \lambda_z D_{n_z,q_z}^T D_{n_z,q_z} \otimes I_{n_x}$.

1.5 Explicit solution

Equation 2 has an explicit solution:

$$\hat{\mathbf{y}} = (W + P_\lambda)^{-1} W \mathbf{y}. \tag{3}$$

Indeed, as a minimum, $\hat{\mathbf{y}}$ satisfies:

$$0 = \frac{\partial}{\partial \boldsymbol{\theta}}\Big|_{\hat{\mathbf{y}}} \left\{ (\mathbf{y} - \boldsymbol{\theta})^T W (\mathbf{y} - \boldsymbol{\theta}) + \boldsymbol{\theta}^T P_\lambda \boldsymbol{\theta} \right\} = -2X^T W (y - \hat{\mathbf{y}}) + 2X^T P_\lambda \hat{\mathbf{y}}.$$

It follows that $(W + P_\lambda)\hat{\mathbf{y}} = W\mathbf{y}$, and if $W + P_\lambda$ is invertible, then $\hat{\mathbf{y}}$ is indeed a solution of Equation 3. When $\lambda \neq 0$, $W + P_\lambda$ is invertible as long as $\mathbf{w}$ has q non-zero elements in the one-dimensional case, and Ω has at least $q_x \times q_z$ non-zero elements distributed over q_x different rows and q_z different columns in the two-dimensional case. These sufficient conditions are always satisfied in practice and will not be demonstrated here.

1.6 Impact of the smoothing parameter(s)

In Equation 1 $(\mathbf{y}-\boldsymbol{\theta})^T W(\mathbf{y}-\boldsymbol{\theta})$ represents a fidelity criterion to the observations and $\boldsymbol{\theta}^T P_\lambda \boldsymbol{\theta}$ represents a smoothness criterion. The relative importance of these criteria is controlled by the parameter (or pair of parameters in the two-dimensional case) λ. Figure 1 provides an illustration of one-dimensional smoothing on mortality data using three different values of the smoothing parameter. The effective degrees of freedom shown in this figure can be calculated by summing the diagonal values of H, the *hat matrix* of the model. The hat matrix satisfies $\hat{\mathbf{y}} = H\mathbf{y}$ and can be identified in Equation 3 as $H = (W + P_\lambda)^{-1}W$. These effective degrees of freedom serve as a non-parametric equivalent of the number of independent parameters in parametric models but can take non-integer values. The concept of degrees of freedom will be discussed in more detail in Section 6 of the paper. It can be observed here that the result of smoothing is highly sensitive to the chosen value of the smoothing parameter. The choice of $\lambda = 10^1$ leads to a highly volatile fit that largely reproduces the sampling fluctuations present in the data. On the other hand, the choice of $\lambda = 10^7$ seems too rigid to capture the underlying pattern of the data. The choice of $\lambda = 10^4$ appears, at first glance, as a satisfactory compromise between these two extremes.

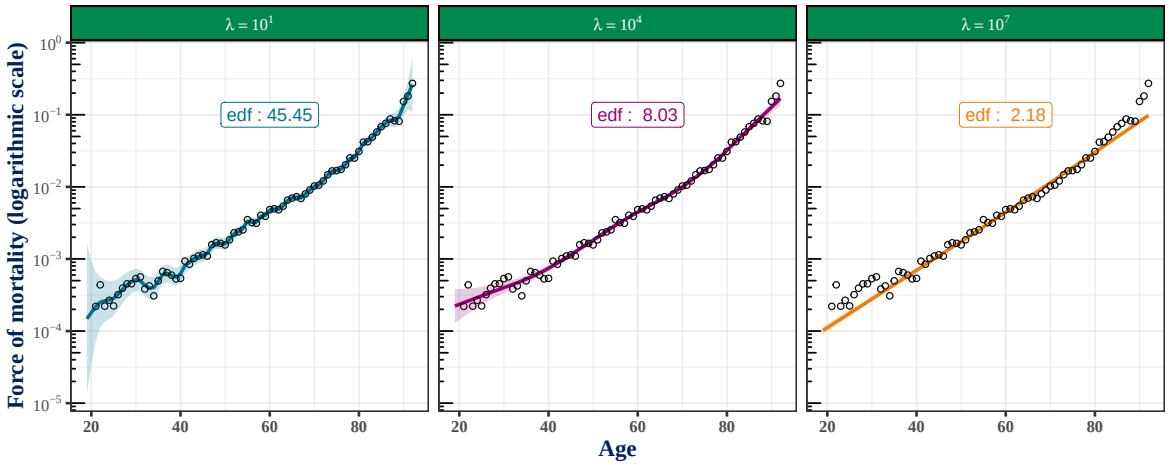

Figure 1: Whittaker-Henderson smoothing applied to a portfolio of synthetic mortality data for 3 choices of the smoothing parameter. The points represent the initial observations, the curves depict the result of the smoothing, and the shaded areas represent the associated credibility intervals. *edf* refers to the effective degrees of freedom of the model.

In the two-dimensional case, two penalties are applied to the model parameters. Figure 2 illustrates the application of smoothing to mortality data for a portfolio of insured individuals with dependence, based on the age of the insured individuals and their duration in the dependent state. Different combinations of the two smoothing parameters in this model are used to produce the 9 smoothers represented in this figure. While some combinations of smoothing parameters

may be deemed implausible, such as those associated with the choice $\lambda_x = 10^0$, it is challenging in the two-dimensional case to determine graphically which combination of parameters is most suitable. The selection of the smoothing parameter will be discussed in detail in Section 5.

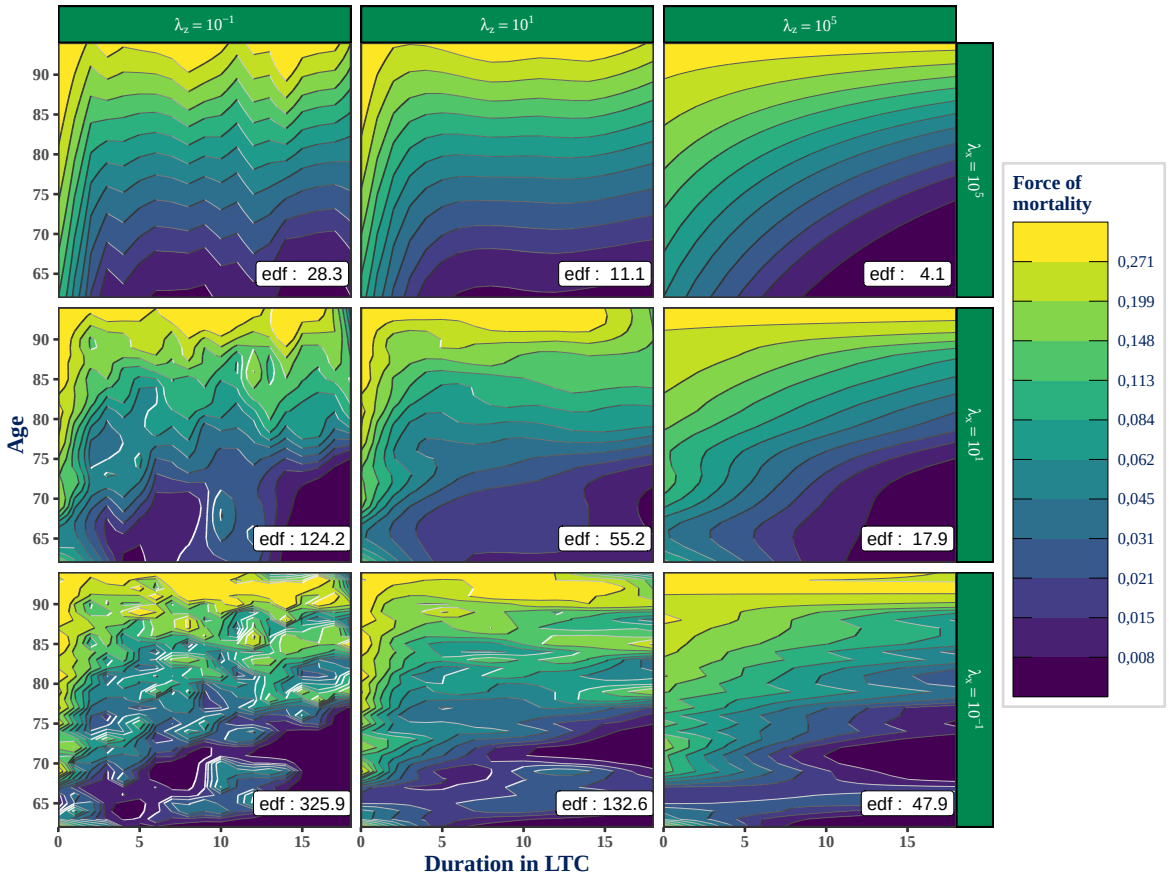

Figure 2: Whittaker-Henderson smoothing applied to a synthetic portfolio of policyholders in long-term care situation for 9 combinations of smoothing parameters. The contour lines and colors depict the surface of the smoothed mortality force as a function of the age of the policyholders and the duration in the long-term care state.

1.7 Plan for the paper

This paper aims to address 6 practical questions associated with the use of Whittaker-Henderson smoothing, each of which will be covered in a dedicated section.

1.7.1 How to measure uncertainty in smoothing results?

We propose to measure the uncertainty in smoothing results based on the available observation volume. In a frequentist approach, the estimator associated with smoothing appears to be biased, which makes it challenging to construct relevant confidence intervals for it. However, smoothing can be naturally interpreted in a Bayesian framework, allowing the construction of credible intervals.

1.7.2 Which observation and weight vectors to use?

Candidates are proposed for the vectors $\mathbf{y}$ and $\mathbf{w}$. By adopting the framework of duration models used for experience rating and considering the maximum likelihood estimator of crude rates under the assumption of piecewise constant transition intensity, we present a couple $(\mathbf{y}, \mathbf{w})$ that satisfies the implicit conditions associated with the use of smoothing in an asymptotic sense.

1.7.3 How to improve the accuracy of smoothing with limited data volume?

The previous theoretical framework is based on an asymptotic approximation, which, in many practical situations where the available data volume is limited, can be biased. We propose a generalization of Whittaker-Henderson smoothing based on exact likelihood rather than its normal approximation. This approach requires iterative solving of penalized likelihood equations. By comparing this exact approach with the normal approximation associated with the initial smoothing, we show that this approximation can sometimes come at the cost of substantial bias in the obtained results.

1.7.4 How to choose the smoothing parameter(s)?

We then address the choice of the smoothing parameter λ to use. This parameter controls the strength of the applied smoothing and has a crucial impact on the obtained results. To select this parameter, we propose maximizing a marginal likelihood using the optimization methods of Brent (1973) and John A. Nelder and Mead (1965), employed in the one-dimensional and two-dimensional case, respectively. These methods are implemented through the `optimize` and `optim` functions in the `stats` library of the statistical programming language R. In the context of the proposed generalization of smoothing, where solving likelihood equations and selecting the smoothing parameter are performed through nested iterations, we compare the precision and computational time of two known strategies called *outer iteration* and *performance iteration*. These strategies differ in the order in which the iterations are nested, and we demonstrate that the latter approach significantly reduces computation time at the cost of negligible loss of precision.

1.7.5 How to improve smoothing performance with a large number of data points?

We propose a practical solution to accelerate smoothing when the number of observations - and thus parameters - is substantial. This situation is common in practice, especially for two-dimensional tables used in the modeling of disability and long-term care risks. To address this issue, an eigendecomposition of the one-dimensional penalization matrices involved in smoothing is used. This decomposition provides a new interpretation of smoothing and opens the door to the use of a rank reduction method to reduce the number of parameters used and the computational time, while incurring minimal loss of precision.

1.7.6 How to extrapolate smoothing results?

Finally, we cover the problem of extrapolation in smoothing, which involves solving a modified version of the initial optimization problem where the fidelity criterion remains unchanged and the smoothness criterion is extended to extrapolated values. In the one-dimensional case, a solution may immediately be obtained. However, in the two-dimensional case, as shown by Carballo, Durban, and Lee (2021), constraints must be imposed on the model to prevent the extrapolation step from altering the results obtained during the initial use of the smoothing.

2 How to measure uncertainty in smoothing results?

Equation 3 indicates that $\mathbb{E}(\hat{\mathbf{y}}) = (W + P_\lambda)^{-1} W \mathbb{E}(\mathbf{y}) \neq \mathbb{E}(\mathbf{y})$ when $\lambda \neq 0$. This implies that penalization introduces a smoothing bias, which prevents the construction of a confidence interval centered around $\mathbb{E}(\mathbf{y})$. Therefore, in this section, we turn to a Bayesian approach where smoothing can be interpreted more naturally.

2.1 Maximum a posteriori

Let us suppose that $\mathbf{y}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, W^-)$ and $\boldsymbol{\theta} \sim \mathcal{N}(0, P_\lambda^-)$. The Bayes' formula allows us to express the posterior likelihood $f(\boldsymbol{\theta}|\mathbf{y})$ associated with these choices in the following form:

$$
\begin{aligned}
f(\boldsymbol{\theta}|\mathbf{y}) &= \frac{f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(y)} \\
&\propto f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \\
&\propto \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\theta})^T W (\mathbf{y}-\boldsymbol{\theta})\right) \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T P_\lambda \boldsymbol{\theta}\right) \\
&\propto \exp\left(-\frac{1}{2}\left[(\mathbf{y}-\boldsymbol{\theta})^T W (\mathbf{y}-\boldsymbol{\theta}) + \boldsymbol{\theta}^T P_\lambda \boldsymbol{\theta}\right]\right).
\end{aligned}
$$

The mode of the posterior distribution, $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}}[f(\boldsymbol{\theta}|\mathbf{y})]$, also known as the maximum a posteriori (MAP) estimate, coincides with the solution $\hat{\mathbf{y}}$ from Equation 2, whose explicit form is given by Equation 3.

2.2 Posterior distribution of $\boldsymbol{\theta}|\mathbf{y}$

A second-order Taylor expansion of the log-posterior likelihood around $\hat{\mathbf{y}} = \hat{\boldsymbol{\theta}}$ gives us:

$$\ln f(\boldsymbol{\theta}|\mathbf{y}) = \ln f(\hat{\boldsymbol{\theta}}|\mathbf{y}) + \frac{\partial \ln f(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}^{T} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \frac{\partial^2 \ln f(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \quad (4)$$

$$\text{where} \quad \frac{\partial \ln f(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0 \quad \text{and} \quad \frac{\partial^2 \ln f(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -(W + P_\lambda).$$

Note that this last derivative no longer depends on $\boldsymbol{\theta}$, and higher-order derivatives beyond 2 are all zero. The Taylor expansion allows for an exact computation of $\ln f(\boldsymbol{\theta}|\mathbf{y})$. By substituting the result back into Equation 4, we obtain:

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{y}) &\propto \exp\left[\ln f(\hat{\boldsymbol{\theta}}|\mathbf{y}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T (W + P_\lambda)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right] \\ &\propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T (W + P_\lambda)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right] \end{aligned}$$

which can immediately be recognized as the density of the $\mathcal{N}(\hat{\boldsymbol{\theta}}, (W + P_\lambda)^{-1})$ distribution.

2.3 Consequence for Whittaker-Henderson smoothing

The assumption $\boldsymbol{\theta} \sim \mathcal{N}(0, P_\lambda^-)$ corresponds to a simple Bayesian formalization of the smoothness criterion. It reflects a (improper) prior belief of the modeler about the underlying distribution of the observation vector $\mathbf{y}$.

The use of Whittaker-Henderson smoothing in the Bayesian framework and the construction of credible intervals are conditioned on the validity of the assumption $\mathbf{y}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, W^-)$. The components of the observation vector should be independent and have known variances. The weight vector $\mathbf{w}$ used should contain the inverses of these variances. Under these assumptions, credible intervals at $100(1-\alpha)\%$ for smoothing can be constructed and take the form:

$$\mathbb{E}(\mathbf{y})|\mathbf{y} \in \left[\hat{\mathbf{y}} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\mathbf{diag}\left\{(W + P_\lambda)^{-1}\right\}}\right] \quad (5)$$

with probability $1 - \frac{\alpha}{2}$ where $\hat{\mathbf{y}} = (W + P_\lambda)^{-1} W\mathbf{y}$ and Φ denotes the cumulative distribution function of the standard normal distribution. According to Marra and Wood (2012), these credible intervals provide satisfactory coverage of the corresponding frequentist confidence intervals and can therefore be used as practical substitutes.

3 Which observation and weight vectors to use?

Section 2 highlighted the need to apply the Whittaker-Henderson smoothing to independent observation vectors $\mathbf{y}$ and weight vectors $\mathbf{w}$ corresponding to the inverses of the variances of the components of $\mathbf{y}$ in order to obtain a measure of uncertainty in the results. In this section, we propose, within the framework of duration models used for constructing experience tables, vectors $\mathbf{y}$ and $\mathbf{w}$ that satisfy these conditions.

3.1 Duration models framework: one-dimensional case

Consider the observation of m individuals in a longitudinal study subject to left truncation and right censoring phenomena. Suppose we want to estimate a single distribution that depends on only one continuous explanatory variable, denoted by x. For illustration purposes, we can consider it as a mortality distribution with the explanatory variable of interest x representing age. Such a distribution is fully characterized by one of the following quantities:

- the cumulative distribution function $F(x)$ or its complement, the survival function $S(x) = 1 - F(x)$,
- the associated probability density function $f(x) = -\frac{\mathrm{d}}{\mathrm{d}x} S(x)$,
- the instantaneous hazard function $\mu(x) = -\frac{\mathrm{d}}{\mathrm{d}x} \ln S(x)$.

Suppose that the considered distribution depends on a vector of parameters $\boldsymbol{\theta}$ that we want to estimate using maximum likelihood. The likelihood associated with the observation of the individuals can be written as follows:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{m} \left[\frac{f(x_i + t_i, \boldsymbol{\theta})}{S(x_i, \boldsymbol{\theta})} \right]^{\delta_i} \left[\frac{S(x_i + t_i, \boldsymbol{\theta})}{S(x_i, \boldsymbol{\theta})} \right]^{1-\delta_i} \tag{6}$$

Where x_i represents the age at the start of observation, t_i represents the observation duration, i.e., the time elapsed between the start and end dates of observation, and δ_i is the indicator of event observation, which takes the value 1 if the event of interest is observed and 0 if the observation is censored. We will not go into the details of calculating these three quantities, which should take into account the individual-specific information such as the subscription date, redemption date if applicable, as well as the global characteristics of the product, such as the presence of a deductible or medical selection phenomenon, and the choice of a restricted observation period due to data quality issues or delays in the reporting of observations. These

factors typically lead to a narrower observation period than the actual presence period of individuals in the portfolio.

The various quantities introduced above are related by the following relationships:

$$S(x) = \exp\left(\int_{u=0}^{x} \mu(u)\mathrm{d}u\right) \quad \text{and} \quad f(x) = \mu(x)S(x).$$

The maximization of the likelihood in Equation 6 is equivalent to the maximization of the associated log-likelihood, which can be rewritten using only the instantaneous hazard function (also known as the mortality rate in the case of the death risk):

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} \left[\delta_i \ln \mu(x_i + t_i, \boldsymbol{\theta}) - \int_{u=0}^{t_i} \mu(x_i + u, \boldsymbol{\theta})\mathrm{d}u\right] \tag{7}$$

To discretize the problem, let us assume that the mortality rate is piecewise constant over one-year intervals between two integer ages. Formally, we have $\mu(x + \epsilon) = \mu(x)$ for all $x \in \mathbb{N}$ and $\epsilon \in [0, 1[$. Furthermore, if $\mathbf{1}$ denotes the indicator function, then for any $0 \leq a < x_{\max}$, we have $\sum_{x=x_{\min}}^{x_{\max}} \mathbf{1}(x \leq a < x + 1) = 1$, where $x_{\min} = \min(\mathbf{x})$ and $x_{\max} = \max(\mathbf{x})$. Equation 7 can be rewritten as:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} \left[\sum_{x=x_{\min}}^{x_{\max}} \delta_i \mathbf{1}(x \leq x_i + t_i < x + 1) \ln \mu(x_i + t_i, \boldsymbol{\theta}) \right.$$
$$\left. - \int_{u=0}^{t_i} \sum_{x=x_{\min}}^{x_{\max}} \mathbf{1}(x \leq x_i + u < x + 1)\mu(x_i + u, \boldsymbol{\theta})\mathrm{d}u\right].$$

The assumption of piecewise constant mortality rate implies that:

$$\mathbf{1}(x \leq x_i + t_i < x + 1) \ln \mu(x_i + t_i, \boldsymbol{\theta}) = \mathbf{1}(x \leq x_i + t_i < x + 1) \ln \mu(x, \boldsymbol{\theta}) \quad \text{and}$$
$$\mathbf{1}(x \leq x_i + u < x + 1)\mu(x_i + u, \boldsymbol{\theta}) = \mathbf{1}(x \leq x_i + u < x + 1) \ln \mu(x, \boldsymbol{\theta}).$$

It is then possible to interchange the two summations to obtain the following expressions:

$$\ell(\boldsymbol{\theta}) = \sum_{x=x_{\min}}^{x_{\max}} [\ln \mu(x,\boldsymbol{\theta}) d(x) - \mu(x,\boldsymbol{\theta}) e_c(x)] \quad \text{where}$$

$$d(x) = \sum_{i=1}^{m} \delta_i \mathbf{1}(x \leq x_i + t_i < x+1) \quad \text{and}$$

$$e_c(x) = \sum_{i=1}^{m} \int_{u=0}^{t_i} \mathbf{1}(x \leq x_i + u < x+1) \mathrm{d}u = \sum_{i=1}^{m} [\min(t_i, x - x_i + 1) - \max(0, x - x_i)]^+$$

by denoting $a^+ = \max(a, 0)$, where $d(x)$ and $e_c(x)$ correspond to the number of observed deaths between ages x and $x+1$ and the sum of observation durations of individuals between these ages, respectively (the latter quantity is also known as central exposure to risk).

3.2 Extension to the two-dimensional case

The extension of the proposed approach to the two-dimensional framework requires only minor adjustments to the previous reasoning. Let $z_{\min} = \min(\mathbf{z})$ and $z_{\max} = \max(\mathbf{z})$. The piecewise constant assumption for the mortality rate needs to be extended to the second dimension. Formally, we now assume that $\mu(x+\epsilon, z+\xi) = \mu(x,z)$ for all pairs $x, z \in \mathbb{N}$ and $\epsilon, \xi \in [0,1[$. The sums involving the variable x are then replaced by double sums considering all combinations of x and z. The log-likelihood is given by:

$$\ell(\boldsymbol{\theta}) = \sum_{x=x_{\min}}^{x_{\max}} \sum_{z=z_{\min}}^{z_{\max}} [\ln \mu(x,z,\boldsymbol{\theta}) d(x,z) - \mu(x,z,\boldsymbol{\theta}) \mathbf{e_c}(x,z)] \quad \text{where}$$

$$d(x,z) = \sum_{i=1}^{m} \delta_i \mathbf{1}(x \leq x_i + t_i < x+1) \mathbf{1}(z \leq z_i + t_i < z+1) \quad \text{and}$$

$$\begin{aligned} \mathbf{e_c}(x,z) &= \sum_{i=1}^{m} \int_{u=0}^{t_i} \mathbf{1}(x \leq x_i + u < x+1) \mathbf{1}(z \leq z_i + u < z+1) \mathrm{d}u \\ &= \sum_{i=1}^{m} [\min(t_i, x+1-x_i, z+1-z_i) - \max(0, x-x_i, z-z_i)]^+ \end{aligned}$$

3.3 Likelihood equations

The choice $\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{exp}(\boldsymbol{\theta})$, which corresponds to considering one parameter per observation, allows us to relate to the Whittaker-Henderson smoothing. Using the exponential function ensures positive values for the estimated mortality rate. The expressions of the likelihood in the one-dimensional or two-dimensional case can then be written in a common vectorized form:

$$\ell(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{d} - \mathbf{exp}(\boldsymbol{\theta})^T \mathbf{e_c} \tag{8}$$

where $\mathbf{d}$ and $\mathbf{e}_c$ represent the vectors of observed deaths and exposures to risk, respectively. The derivatives of the likelihood function for this model are given by:

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = [\mathbf{d} - \mathbf{exp}(\boldsymbol{\theta}) \odot \mathbf{e_c}] \quad \text{and} \quad \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = -\text{Diag}(\mathbf{exp}(\boldsymbol{\theta}) \odot \mathbf{e_c}). \tag{9}$$

Note that these likelihood equations are exactly what we would obtain by assuming that the observed numbers of deaths, conditional on the observed exposures to risk $\mathbf{e}_c$, follow Poisson distributions with parameters $\boldsymbol{\mu}(\boldsymbol{\theta}) \odot \mathbf{e}_c$. The model presented here has many similarities with a Poisson GLM (John Ashworth Nelder and Wedderburn 1972). However, the initial assumptions are not the same for both models.

These likelihood equations have an explicit solution given by $\hat{\boldsymbol{\theta}} = \ln(\mathbf{d}/\mathbf{e}_c)$. This model, which treats each age independently, is known as the crude rates estimator. The properties of the maximum likelihood estimator imply that asymptotically $\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, W_{\hat{\boldsymbol{\theta}}}^{-1})$, where $W_{\hat{\boldsymbol{\theta}}}$ is a diagonal matrix with elements $\mathbf{exp}(\hat{\boldsymbol{\theta}}) \odot \mathbf{e}_c = (\mathbf{d}/\mathbf{e}_c) \odot \mathbf{e}_c = \mathbf{d}$. It should be noted that the asymptotic nature and the validity of this approximation are related to the number of individuals m in the portfolio and not the size n of the vectors $\mathbf{d}$ and $\mathbf{e}_c$.

Thus, we have shown that in the framework of duration models, using the crude rates estimator, asymptotically $\ln(\mathbf{d}/\mathbf{e}_c) \sim \mathcal{N}(\ln \boldsymbol{\mu}, W^{-1})$, where $W = \text{Diag}(\mathbf{d})$. This justifies applying the Whittaker-Henderson smoothing to the observation vector $\mathbf{y} = \ln(\mathbf{d}/\mathbf{e}_c)$ and weight vector $\mathbf{w} = \mathbf{d}$. According to the results from Section 2, we obtain the following associated credible intervals:

$$\ln \boldsymbol{\mu} | \mathbf{d}, \mathbf{e_c} \in \left[\hat{\boldsymbol{\theta}} \pm \Phi^{-1} \left(1 - \frac{\alpha}{2}\right) \sqrt{\mathbf{diag}\left\{(\text{Diag}(\mathbf{d}) + P_\lambda)^{-1}\right\}} \right]$$

where $\hat{\boldsymbol{\theta}} = (\text{Diag}(\mathbf{d}) + P_\lambda)^{-1} \text{Diag}(\mathbf{d})[\ln(\mathbf{d}) - \ln \mathbf{e}_c]$. Results on $\boldsymbol{\mu}$ can be obtained directly by exponentiating the above expression.

4 How to improve the accuracy of smoothing with limited data volume?

4.1 Generalized Whittaker-Henderson smoothing

The approach described in Section 3.3 relies on the asymptotic properties of the maximum likelihood estimator, which provide a theoretical framework for applying smoothing using the

crude rates estimator. However, the validity of these asymptotic properties in practice, where we have a limited number of observations, may be questioned.

Another approach is to apply the Bayesian reasoning presented in Section 2 directly to the likelihood of Equation 8. Let us assume again that $\boldsymbol{\theta} \sim \mathcal{N}(0, P_\lambda^-)$ and write, using Bayes' theorem:

$$f(\boldsymbol{\theta}|\mathbf{d}, \mathbf{e_c}) = \frac{f(\mathbf{d}, \mathbf{e_c}|\boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\mathbf{d}, \mathbf{e_c})} \propto f(\mathbf{d}, \mathbf{e_c}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \propto \exp\left[\ell(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T P_\lambda \boldsymbol{\theta}\right].$$

The quantity $\ell_P(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta})^T P_\lambda(\boldsymbol{\theta})$ will be referred to as the penalized likelihood. The maximum a posteriori $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} f(\boldsymbol{\theta}|\mathbf{d}, \mathbf{e_c})$ corresponds, once again, to the maximum of the penalized likelihood $\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell_P(\boldsymbol{\theta})$.

A second-order Taylor expansion of the posterior log-likelihood around $\hat{\boldsymbol{\theta}}$ yields:

$$\ln f(\boldsymbol{\theta}|\mathbf{d}, \mathbf{e_c}) \simeq \ln f(\hat{\boldsymbol{\theta}}|\mathbf{d}, \mathbf{e_c}) + \left.\frac{\partial \ln f(\boldsymbol{\theta}|\mathbf{d}, \mathbf{e_c})}{\partial \boldsymbol{\theta}}\right|^T_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \left.\frac{\partial^2 \ln f(\boldsymbol{\theta}|\mathbf{d}, \mathbf{e_c})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{10}$$

$$\text{avec ici} \quad \left.\frac{\partial \ln f(\boldsymbol{\theta}|\mathbf{d}, \mathbf{e_c})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0 \quad \text{and} \quad \left.\frac{\partial^2 \ln f(\boldsymbol{\theta}|\mathbf{d}, \mathbf{e_c})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -(W_{\hat{\boldsymbol{\theta}}} + P_\lambda).$$

where $W_{\hat{\boldsymbol{\theta}}} = \text{Diag}(\mathbf{exp}(\hat{\boldsymbol{\theta}}) \odot \mathbf{e_c})$.

Unlike the normal case studied in Section 2, the higher-order derivatives of the posterior log-likelihood are not zero, and the Equation 10 represents an approximation of the posterior log-likelihood known as the Laplace approximation, which becomes more accurate as the number of observations increases. Asymptotically:

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{d}, \mathbf{e_c}) &\propto \exp\left[\ell_P(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T (W_{\hat{\boldsymbol{\theta}}} + P_\lambda)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right] \\ &\propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T (W_{\hat{\boldsymbol{\theta}}} + P_\lambda)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right] \end{aligned}$$

and the posterior distribution is thus equivalent to $\mathcal{N}(\hat{\boldsymbol{\theta}}, (W_{\hat{\boldsymbol{\theta}}} + P_\lambda)^{-1})$. Since $\boldsymbol{\theta} = \ln \boldsymbol{\mu}$, this result allows for the construction of asymptotic credible intervals at $100(1-\alpha)$ for $\ln \boldsymbol{\mu}|\mathbf{d}, \mathbf{e_c}$, in the form:

$$\ln \boldsymbol{\mu}|\mathbf{d}, \mathbf{e_c} \in \left[\hat{\boldsymbol{\theta}} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\mathbf{diag}\left\{(W_{\hat{\boldsymbol{\theta}}} + P_\lambda)^{-1}\right\}}\right].$$

Unlike Equation 9, it is not possible to determine the maximum of the penalized likelihood $\hat{\boldsymbol{\theta}}$ explicitly. However, the Newton algorithm allows for a numerical solution of the likelihood equations by constructing a sequence of estimators $(\boldsymbol{\theta}_k)_{k\geq 0}$ that converges to $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell_P(\boldsymbol{\theta})$.

These estimators are recursively defined as:

$$\begin{aligned}\boldsymbol{\theta}_{k+1} &= \boldsymbol{\theta}_k - \left(\frac{\partial^2 \ell_P}{\partial \boldsymbol{\theta}^2}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k} \right)^{-1} \frac{\partial \ell_P}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k} \\ &= \boldsymbol{\theta}_k + (W_k + P_\lambda)^{-1}(\mathbf{d} - \mathbf{exp}(\boldsymbol{\theta}_k) \odot \mathbf{e_c} - P_\lambda \boldsymbol{\theta}_k)] \\ &= (W_k + P_\lambda)^{-1} W_k \mathbf{z}_k\end{aligned}$$

by denoting $W_k = \text{Diag}(\mathbf{exp}(\boldsymbol{\theta}_k) \odot \mathbf{e_c})$ and $\mathbf{z}_k = \boldsymbol{\theta}_k + W_k^{-1}[\mathbf{d} - \mathbf{exp}(\boldsymbol{\theta}_k) \odot \mathbf{e_c}]$. An interesting initialization choice for the algorithm starts with the crude rates estimator, setting $\boldsymbol{\theta}_0 = \ln(\mathbf{d}/\mathbf{e_c})$, from which we derive $W_0 = \text{Diag}(\mathbf{d})$ and $z_0 = \ln(\mathbf{d}/\mathbf{e_c})$.

With this choice $\boldsymbol{\theta}_1 = (W_0 + P_\lambda)^{-1} W_0 z_0$ corresponds to the solution of the original Whittaker-Henderson smoothing. Subsequent iterations can also be seen as successive applications of the original smoothing to pseudo-observation vectors $\mathbf{z}_k$ and associated weight vectors $w_k = \mathbf{diag}(W_k)$, adjusted at each iteration. The proposed approach can thus be considered as an iterative variant of the original Whittaker-Henderson smoothing, where the choice of observation and weight vectors is refined at each step. The connection between the approach presented in this section and the original smoothing is of the same nature as that between a linear model and its generalized version. Therefore, we will refer to this approach as *generalized Whittaker-Henderson smoothing* throughout the rest of the paper. Algorithm 1 provides an implementation of this new version of the smoothing.

4.2 Impact of the normal approximation of the original Whittaker-Henderson smoothing

We have shown that the original Whittaker-Henderson smoothing can be interpreted as a normal approximation of a penalized likelihood maximization problem. To assess the impact of this approximation in practice, we rely on six simulated datasets that capture the characteristics of real biometric risks. These datasets include:

- In the one-dimensional case, three annuity portfolios with 20,000, 100,000, and 500,000 policyholders, respectively.
- In the two-dimensional case, three annuity portfolios for policyholders in a long-term care situation, with 1,000, 5,000, and 25,000 dependent policyholders, respectively.

Algorithm 1: Iterative solution of generalized Whittaker-Henderson smoothing

inputs : $\mathbf{d}$ and $\mathbf{e_c}$
outputs : $\hat{\boldsymbol{\theta}}$ and $\mathrm{Var}(\hat{\boldsymbol{\theta}})$
parameters: λ, q, $\epsilon_\ell = 10^{-8}$
begin
  Construct the penalty matrix P_λ based on the difference matrices of order q.
  $k \leftarrow 0$
  $\boldsymbol{\theta}_0 \leftarrow \ln(\mathbf{d}/\mathbf{e_c})$
  $l_0 \leftarrow -\infty, \quad \mathrm{cond}_\ell \leftarrow \mathrm{true}$
  while *cond*$_\ell$ **do**
    $W_k \leftarrow \mathrm{Diag}(\mathbf{exp}(\boldsymbol{\theta}_k) \odot \mathbf{e_c})$
    $\mathbf{z}_k \leftarrow \boldsymbol{\theta}_k + W_k^{-1}[\mathbf{d} - \mathbf{exp}(\boldsymbol{\theta}_k) \odot \mathbf{e_c}]$
    $\Psi_{k+1} \leftarrow (W_k + P_\lambda)^{-1}$
    $\boldsymbol{\theta}_{k+1} \leftarrow \Psi_{k+1} W_k \mathbf{z}_k$
    $l_{k+1} \leftarrow \ell_P(\boldsymbol{\theta}_{k+1}), \quad \mathrm{cond}_\ell \leftarrow l_{k+1} - l_k \geq \epsilon_\ell \times \mathrm{sum}(d)$
    $k \leftarrow k + 1$
  $\hat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}_k, \quad \mathrm{Var}(\hat{\boldsymbol{\theta}}) \leftarrow \Psi_k$

For each of these portfolios, we first calculate the observed event counts $\mathbf{d}$ and the central exposure to risk $\mathbf{e_c}$ based on discretized explanatory variables of interest: age x for the one-dimensional case and combinations (x, z) of age and long-term care duration for the two-dimensional case. We then apply Whittaker-Henderson smoothing to the vectors $\mathbf{y} = \ln(\mathbf{d}/\mathbf{e_c})$ and $\mathbf{w} = \mathbf{d}$ on one hand, or directly to the vectors $\mathbf{d}$ and $\mathbf{e_c}$ within the framework of the generalization proposed in Section 4 on the other hand. The parameter λ used is determined for each example using a method that will be detailed in Section 5. Here, we simply specify that for each portfolio, the same parameter λ is used for both approaches, which corresponds to the same prior belief on the vector $\boldsymbol{\theta} = \ln \boldsymbol{\mu}$.

The two approaches aiming to estimate the vector $\boldsymbol{\theta} = \ln \boldsymbol{\mu}$ can be directly compared using the parameters $\hat{\boldsymbol{\theta}}_{\mathrm{norm}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$, estimated respectively by the normal approximation of the smoothing and by the maximization of the penalized likelihood. Since the first estimator can be seen as an approximate version of the second, we propose to use the error measure:

$$\Delta(\boldsymbol{\theta}) = \frac{\ell_P(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) - \ell_P(\boldsymbol{\theta})}{\ell_P(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) - \ell_P(\hat{\boldsymbol{\theta}}_\infty)} \tag{11}$$

where $\hat{\boldsymbol{\theta}}_\infty$ corresponds to the parameter vector that maximizes the penalized likelihood with an infinite penalty. It will become apparent in Section 6.2 that this choice corresponds to the polynomial of degree $q-1$ that maximizes the likelihood. By definition, $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}}\, \ell_P(\boldsymbol{\theta})$, and thus $\Delta(\boldsymbol{\theta}) \geq 0$ for any vector $\boldsymbol{\theta}$. Furthermore, $\Delta(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) = 0$ and $\Delta(\hat{\boldsymbol{\theta}}_\infty) = 1$. A model

Table 1: Impact of the normal approximation associated with Whittaker smoothing on a selection of 6 fictional insurance portfolios.

Portfolio	Head count	Relative error on penalized deviance
One-dimensional	20 000	7,52 %
One-dimensional	100 000	0,23 %
One-dimensional	500 000	0,09 %
Two-dimensional	1 000	1 784,54 %
Two-dimensional	5 000	122,87 %
Two-dimensional	25 000	6,95 %

for which $\Delta(\boldsymbol{\theta}) > 1$ can be considered as having no practical interest, as it is less probable than a simple polynomial fit according to the prior belief used. The indicator Δ constructed in this way allows interpreting the quality of the used approximation within certain limits. For the purposes of this paper, a difference of less than 1% is considered negligible, while a difference greater than 10% is considered prohibitive. Differences between 1% and 10% will be interpreted more nuanced. This choice is obviously arbitrary.

The discrepancies obtained for each of the 6 portfolios studied are presented in Table 1. These discrepancies naturally decrease with the size of the portfolio, as the validity of the normal approximation increases. In the one-dimensional case, the discrepancy is modest for the two largest portfolios but significant for the smallest one. In the two-dimensional case, the discrepancy is significant for the largest portfolio and prohibitive for the two smallest ones. This is because using the observed death vector as the weight vector introduces a bias by overweighting observations with the highest crude rate and underweighting those with the lowest crude rate. In light of these results, it seems appropriate to prioritize the use of the generalized Whittaker-Henderson smoothing, especially since its implementation does not pose significant practical difficulties.

5 How to choose the smoothing parameter(s)?

This section is dedicated to the choice of the smoothing parameter λ in the one-dimensional case or the pair $\lambda = (\lambda_x, \lambda_z)$ in the two-dimensional case. As shown in Figures 1 and 2, this choice strongly impacts the obtained results. Ideally, the selection of the smoothing parameter is based on the optimization of a statistical criterion, typically belonging to one of two major families.

On one hand, there are criteria based on the minimization of the model's prediction error, among which the Akaike information criterion (AIC, Akaike 1973) and the generalized cross-validation (GCV, Wahba 1980) are included. The Bayesian information criterion (BIC, Schwarz 1978) has a similar form to AIC, although its theoretical justification is very different, and thus it can be considered part of this group.

On the other hand, there are criteria based on the maximization of a likelihood function known as the marginal likelihood. This type of criterion was introduced by Patterson H. D. (1971) in the Gaussian case, initially under the name of restricted likelihood (REML), and used by Anderssen and Bloomfield (1974) for the selection of smoothing parameters. Wahba (1985) and Kauermann (2005) show that criteria minimizing prediction error have the best asymptotic performance, but their convergence to the optimal smoothing parameters is slower. For finite sample sizes, criteria based on the maximization of a likelihood function are considered a more robust choice by many authors, such as Reiss and Todd Ogden (2009) or Wood (2011).

Figure 3 shows the values of GCV and marginal likelihood as a function of the smoothing parameter for the data used to produce Figure 1. When the marginal likelihood exhibits a clearly defined optimum, GCV has two optima, one of which coincides with the marginal likelihood's optimum. The second optimum results in a model with nearly 45 degrees of freedom, which is not a plausible choice for the underlying mortality risk. For these reasons, we will favor the marginal likelihood as the selection criterion, especially since this choice naturally fits into the Bayesian framework introduced in Sections 2 and 4.

We will first discuss the selection in the case of the original Whittaker-Henderson smoothing and then extend the selection to the generalized case introduced in Section 4, introducing two competing approaches that will be subsequently compared.

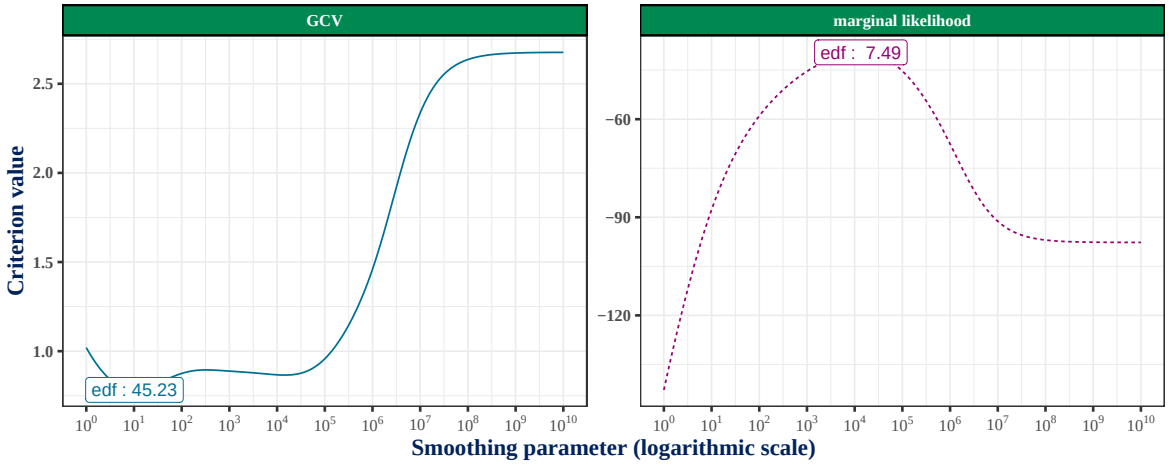

Figure 3: Comparison, in the context of one-dimensional Whittaker-Henderson smoothing, of the Generalized Cross-Validation (GCV) criterion and the marginal likelihood, as selection criteria for the smoothing parameter.

5.1 Selection in the context of the original smoothing

Let us start from the notations and assumptions from Section 2, namely $\mathbf{y}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, W^{-})$ and $\boldsymbol{\theta}|\lambda \sim \mathcal{N}(0, P_{\lambda}^{-})$. In a purely Bayesian approach, it would be necessary to define a prior

distribution on λ and then estimate the posterior distribution of each parameter vector using methods such as Markov Chain Monte Carlo. The empirical Bayesian approach we adopt seeks to find the value of λ that maximizes the marginal likelihood:

$$\mathcal{L}^m_{\text{norm}}(\lambda) = f(\mathbf{y}|\lambda) = \int f(\mathbf{y}, \boldsymbol{\theta}|\lambda) \mathrm{d}\boldsymbol{\theta} = \int f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\lambda) \mathrm{d}\boldsymbol{\theta}.$$

This corresponds to the maximum likelihood method applied to the smoothing parameter. Let us explicitly rewrite the expressions of $f(\mathbf{y}|\boldsymbol{\theta})$ and $f(\boldsymbol{\theta}|\lambda)$ introduced in Section 2:

$$f(\mathbf{y}|\boldsymbol{\theta}) = \sqrt{\frac{|W|_+}{(2\pi)^{n_*}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\theta})^T W (\mathbf{y}-\boldsymbol{\theta})\right)$$
$$f(\boldsymbol{\theta}|\lambda) = \sqrt{\frac{|P_\lambda|_+}{(2\pi)^{p-q}}} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T P_\lambda \boldsymbol{\theta}\right)$$

where $|A|_+$ denotes the product of the non-zero eigenvalues of A, n_* is the number of non-zero diagonal elements of W, and q is the number of zero eigenvalues of P_λ ($q = q_x \times q_z$ in the two-dimensional case). Based on the Taylor expansion performed in Section 2, let us recall that:

$$\ln f(\mathbf{y}, \boldsymbol{\theta}|\lambda) = \ln f(\mathbf{y}, \hat{\boldsymbol{\theta}}_\lambda|\lambda) + \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_\lambda)^T (W + P_\lambda)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_\lambda)$$

which leads to:

$$\begin{aligned}
\mathcal{L}^m_{\text{norm}}(\lambda) &= \int \exp[\ln f(\mathbf{y}, \boldsymbol{\theta}|\lambda)] \mathrm{d}\boldsymbol{\theta} \\
&= f(\mathbf{y}, \hat{\boldsymbol{\theta}}_\lambda|\lambda) \int \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_\lambda)^T (W + P_\lambda)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_\lambda)\right] \mathrm{d}\boldsymbol{\theta} \\
&= f(\mathbf{y}, \hat{\boldsymbol{\theta}}_\lambda|\lambda) \sqrt{\frac{(2\pi)^p}{|W + P_\lambda|}} \\
&= f_{\mathbf{y}}(\mathbf{y}|\hat{\boldsymbol{\theta}}_\lambda) f_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_\lambda|\lambda) \sqrt{\frac{(2\pi)^p}{|W + P_\lambda|}} \\
&= \sqrt{\frac{|W|_+ |P_\lambda|_+}{(2\pi)^{n_*-q}|W + P_\lambda|}} \exp\left(-\frac{1}{2}\left[(\mathbf{y}-\hat{\boldsymbol{\theta}}_\lambda)^T W (\mathbf{y}-\hat{\boldsymbol{\theta}}_\lambda) + \hat{\boldsymbol{\theta}}_\lambda^T P_\lambda \hat{\boldsymbol{\theta}}_\lambda\right]\right).
\end{aligned}$$

The associated log-likelihood can be expressed as follows:

$$\begin{aligned}
\ell^m_{\text{norm}}(\lambda) = -\frac{1}{2}\Big[&(\mathbf{y}-\hat{\boldsymbol{\theta}}_\lambda)^T W (\mathbf{y}-\hat{\boldsymbol{\theta}}_\lambda) + \hat{\boldsymbol{\theta}}_\lambda^T P_\lambda \hat{\boldsymbol{\theta}}_\lambda \\
&- \ln|W|_+ - \ln|P_\lambda|_+ + \ln|W + P_\lambda| + (n_* - q)\ln(2\pi)\Big].
\end{aligned} \tag{12}$$

Once the selection of λ has been determined using the estimator $\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, \ell^m_{\text{norm}}(\lambda)$, the lack of an explicit solution to Equation 12 forces us to resort to numerical methods for its resolution. The Newton algorithm could once again be employed here and is a robust choice. This approach was notably adopted by Wood (2011). However, explicitly calculating the derivatives of the likelihood ℓ^m_{norm} is rather difficult from an operational perspective. Instead, we will use general heuristics such as those provided by Brent (1973) and John A. Nelder and Mead (1965), which are applicable to any sufficiently smooth function. These heuristics do not require derivative calculations and are implemented in the `optimize` and `optim` functions of the statistical programming language R.

Computing the marginal likelihood using Equation 12 requires the estimation of various intermediate quantities, with the most demanding being the estimation of $\hat{\boldsymbol{\theta}}_\lambda$. This estimation is based on Equation 3 and ideally involves the inversion of the symmetric matrix $W + P_\lambda$ through a Cholesky decomposition. The calculation of $\ln|W + P_\lambda|$ is immediate for the triangular matrices resulting from this decomposition. Note that since the matrix P_λ is a linear combination of the matrix $D^T_{n,q} D_{n,q}$ in the one-dimensional case, and the matrices $I_{n_z} \otimes D^T_{n_x,q_x} D_{n_x,q_x}$ and $D^T_{n_z,q_z} D_{n_z,q_z} \otimes I_{n_x}$ in the two-dimensional case, it is only necessary to form these matrices once. Furthermore, the calculation of $\ln|P_\lambda|_+$ can be efficiently performed by computing the eigendecomposition of the matrix $D^T_{n,q} D_{n,q}$ (or the matrices $D^T_{n_x,q_x} D_{n_x,q_x}$ and $D^T_{n_z,q_z} D_{n_z,q_z}$ in the two-dimensional case). This calculation then only requires multiplying these eigenvalues by the corresponding element of λ and taking the logarithm. Finally, since the terms $\ln|W|_+$ and $(n_* - q)\ln(2\pi)$ do not depend on λ, they can be ignored altogether.

5.2 Selection in the generalized smoothing framework: *outer iteration* approach

It is possible to extend the previous approach to the framework of penalized likelihood introduced in Section 4. The Taylor expansion used in Section 5.1 can be applied in this context and yields:

$$
\begin{aligned}
\mathcal{L}^m_{\text{ML}}(\lambda) &= \int \exp[\ln f(\mathbf{d}, \mathbf{e_c}, \boldsymbol{\theta}|\lambda)]\mathrm{d}\boldsymbol{\theta} \\
&\simeq f(\mathbf{d}, \mathbf{e_c}, \hat{\boldsymbol{\theta}}_\lambda|\lambda) \int \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_\lambda)^T (W_{\hat{\boldsymbol{\theta}}_\lambda} + P_\lambda)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_\lambda)\right) \mathrm{d}\boldsymbol{\theta} \\
&\simeq f(\mathbf{d}, \mathbf{e_c}, \hat{\boldsymbol{\theta}}_\lambda|\lambda) \sqrt{\frac{(2\pi)^p}{|W_{\hat{\boldsymbol{\theta}}_\lambda} + P_\lambda|}} \\
&\simeq l(\hat{\boldsymbol{\theta}}_\lambda) f_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_\lambda|\lambda) \sqrt{\frac{(2\pi)^p}{|W_{\hat{\boldsymbol{\theta}}_\lambda} + P_\lambda|}} \\
&\simeq l(\hat{\boldsymbol{\theta}}_\lambda) \sqrt{\frac{(2\pi)^q |P_\lambda|_+}{|W_{\hat{\boldsymbol{\theta}}_\lambda} + P_\lambda|}} \exp\left(-\frac{1}{2}\hat{\boldsymbol{\theta}}^T_\lambda P_\lambda \hat{\boldsymbol{\theta}}_\lambda\right).
\end{aligned}
$$

This gives us the expression for the likelihood in this framework:

$$\ell^m_{\mathrm{ML}}(\lambda) \approx l(\hat{\boldsymbol{\theta}}_\lambda) - \frac{1}{2}\left[\hat{\boldsymbol{\theta}}_\lambda^T P_\lambda \hat{\boldsymbol{\theta}}_\lambda - \ln|P_\lambda|_+ + \ln|W_{\hat{\boldsymbol{\theta}}_\lambda} + P_\lambda| - q\ln(2\pi)\right]. \tag{13}$$

Unlike the normal case, the Taylor expansion does not allow for an exact calculation of the marginal likelihood but provides an approximation known as the Laplace approximation, whose validity depends on the number of available observations. Equation 13 does not have an explicit solution and requires numerical resolution. The calculation of the Laplace approximation of the marginal likelihood is done in a similar manner to Equation 12 from Section 5.1, with the only notable difference being that the vector $\hat{\boldsymbol{\theta}}_\lambda$ must be estimated iteratively using Algorithm 1. This leads to Algorithm 2 for selecting the smoothing parameter which includes two nested iterative calculations: the calculation of $\hat{\boldsymbol{\theta}}_\lambda$, for a fixed λ, using the Newton algorithm, and the search for the maximum of ℓ^m_{ML} using heuristics such as the Brent and Nelder-Mead algorithms in the one-dimensional and two-dimensional case, respectively. Since the selection of λ corresponds to the outer iterative loop here, and to distinguish it from the approach introduced in the next section, we will refer to it as the *outer iteration* approach.

Algorithm 2: Parameter selection for generalized Whittaker-Henderson smoothing - outer iteration approach.

inputs : $\mathbf{d}$ and $\mathbf{e_c}$
outputs : $\hat{\lambda}$, $\hat{\boldsymbol{\theta}}_{\hat{\lambda}}$ and $\mathrm{Var}(\hat{\boldsymbol{\theta}}_{\hat{\lambda}})$
parameters: q, $\epsilon_\ell = 10^{-8}$, $\epsilon_{\mathrm{ml}} = 10^{-8}$
begin
- $k \leftarrow 0$
- $\mathrm{ml}_0 \leftarrow -\infty$, $\quad \mathrm{cond}_{\mathrm{ml}} \leftarrow$ true
- **while** *cond*$_{ml}$ **do**
  - If $k = 0$, choose an arbitrary value λ_0 for the smoothing parameter(s); otherwise, choose the next value λ_k using the selected heuristic
  - Use Algorithm 1 to determine the vector $\boldsymbol{\theta}_{\lambda_k}$ associated with the choice of λ_k with an accuracy greater than ϵ_ℓ
  - Calculate the marginal likelihood $\ell^m_{\mathrm{ML}}(\lambda_k)$ associated with the choice of λ_k using the intermediate quantities calculated during the estimation of $\boldsymbol{\theta}_{\lambda_k}$
  - $\mathrm{ml}_{k+1} \leftarrow \ell^m_{\mathrm{ML}}(\lambda_k)$, $\quad \mathrm{cond}_{\mathrm{ml}} \leftarrow \mathrm{ml}_{k+1} - \mathrm{ml}_k \geq \epsilon_{\mathrm{ml}} \times \mathrm{sum}(d)$
  - $k \leftarrow k + 1$
- $\hat{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}}_{\hat{\lambda}}$, $\quad \mathrm{Var}(\hat{\boldsymbol{\theta}}) \leftarrow \Psi_k$

5.3 Selection in the generalized smoothing framework: *performance iteration* approach

An alternative to the approach developed in Section 5.2 is to start from Algorithm 1 and notice that at each step of the algorithm, the solved equations coincide with those obtained by assuming that $\mathbf{z}_k|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, W_k^{-1})$ and $\boldsymbol{\theta} \sim \mathcal{N}(0, P_\lambda^-)$. Therefore, it is possible to rely on the methodology proposed in Section 5.1 and numerically estimate, at the beginning of each iteration of Algorithm 1, the smoothing parameter that maximizes the marginal likelihood ℓ^m_{norm} in which $\mathbf{y}$ would be replaced by $\mathbf{z}_k$ and W by W_k. This is what Algorithm 3 proposes. A justification for this approach is given by (Wood 2006, p149). From a practical point of view, this approach conceptually reverses the nesting order of the two types of iterations and can, in some cases, provide a considerable time gain. It was introduced by Gu (1992) as the "performance (oriented) iteration" approach. However, unlike the *outer iteration* approach, the convergence of the *performance iteration* approach cannot be guaranteed. Indeed, the penalized likelihood calculated in Algorithm 3 is not directly comparable between iterations because it is based on a different smoothing parameter and, therefore, a different prior. There are situations in which this algorithm may not converge. However, this difficulty seems to be more of a theoretical nature than a practical one for the optimization problem addressed in this paper.

Algorithm 3: Parameter selection for generalized Whittaker-Henderson smoothing - performance iteration approach

inputs : $\mathbf{d}$ and $\mathbf{e_c}$
outputs : $\hat{\lambda}$, $\hat{\boldsymbol{\theta}}_{\hat{\lambda}}$ and $\text{Var}(\hat{\boldsymbol{\theta}}_{\hat{\lambda}})$
parameters: q, $\epsilon_\ell = 10^{-8}$, $\epsilon_{\text{ml}} = 10^{-8}$
begin
- $k \leftarrow 0$
- $\boldsymbol{\theta}_0 \leftarrow \ln(\mathbf{d}/\mathbf{e_c})$
- $l_0 \leftarrow -\infty$, $\quad \text{cond}_\ell \leftarrow \text{true}$
- **while** *cond*$_\ell$ **do**
  - $W_k \leftarrow \text{Diag}(\mathbf{exp}(\boldsymbol{\theta}_k) \odot \mathbf{e_c})$
  - $\mathbf{z}_k \leftarrow \boldsymbol{\theta}_k + W_k^{-1}[\mathbf{d} - \mathbf{exp}(\boldsymbol{\theta}_k) \odot \mathbf{e_c}]$
  - Estimate the parameter λ_k maximizing the regularized marginal likelihood ℓ^m_{norm} associated with the observation vector $\mathbf{z}k$ and the weight matrix W_k, using the selected heuristic, with an accuracy greater than $\epsilon_{\text{ml}} \times \text{sum}(d)$
  - $\Psi_{k+1} \leftarrow (W_k + P_{\lambda_k})^{-1}$
  - $\boldsymbol{\theta}_{k+1} \leftarrow \Psi_{k+1} W_k \mathbf{z}_k$
  - $l_{k+1} \leftarrow \ell_P(\boldsymbol{\theta}_{k+1}|\lambda_k)$, $\quad \text{cond}_\ell \leftarrow l_{k+1} - l_k \geq \epsilon_\ell \times \text{sum}(d)$
  - $k \leftarrow k + 1$
- $\hat{\lambda} \leftarrow \lambda_k$; Use Algorithm 1 to obtain the values of $\hat{\boldsymbol{\theta}}_{\hat{\lambda}}$ and $\text{Var}(\hat{\boldsymbol{\theta}}_{\hat{\lambda}})$ corresponding to the parameter $\hat{\lambda}$

5.4 Comparison of *outer iteration* and *performance iteration* approaches

In Sections 5.2 and 5.3, we introduced two alternatives for the selection of parameters in generalized Whittaker-Henderson smoothing. This section aims to study the consequences of choosing one approach over the other in terms of result accuracy and computational time. While the objective of Section 4.2 was to highlight a bias in the asymptotic approximation associated with the original Whittaker-Henderson smoothing for small portfolios, here we aim to measure these impacts more precisely. We limit our analysis to intermediate-sized portfolios as presented in Section 4.2 (i.e., 100,000 rows for the one-dimensional case and 5,000 rows for the two-dimensional case), but we use 100 replicates of each portfolio, generated from the same mortality and censoring laws, to increase the robustness of our analysis.

The *outer iteration* and *performance iteration* approaches can be directly compared based on the selected parameter λ and using the marginal likelihood $\ell^m_{\text{ML}}(\lambda)$. Similar to Section 4.2, we define the error criterion as:

$$\Delta(\lambda) = \frac{\ell^m_{\text{ML}}(\hat{\lambda}_{\text{outer}}) - \ell^m_{\text{ML}}(\lambda)}{\ell^m_{\text{ML}}(\hat{\lambda}_{\text{outer}}) - \ell^m_{\text{ML}}(\infty)} \tag{14}$$

where $\ell^m_{\text{ML}}(\infty)$ corresponds to the approximation of the marginal likelihood associated with the choice of an infinite smoothing parameter. Again, the defined Δ satisfies the properties $\Delta(\lambda_{\text{outer}}) = 0$ and $\Delta(\infty) = 1$. This criterion is interpreted in the same way as the one given by Equation 11.

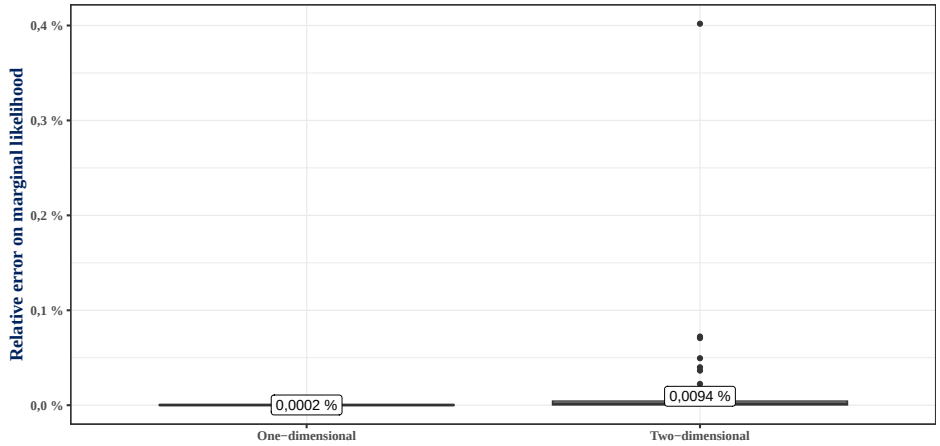

Figure 4: Distribution of the difference between the two presented approaches for selecting the smoothing parameter, for 100 simulated portfolios, in the one-dimensional and two-dimensional case.

Figure 4 represents the empirical distribution of the difference between the *outer iteration* and *performance iteration* approaches, and Figure 5 represents the distribution of the associated computation time for each approach. The use of the *performance iteration* approach results in a difference below 1% for the 100 simulated datasets, both in the one-dimensional and two-dimensional case. The associated computation time is reduced by a factor of 2 in the one-dimensional case and nearly 4 in the two-dimensional case. Considering the significant computation times in the latter case, the use of the *performance iteration* approach presents a real operational advantage.

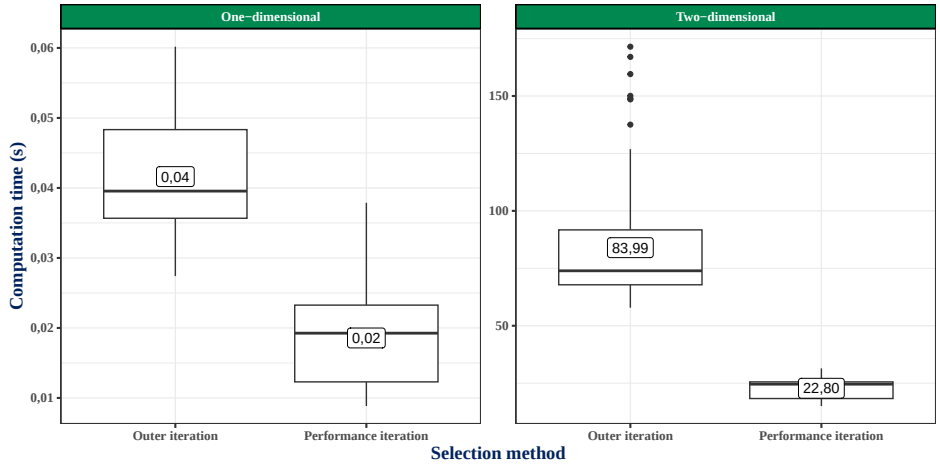

Figure 5: Distribution of the computation time associated with the two presented approaches for selecting the smoothing parameter, for 100 simulated portfolios, in the one-dimensional and two-dimensional case.

6 How to improve smoothing performances with a large number of data points?

6.1 Motivation

Whittaker-Henderson smoothing is considered to be full-rank smoothing as it contains as many parameters as there are observation points. This characteristic allows it to faithfully reproduce any input signal, provided that a sufficient number of observations are available. More formally, the estimator associated with smoothing is asymptotically unbiased since:

$$\mathbb{E}(\hat{\mathbf{y}}) = (W + P_\lambda)^{-1} W \mathbb{E}(\mathbf{y}) \underset{m \to \infty}{\to} \mathbb{E}(\mathbf{y})$$

where m represents the number of observed individuals and is involved in the matrix W. The downside is that it may become impractical in presence of many observations points n. Indeed,

the algorithms presented in Section 5 require the inversion of the $n \times n$ matrix $W + P_\lambda$, an operation with a time complexity of $O(n^3)$, which needs to be repeated at each iteration of the algorithm, regardless of the chosen method.

For biometric risks that depend solely on age, discretized on an annual basis, the number of observations rarely exceeds 100, and computation time is not a significant concern. However, in the two-dimensional case, the number of observations can take much larger values in several practical cases, including:

- For the disability risk, it is necessary to construct disability survival tables for entry ages ranging from 18 to 62 and exit ages ranging from the entry age to 62. This represents $[(62 - 18) \times (62 - 18 + 1)/2 = 990$ observations.
- Also for the disability risk, it is necessary to construct transition tables from incapacity to disability for entry ages ranging from 18 to 67 and monthly incapacity durations ranging from 0 to 36 months. This represents $(67 - 18) \times (36 - 0) = 1,764$ observations.
- For the long-term care risk, as the coverage is lifelong, it is necessary to construct tables for all ages and long-term care durations encountered in the data. Assuming observation ages ranging from 50 to 110 and durations ranging from 0 to 20 years, this already represents $(110 - 50) \times (20 - 0) = 1,200$ observations.

The practical applications described above thus require computation times on the order of several minutes for each smoothing application. When smoothing is intended to be used repeatedly, for example, in simulation contexts, reducing the computation time becomes an operational challenge.

6.2 Smoothing and eigendecomposition

The eigendecomposition of the penalization matrix is key to a better understanding of smoothing and allows for an approximation of the smoothing problem with $p < n$ parameters. Let us consider the one-dimensional case and write the decomposition for the symmetric matrix $D_{n,q}^T D_{n,q}$. It takes the form $D_{n,q}^T D_{n,q} = U\Sigma U^T$, where Σ is a diagonal matrix containing the eigenvalues of $D_{n,q}^T D_{n,q}$ and U is an orthogonal matrix such that $U^T U = UU^T = I_n$. Let us perform the reparameterization $\boldsymbol{\beta} = U^T\boldsymbol{\theta} \Leftrightarrow \boldsymbol{\theta} = U\boldsymbol{\beta}$. The smoothness criterion becomes:

$$\boldsymbol{\theta}^T P_\lambda \boldsymbol{\theta} = (U\boldsymbol{\beta})^T P_\lambda (U\boldsymbol{\beta}) = \lambda \boldsymbol{\beta}^T U^T U \Sigma U^T U \boldsymbol{\beta} = \lambda \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}. \tag{15}$$

and Equation 2 takes the form:

$$\hat{\mathbf{y}} = U\hat{\boldsymbol{\beta}} \quad \text{where} \quad \hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ (\mathbf{y} - U\boldsymbol{\beta})^T W (\mathbf{y} - U\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta} \right\}. \tag{16}$$

In the original formulation of smoothing, the parameters directly corresponded to the smoothed values. With this new parameterization, $\boldsymbol{\beta}$ can be interpreted as a vector of coordinates in the basis of eigenvectors of P_λ, providing a decomposition of the signal into components that are more or less smooth according to the employed penalization. Figure 6 represents 8 of the eigenvectors associated with $q = 2$ for a basis of size $n = 74$, which corresponds to the number of observations in the mortality portfolio used, notably in Figure 1. The eigenvalues associated with the first q eigenvalues are zero. This result can be easily explained by noting that $D_{n,q}$ is a matrix of dimensions $(n - q) \times n$ and rank $n - q$.

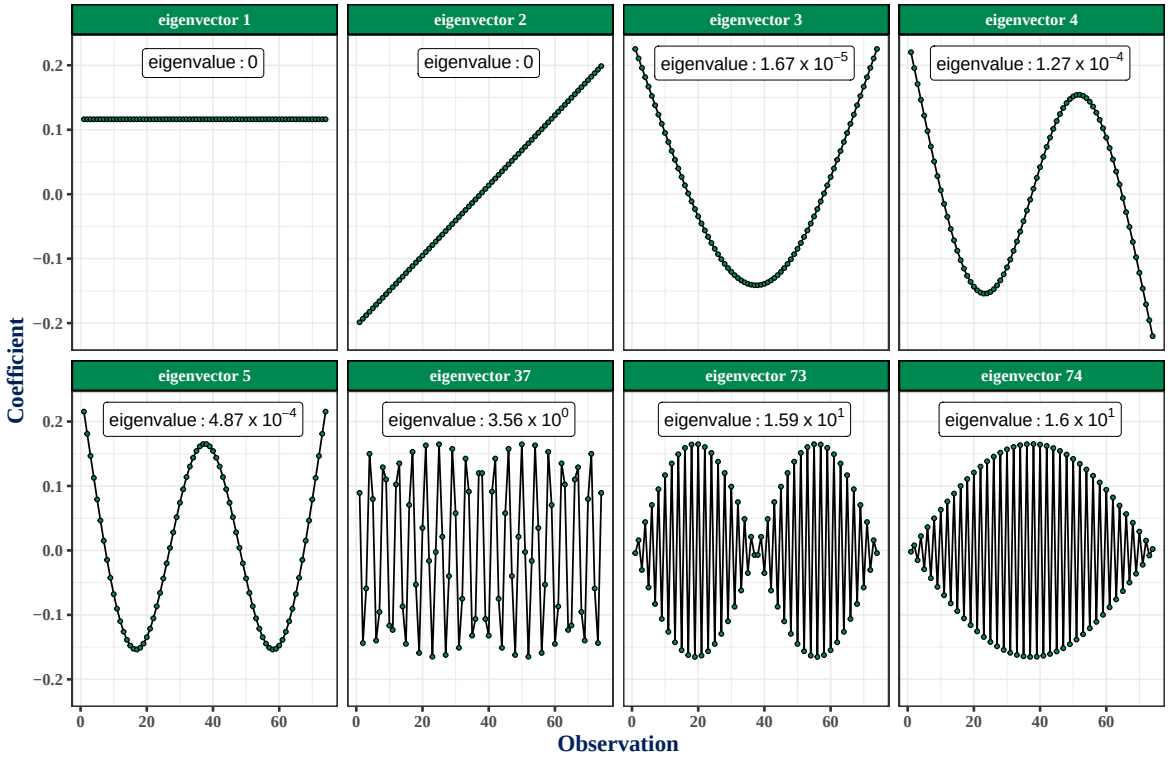

Figure 6: Graphical representation of a subset of columns from a basis of eigenvectors associated with $n = 74$ and $q = 2$ in the one-dimensional case. The 8 vectors in the subset are displayed in order of increasing penalization.

By using the fact that $U^{-1} = U^T$ and making the connection with Equation 3, we obtain the explicit solution:

$$\hat{\mathbf{y}} = U(U^T W U + S_\lambda)^{-1} U^T W \mathbf{y} \quad \text{where} \quad S_\lambda = \lambda \Sigma \tag{17}$$

In order to interpret Equation 17, let us consider the special case where all weights are equal to 1, and we have:

$$\hat{\mathbf{y}} = U(U^TU + \lambda\Sigma)^{-1}U^T\mathbf{y} = U(I_n + \lambda\Sigma)^{-1}U^T\mathbf{y}$$

The transformation from $\mathbf{y}$ to $\hat{\mathbf{y}}$ can then be seen as a three-step process, reading the equation from right to left:

1. Decomposition of the signal $\mathbf{y}$ in the basis of eigenvectors through the left multiplication by U^T.
2. Attenuation of the signal components based on the eigenvalues associated with these components. If we denote $s = \text{diag}(\Sigma)$, then $(I_n + \lambda\Sigma)^{-1} = \text{Diag}[1/(1 + \lambda s)]$. After the left multiplication by $(I_n + \lambda\Sigma)^{-1}$, each component is divided by a coefficient $1 + \lambda s \geq 1$. This attenuation coefficient increases linearly with λ, but at different rates for each eigenvalue.
3. Reconstruction of the attenuated signal in the canonical basis through the left multiplication by U.

In the presence of non-unit weights, things are not as straightforward since U^TWU is no longer a diagonal matrix. However, it is still possible to visualize the effect of smoothing by looking at the matrix $F = (U^TWU + S_\lambda)^{-1}U^TWU$. In fact:

$$U^T\hat{\mathbf{y}} = U^TU\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} = (U^TWU + S_\lambda)^{-1}U^TW\mathbf{y} = (U^TWU + S_\lambda)^{-1}U^TWUU^T\mathbf{y} = FU^T\mathbf{y}$$

Since the vectors $U^T\mathbf{y}$ and $\hat{\boldsymbol{\beta}} = U^T\hat{\mathbf{y}}$ represent the coordinates of the signals $\mathbf{y}$ and $\hat{\mathbf{y}}$ in the basis of eigenvectors of $D_{n,q}^TD_{n,q}$, F corresponds to a coordinate transformation matrix playing a similar role for the parameters as the hat matrix $H = U(U^TWU + S_\lambda)^{-1}U^TW$ does for the observations. The diagonal values of F can be interpreted as the effective degrees of freedom associated with each eigenvector after smoothing. It can be verified that:

$$\text{tr}(F) = \text{tr}[(U^TWU + S_\lambda)^{-1}U^TWU] = \text{tr}[U(U^TWU + S_\lambda)^{-1}U^TW] = \text{tr}(H)$$

which means that the sum of the effective degrees of freedom remains the same whether it is counted per observation or per parameter.

Figure 7 represents the effective degrees of freedom per parameter in the previous illustration of smoothing. The first q eigenvectors are never penalized, so their effective degrees of freedom are always equal to 1, regardless of the smoothing parameter used. The other eigenvectors have strictly decreasing effective degrees of freedom with λ. These degrees of freedom are generally decreasing with increasing eigenvalues of $D_{n,q}^TD_{n,q}$, although in the presence of non-unit weights and for small values of λ, this may not always be the case.

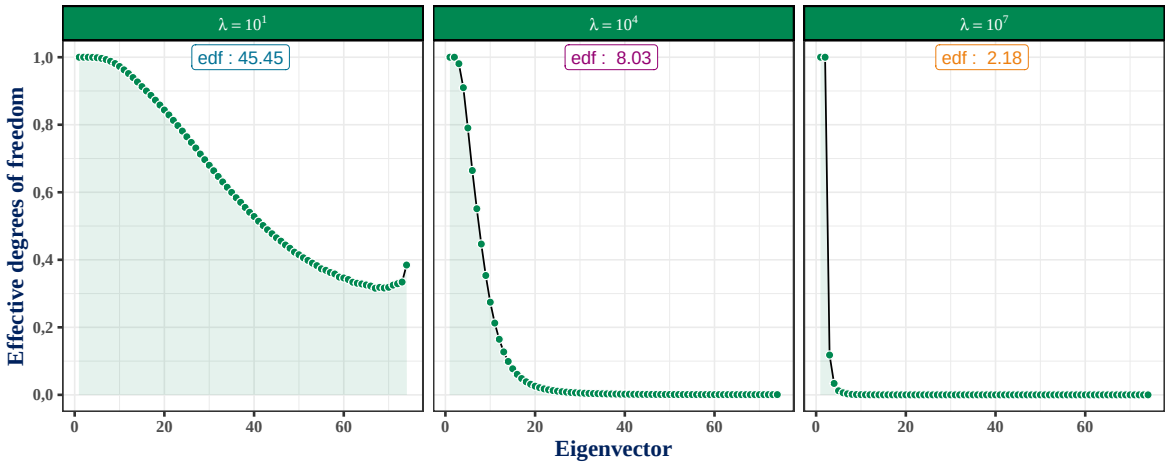

Figure 7: Distribution of residual degrees of freedom per eigenvector after applying the Whittaker-Henderson smoothing, in the one-dimensional case, for different values of the smoothing parameter.

6.3 Extension to the two-dimensional case

In the two-dimensional case, we have $P_\lambda = \lambda_x I_{n_z} \otimes D^T_{n_x,q_x} D_{n_x,q_x} + \lambda_z D^T_{n_z,q_z} D_{n_z,q_z} \otimes I_{n_x}$. Similar to the one-dimensional case, we can perform the eigendecomposition of the matrices $D^T_{n_x,q_x} D_{n_x,q_x}$ and $D^T_{n_z,q_z} D_{n_z,q_z}$, yielding $D^T_{n_x,q_x} D_{n_x,q_x} = U_x \Sigma_x U_x^T$ and $D^T_{n_z,q_z} D_{n_z,q_z} = U_z \Sigma_z U_z^T$. Let us define $U = U_z \otimes U_x$ and perform the reparameterization $\boldsymbol{\beta} = U^T \boldsymbol{\theta} \Leftrightarrow \boldsymbol{\theta} = U\boldsymbol{\beta}$. By leveraging the properties of the Kronecker product, we can rewrite the smoothness criterion in a simplified form:

$$
\begin{aligned}
\boldsymbol{\theta}^T P_\lambda \boldsymbol{\theta} &= (U\boldsymbol{\beta})^T P_\lambda (U\boldsymbol{\beta}) \\
&= (U\boldsymbol{\beta})^T (\lambda_x I_{n_z} \otimes U_x \Sigma_x U_x^T + \lambda_z U_z \Sigma_z U_z^T \otimes I_{n_x})(U\boldsymbol{\beta}) \\
&= \boldsymbol{\beta}^T (U_z \otimes U_x)^T (\lambda_x I_{n_z} \otimes U_x \Sigma_x U_x^T + \lambda_z U_z \Sigma_z U_z^T \otimes I_{n_x})(U_z \otimes U_x)\boldsymbol{\beta} \\
&= \boldsymbol{\beta}^T (U_z^T \otimes U_x^T)(\lambda_x I_{n_z} \otimes U_x \Sigma_x U_x^T + \lambda_z U_z \Sigma_z U_z^T \otimes I_{n_x})(U_z \otimes U_x)\boldsymbol{\beta} \\
&= \boldsymbol{\beta}^T (\lambda_x U_z^T U_z \otimes U_x^T U_x \Sigma_x U_x^T U_x + \lambda_z U_z^T U_z \Sigma_z U_z^T U_z \otimes U_x^T U_x)\boldsymbol{\beta} \\
&= \boldsymbol{\beta}^T (\lambda_x I_{n_z} \otimes \Sigma_x + \lambda_z \Sigma_z \otimes I_{n_x})\boldsymbol{\beta}.
\end{aligned}
$$

This leads to an alternative formulation of the optimization problem:

$$
\hat{\mathbf{y}} = U\hat{\boldsymbol{\beta}} \quad \text{where} \quad \hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ (\mathbf{y} - U\boldsymbol{\beta})^T W (\mathbf{y} - U\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T (\lambda_x I_{n_z} \otimes \Sigma_x + \lambda_z \Sigma_z \otimes I_{n_x})\boldsymbol{\beta} \right\}. \tag{18}
$$

The solution to the smoothing problem, as in the one-dimensional case, is given by:

$$\hat{\mathbf{y}} = U(U^T W U + S_\lambda)^{-1} U^T W \mathbf{y} \quad \text{where} \quad S_\lambda = \lambda_x I_{n_z} \otimes \Sigma_x + \lambda_z \Sigma_z \otimes I_{n_x}. \tag{19}$$

Figure 8 represents the residual degrees of freedom associated with each parameter after applying the smoothing, in the two-dimensional case, for different combinations of the smoothing parameters. Similar to the one-dimensional case, these degrees of freedom decrease as the smoothing parameters increase and are particularly small for higher eigenvalues. The eigenvectors are sorted in ascending order of eigenvalues for each one-dimensional penalty matrix $D^T_{n_x,q_x} D_{n_x,q_x}$ and $D^T_{n_z,q_z} D_{n_z,q_z}$.

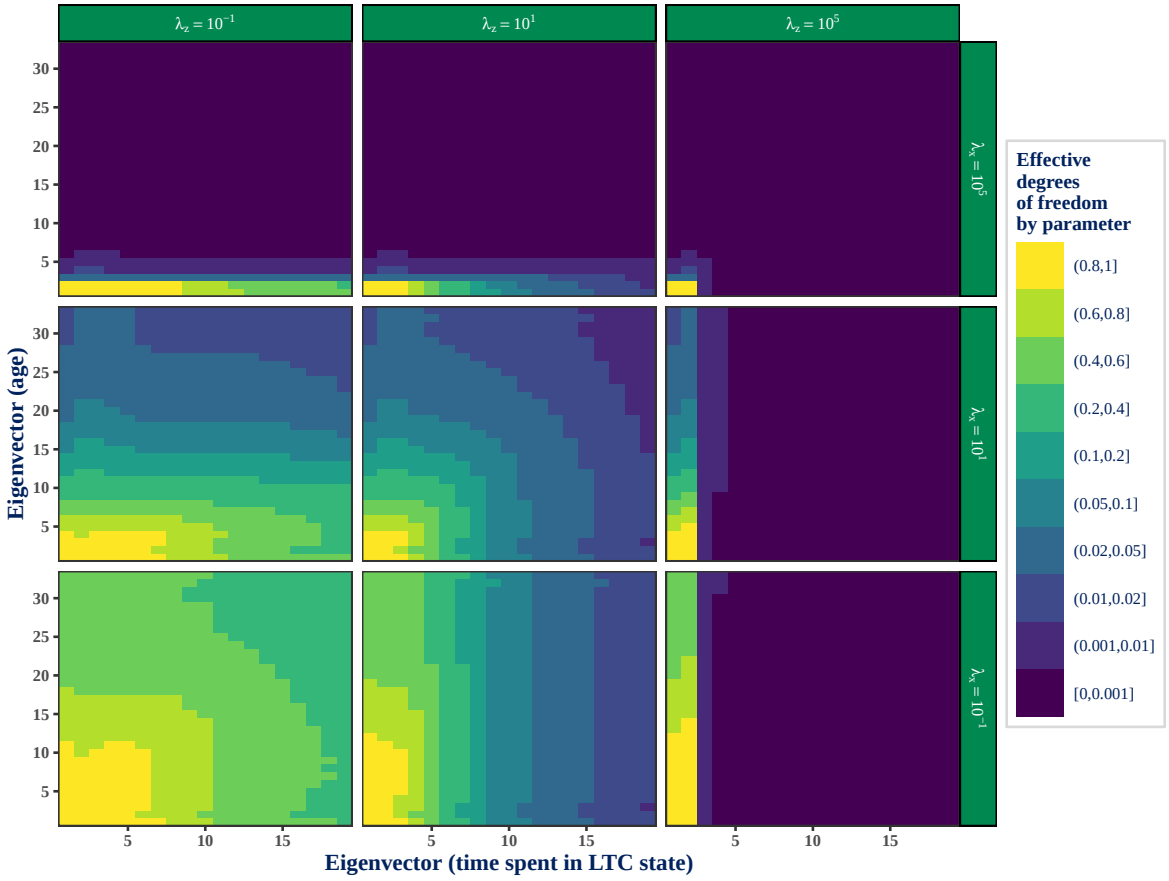

Figure 8: Distribution of residual degrees of freedom of the model by eigenvector after applying Whittaker-Henderson smoothing in the two-dimensional case, for different combinations of smoothing parameters.

6.4 Eigendecomposition and rank reduction

In addition to providing a more intuitive interpretation of the smoothing effect, the eigendecomposition of the penalty matrix is key to reducing the dimension of the optimization problem associated with it. Figures 7 and 8 show that eigenvectors associated with higher eigenvalues are more penalized by the smoothing, to the extent that a large majority of the eigenvectors represented in these figures have residual degrees of freedom very close to 0 for intermediate values of λ. This suggests that if we simply remove the parameters associated with these eigenvectors from the model, the smoothing results would be minimally affected. We propose setting the coordinates associated with the $n - p$ largest eigenvectors to 0, thus retaining only a reduced number $p < n$ of parameters to estimate.

In the one-dimensional case, this translates to using instead of $\hat{\mathbf{y}}$ the approximate estimator:

$$\hat{\mathbf{y}}_p = U_p(U_p^T W U_p + \lambda \Sigma_p)^{-1} U_p^T W \mathbf{y} \tag{20}$$

where U_p contains the first p columns of U, and Σ_p is a diagonal matrix constructed from the p smallest eigenvalues of Σ. In the two-dimensional case, we use:

$$\hat{\mathbf{y}}_{p_x,p_z} = U_{p_x,p_z}(U_{p_x,p_z}^T W U_{p_x,p_z} + \lambda_x I_{p_z} \otimes \Sigma_{x,p_x} + \lambda_z \Sigma_{z,p_z} \otimes I_{p_x})^{-1} U_{p_x,p_z}^T W \mathbf{y} \tag{21}$$

where $U_{p_x,p_z} = U_{z,p_z} \otimes U_{x,p_x}$, U_{x,p_x} (resp. U_{z,p_z}) contains the first p_x (resp. p_z) columns of U_x (resp. U_z), and Σ_{x,p_x} (resp. Σ_{z,p_z}) corresponds to a diagonal matrix constructed from the p_x (resp. p_z) smallest eigenvalues of Σ_x (resp. Σ_z).

In the two-dimensional case, there are several possible strategies for choosing the pair (p_x, p_z). One simple solution is to set an upper bound $p_{\max}$ on the number of parameters to be retained and choose p_x and p_z proportionally to the numbers n_x and n_z of categories in each dimension. Let $\kappa = \sqrt{p_{\max}/n_x n_z}$, and choose $p_x = \lfloor \min(\kappa, 1) n_x \rfloor$ and $p_z = \lfloor \min(\kappa, 1) n_z \rfloor$. By construction, the pair $(p_x, p_z) \in \mathbb{N}^2$ defined in this way satisfies the condition $p_x \times p_z \leq \kappa^2 n_x n_z = p_{\max}$. Other strategies may be more effective. If it is known, for example, that the phenomenon under study exhibits greater smoothness in one dimension or the other, it may be interesting to retain a smaller proportion of components associated with that dimension.

The results presented here can be directly applied to generalized Whittaker-Henderson smoothing by replacing $\mathbf{y}$ with $\mathbf{z}_k$ and W with W_k in Equations 20 and 21. This approach allows for reducing the rank of the optimization problem to be solved and can be used at all stages of the algorithms presented. However, we limit its use to the selection of the smoothing parameter and prefer, once the optimal parameter is chosen, to use the estimator $\hat{\mathbf{y}}$ (or $\hat{\boldsymbol{\theta}}$) associated with the full-rank problem in Equation 3 or Algorithm 1.

6.5 Impact of rank reduction method

Here, we aim to assess the impact of using the rank-reduced estimators defined in Section 6.4 on the smoothing results and computation time. To do so, we start from the 100 replicate datasets (one-dimensional and two-dimensional) used in Section 5.4. Since the use of rank reduction is limited to the selection of the smoothing parameter, and the final result is always obtained from the full-rank estimator, we can again use the error measure defined by Equation 14 to quantify the impact of the rank reduction method on the smoothing results. The quantity $\ell^m_{\mathrm{ML}}(\lambda)$ involved in this formula will be computed based on the values of λ obtained from the *performance iteration* approach for different values of the number of retained components p. The quantities $\ell^m_{\mathrm{ML}}(\lambda_{\mathrm{outer}})$ and $\ell^m_{\mathrm{ML}}(\infty)$ appearing in the expression of the criterion will always be calculated using the *outer iteration* method and retaining all components in order to capture potential cross-effects of the choice of method and the application of rank reduction.

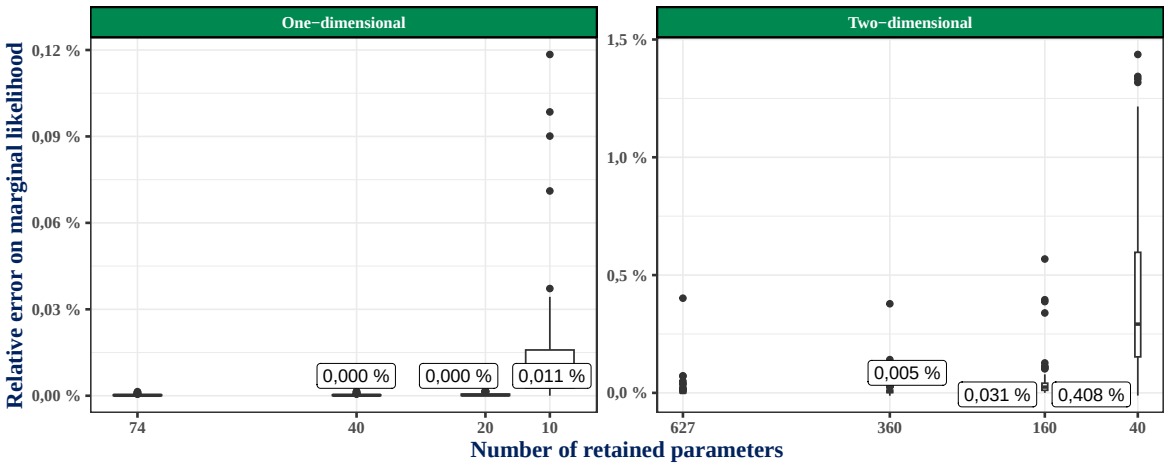

Figure 9: Distribution of the discrepancies induced by the proposed rank reduction method, depending on the number of retained eigenvectors during the selection of the smoothing parameter, in the one-dimensional and two-dimensional case.

Figure 9 represents the impact of rank reduction on the smoothing results as a function of the number of retained components. In the one-dimensional case, reducing the number of retained eigenvectors from 74 to 10 has only a marginal impact on the results obtained. However, as shown in Figure 10, it reduces the computation time by a factor of 3. In the two-dimensional case, reducing the number of components to 160 reduces the computation time by a factor close to 10, with imperceptible differences in the obtained results. The proposed rank reduction method, in combination with the *performance iteration* approach, thus allows, in the two-dimensional case, by considering only 160 components, to reduce the computation time by a factor of 40 at the cost of an average error of 0.03%, which exceeds 0.5% in only one of the 100 simulated datasets considered. This solution offers significant operational advantages, especially for applications mentioned in Section 6.1 where computation times can be much

more significant than in the analysis presented here.

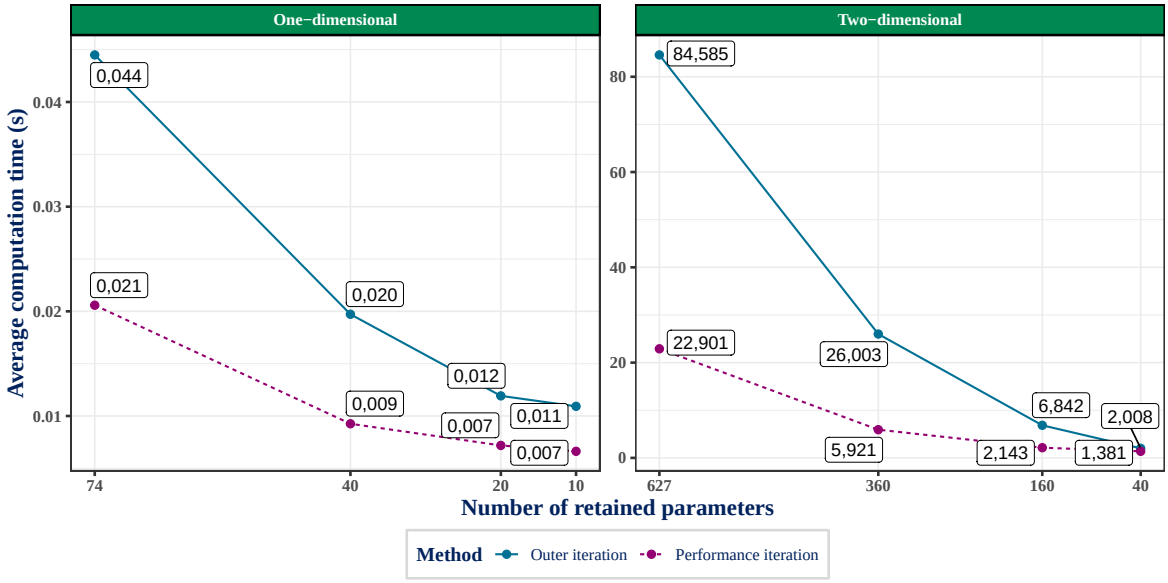

Figure 10: Distribution of computation time for the proposed rank reduction method, depending on the number of retained eigenvectors during the selection of the smoothing parameter, in the one-dimensional and two-dimensional case.

7 How to extrapolate the smoothing?

7.1 Defining the (unconstrained) extrapolation of the smoothing

Let $\hat{\mathbf{y}}$ be the result of Whittaker-Henderson smoothing for an observation vector $\mathbf{y}$, which corresponds in the one-dimensional case to an explanatory variable vector $\mathbf{x}$ or in the two-dimensional case to a combination of explanatory variables $\mathbf{x}$ and $\mathbf{z}$. Suppose we have an observation vector $\mathbf{x}_+$ or combinations of vectors $\mathbf{x}_+$ and $\mathbf{z}_+$ for which we want to make predictions using the model. Since Whittaker-Henderson smoothing applies only to evenly spaced observations, without loss of generality, let $\mathbf{x}_+$ and $\mathbf{z}_+$ be sequences of consecutive integers such that $\mathbf{x} \subset \mathbf{x}_+$ and $\mathbf{z} \subset \mathbf{z}_+$. In the one-dimensional case, let n_+ be the length of $\mathbf{x}_+$, and in the two-dimensional case, let n_{x+} and n_{z+} be the respective lengths of $\mathbf{x}_+$ and $\mathbf{z}_+$, with $n_+ = n_{x+} \times n_{z+}$. Let C_x (resp. C_z) be a matrix of size $n_x \times n_{x+}$ (resp. $n_z \times n_{z+}$) defined as $C_j = (0_{n_j,\min j - \min j_+} | I_{n_j,n_j} | 0_{n_j,\max j_+ - \max j})$ for $j \in \{x, z\}$, and let us note:

$$C = \begin{cases} C_x & \text{in the one-dimensional case} \\ C_z \otimes C_x & \text{in the two-dimensional case} \end{cases}$$

The matrix C of size $n \times n_+$ defined in this way has the following properties:

- For any vector $\mathbf{y}_+$ of size n_+, $C\mathbf{y}_+$ is a vector of size n containing only the n values of $\mathbf{y}_+$ matching the positions of the observations in the initial smoothing,
- $C^T\mathbf{y}$ is a vector of size n_+ containing the values of $\mathbf{y}$ at the positions corresponding to the initial observations and zeros elsewhere,
- $CC^T = I_n$.

The extrapolation of Whittaker-Henderson smoothing can now be defined as finding the solution $\hat{\mathbf{y}}_+$ to the extended optimization problem:

$$\hat{\mathbf{y}}_+ = \underset{\boldsymbol{\theta}_+}{\operatorname{argmin}} \left\{ (\mathbf{y}_+ - \boldsymbol{\theta}_+)^T W_+ (\mathbf{y}_+ - \boldsymbol{\theta}_+) + \boldsymbol{\theta}_+^T P_+ \boldsymbol{\theta}_+ \right\} \tag{22}$$

where $\mathbf{y}_+ = C^T\mathbf{y}$, $W_+ = C^T W C$, and

$$P_+ = \begin{cases} \lambda D_{n_+,q}^T D_{n_+,q} & \text{in the one-dimensional case} \\ \lambda_x I_{z+} \otimes D_{n_{x+},q_x}^T D_{n_{x+},q_x} + \lambda_z D_{n_{z+},q_z}^T D_{n_{z+},q_z} \otimes I_{x+} & \text{in the two-dimensional case} \end{cases}$$

In this last expression, λ, λ_x and λ_z are the parameters chosen for the smoothing. Similar to the initial smoothing, Equation 22 involves a fidelity criterion and a smoothness criterion. It should be noted that:

$$(\mathbf{y}_+ - \boldsymbol{\theta}_+)^T W_+ (\mathbf{y}_+ - \boldsymbol{\theta}_+) = (C^T\mathbf{y} - \boldsymbol{\theta}_+)^T C^T W C (C^T\mathbf{y} - \boldsymbol{\theta}_+) = (\mathbf{y} - \boldsymbol{\theta})^T W (\mathbf{y} - \boldsymbol{\theta})$$

hence the fidelity criterion remains unchanged compared to the initial estimation problem. This is consistent with the fact that extrapolation does not involve any additional data. The smoothness criterion, however, applies to all elements of the vector $\boldsymbol{\theta}_+$, whether initially present or not.

7.2 Unconstrained solution for the one-dimensional case

The solution to the extended optimization problem of Equation 22 is readily obtained by taking the derivatives in $\boldsymbol{\theta}_+$ and setting them to 0, as in Section 1.5. This yields the solution:

$$\hat{\mathbf{y}}_+ = (W_+ + P_+)^{-1} W_+ \mathbf{y}_+ \quad \text{where} \quad \mathbf{y}_+ = C^T\mathbf{y} \quad \text{and} \quad W_+ = C^T W C \tag{23}$$

Let us further assume that $\mathbf{y}_+ | \boldsymbol{\theta}_+ \sim \mathcal{N}(\boldsymbol{\theta}_+, W_+^-)$ where $\boldsymbol{\theta}_+ \sim \mathcal{N}(0, P_+^-)$ and procede as in Section 2 to obtain the following credible intervals:

$$\mathbb{E}(\mathbf{y}_+)|\mathbf{y}_+ \in \left[(W_+ + P_+)^{-1} W_+ \mathbf{y}_+ \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\mathbf{diag}\left\{(W_+ + P_+)^{-1}\right\}}\right] \tag{24}$$

To get a better understanding about how the variance-covariance matrix $\Psi_+ = (W_+ + P_+)^{-1}$ for the unconstrained extrapolation problem of Equation 22 is related to the variance-covariance matrix $\Psi = (W + P_\lambda)^{-1}$ of the original smoothing problem, let us introduce:

$$\overline{C}_j = \begin{bmatrix} I_{\min j_+ - \min j, \min j_+ - \min j} & 0_{\min j_+ - \min j, n_j} & 0_{\min j_+ - \min j, \max j_+ - \max j} \\ 0_{\max j_+ - \max j, \min j_+ - \min j} & 0_{\max j_+ - \max j, n_j} & I_{\max j_+ - \max j, \max j_+ - \max j} \end{bmatrix} \quad \text{for} \quad j \in \{x, z\},$$

$$\overline{C} = \begin{cases} \overline{C}_x & \text{in the one-dimensional case} \\ \overline{C}_z \otimes \overline{C}_x & \text{in the two-dimensional case} \end{cases} \quad \text{and} \quad Q = \begin{bmatrix} C \\ \overline{C} \end{bmatrix}.$$

With this definition, Q is a permutation matrix of size $n_+ \times n_+$ which select the rows whose indices correspond to those of the initial observation positions in the extended observation vector and put them first. It also verifies $Q^{-1} = Q^T = Q$.

In the unidimensional case, the extended difference matrix $D_{n_+,q}$ takes the form:

$$D_{n_+,q} = \begin{bmatrix} D_{2-} & D_{1-} & 0 \\ 0 & D_{n,q} & 0 \\ 0 & D_{1+} & D_{2+} \end{bmatrix} = Q \begin{bmatrix} D_{n,q} & 0 & 0 \\ D_{1-} & D_{2-} & 0 \\ D_{1+} & 0 & D_{2+} \end{bmatrix} Q = Q \begin{bmatrix} D_{n,q} & 0 \\ D_1 & D_2 \end{bmatrix} Q$$

$$\text{where} \quad D_1 = \begin{bmatrix} D_{1-} \\ D_{1+} \end{bmatrix} \quad \text{and} \quad D_2 = \begin{bmatrix} D_{2-} & 0 \\ 0 & D_{2+} \end{bmatrix}$$

Detailed expressions of matrices D_{1+} and D_{2+} for the most common values of q may be found in Carballo et al. (2021) where they are simply noted D_1 and D_2. To cover extrapolation on both sides of the initial observation vector, matrices D_{1-} and D_{2-} were introduced. Those may be simply obtained by taking the transpose of adequate size D_1 and D_2 matrices found in Carballo et al. (2021). The extended weight and penalization matrices thus take the form:

$$W_+ = Q \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} Q \quad \text{and} \quad P_+ = D_{n_+,q}^T D_{n_+,q} = \lambda Q \begin{bmatrix} P_\lambda + P_+^{11} & P_+^{12} \\ P_+^{21} & P_+^{22} \end{bmatrix} Q \tag{25}$$

where $P_+^{11} = \lambda D_1^T D_1$, $P_+^{12} = \lambda D_1^T D_2$, $P_+^{21} = \lambda D_2^T D_1$ and $P_+^{22} = \lambda D_2^T D_2$.

Applying the formulas for the inversion of a symmetric matrix partitioned with 2×2 blocks to $Q(W_+ + P_+)Q$, we obtain a more detailed expression for Ψ_+:

$$\Psi_+ = Q \begin{bmatrix} \Psi_+^{11} & -\Psi_+^{11} P_+^{12} (P_+^{22})^{-1} \\ -(P_+^{22})^{-1} P_+^{21} \Psi_+^{11} & (P_+^{22})^{-1} + (P_+^{22})^{-1} P_+^{21} \Psi_+^{11} P_+^{12} (P_+^{22})^{-1} \end{bmatrix} Q. \tag{26}$$

where $\Psi_+^{11} = [W + P_\lambda + P_+^{11} - P_+^{12}(P_+^{22})^{-1}P_+^{21}]^{-1}$.

Let us denote by *initial positions coefficients* the subvector $C\hat{\mathbf{y}}_+$ that correspond to the coefficients at the position of the initial observations in $\hat{\mathbf{y}}_+$ and by *new positions coefficients* the remaining coefficients $\overline{C}\hat{\mathbf{y}}_+$. The *initial positions coefficients* may be recovered as $C\hat{\mathbf{y}}_+ = C\Psi_+ C^T W_+ y_+ = \Psi_+^{11} Wy$. This does not simplify to $\Psi Wy = \hat{\mathbf{y}}$ for all weighted observation vectors Wy unless $\Psi_+^{11} = \Psi$. Indeed, the *initial positions coefficients* are chosen to optimize the overall smoothness of the extrapolated coefficient vectors and note just the smoothness of $C\hat{\mathbf{y}}_+$. Besides, the expression of Ψ_+^{22}, it contains two terms: an *innovation error* $(P_+^{22})^{-1}$ associated with the prior on the *new positions coefficients* while $(P_+^{22})^{-1}P_+^{21}\Psi_+^{11}P_+^{12}(P_+^{22})^{-1}$ represents an additional uncertainty on the *new positions coefficients* caused by the uncertainty on the *initial positions coefficients*.

In the one-dimensional case, D_2 is a block-diagonal matrix of triangular matrices with non-zero diagonal elements and is thus non-singular. Hence $P_+^{11} - P_+^{12}(P_+^{22})^{-1}P_+^{21} = D_1^T D_1 - D_1^T D_2 (D_2^T D_2)^{-1} D 2^T D_1 = 0$ which means that $\Psi_+^{11} = (W + P_\lambda)^{-1} = \Psi$ and $C\hat{\mathbf{y}}_+ = \hat{\mathbf{y}}$. Therefore, in the one-dimensional case, the solution given by Equation 22 preserves the values from the original fit. As shown by Carballo et al. (2021), it is indeed always possible in the case of penalizations based on difference matrices to pick the *new positions coefficients* so that the smoothness criterion does not increase or in other words to find a perfectly smooth extrapolation for the fit. Figure 11 shows the extrapolation associated with the Whittaker-Henderson smoothing applied to the data use to produce Figure 1.

7.3 Constrained solution for the two-dimensional case

In the two-dimensional case, while the extended penalization matrix P_+ still takes the form of Equation 25, expressions of P_+^{11}, P_+^{12}, P_+^{21} and P_+^{22} are more complex. In particular, $P_+^{11} - P_+^{12}(P_+^{22})^{-1}P_+^{21} \neq 0$ which implies that $\Psi_+^{11} \neq \Psi$ and $C\hat{\mathbf{y}}_+ \neq \hat{\mathbf{y}}$. Solving Equation 23 thus leads to a change in the value of the *initial positions coefficients* compared to the coefficients obtained during the initial smoothing as shown by Carballo, Durban, and Lee (2021). Indeed, the smoothness criterion includes penalizations on both rows and columns and it is no longer possible, as in the one-dimensional case, to extrapolate the fit without increasing that criterion. As the smoothness criterion carries more weight in the extrapolation problem compared to the original problem, the optimal solution to the extended optimization problem will compromise on the fidelity to the initial observations in order to improve the overall smoothness.

To obtain an estimator $\hat{\mathbf{y}}_+^*$ that minimizes the penalized regression problem under the constraint of preserving the initial coefficients, *i.e.* $C\hat{\mathbf{y}}_+^* = \hat{\mathbf{y}}$, we follow the approach proposed by Carballo,

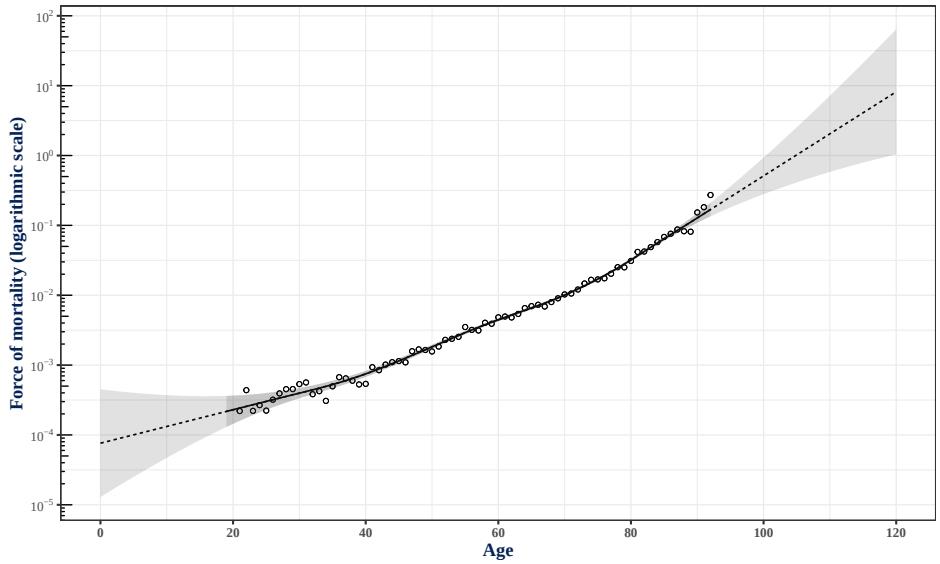

Figure 11: Extrapolation of Whittaker-Henderson smoothing in the one-dimensional case. The smoother extrapolation is achieved, on both sides of the smoothing, using a polynomial of degree q - 1 (in this case a straight line as q = 2).

Durban, and Lee (2021) and introduce the Lagrange multiplier $\boldsymbol{\omega}$. The associated constrained extended optimization problem is now written as:

$$(\hat{\mathbf{y}}_+^*, \hat{\boldsymbol{\omega}}) = \underset{\boldsymbol{\theta}_+^*, \boldsymbol{\omega}}{\operatorname{argmin}} \left\{ (\mathbf{y}_+ - \boldsymbol{\theta}_+^*)^T W_+ (\mathbf{y}_+ - \boldsymbol{\theta}_+^*) + \boldsymbol{\theta}_+^{*T} P_+ \boldsymbol{\theta}_+^* + 2\boldsymbol{\omega}^T (C\boldsymbol{\theta}_+^* - \hat{\mathbf{y}}) \right\}. \tag{27}$$

Taking the partial derivatives of Equation 27 with respect to $\boldsymbol{\theta}_+^*$ and $\boldsymbol{\omega}$ gives:

$$\frac{\partial}{\partial \boldsymbol{\theta}_+^*} \left\{ (\mathbf{y}_+ - \boldsymbol{\theta}_+^*)^T W_+ (\mathbf{y}_+ - \boldsymbol{\theta}_+^*) + \boldsymbol{\theta}_+^{*T} P_+ \boldsymbol{\theta}_+^* + 2\boldsymbol{\omega}^T (C\boldsymbol{\theta}_+^* - \hat{\mathbf{y}}) \right\} = -2W_+ (\mathbf{y}_+ - \boldsymbol{\theta}_+^*) + 2P_+ \boldsymbol{\theta}_+^* + 2\boldsymbol{\omega}^T C$$
$$\frac{\partial}{\partial \boldsymbol{\omega}} \left\{ (\mathbf{y}_+ - \boldsymbol{\theta}_+^*)^T W_+ (\mathbf{y}_+ - \boldsymbol{\theta}_+^*) + \boldsymbol{\theta}_+^{*T} P_+ \boldsymbol{\theta}_+^* + 2\boldsymbol{\omega}^T (C\boldsymbol{\theta}_+^* - \hat{\mathbf{y}}) \right\} = 2(C\boldsymbol{\theta}_+^* - \hat{\mathbf{y}})$$

Setting these derivatives to zero yields the linear system:

$$\begin{bmatrix} W_+ + P_+ & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{y}}_+^* \\ \hat{\boldsymbol{\omega}} \end{bmatrix} = \begin{bmatrix} W_+ \mathbf{y}_+ \\ \hat{\mathbf{y}} \end{bmatrix}$$

The solution for $\hat{\mathbf{y}}_+^*$ can be derived using formulas for the inversion of a symmetric partitioned matrix with 2×2 blocks:

$$\begin{aligned}\hat{\mathbf{y}}_+^* =&(W_+ + P_+)^{-1}\left\{I - C^T[C(W_+ + P_+)^{-1}C^T]^{-1}C(W_+ + P_+)^{-1}\right\}W_+\mathbf{y}_+ \\ &+ (W_+ + P_+)^{-1}C^T[C(W_+ + P_+)^{-1}C^T]^{-1}\hat{\mathbf{y}}\end{aligned}$$

Since $W_+ = C^T W C$, the first term is actually zero, and this expression simplifies to:

$$\hat{\mathbf{y}}_+^* = (W_+ + P_+)^{-1}C^T[C(W_+ + P_+)^{-1}C^T]^{-1}\hat{\mathbf{y}} = Q\begin{bmatrix} I \\ \Psi^{21}(\Psi^{11})^{-1}\end{bmatrix}\hat{\mathbf{y}} = Q\begin{bmatrix} I \\ -(P_+^{22})^{-1}P_+^{21}\end{bmatrix}\hat{\mathbf{y}} \tag{28}$$

which is a linear transformation of $\hat{\mathbf{y}}$. Defining $A_+^* = Q\begin{bmatrix} I \\ -(P_+^{22})^{-1}P_+^{21}\end{bmatrix}$, a natural candidate for the variance-covariance of $\mathbf{y}_+^*|\boldsymbol{\theta}_+$ is given by:

$$A_+^*\Psi A_+^{*T} = Q\begin{bmatrix} \Psi & -\Psi P_+^{12}(P_+^{22})^{-1} \\ -(P_+^{22})^{-1}P_+^{21}\Psi & (P_+^{22})^{-1}P_+^{21}\Psi P_+^{12}(P_+^{22})^{-1}\end{bmatrix}Q \tag{29}$$

Equation 29 is very similar to Equation 26 with however two differences. First, every occurence of Ψ_+^{11} is replaced by Ψ. This is consistent with the constraint that the *initial positions coefficients* are forced to take the value of the coefficients used during the initial smoothing. Second, as the solution to the constrained extended optimization problem of Equation 27 was expressed as a linear transformation of $\hat{\mathbf{y}}$, Equation 29 is missing the *innovation error* term $(P_+^{22})^{-1}$ associated with the prior on the *new positions coefficients*. Not including this term would be tantamount to considering that $\boldsymbol{\theta}_+$ has some degree of wigglyness in the region of the initial data but is perfectly smooth everywhere else. Adding the *innovation error* back, we obtain the following variance-covariance matrix for the constrained optimization problem:

$$\Psi_+^* = Q\begin{bmatrix} \Psi & -\Psi P_+^{12}(P_+^{22})^{-1} \\ -(P_+^{22})^{-1}P_+^{21}\Psi & (P_+^{22})^{-1} + (P_+^{22})^{-1}P_+^{21}\Psi P_+^{12}(P_+^{22})^{-1}\end{bmatrix}Q \tag{30}$$

which still verifies

$$\Psi_+^* W_+\mathbf{y}_+ = \Psi_+^* C W\mathbf{y} = Q\begin{bmatrix} I \\ -(P_+^{22})^{-1}P_+^{21}\end{bmatrix}\Psi W\mathbf{y} = \hat{\mathbf{y}}_+^*.$$

The associated credible intervals are readily obtained as:

$$\mathbb{E}(\mathbf{y}_+)|\mathbf{y}_+ \in \left[\hat{\mathbf{y}}_+^* \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\mathbf{diag}(\Psi_+^*)}\right]$$

with probability $1 - \frac{\alpha}{2}$, where Φ denotes the cumulative distribution function of the standard normal distribution.

Figure 12 represents the results of the constrained extrapolation presented in this section, with the associated standard deviation which accounts for the *innovation error*. It is based on the data used to generate Figure 2. Without the dotted lines marking the boundaries of the initial smoothing region, it would not be possible to tell from those plots where the extrapolation starts, which is precisely the goal of this procedure. Figure 13 represents the ratio between the mortality rates obtained from the unconstrained solution of Equation 23 and the constrained solution of Equation 28. The unconstrained solution shows significant discrepancies both in the initial smoothing region and the extrapolated region compared to the constrained solution. Finally, Figure 14 represents ratios between standard deviation derived using the three presented extrapolation methods. The denominator is the constrained extrapolation method which accounts for the *innovation error*, which is used as a reference, while the numerator shows the unconstrained method as well as the constrained method which ignores the *innovation error*. The constrained extrapolation method which includes the *innovation error* has a higher standard deviation compared to the unconstrained method as it is always either equally or less smooth.

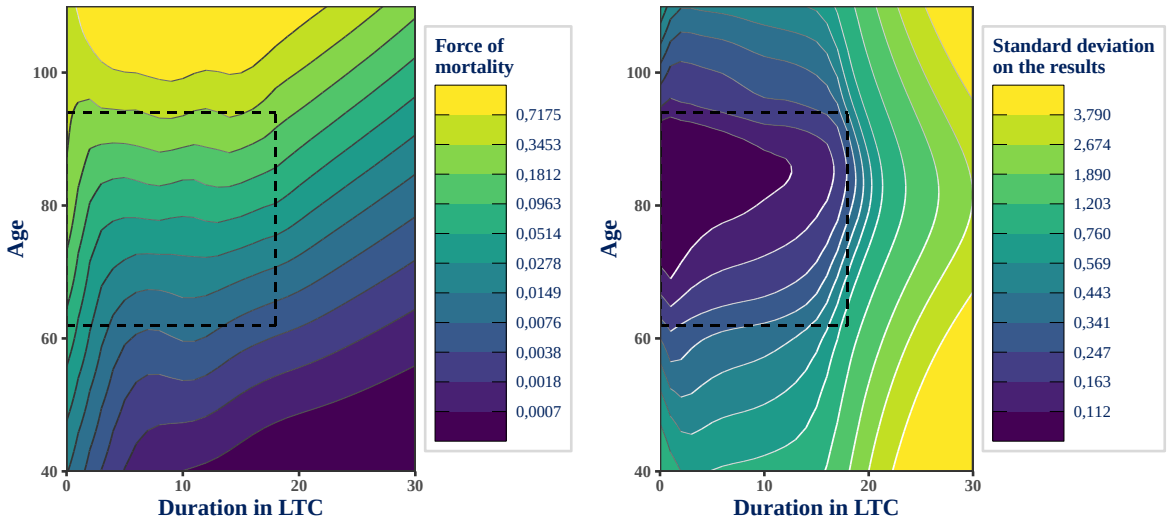

Figure 12: Constrained extrapolation of Whittaker-Henderson smoothing in the two-dimensional case. The contour lines of mortality rates, and the associated standard deviation, are depicted. The dotted lines delimitate the boundaries boundaries of the initial smoothing region.

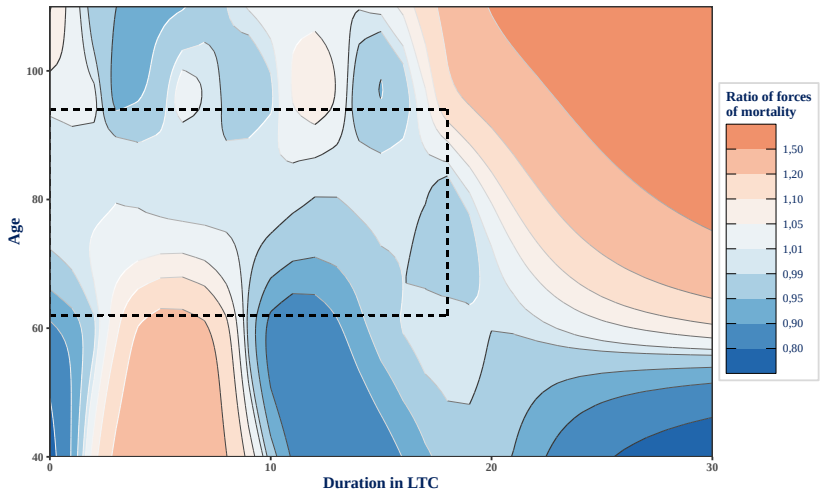

Figure 13: Ratio of mortality rates resulting from the extrapolation of the two-dimensional Whittaker-Henderson smoothing. The numerator corresponds to the unconstrained extrapolation and the denominator represents the constrained extrapolation presented in Figure 12.

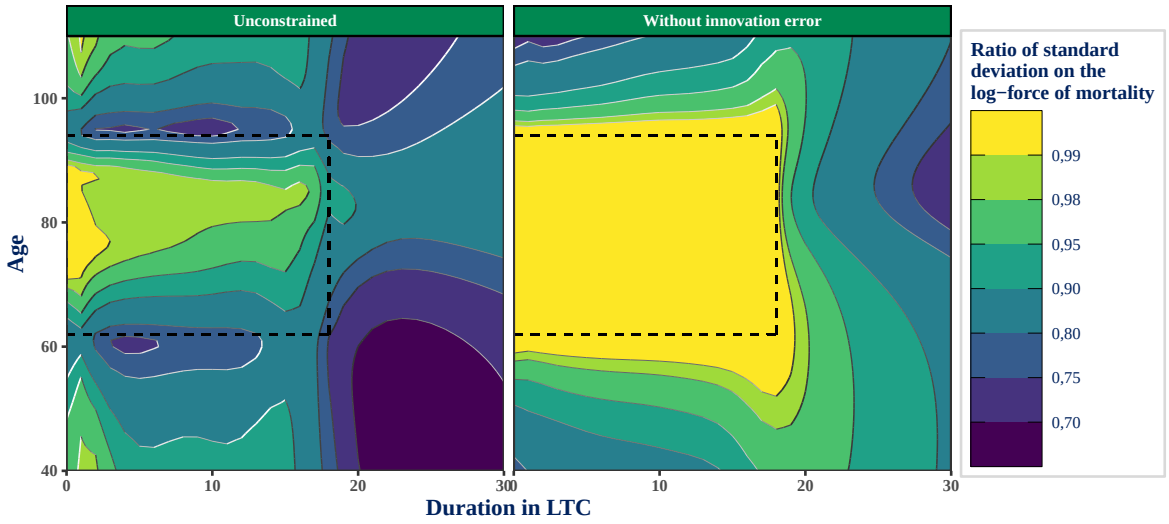

Figure 14: Ratio of standard deviation of log-mortality rates resulting from the extrapolation of the two-dimensional Whittaker-Henderson smoothing. The numerator on the left plot corresponds to the unconstrained extrapolation and on the right plot to the constrained extrapolation that does not account for the *innovation error*. The denominator represents the constrained extrapolation presented in Figure 12.

8 Discussion

In this paper, we have shown that Whittaker-Henderson smoothing, introduced in 1922, can be naturally interpreted in a Bayesian framework, which allows for the construction of credible intervals for it, provided that the observations are independent and their variances are known and used as weights. By linking it to the framework of duration models, we have demonstrated that in the construction of survival laws from experience, smoothing can be applied to the vector of raw exit rates using the observed number of exits as the weight vector. This is justified by the asymptotic properties of the maximum likelihood estimator of the raw exit rates. We have then established that the use of this asymptotic property comes at the cost of precision loss, especially when the number of observations is limited, and that a more precise iterative version of the smoothing can be proposed by directly solving penalized likelihood equations.

Furthermore, we have introduced an empirical Bayesian approach for selecting the smoothing parameter, based on maximizing a marginal likelihood function. The parameter selection introduces an iterative process that is added to the one generated by solving the penalized likelihood equations for a fixed smoothing parameter. In this case, there are two main possible approaches for parameter selection, depending on the order in which the iterations are nested: the *outer iteration* and the *performance iteration* approaches. The latter has significantly lower computation time compared to the former, at the expense of lacking theoretical convergence guarantees, which does not seem to pose a problem in practice. We have also introduced a rank reduction method based on the eigendecomposition of one-dimensional penalty matrices, which accelerates the selection of the smoothing parameter - our study on simulated data showed a 40-fold reduction in computation time - with negligible loss of precision.

Finally, we have addressed the issue of extrapolation of smoothing and shown that it requires solving a new optimization problem. Extrapolation is straightforward in the one-dimensional case. However, in the two-dimensional case constraints needs to be imposed in order to preserve the values of the coefficients obtained during the initial smoothing step.

Whittaker-Henderson smoothing has been used for nearly a century by actuaries, with little change in the approach. In parallel, statistical theory on smoothing methods has undergone numerous developments, particularly in the past 30 years with the emergence of generalized additive models, covered notably by Hastie (2017) and Wood (2006). Our goal in this paper was to bridge the gap between these two perspectives. We also created an R package named `WH` which implements all the steps mentioned through the paper and should be straightforward to use. The results derived in the paper are directly applicable to other types of smoothing, such as the widely used P-splines smoothing by Eilers and Marx (1996). Compared to Whittaker-Henderson smoothing, P-splines naturally consider fewer parameters than observations, removing the need for the method introduced in Section 6.

While Whittaker-Henderson smoothing is applicable to both the one-dimensional and two-dimensional case, it does have limitations. Firstly, it requires regularly spaced observations. This aligns well with the format of life insurance pricing and reserving assumptions, which

traditionally include age and duration spent in certain states that trigger indemnization. However, this can be a limit when the exits are not evenly distributed. For example, in the case of disability and long-term care risks, most exits occur in the first few months following entry into the state. In such situations, using a spline basis that is arranged to prioritize areas with more observations, as proposed by Wood (2017), could yield better results. Lastly, Whittaker-Henderson smoothing does not allow for the incorporation of additional explanatory variables in the experience table, starting with gender, which plays a major role in most biometric risks. Fortunately, it is possible to introduce these additional variables as random effects in the model by adopting a Smoothing Splines ANOVA approach, as described in Lee and Durban (2011) and Gu (2013), yielding results that are both precise and robust compared to a unisex or stratified model.

References

Akaike, Hirotsugu. 1973. "Information Theory and an Extension of the Maximum Likelihood Principle." In *2nd International Symposium on Information Theory, 1973*.

Anderssen, RS, and Peter Bloomfield. 1974. "A Time Series Approach to Numerical Differentiation." *Technometrics* 16 (1): 69–75.

Brent, Richard P. 1973. "Algorithms for Minimization Without Derivatives, Chap. 4." Prentice-Hall, Englewood Cliffs, NJ.

Carballo, Alba, Maria Durban, Göran Kauermann, and Dae-Jin Lee. 2021. "A General Framework for Prediction in Penalized Regression." *Statistical Modelling* 21 (4): 293–312.

Carballo, Alba, Maria Durban, and Dae-Jin Lee. 2021. "Out-of-Sample Prediction in Multidimensional p-Spline Models." *Mathematics* 9 (15): 1761.

Eilers, Paul H. C., and Brian D. Marx. 1996. "Flexible Smoothing with B-Splines and Penalties." *Statistical Science* 11 (2): 89–102.

Gu, Chong. 1992. "Cross-Validating Non-Gaussian Data." *Journal of Computational and Graphical Statistics* 1 (2): 169–79.

———. 2013. *Smoothing Spline ANOVA Models*. Vol. 297. Springer.

Hastie, Trevor J. 2017. "Generalized Additive Models." In *Statistical Models in s*, 249–307. Routledge.

Henderson, Robert. 1924. "A New Method of Graduation." *Transactions of the Actuarial Society of America* 25: 29–40.

Kauermann, Göran. 2005. "A Note on Smoothing Parameter Selection for Penalized Spline Smoothing." *Journal of Statistical Planning and Inference* 127 (1-2): 53–69.

Lee, Dae-Jin, and Maria Durban. 2011. "P-Spline ANOVA-Type Interaction Models for Spatio-Temporal Smoothing." *Statistical Modelling* 11 (1): 49–69.

Marra, Giampiero, and Simon N Wood. 2012. "Coverage Properties of Confidence Intervals for Generalized Additive Model Components." *Scandinavian Journal of Statistics* 39 (1): 53–74.

Nelder, John A, and Roger Mead. 1965. "A Simplex Method for Function Minimization." *The Computer Journal* 7 (4): 308–13.

Nelder, John Ashworth, and Robert WM Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society: Series A (General)* 135 (3): 370–84.

Patterson H. D., Thompson R. 1971. "Recovery of Inter-Block Information When Block Sizes Are Unequal." *Biometrika* 58: 545–54.

Reiss, Philip T, and R Todd Ogden. 2009. "Smoothing Parameter Selection for a Class of Semiparametric Linear Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71 (2): 505–23.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics*, 461–64.

Wahba, Grace. 1980. *Spline Bases, Regularization, and Generalized Cross Validation for Solving Approximation Problems with Large Quantities of Noisy Data*. University of WISCONSIN.

———. 1985. "A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem." *The Annals of Statistics*, 1378–1402.

Whittaker, Edmund T. 1922. "On a New Method of Graduation." *Proceedings of the Edinburgh Mathematical Society* 41: 63–75.

Wood, Simon N. 2006. *Generalized Additive Models: An Introduction with r*. chapman; hall/CRC.

———. 2011. "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (1): 3–36.

———. 2017. "P-Splines with Derivative Based Penalties and Tensor Product Smoothing of Unevenly Distributed Data." *Statistics and Computing* 27: 985–89.