Astin Bulletin (2026), **56**, pp. 1–31
doi:10.1017/asb.2025.10061

International Actuarial Association
Association Actuarielle Internationale

RESEARCH ARTICLE

Whittaker–Henderson smoothing revisited: A modern statistical framework for practical use

Guillaume Biessy, PhD[1,2]

[1]LinkPact, Paris, 75015, France
[2]Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, Paris, 75005, France
Email: guillaume.biessy78@gmail.com

Received: 17 October 2024; **Revised:** 23 July 2025; **Accepted:** 24 July 2025; **First published online:** 3 September 2025

Keywords: smoothing; duration models; experience tables; maximum likelihood; generalized additive models; empirical Bayes approach; marginal likelihood; LAML; extrapolation

Abstract

Introduced over a century ago, Whittaker–Henderson smoothing remains widely used by actuaries in constructing one-dimensional and two-dimensional experience tables for mortality, disability, and other life insurance risks. In this paper, we reinterpret this smoothing technique within a modern statistical framework and address six practically relevant questions about its use. First, we adopt a Bayesian perspective on this method to construct credible intervals. Second, in the context of survival analysis, we clarify how to choose the observation and weight vectors by linking the smoothing technique to a maximum likelihood estimator. Third, we improve accuracy by relaxing the method's reliance on an implicit normal approximation. Fourth, we select the smoothing parameters by maximizing a marginal likelihood function. Fifth, we improve computational efficiency when dealing with numerous observation points and consequently parameters. Finally, we develop an extrapolation procedure that ensures consistency between estimated and predicted values through constraints.

Notations

In this paper, vectors are denoted in boldface 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 associated with $\mathbf{y}$, $\mathbf{diag}(A)$ represents the diagonal of matrix A, and $\mathrm{Diag}(\mathbf{y})$ is the diagonal matrix such that $\mathbf{diag}(\mathrm{Diag}(\mathbf{y})) = \mathbf{y}$. The sum of the diagonal elements of A is denoted as $\mathrm{tr}(A)$ and its transpose as A^T. In the case where A is invertible, A^{-1} denotes its inverse and $|A|$ denotes the product of the eigenvalues of A. For a non-invertible matrix A, A^- refers to the Moore–Penrose pseudo-inverse 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 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 leaving the zero eigenvalues unchanged, the pseudo-inverse 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 (element-wise) product is denoted as $A \odot B$. $\lfloor x \rfloor$ denotes the greatest integer less than or equal to $x \in \mathbb{R}$. Finally, the symbol $\propto$ denotes proportionality between the expressions on both sides.

1. Introduction

Whittaker–Henderson (WH) smoothing is a graduation method designed to mitigate the effects of sampling fluctuations in a vector of evenly spaced discrete observations. Although this method was originally proposed by Bohlmann (1899), it is named after Whittaker (1923), who applied it to graduate mortality tables, and Henderson (1924), who popularized it among actuaries in the United States. The method

© The Author(s), 2025. Published by Cambridge University Press on behalf of The International Actuarial Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

was later extended to two dimensions by Knorr (1984). WH smoothing may be used to build experience tables for a broad spectrum of life insurance risks, such as mortality, disability, long-term care, lapse, mortgage default, and unemployment. We begin with a brief overview of the method before outlining the structure and main contributions of the paper.

1.1. A brief reminder of WH smoothing mathematical formulation

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 WH smoothing is given by

$$\hat{\mathbf{y}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\{F(\mathbf{y}, \mathbf{w}, \boldsymbol{\theta}) + R_{\lambda,q}(\boldsymbol{\theta})\} \tag{1.1}$$

where:

- $F(\mathbf{y}, \mathbf{w}, \boldsymbol{\theta}) = \sum_{i=1}^{n} w_i (y_i - \theta_i)^2$ represents a fidelity criterion with respect 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, $\lambda \geq 0$ is a smoothing parameter and Δ^q denotes the forward difference operator of order q, such that for any $i \in \{1, \dots, n-q\}$:

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

Define $W = \operatorname{Diag}(\mathbf{w})$, 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 first- and second-order difference matrices are given by

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

while higher-order difference matrices follow the recursive formula $D_{n,q} = D_{n-1,q-1} D_{n,1}$. The fidelity and smoothness criteria can be rewritten with matrix notations as

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

and the WH smoothing estimator thus becomes

$$\hat{\mathbf{y}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ (\mathbf{y} - \boldsymbol{\theta})^T W (\mathbf{y} - \boldsymbol{\theta}) + \boldsymbol{\theta}^T P_\lambda \boldsymbol{\theta} \right\} \tag{1.2}$$

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

The two-dimensional case

In the two-dimensional case, consider a matrix Y of observations and a matrix Ω of non-negative weights, both of dimensions $n_x \times n_z$. The WH smoothing estimator solves:

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

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 with respect to the observations,

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

- $R_{\lambda,q}(\Theta) = \lambda_x \sum_{j=1}^{n_z} \sum_{i=1}^{n_x - q_x} (\Delta^{q_x} \Theta_{\bullet,j})_i^2 + \lambda_z \sum_{i=1}^{n_x} \sum_{j=1}^{n_z - q_z} (\Delta^{q_z} \Theta_{i,\bullet})_j^2$ is a smoothness criterion with $\lambda = (\lambda_x, \lambda_z)$.

This latter criterion adds row-wise and column-wise regularization criteria to Θ, with respective orders q_x and q_z, weighted by non-negative smoothing parameters λ_x and λ_z. In matrix notation, let $\mathbf{y} = \mathbf{vec}(Y)$, $\mathbf{w} = \mathbf{vec}(\Omega)$, and $\boldsymbol{\theta} = \mathbf{vec}(\Theta)$ as the vectors obtained by stacking the columns of the matrices Y, Ω, and Θ, respectively. Additionally, denote $W = \mathrm{Diag}(\mathbf{w})$ and $n = n_x \times n_z$. The fidelity and smoothness criteria become

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

and the associated estimator also takes the form of Equation 1.2 except in this case

$$P_\lambda = \lambda_x I_{n_z} \otimes D_{n_x,q_x}^T D_{n_x,q_x} + \lambda_z D_{n_z,q_z}^T D_{n_z,q_z} \otimes I_{n_x}.$$

Extension to higher dimensions is straightforward and not discussed here.

An explicit solution

If $W + P_\lambda$ is invertible, Equation (1.2) admits the closed-form solution:

$$\hat{\mathbf{y}} = (W + P_\lambda)^{-1} W \mathbf{y}. \tag{1.3}$$

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

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

It follows that $(W + P_\lambda)\hat{\mathbf{y}} = W\mathbf{y}$, proving Equation (1.3). If $\lambda \neq 0$, $W + P_\lambda$ 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 spread across q_x different rows and q_z different columns in the two-dimensional case. These conditions are always met in real datasets.

1.2. Structure of the paper

Introduced a century ago, WH smoothing remains widely used by actuaries, particularly in France and North America (Canadian Institute of Actuaries, 2017; Society of Actuaries, 2018). Other non-parametric smoothing methods have since emerged, notably spline-based techniques (Reinsch, 1967), which gained even greater popularity with P-splines (Eilers and Marx, 1996). A broader overview of alternative smoothers is available in Wood (2017, chap. 5).

For evenly spaced discrete observations, WH smoothing may be considered a particular case of P-splines with degree-zero splines and identity model matrix. Its appeal lies in its simplicity: no selection of knots, parameters equal to fitted values, and shape controlled solely via penalization. However, it involves more parameters than low-rank smoothers, making it more computationally intensive.

Originally proposed as an empirical alternative to polynomial regression and weighted averages, WH smoothing offered key benefits noted by Whittaker (1923): first q moment preservation, adjustable smoothing parameters, and robustness at boundaries. While smoothing theory has evolved – particularly via generalized additive models (Hastie and Tibshirani, 1990), use of WH smoothing by actuaries remains largely unchanged. This paper reinterprets WH within modern statistical theory to bridge that gap and address six practical questions, each discussed in a dedicated section.

How to measure uncertainty in smoothing results?

We propose a method to quantify the uncertainty in WH smoothing based on data volume, a topic that has received little attention in the literature. In a frequentist framework, the WH estimator is biased, which

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

complicates the construction of valid confidence intervals for finite samples. However, under certain conditions, WH smoothing can be viewed as a Bayesian model, enabling the derivation of credible intervals. This Bayesian interpretation was originally suggested by Whittaker (1923) as a justification for the method and formally revisited decades later by Taylor (1992). In this section, we build on that equivalence to derive credible intervals for WH smoothing.

Which observation and weight vectors to use?

For the Bayesian interpretation of WH smoothing discussed in Section 2 to hold, it must be applied to a vector $\mathbf{y}$ of independent, normally distributed observations with known variances. The weight vector $\mathbf{w}$ should then contain the inverse variances (up to a constant), as noted by Taylor (1992) and Verrall (1993). We show that, under piecewise constant transition intensities in duration models, the maximum likelihood estimator of crude rates produces vectors $(\mathbf{y}, \mathbf{w})$ that asymptotically meet these conditions. This, combined with the results from the previous section, offers a statistical foundation for the use of WH smoothing in constructing experience tables for life insurance risks.

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

The standard approach applies WH smoothing to crude rate estimates, assuming they are asymptotically normal. However, this assumption often breaks down in practice when data are limited, making the method unreliable in such cases. Following Verrall (1993), we propose a generalization of WH smoothing that replaces the two-step procedure with the direct maximization of a penalized log-likelihood. Instead of smoothing pre-estimated rates, this method works directly with aggregated event and exposure counts. The estimation is performed iteratively using the PIRLS algorithm. We evaluate both methods on simulated datasets reflecting typical life insurance portfolios. Results show that, in smaller samples, the normal approximation in the traditional method introduces notable bias. This supports the use of the generalized approach – based on penalized log-likelihood – as a more robust alternative when data are limited.

How to select the smoothing parameters?

We now turn to the crucial choice of the smoothing parameter λ, which has long been left to actuarial judgement. Giesecke and Center (1981) suggested choosing λ so that the variance of the smoothed results matches the average variance of a Chi-square statistic, but uses $n - q$ as degrees of freedom, thus ignoring the reduction in effective model dimension due to penalization. Brooks *et al.* (1988) minimized the global cross-validation criterion introduced by Wahba (1980), though this can result in severe undersmoothing as noted by Wood (2011).

We instead propose to select λ by maximizing a marginal likelihood function, a method first introduced by Patterson and Thompson (1971) and later applied to smoothing parameter selection by Anderssen and Bloomfield (1974). This approach is consistent with the Bayesian framework discussed earlier and performs well in small samples, as shown by Reiss and Todd Ogden (2009). This marginal likelihood function has a closed-form expression and can be maximized numerically. For the proposed generalization of WH smoothing, the marginal likelihood is no longer available in closed form. Instead, we rely on the Laplace approximation of the marginal likelihood (LAML), which can be maximized numerically. As both solving likelihood equations and selecting the optimal smoothing parameter are iterative processes, we explore different ways of nesting these iterations. We compare three nesting strategies combined with three numerical optimization algorithms for maximizing the marginal likelihood or LAML. Simulation results show that all strategies have near-optimal accuracy, with the fastest performance achieved using the outer iteration strategy combined with the Newton algorithm

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

How to improve smoothing computational efficiency?

When the number of observations – and thus parameters – is large, the computational cost of WH smoothing becomes a major challenge. This is particularly relevant in actuarial contexts, such as smoothing two-dimensional tables for disability or long-term care modelling. Beyond actuarial applications, WH smoothing is also widely used in economics for long time series, where it is known as the Hodrick-Prescott filter (Hodrick and Prescott, 1997). Although fast algorithms have been developed to exploit the structure of the penalization matrix (e.g., Weinert, 2007; Cornea-Madeira, 2017) they are typically limited to the one-dimensional case and cannot be directly extended to two dimensions.

After briefly outlining the main computational steps of (generalized) WH smoothing-including smoothing parameter selection via marginal likelihood or LAML-and their leading-order costs, we introduce two complementary strategies to reduce the computational burden:

1. Banded matrix exploitation: WH smoothing involves model and penalization matrices with banded structure. Taking advantage of this structure greatly accelerates key computations.
2. Reduced-rank basis via natural parametrization: Building on the work of Demmler and Reinsch (1975), we apply an eigendecomposition to the one-dimensional penalization matrices and drop components associated with the largest eigenvalues, which reduces the problem size. In two dimensions, we further improve efficiency using the generalized linear array model (GLAM) framework (Currie *et al*., 2006) which leverages the rectangular shape of the data.

In the two-dimensional case, we compare these strategies with a cubic P-spline alternative using simulated datasets. Results show that the banded implementation reduces computation time by up to a factor of 25. The reduced-rank approach brings further gains – up to a factor of 250 – at the cost of a slight reduction in accuracy. Its performance is comparable to P-spline smoothing with a cubic basis of similar size.

How to extrapolate smoothing results?

We conclude by addressing how to extrapolate smoothing results. Semi-parametric models like WH and P-splines can extrapolate beyond the observed data – similar to parametric models – but this feature is often overlooked in actuarial practice. The existing literature is limited and mostly focused on mortality forecasting.

Currie *et al*. (2004) uses P-splines to fit and forecast mortality rates by treating the extrapolated positions as zero-weight observations (see also Delwarde *et al*., 2007; Currie, 2013). While this works well in one dimension, Carballo *et al*. (2021) showed that it distorts the fit in two dimensions. To fix this, they proposed adding constraints to preserve the values that would result from fitting the observed data alone.

However, their approach to confidence intervals overlooks potential innovation error beyond the observed data, effectively treating the extrapolated process as perfectly smooth. In contrast, we propose an approach that derives credible intervals for extrapolated values, accounting for the underlying variability beyond the observed data range.

2. How to measure uncertainty in smoothing results?

The explicit solution given by Equation (1.3) 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 confidence intervals for finite samples centred on $\mathbb{E}(\mathbf{y})$. Therefore, in this section, we turn to a Bayesian framework where smoothing can be interpreted more naturally.

2.1. Maximum a posteriori estimate

Suppose that $\mathbf{y} \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 W^-)$ and $\boldsymbol{\theta} \sim \mathcal{N}(0, \sigma^2 P_\lambda^-)$ for some $\sigma > 0$. The Bayes formula allows us to express the posterior likelihood $f(\boldsymbol{\theta} \mid \mathbf{y})$ associated with these choices in the following form:

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

$$f(\boldsymbol{\theta} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2\sigma^2}\left[(\mathbf{y}-\boldsymbol{\theta})^T W(\mathbf{y}-\boldsymbol{\theta}) + \boldsymbol{\theta}^T P_\lambda \boldsymbol{\theta}\right]\right).$$

Hence the mode of the posterior distribution, $\hat{\boldsymbol{\theta}} = \operatorname{argmax}[f(\boldsymbol{\theta} \mid \mathbf{y})]$, also known as the *maximum a posteriori* (MAP) estimate, coincides with the solution $\hat{\mathbf{y}}$ from Equation (1.2), whose explicit form is given by Equation (1.3).

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

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

$$\ln f(\boldsymbol{\theta} \mid \mathbf{y}) = \ln f(\hat{\boldsymbol{\theta}} \mid \mathbf{y}) + \left.\frac{\partial \ln f(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta}}\right|^T_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T \left.\frac{\partial^2 \ln f(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \tag{2.1}$$

$$\text{where} \quad \left.\frac{\partial \ln f(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0 \quad \text{and} \quad \left.\frac{\partial^2 \ln f(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\frac{1}{\sigma^2}(W + P_\lambda).$$

As this last derivative no longer depends on $\boldsymbol{\theta}$, higher-order derivatives are all zero. The Taylor expansion allows for an exact computation of $\ln f(\boldsymbol{\theta} \mid \mathbf{y})$. Substituting the result back into Equation (2.1) yields:

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

which can immediately be recognized as the density of the $\mathcal{N}(\hat{\boldsymbol{\theta}}, \sigma^2(W + P_\lambda)^{-1})$ distribution.

2.3. Consequence for the WH smoothing

The prior $\boldsymbol{\theta} \sim \mathcal{N}(0, \sigma^2 P_\lambda^-)$ provides a Bayesian interpretation of the smoothness penalty, expressing an (improper) prior belief about the structure of $\mathbf{y}$.

This Bayesian framework and the resulting credible intervals rely on the assumption that $\mathbf{y} \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 W^-)$, meaning that the components of $\mathbf{y}$ are independent with known variances (up to a constant σ^2). The weight vector $\mathbf{w}$ must then be proportional to the inverse variances, not chosen empirically. If σ^2 is known, $100(1-\alpha)\%$ credible intervals take the form:

$$\mathbb{E}(\mathbf{y}) \mid \mathbf{y} \in \left[\hat{\mathbf{y}} \pm \Phi^{-1}(1-\alpha/2)\sqrt{\sigma^2 \mathbf{diag}\left\{(W + P_\lambda)^{-1}\right\}}\right] \tag{2.2}$$

where $\hat{\mathbf{y}} = (W + P_\lambda)^{-1} W\mathbf{y}$ and Φ is the cumulative distribution function for the standard normal distribution. According to Marra and Wood (2012), such intervals have good Frequentist coverage.

If σ^2 is unknown, it can be estimated as

$$\hat{\sigma}^2 = \frac{(\mathbf{y}-\hat{\mathbf{y}})^T W(\mathbf{y}-\hat{\mathbf{y}})}{n - \operatorname{tr}(H)} \quad \text{where} \quad H = (W + P_\lambda)^{-1} W.$$

In that case, σ^2 is replaced by $\hat{\sigma}^2$ and the normal distribution in Equation (2.2) by the Student t - distribution with $n - \operatorname{tr}(H)$ degrees of freedom.

3. Which observation and weight vectors to use?

Section 2 highlighted that WH smoothing may be interpreted in a robust statistical framework when applied to a vector $\mathbf{y}$ of independent, normally distributed observations with known variances, and a weight vector $\mathbf{w}$ proportional to the inverses of those variances. In this section, we propose, within the

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

framework of duration models used for constructing experience tables for life insurance risks, vectors $\mathbf{y}$ and $\mathbf{w}$ that satisfy these conditions.

3.1. Survival analysis framework

We consider a longitudinal follow-up of m individuals, subject to left truncation and non-informative right censoring, and aim to estimate a distribution governed by a continuous explanatory variable x (e.g. age). Let μ denote the hazard function, also known as the force of mortality in the study of the death risk. Under standard survival analysis assumptions, the log-likelihood takes the following continuous-time form:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} \left[\delta_i \ln \mu(x_i + t_i, \boldsymbol{\theta}) - \int_{u=0}^{t_i} \mu(x_i + u, \boldsymbol{\theta}) \mathrm{d}u \right]. \tag{3.1}$$

Here x_i is the age at the start of observation, t_i is the follow-up duration for individual i and δ_i is an event indicator: 1 if the event is observed and 0 if censored.

Although model estimation can be based on direct maximization of Equation (3.1), this approach scales poorly with large m and generally requires numerical integration – except in simple parametric cases. We instead adopt a discrete approximation by assuming the hazard rate is piecewise constant over one-year intervals:

$$\mu(x + \epsilon) = \mu(x) \quad \text{for all} \quad x \in \mathbb{N}, \epsilon \in [0, 1).$$

Under this assumption, the log-likelihood simplifies to a sum over discrete ages:

$$\ell(\boldsymbol{\theta}) = \sum_{x=x_{\min}}^{x_{\max}} \ln \mu(x, \boldsymbol{\theta}) d(x) - \mu(x, \boldsymbol{\theta}) e_c(x). \tag{3.2}$$

Here $d(x)$ is the number of observed events at age x and $e_c(x)$ is the central exposure to risk, that is, the total duration individuals are observed at age x.

This discretization, first introduced by Hoem (1971), is widely used in actuarial science. Its advantages are underlined for example in Gschlössl *et al*. (2011). It extends naturally to the two-dimensional case by assuming $\mu(x + \epsilon, z + \xi) = \mu(x, z)$ and summing over (x, z) pairs.

Details on the derivation of Equations (3.1) and (3.2), along with the computation of central exposures and event counts, are provided in Section A of the Supplementary Materials.

3.2. Likelihood equations

Assuming one parameter per observation and using the exponential link $\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{exp}(\boldsymbol{\theta})$, we recover the crude rates estimator, which models each age (or age pair) independently. The exponential link ensures positive hazard rates. The log-likelihood, in both one- and two-dimensional cases, takes the vectorized form:

$$\ell(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{d} - \mathbf{exp}(\boldsymbol{\theta})^T \mathbf{e_c} \tag{3.3}$$

where $\mathbf{d}$ and $\mathbf{e}_c$ are the vectors of observed deaths and central exposures.

The derivatives of this likelihood are

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \mathbf{d} - \mathbf{exp}(\boldsymbol{\theta}) \odot \mathbf{e_c} \quad \text{and} \quad \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = -\mathrm{Diag}(\mathbf{exp}(\boldsymbol{\theta}) \odot \mathbf{e_c}). \tag{3.4}$$

These equations correspond to those of a Poisson GLM (Nelder and Wedderburn, 1972) with mean $\boldsymbol{\mu}(\boldsymbol{\theta}) \odot \mathbf{e}_c$, although derived under different assumptions.

The model admits the closed-form solution $\hat{\boldsymbol{\theta}} = \ln(\mathbf{d}/\mathbf{e}_c)$. Under standard regularity conditions, the maximum likelihood estimator satisfies $\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, W_{\hat{\boldsymbol{\theta}}}^{-1})$, with $W_{\hat{\boldsymbol{\theta}}} = \mathrm{Diag}(\mathbf{d})$.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

Notably, this asymptotic approximation depends on the number of individuals m and not the dimension n of the aggregated vectors.

3.3. Consequence for the WH smoothing

We conclude that, under the duration model framework and using crude rates, the log-estimate $\ln(\mathbf{d}/\mathbf{e}_c)$ is asymptotically normal:

$$\ln(\mathbf{d}/\mathbf{e}_c) \sim \mathcal{N}(\ln \boldsymbol{\mu}, W^{-1}) \quad \text{with} \quad W = \mathrm{Diag}(\mathbf{d}).$$

This justifies applying WH smoothing to the observation vector $\mathbf{y} = \ln(\mathbf{d}/\mathbf{e}_c)$ with weight vector $\mathbf{w} = \mathbf{d}$. Using results from Section 2, and $\sigma^2 = 1$, the credible intervals for $\ln \boldsymbol{\mu}$ are

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

with $\hat{\boldsymbol{\theta}} = (W + P_\lambda)^{-1} W(\ln \mathbf{d} - \ln \mathbf{e}_c)$. Credible intervals for $\boldsymbol{\mu}$ itself are then obtained by exponentiating the bounds.

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

4.1. Generalized Whittaker–Henderson smoothing

The approach described in Section 3.2 assumes that the crude rates estimator is asymptotically normal, justifying the application of WH smoothing to its logarithm. However, with limited data, this approximation may introduce significant bias. We therefore propose an alternative based directly on the exact likelihood in Equation (3.3). Applying the Bayesian framework from Section 2 and assuming $\boldsymbol{\theta} \sim \mathcal{N}(0, P_\lambda^-)$, Bayes' theorem gives

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

We define the penalized log-likelihood as $\ell_P(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \boldsymbol{\theta}^T P_\lambda \boldsymbol{\theta}/2$. The maximum a posteriori estimate is the maximizer of ℓ_P.

Using a second-order Taylor expansion of the posterior log-likelihood around $\hat{\boldsymbol{\theta}}$ leads to the Laplace approximation:

$$f(\boldsymbol{\theta} \mid \mathbf{d}, \mathbf{e_c}) \approx \mathcal{N}(\hat{\boldsymbol{\theta}}, (W_{\hat{\theta}} + P_\lambda)^{-1}) \tag{4.1}$$

where $W_{\hat{\theta}} = \mathrm{Diag}(\mathbf{exp}(\hat{\boldsymbol{\theta}}) \odot \mathbf{e_c})$. Unlike the normal case studied in Section 2, the higher-order derivatives of the posterior log-likelihood are not zero, and Equation (4.1) only provides an approximation of the posterior log-likelihood, which yields asymptotic credible intervals:

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

Unlike the closed-form estimator in Equation (3.4), no analytical solution for $\hat{\boldsymbol{\theta}}$ exists here. We solve numerically using Newton's algorithm, which iteratively updates:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + (W_k + P_\lambda)^{-1}(\mathbf{d} - \mathbf{exp}(\boldsymbol{\theta}_k) \odot \mathbf{e_c} - P_\lambda \boldsymbol{\theta}_k)]$$

with $W_k = \mathrm{Diag}(\mathbf{exp}(\boldsymbol{\theta}_k) \odot \mathbf{e_c})$. The update can be rewritten as

$$\boldsymbol{\theta}_{k+1} = (W_k + P_\lambda)^{-1} W_k \mathbf{z}_k \quad \text{where} \quad \mathbf{z}_k = \boldsymbol{\theta}_k + W_k^{-1}[\mathbf{d} - \mathbf{exp}(\boldsymbol{\theta}_k) \odot \mathbf{e_c}].$$

Initializing with the crude rates estimator $\boldsymbol{\theta}_0 = \ln(\mathbf{d}/\mathbf{e_c})$ implies $W_0 = \mathrm{Diag}(\mathbf{d})$ and $z_0 = \ln(\mathbf{d}/\mathbf{e_c})$, so the first iteration recovers the classical WH smoothing result.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

Table 1. *Key figures associated with the 6 simulated datasets.*

Portolio type	Dimensions	Head count	Exposure count	Death count
Annuity	45	20,000	136,524	1,722
Annuity	45	100,000	679,728	8,452
Annuity	45	500,000	3,405,892	42,499
LTC	30×15	20,000	8,115	1,888
LTC	30×15	100,000	40,004	9,281
LTC	30×15	500,000	202,666	47,358

Subsequent iterations refine the observation and weight vectors. This process can thus be interpreted as an iterative generalization of WH smoothing, akin to how generalized linear models extend linear models.

We refer to this method as generalized WH smoothing. The iterative estimation algorithm described above corresponds to the penalized iteratively reweighted least squares (PIRLS) algorithm, widely used for fitting generalized additive models.

This framework naturally extends to other exponential family distributions, such as the binomial case suggested in Verrall (1993), by adapting the likelihood, link function, weight matrix, and working vector. However, we advocate for the Poisson-like likelihood of Equation (3.3), which offers several advantages: it generalizes to competing risks, supports multiplicative covariate effects via the log link and allows the use of an external reference table as a multiplicative offset.

4.2. Impact of the normal approximation in the original smoothing

As discussed in Section 3, classical WH smoothing can be viewed as an approximation to a penalized likelihood maximization, relying on a crude rate estimator assumed to be asymptotically normal. To assess the practical consequences of this approximation, we conduct an empirical comparison based on six simulated datasets reflecting the typical structure and volume of real insurance portfolios:

- The first three datasets simulate annuity portfolios with 20,000, 100,000, and 500,000 policyholders. The sole covariate is age, ranging from 50 to 95.
- The next three mimic long-term care (LTC) portfolios of the same sizes. Modelling of LTC typically relies on the illness-death model (Fix and Neyman, 1951; Clifford, 1977). To get a two-dimensional illustration we focus on the transition between the disabled and dead states (the two other transitions would provide additional one-dimensional examples). Two covariates are used: age (70–100) and duration in LTC (0–15 years).

Each dataset consists of individual-level longitudinal data, from which we derive event counts $\mathbf{d}$ and exposures $\mathbf{e}_c$, aggregated by age x (for annuities) of by (x, z) pairs (for LTC). All datasets within each group share the same underlying structure and differ only in size. Key dataset statistics are provided in Table 1 and additional details about how those datasets were generated are provided in Section B of the Supplementary Materials.

We apply two methods:

1. Original WH smoothing using $\mathbf{y} = \ln(\mathbf{d}/\mathbf{e}_c)$ and weights $\mathbf{w} = \mathbf{d}$ as in Section 3.
2. Generalized WH smoothing, using the likelihood formulation of Section 4.

Both methods use the same smoothing parameter(s) λ, to ensure that prior assumptions on $\boldsymbol{\theta} = \ln \boldsymbol{\mu}$ are held constant. We fix the penalty order at $q = 2$, corresponding to second-order differences.

As both estimators target $\boldsymbol{\theta}$, we compare them using the following relative error metric:

$$\Delta(\boldsymbol{\theta}) = \frac{\ell_P(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) - \ell_P(\boldsymbol{\theta})}{\ell_P(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) - \ell_P(\hat{\boldsymbol{\theta}}_{\infty})}. \tag{4.2}$$

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

Table 2. *Impact of the approximation from the original WH smoothing on the 6 simulated datasets.*

Portolio type	Head count	Relative error	SMR
Annuity	20,000	1,91%	99,19%
Annuity	100,000	0,02%	99,89%
Annuity	500,000	0,00%	99,99%
LTC	20,000	93,27%	86,86%
LTC	100,000	5,12%	97,59%
LTC	500,000	0,24%	99,56%

Here $\hat{\boldsymbol{\theta}}_\infty$ maximizes the penalized likelihood, while $\hat{\boldsymbol{\theta}}_\infty$ corresponds to the solution with $\lambda \to \infty$, which we later show to be the degree-$(q-1)$ polynomial that maximizes the likelihood. By construction:

$$\Delta(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) = 0, \quad \Delta(\hat{\boldsymbol{\theta}}_\infty) = 1, \quad \text{and} \quad \Delta(\boldsymbol{\theta}) \geq 0.$$

A model with $\Delta(\boldsymbol{\theta}) > 1$ performs worse than a simple polynomial fit under the prior.

Table 2 presents the values of $\Delta(\hat{\boldsymbol{\theta}}_{\mathrm{norm}})$ across the six datasets. As expected, discrepancies decrease with portfolio size. For annuities, the approximation performs reasonably well even at smaller scales. In contrast, for LTC, it yields substantial errors, except for the largest portfolio.

One explanation, supported by the standardized mortality ratio (SMR) also provided in Table 2, is the positive correlation between observed event counts and their use as weights. This causes high crude rates to be overweighted, and low rates to be underweighted – introducing systematic overestimation of mortality rates. This bias is more severe in the LTC case where the observed deaths by data point is lower. In contrast, generalized WH smoothing preserves total event counts by construction, always yielding an SMR of exactly 100%. These results support adopting generalized WH smoothing in most practical settings. It retains the advantages of the original method while offering improved accuracy – even in small samples – and remains straightforward to implement.

5. How to select the smoothing parameters?

5.1. Impact of smoothing parameter choice

In the one-dimensional case, WH smoothing involves a single smoothing parameter λ; in two dimensions a pair $\lambda = (\lambda_x, \lambda_z)$. These parameters govern the trade-off between fidelity to the data and smoothness of the estimate, as defined in Equation (1.1).

Figure 1 illustrates this effect in a one-dimensional annuity dataset (100,000 policyholders, see Section 4.2), with three values of λ. The effective degrees of freedom (edf), computed as the trace of the hat matrix $H = (W + P_\lambda)^{-1} W$, are shown for each curve. This quantity serves as a non-parametric analog of the number of free parameters in classical models and can take fractional values.

As shown, a low value $\lambda = 10^1$ yields an overfitted result that mirrors sampling noise, while a high value $\lambda = 10^7$ oversmooths and obscures the underlying trend. A mid-range value $\lambda = 10^4$ appears visually balanced. However, selecting a smoothing parameter by eye is unreliable: small-sample variability at the extremes of the age range can easily be mistaken for meaningful patterns.

The two-dimensional case further illustrates this difficulty. Figure 2 presents the smoothed transition rates from disability to death in an LTC portfolio (100,000 policyholders), using 9 combinations of (λ_x, λ_z). Choosing an appropriate parameter pair visually becomes nearly impossible, reinforcing the need for a data-driven statistical selection criterion.

5.2. Statistical criteria for parameter selection

Smoothing parameter selection typically relies on two classes of statistical criteria:

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

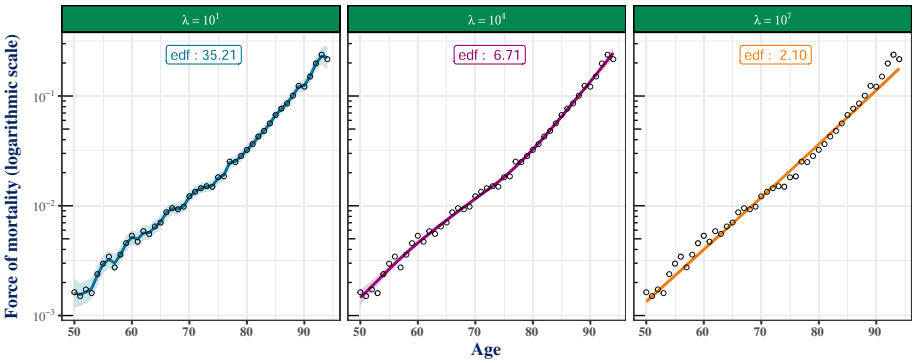


Figure 1. *WH smoothing on a synthetic annuity portfolio with 3 smoothing levels. Dots: crude rates; curves: smoothed estimates; shaded areas: credibility intervals. edf: effective degrees of freedom.*

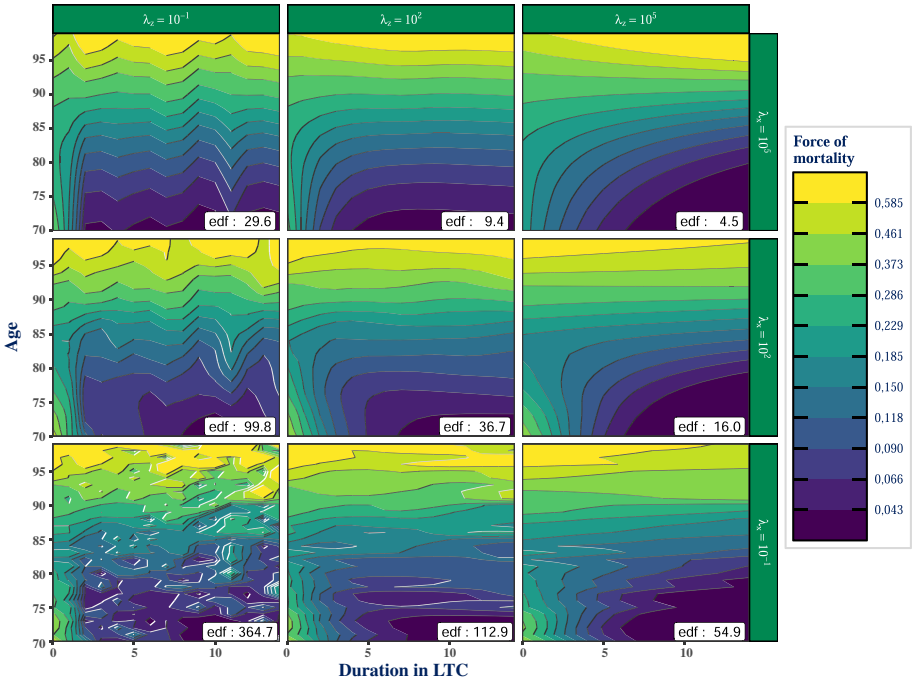


Figure 2. *WH smoothing applied to disability-to-death transitions in an LTC portfolio, using 9 combinations of smoothing parameters. Contour lines and colours show the smoothed mortality surface by age and LTC duration.*

1. Prediction-based criteria, which aim to minimize prediction error, such as the Akaike Information Criterion (AIC) (Akaike, 1973), and generalized cross-validation (GCV) (Wahba, 1980);
2. Likelihood-based criteria, which maximize the marginal likelihood – an approach introduced by Patterson and Thompson (1971) (under the name REML in the Gaussian case) and adapted to smoothing by Anderssen and Bloomfield (1974).

While prediction-based criteria have desirable asymptotic properties (Wahba, 1985; Kauermann, 2005), their convergence towards optimal smoothing parameters can be slow. In contrast, marginal

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

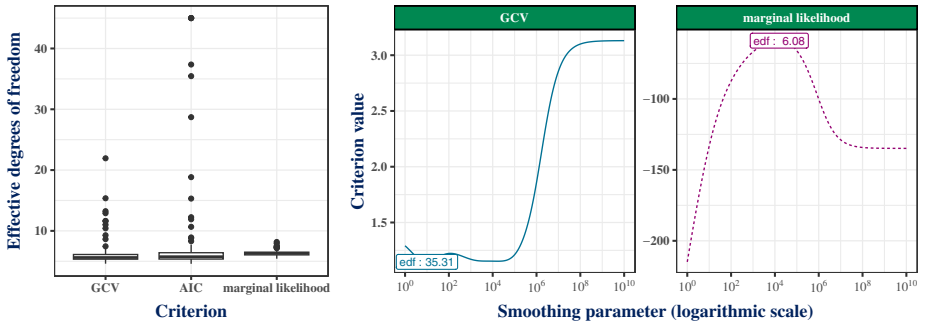


Figure 3. *Comparison of criteria for selecting the smoothing parameter in one-dimensional WH smoothing. Left: distribution of effective degrees of freedom under AIC, GCV, and marginal likelihood across 100 replicates. Right: GCV and marginal likelihood values for one replicate as functions of the smoothing parameter.*

likelihood criteria tend to perform more robustly in finite samples (Reiss and Todd Ogden, 2009; Wood, 2011).

To illustrate this, we apply AIC, GCV, and marginal likelihood to 100 replicates of the annuity portfolio with 100,000 policyholders (see Section 4.2). For each replicate, we select the optimal smoothing parameter and compute the corresponding effective degrees of freedom.

As shown in the left side of Figure 3, marginal likelihood produces stable and coherent degrees of freedom across replicates, whereas AIC and especially GCV often yield overly complex models. On the right, we plot the GCV and marginal likelihood profiles for a single replicate: marginal likelihood exhibits a well-defined maximum, while GCV presents two local minima. One aligns with the marginal likelihood optimum, but the global minimum corresponds to a model with 35 degrees of freedom – an implausibly complex mortality curve.

These observations support the use of marginal likelihood over prediction-based criteria, especially in actuarial applications where robustness is key. Moreover, this choice aligns naturally with the Bayesian framework introduced in Sections 2–4.

We now detail its implementation – first for the original WH smoothing, then for the generalized setting – introducing three optimization strategies and three numerical algorithms and comparing their respective performances.

5.3 Selection in the original smoothing

We consider again the normal framework from Section 2, where $\mathbf{y} \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 W^-)$ and $\boldsymbol{\theta} \mid \lambda \sim \mathcal{N}(0, \sigma^2 P_\lambda^-)$. In the empirical Bayes approach, the smoothing parameter λ is estimated by maximizing the marginal likelihood:

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

This is simply the maximum likelihood method applied to the smoothing parameter, treated as deterministic but unknown. A closed-form expression for this integral can be derived using standard Gaussian identities (see Section C of the Online Supplementary Materials), yielding the marginal log-likelihood:

$$\ell^m_{\text{norm}}(\lambda) = -\frac{1}{2}\left[(\mathbf{y} - \hat{\boldsymbol{\theta}}_\lambda)^T W (\mathbf{y} - \hat{\boldsymbol{\theta}}_\lambda)/\sigma^2 + \hat{\boldsymbol{\theta}}_\lambda^T P_\lambda \hat{\boldsymbol{\theta}}_\lambda/\sigma^2 + \ln|W + P_\lambda| - \ln|P_\lambda|_+ + C\right].$$

where $\hat{\boldsymbol{\theta}}_\lambda = (W + P_\lambda)^{-1} W\mathbf{y}$, and $C = -\ln|W|_+^+ + (n_* - q)\ln(2\pi\sigma^2)$ is a constant independent of λ. This function is maximized numerically to obtain $\hat{\lambda}_{\text{norm}}$.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

5.4. Selection in the generalized smoothing

The empirical Bayes approach introduced in the normal framework can be extended to the generalized smoothing framework developed in Section 4. While no closed-form expression exists for the marginal likelihood in this context, it can be approximated using a second-order Taylor expansion of the log-posterior density around its maximum $\hat{\boldsymbol{\theta}}_\lambda$ – similarly to what was done in the normal case. This yields the so-called Laplace approximation of the marginal likelihood (LAML), defined as

$$\ell^m_{\text{LAML}}(\lambda) = \ell(\hat{\boldsymbol{\theta}}_\lambda) - \frac{1}{2}\left[\hat{\boldsymbol{\theta}}_\lambda^T P_\lambda \hat{\boldsymbol{\theta}}_\lambda + \ln|W_\lambda + P_\lambda| - \ln|P_\lambda|_+ - q\ln(2\pi)\right]$$

where $W_\lambda = \text{Diag}(\exp(\hat{\boldsymbol{\theta}}_\lambda) \odot \mathbf{e}_c)$ and $\ell(\hat{\boldsymbol{\theta}}_\lambda)$ is the log-likelihood evaluated at the penalized MLE. The detailed derivation of the Laplace approximation in this setting is provided in Section C of the Supplementary Materials. This approximation plays a central role in the automatic selection of the smoothing parameter λ in the generalized WH smoothing framework. As in the normal case, the marginal likelihood $\ell^m_{\text{LAML}}(\lambda)$ must be maximized numerically. However, a key distinction is that the penalized likelihood maximizer $\hat{\boldsymbol{\theta}}_\lambda$ now depends on λ and must be recomputed at each iteration via the PIRLS algorithm. This leads to a two-level optimization procedure:

- an inner loop estimating $\hat{\boldsymbol{\theta}}_\lambda$ for fixed λ using PIRLS;
- and an outer loop optimizing $\ell^m_{\text{LAML}}(\lambda)$ with respect to λ.

This outer iteration approach is the most principled method for smoothing parameter selection in this setting.

Alternative strategies have been proposed to reduce computational burden. The first one, known as performance-oriented iteration, was introduced by Gu (1992) and relies on the observation that, at each PIRLS step, the working response vector $\mathbf{z}_k$ can be treated as approximately normal: $\mathbf{z}_k \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, W_k^{-1})$. Assuming W_k independent of λ, the marginal likelihood can be maximized within each PIRLS step using the normal approximation methodology of Section 5.3, with $\mathbf{y}$ replaced by $\mathbf{z}_k$ and W by W_k. This effectively reverses the nesting structure, potentially saving computational time when updating λ is less costly than recomputing a PIRLS step. A formal justification of the method is provided by Wood (2017, 149), which emphasizes that it does not actually require $\mathbf{z}_k$ to have a normal distribution to be well founded.

A third and even simpler strategy is the alternate iteration approach, used for instance by Wood *et al.* (2017). It consists in alternating updates of $\boldsymbol{\theta}$ (via PIRLS) and λ (via approximate marginal likelihood), without fully optimizing either at each step. This relies on the empirical observation that a coarse update of λ may suffice, as the marginal likelihood surface changes between iterations.

Despite their efficiency, both performance-oriented and alternate iteration approaches lack formal convergence guarantees. Unlike outer iteration, they operate on different smoothing parameters at each step, rendering penalized likelihood values non-comparable across iterations. Moreover, they do not track the value of $\ell^m_{\text{LAML}}(\lambda)$ during the optimization, making it harder to assess convergence or apply step-length controls.

Detailed algorithmic formulations of all three strategies in the generalized WH smoothing framework are provided in Section D of the Supplementary Materials.

5.5. Algorithms for the maximization of the marginal likelihood

Several algorithms can be used to maximize the marginal likelihood or its Laplace approximation (LAML). It is generally preferable to apply these algorithms to the logarithm of the smoothing parameters, for three main reasons:

1. It ensures positivity of the smoothing parameters;
2. It simplifies the expressions of derivatives, when required;

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

3. It allows more uniform coverage of the range of interest (e.g. from $\lambda = 10^1$ to 10^7, as in Figure 1, differences of comparable magnitude occur on a logarithmic scale).

Derivative-free heuristics

A first, operationally simple option is to use general-purpose derivative-free optimization methods:

- Brent's method (Brent, 1973) in the one-dimensional case;
- The Nelder–Mead simplex algorithm (Nelder and Mead, 1965) in higher dimensions.

These are readily available in base `R` via the `optimize` and `optim` functions. They only require evaluating the marginal likelihood or LAML at each step, which is computationally inexpensive. However, they typically require more iterations to converge and cannot be combined with the alternate iteration approach, as they do not guarantee systematic improvement of the criterion at each step.

Generalized Fellner–Schall method

A more specialized algorithm is the generalized Fellner–Schall method, based on ideas from Fellner (1986) and Schall (1991), and adapted for smoothing parameter selection in multidimensional generalized linear models by Rodríguez-Álvarez *et al*. (2015). It may be summarized by the update formula:

$$\lambda_j^{\text{next}} = \frac{\text{tr}(P_\lambda^- P_j) - \text{tr}[(X^T W X + P_\lambda)^{-1} P_j]}{\hat{\boldsymbol{\beta}}_\lambda^T P_j \hat{\boldsymbol{\beta}}_\lambda} \lambda_j^{\text{current}} \quad \text{for} \quad j \in \{x, z\}. \tag{5.1}$$

where for WH smoothing $P_j = D_{n_+,q}^T D_{n_+,q}$ in the one-dimensional case and P_x (resp. P_z) is the marginal penalization matrix $I_{n_z} \otimes D_{n_x,q_x}^T D_{n_x,q_x}$ (resp. $D_{n_z,q_z}^T D_{n_z,q_z} \otimes I_{n_x}$) in the two-dimensional case.This update can be interpreted more intuitively as

$$\hat{\boldsymbol{\beta}}_\lambda^T (\lambda_j^{\text{next}} P_j) \hat{\boldsymbol{\beta}}_\lambda = \text{tr}[P_\lambda^- \lambda_j^{\text{current}} P_j - (X^T W X + P_\lambda)^{-1} \lambda_j^{\text{current}} P_j]$$

where the right-hand side corresponds to an effective degrees of freedom associated with $\lambda_j^{\text{current}} P_j$, and the left-hand side to a squared error, normalized by the updated penalty precision. This makes λ_j^{next} resemble a REML-based estimator for the inverse variance. More details may be found in Rodríguez-Álvarez *et al*. (2019). This method:

- May be combined with any of the three iteration nesting schemes (outer, performance, and alternate);
- Does not require explicit derivative computations;
- Converges towards an approximate maximum of LAML in the generalized case, since it ignores the dependence of W on λ;
- Tends to take longer steps than EM-like algorithms (Dempster *et al*., 1977), but shorter than Newton updates (see Wood and Fasiolo 2017, which also provides a thorough justification for the method).

Newton algorithm

A third option is the Newton method, which involves computing both the first and second derivatives of the marginal likelihood (or LAML) with respect to $\ln \lambda$. Full derivations are provided in Wood (2011), which covers a more general case. The method applies in both the normal and generalized cases, but in the latter, derivative expressions are more complex due to the dependence of W on λ. The Newton algorithm is fast and precise and applicable to all three nesting strategies. The downside is the operational complexity associated with this method, especially in the generalized case.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

5.6. Performance comparison

Sections 5.4 and 5.5 introduced eight combinations of nesting strategies and optimization algorithms applicable to the generalized WH smoothing. We now assess the potential convergence issues and approximation errors associated with each of them.

This analysis is based on 100 replicates of the simulated annuity and LTC portfolios with 100,000 policyholders, as described in Section 4.2. For each replicate and each method combination, we compute the LAML at the selected smoothing parameters and compare this value to the (approximate) optimal value obtained across all combinations, denoted $\hat{\lambda}_{\text{opt}}$.

To quantify the discrepancy, we define the relative error:

$$\Delta(\lambda) = \frac{\ell^m_{\text{LAML}}(\hat{\lambda}_{\text{opt}}) - \ell^m_{\text{LAML}}(\lambda)}{\ell^m_{\text{LAML}}(\hat{\lambda}_{\text{opt}}) - \ell^m_{\text{LAML}}(\infty)} \tag{5.2}$$

where $\ell^m_{\text{LAML}}(\infty)$ corresponds to the LAML value when using an infinite smoothing penalty, that is, the overly smooth baseline. By construction, $\Delta(\lambda) \geq 0$ for all tested methods, with $\Delta(\hat{\lambda}_{\text{opt}}) = 0$ and $\Delta(\infty) = 1$. In the two-dimensional setting, we also compare average computation time for each method across replicates.

Results are summarised in Figure 4. The top panel displays the relative error $\Delta(\lambda)$ (capped below 10^{-10} for readability). In the outer iteration framework:

- The Newton method consistently achieves relative errors below 10^{-10};
- Brent and Nelder–Mead heuristics yield slightly higher errors but remain below 10^{-7};
- The generalized Fellner–Schall method produces higher errors, but still below 10^{-5} and negligible in practice.

In the performance and alternate iteration frameworks, all methods yield similar errors, consistently below 10^{-5}, with no convergence issues observed in any replicate. These findings suggest that method selection can be guided by practical considerations such as speed and implementation ease.

The bottom panel of Figure 4 compares computation times (relative to the Nelder–Mead + outer iteration baseline):

- In the outer iteration framework, the Newton method is the fastest, followed by the Fellner-Schall approach;
- All outer iteration variants are faster than their performance or alternate counterparts.

This is unsurprising, as PIRLS steps are particularly lightweight in WH smoothing (where the model matrix is the identity). However, alternate strategies may remain useful for more general cases like those described in Section 6.4.

For reference, the average time required for a single iteration using Nelder–Mead in the 2D outer iteration case is approximately 1.68 s (versus 5 ms in the 1D case).

6. How to improve smoothing computational efficiency?

6.1. Motivation

WH smoothing is a full-rank method, meaning that it includes as many parameters as there are observation points. This feature ensures a high degree of flexibility, allowing the estimator to closely track the input signal when sufficient data are available. Formally, WH smoothing is asymptotically unbiased since:

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

where m denotes the number of observed individuals, which influences the matrix W.

However, this flexibility comes at a computational cost. Some key operations, such as (implicit) matrix inversions, scale cubically with the number of parameters. As a result, WH smoothing may

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

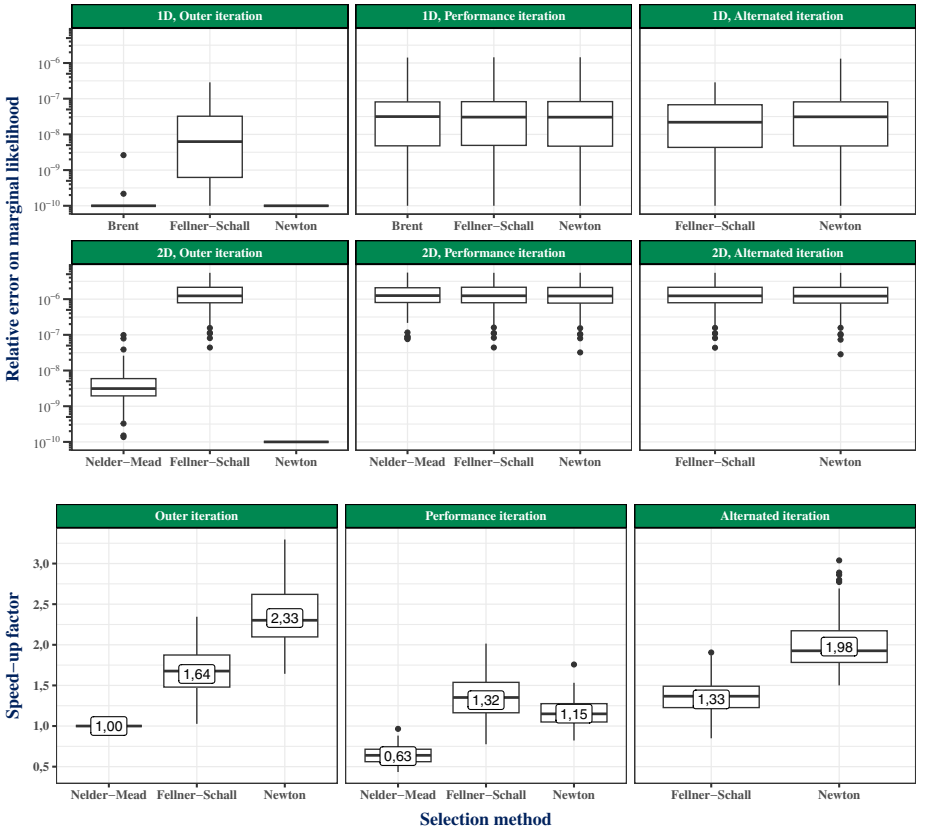


Figure 4. *Comparison of the 8 nesting strategy and algorithm combinations in the 1D and 2D simulated cases. Top: relative error on the LAML (log scale). Bottom: improvement in average computation time compared to the Nelder-Mead + outer iteration reference.*

become impractical with large number of combinations or when applied repeatedly (e.g. in simulations or bootstraps).

In one-dimensional settings, such as age-only models with annual discretization, the number of points rarely exceeds 100, and computation time is negligible. In contrast, two-dimensional use cases – common in insurance – can lead to substantially larger datasets:

- Disability tables in France must cover entry ages from 18 to 61 and exit ages up to 62, resulting in $(62 - 18) \times (62 - 18 + 1)/2 = 990$ combinations.
- Transition tables from short-term incapacity to disability involve entry ages from 18 to 67 and monthly durations from 0 to 36 months, yielding $(67 - 18) \times 36 = 1,764$ combinations.
- Long-term care (LTC) models require coverage over ages 50–100 and durations from 0 to 20 years, totalling $(100 - 50) \times (20 - 0) = 1000$ combinations (in practice, this number may be lower due to data sparsity).

In such settings, computing WH smoothing – especially when paired with smoothing parameter selection – can take several minutes per application, limiting usability in iterative contexts.

To address this limitation, we now analyse the computational complexity of the main steps in WH smoothing and smoothing parameter selection then introduce two complementary strategies to reduce computation time:

- A structural optimization that exploits the specific form of WH penalization matrices;

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

- A reduced-rank approximation that lowers the number of parameters while minimizing bias compared to the full-rank estimator.

Finally, we benchmark these strategies in terms of runtime and accuracy using 100 replicates of the mid-size annuity and LTC portfolios from Section 4.2. The structural optimization is compared to the original WH method, while the reduced-rank approximation is evaluated against the original method, the structural optimization, and a reference P-spline smoothing approach.

6.2. Practical computation for penalized smoothers

WH smoothing belongs to a broader family of penalized smoothing methods that produce estimates of the form:

$$\hat{\mathbf{y}}_\lambda = X\hat{\boldsymbol{\beta}}_\lambda \quad \text{where } \boldsymbol{\beta}_\lambda \text{ solves } (X^T WX + P_\lambda)\hat{\boldsymbol{\beta}}_\lambda = X^T W\mathbf{y}.$$

Here, X and P_λ denote the model and penalization matrices of size $n \times p$ and $p \times p$, respectively, and W is a diagonal matrix of positive weights of size $n \times n$.

Computational steps

The computation of $\hat{\mathbf{y}}_\lambda$ for a given λ typically involves the following steps:

1. Absorb the weights in the model matrix and observation vector, forming $W^{1/2}X$ and $W^{1/2}\mathbf{y}$, which requires $O(n^2)$ and $O(n)$ operations, respectively (multiplying each row of X and each element of $\mathbf{y}$ by the corresponding element of $\mathbf{w}$).
2. Form the matrix P_λ. The cost of this operation is typically $O(p^2)$ in the general case.
3. Form the matrix $X^T WX$ and the vector $X^T W\mathbf{y}$, which requires up to $O(np^2)$ and $O(np)$ operations respectively.
4. Add together $X^T WX$ and P_λ which requires $O(p^2)$ operations in the general case.
5. Compute the Cholesky decomposition $X^T WX + P_\lambda = R^T R$ at a cost of $O(p^3)$.
6. Obtain $\hat{\boldsymbol{\beta}}_\lambda$ by forward-backward substitution, first solving $R^T\mathbf{u} = X^T W\mathbf{y}$ then $R\hat{\boldsymbol{\beta}}_\lambda = \mathbf{u}$ with an associated cost of $O(p^2)$ for each system.
7. Compute $\hat{\mathbf{y}}_\lambda = X\hat{\boldsymbol{\beta}}_\lambda$ at a cost of $O(np)$.

As an alternative to Cholesky, QR decomposition may be used for greater numerical stability (see Golub and Van Loan, 2013). It applies to the weighted design matrix stacked with a matrix B such that $B^T B = P_\lambda$.

Simplifications for WH smoothing

In WH smoothing, $X = I_n$, which simplifies computations:

- Step 7 is unnecessary, as well as the first part of step 3.
- $X^T W\mathbf{y} = W\mathbf{y}$ (step 3) is computed in $O(n)$ by multiplying $\mathbf{w}$ and $\mathbf{y}$.
- $X^T WX + P_\lambda = W + P_\lambda$ (step 4) is also computed in $O(n)$ by adding the vector $\mathbf{w}$ to the leading diagonal of P_λ.

Generalized WH smoothing with outer iteration

When using the outer iteration approach (see Section 5), each candidate λ requires a full PIRLS cycle to estimate $\hat{\boldsymbol{\theta}}_\lambda$, with new working vector $\hat{\mathbf{z}}_\lambda^k$ and weight matrix W_λ^k. Steps 1–6 above are repeated until convergence of the PIRLS algorithm, which may be assessed by monitoring the changes in penalized deviance. The deviance may be computed at a $O(n)$ cost. For penalization based on differences matrices,

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

computation of $\hat{\boldsymbol{\beta}}_\lambda{}^T P_\lambda \hat{\boldsymbol{\beta}}_\lambda$ should be based on the expression of $R_{\lambda,q}$ provided in Section 1.1 for an associated cost of $O(qp)$. In addition, PIRLS iterations for each new λ can be initialized using the previous estimate of $\hat{\mathbf{y}}_\lambda$ for faster convergence.

LAML computation

Once the deviance is known, computing the marginal likelihood/LAML also requires:

- $\ln|X^T WX + P_\lambda|$, which may be computed at a cost of $O(p)$ from the leading diagonal of the Cholesky/QR factor R computed at step 5 in the derivation of $\hat{\mathbf{y}}_\lambda$.
- $\ln|P_\lambda|_+$, which may be obtained from the eigenvalues of the penalization matrix: Section 6.4 shows that in the two-dimensional case, it can be computed via eigendecomposition of $D^T_{p_x,q_x} D_{p_x,q_x}$ and $D^T_{p_z,q_z} D_{p_z,q_z}$, performed only once, at a $O(p_x^3 + p_z^3)$ cost. Computation of $\ln|P_\lambda|_+$ then only requires scaling the eigenvalues for a cost of $O(p)$.

Algorithm-specific computations

Brent and Nelder–Mead require only marginal likelihood/LAML evaluations.

The generalized Fellner–Schall algorithm relies on the update formula of Equation (5.1):

$$\lambda_j^{\text{next}} = \frac{\operatorname{tr}(P_\lambda^- P_j) - \operatorname{tr}[(X^T WX + P_\lambda)^{-1} P_j]}{\hat{\boldsymbol{\beta}}_\lambda^T P_j \hat{\boldsymbol{\beta}}_\lambda} \lambda_j^{\text{current}} \quad \text{for} \quad j \in \{x, z\}.$$

Evaluation of $\operatorname{tr}(P_\lambda^- P_j)$ does not require any matrix product. In the one-dimensional case it is simply $(p-q)/\lambda$ while in the two-dimensional case it may be obtained directly at a $O(p)$ cost using the eigenvalues of the aforementioned penalization matrices P_j. Evaluation of $\operatorname{tr}[(X^T WX + P_\lambda)^{-1} P_j]$ may use the identity $\operatorname{tr}(AB) = \sum_{i,j} A_{ij} B_{ji}$ and therefore be computed at an $O(p^2)$ cost if the matrix $(X^T WX + P_\lambda)^{-1}$ and P_j are available. Computation of $V = (X^T WX + P_\lambda)^{-1}$ is done by first solving for the inverse $K = R^{-1}$ of the Cholesky/QR factor and then forming V as KK^T. Both operations have a $O(p^3)$ cost.

Newton method also requires computation of V, as well as several matrix products involving the penalization matrix P_j. For example, the second derivatives of marginal likelihood require $\operatorname{tr}[VP_jVP_k]$ terms, and the second derivatives of marginal likelihood require $\operatorname{tr}[V(X(\partial W/\partial\rho_j)X + P_j)V(X(\partial W/\partial\rho_k)X + P_k)]$ terms where $\rho_j = \ln(\lambda_j)$, $j = k = x$ in the one-dimensional case and $\{j,k\} \in \{x,z\}$ in the two-dimensional case. The identity $\operatorname{tr}(AB) = \sum_{i,j} A_{ij} B_{ji}$ can also be used in this case but matrix products VP_j or $V[X(\partial W/\partial\rho_j)X + P_j]$ still need to be explicitly computed, for a respective cost of $O(p^3)$ and $O(np^2)$ each.

These additional computations make Newton updates more expensive than generalized Fellner–Schall updates, but they generally yield faster convergence and higher precision (see Section 5.6).

6.3. Banded optimization for WH smoothing

We now consider how to exploit the banded structure of the penalization matrix in WH smoothing. This structure enables significant computational gains, especially when dealing with large number of observations. Throughout this section, we assume $X = I_n$ and $p = n$, which holds for both the original and generalized WH smoothing.

One-dimensional case

In one dimension, the penalization matrix takes the form: $P_\lambda = \lambda D^T_{n,q} D_{n,q}$. This matrix is symmetric and banded with bandwidth q. As a consequence:

1. Compact storage: P_λ can be stored in a compact form with dimensions $n \times (q+1)$ and updated for new λ at a cost of $O(qn)$. The matrix $W + P_\lambda$ shares this structure.
2. Efficient Cholesky decomposition: the Cholesky factor R of $W + P_\lambda$ can be computed in $O(q^2 n)$ instead of $O(n^3)$, and R is also banded with the same bandwidth.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

3. Efficient back-substitution: computing $\hat{\mathbf{y}}_\lambda = \hat{\boldsymbol{\beta}}_\lambda$ using R is now $O(qn)$ instead of $O(n^2)$.
4. Efficient inversion of R: the inverse $K = R^{-1}$ costs $O(qn^2)$, an improvement over the $O(n^3)$ cost for dense matrices.

However, K is a dense triangular matrix, meaning the computation of $V = KK^T$ remains a $O(n^3)$ operation. Fortunately, the generalized Fellner–Schall algorithm only requires the diagonal of V, which can be obtained from K in $O(n^2)$, since:

$$\lambda_j[\mathrm{tr}(P_\lambda^- P_j) - \mathrm{tr}(VP_j)] = (n-q) - (n - \mathrm{tr}[VW]) = \mathbf{diag}(V)^T\mathbf{w} - q.$$

Furthermore, the Newton algorithm benefits as well: the trace terms involving VP_λ or $V(\partial W/\partial \ln\lambda + P_\lambda)$ are based on banded matrices, making those products computable in $O(qn^2)$ instead of $O(n^3)$.

Two-dimensional case

In two dimensions, the penalization matrix is $P_\lambda = \lambda_x P_x + \lambda_z P_z$ where:

- $P_x = I_{n_z} \otimes D^T_{n_x,q_x} D_{n_x,q_x}$
- $P_z = D^T_{n_z,q_z} D_{n_z,q_z} \otimes I_{n_x}$.

This structure has the following key properties:

- Both matrices are made of $n_z \times n_z$ square blocks of dimensions $n_x \times n_x$ each.
- P_x is block diagonal with n_z identical $n_x \times n_x$ banded blocks (bandwidth q_x).
- P_z is block banded with bandwidth q_z. Each block is a scaled identity matrix.
- As a whole, P_z and P_λ may be viewed as banded matrices with bandwidth $q = q_z \times n_x$.

This implies that all statements made in the one-dimensional case carry over to the two-dimensional case with this value of q. It also suggests that, if $(q_x+1)/n_x < (q_z+1)/n_z$, dimensions x and z should be permuted before applying WH smoothing for maximal efficiency.

As in the one-dimensional case, the generalized Fellner–Schall update formula does not require the full computation of V. Indeed, to compute $\mathrm{tr}[VP_x]$ and $\mathrm{tr}[VP_z]$, we only need access to elements of V for which either P_x or P_z is non-zero. From what precedes, P_x has bandwidth q_x while P_z only contains q_z non-zero diagonals on each side of the leading diagonal. As V is symmetric, we only need to compute $q_x + q_z + 1$ diagonals of V for an associated cost of $O([q_x+q_z]n^2)$ instead of $O(n^3)$.

With the Newton method, while computing $V = KK^T$ still incurs a $O(n^3)$ cost, matrix multiplications like VP_j or $V(\partial W/\partial\rho_j + P_j)$ can be performed block-wise. It may easily be checked, for example, that the products VP_x and $V(\partial W/\partial\rho_x + P_x)$ have a cost of $O(q_x n^2)$, while the products VP_z and $V(\partial W/\partial\rho_z + P_z)$ have a cost of $O(q_z n^2)$.

Summary of complexity gains

Thanks to the banded structure, most computations involved in WH smoothing can be accelerated by a factor of $n/(q+1)$ in the 1D case and $\max(n_x/(q_z+1), n_z/(q_x+1))$ in the 2D case. There are 3 notable exceptions:

- Cholesky decomposition is improved from $O(n^3)$ to $O(q^2 n)$ – a quadratic speed-up.
- Computation of $V = KK^T$ remains $O(n^3)$.
- Some matrix products required by Newton method get a full $n/(q_z+1)$ or $n/(q_z+1)$ speed-up in the 2D case.

Table 3 summarizes theoretical complexities across different frameworks, including a typical generalized additive model framework for which the penalization matrix is diagonal. This last framework is used by the rank-reduced WH smoothing approach introduced next, as well as the P-spline alternative used for comparison.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

Table 3. *Compared theoretical leading-order costs associated with the key steps in smoothing computations for several frameworks. All cells should be read as $O(\ldots)$.*

Computation	Dense	Banded	Rank-reduced
$X^T W X$	$\emptyset$	$\emptyset$	np^2
$X^T W z$	n	n	np
P_λ	n^2	qn	p
$X^T W X + P_\lambda$	n	n	p
R	n^3	$q^2 n$	p^3
$\hat{\boldsymbol{\beta}}_\lambda$	n^2	qn	p^2
$\hat{\mathbf{y}}_\lambda = X\hat{\boldsymbol{\beta}}_\lambda$	$\emptyset$	$\emptyset$	np
ML/LAML	qn	qn	n
$K = R^{-1}$	n^3	qn^2	p^3
$V = KK^T$	n^3	n^3	p^3
Brent/Nelder–Mead	n^3	$q^2 n$	p^3
Fellner–Schall	n^3	qn^2	p^3
Newton	n^3	n^3	p^3

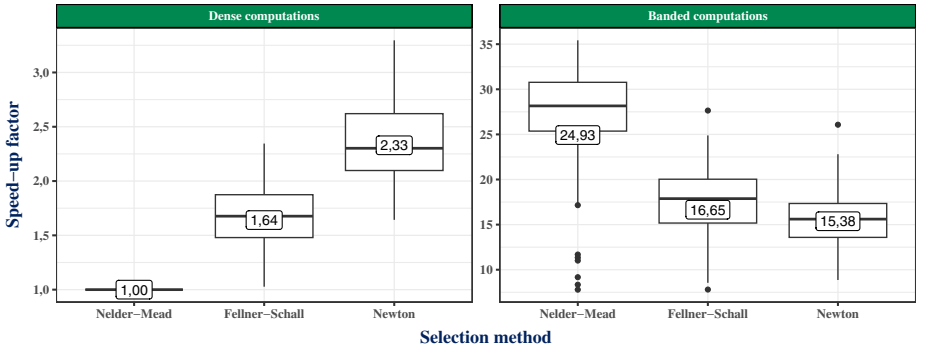


Figure 5. *Computation time comparison for 2D generalized WH smoothing with outer iteration. The speed-up factor is computed relative to the original dense method using the Nelder–Mead algorithm.*

Empirical gains

Figure 5 compares actual computation times of WH smoothing (two-dimensional, outer iteration), showing that adapting the implementation to exploit banded structures results in large speed gains:

- The Nelder–Mead method benefits the most, with a 25 × speedup compared to dense computation.
- Newton and Fellner–Schall methods see 6.6 × and 10 × improvements, respectively, making them fall behind the Nelder–Mead method.

As a final advantage, Brent and Nelder–Mead heuristic methods rely solely on banded matrices that can be stored as compact matrices of dimensions $(q+1) \times n$, adding further efficiency.

6.4. Natural parameterization and rank reduction of WH smoothing

Demmler and Reinsch (1975) proposed a natural parameterization for penalized smoothers using the eigendecomposition of the penalization matrix. This provides both an intuitive interpretation of the smoothing mechanism and a foundation for dimension reduction via rank-restricted estimation.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

One-dimensional case

In one dimension, let $D_{n,q}^T D_{n,q} = U\Sigma U^T$ be the eigendecomposition of the penalty matrix, where U is orthogonal and Σ diagonal with non-negative eigenvalues. A change of variable $\boldsymbol{\theta} = U\boldsymbol{\beta}$ transforms the WH optimization into:

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

yielding the solution:

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

This formulation shows that WH smoothing decomposes the signal into eigenvector components and attenuates each according to the associated eigenvalue – the higher the eigenvalue, the stronger the shrinkage.

We refer to Section E of the Online Supplementary Materials for graphical illustrations of:

- the basis eigenvectors of $D_{n,q}^T D_{n,q}$;
- the evolution of their effective degrees of freedom under smoothing.

These figures show that only the first few components retain substantial degrees of freedom under moderate smoothing, motivating dimensionality reduction.

Two-dimensional case

In two dimensions, the penalization matrix takes the form:

$$P_\lambda = \lambda_x I_{n_z} \otimes D_{n_x,q_x}^T D_{n_x,q_x} + \lambda_z D_{n_z,q_z}^T D_{n_z,q_z} \otimes I_{n_x},$$

with eigendecompositions $D_{n_x,q_x}^T D_{n_x,q_x} = U_x \Sigma_x U_x^T$ and $D_{n_z,q_z}^T D_{n_z,q_z} = U_z \Sigma_z U_z^T$.

Define $U = U_z \otimes U_x$, and $\boldsymbol{\theta} = U\boldsymbol{\beta}$. Then the WH estimate becomes:

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

As in the one-dimensional case, this representation reveals how smoothing operates via coordinate-wise shrinkage in the eigenbasis. Section E of the Online Supplementary Materials displays the corresponding per-parameter effective degrees of freedom.

Rank reduction strategy

Inspection of the effective degrees of freedom reveals that many components are heavily shrunk, especially those associated with high eigenvalues. This suggests reducing the dimension by keeping only the $p < n$ components with the lowest eigenvalues.

In the one-dimensional case, the reduced-rank approximation is

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

where U_p and Σ_p consist of the first p eigenvectors and their corresponding eigenvalues, respectively.

In the two-dimensional case, we retain p_x and p_z eigenvectors in each dimension and use:

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

with $U_{p_x,p_z} = U_{z,p_z} \otimes U_{x,p_x}$. In that case, given a target number of parameters $p_{\max}$, we propose selecting (p_x, p_z) such that $p_x p_z \leq p_{\max}$ and $p_x/n_x \approx p_z/n_z$ using the rule:

$$\kappa = \sqrt{p_{\max}/n_x n_z}, \quad p_x = \lfloor \min(\kappa, 1) n_x \rfloor, \quad p_z = \lfloor \min(\kappa, 1) n_z \rfloor.$$

Adaptations for generalized WH smoothing follow by replacing $(\mathbf{y}, W)$ with $(\mathbf{z}_k, W_k)$ in the above expressions.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

Efficient computation via GLAM

Currie *et al*. (2006) propose a general framework, GLAM, that exploits Kronecker structure for efficient computations. In our context, the model matrix U_{p_x,p_z} inherits a Kronecker product form, allowing operations that rely on this matrix to be executed dimension-wise without explicit construction of the full matrix. This significantly reduces memory use and computation time in the two-dimensional rank-reduced WH framework.

Impact of using the rank-reduced basis

We now evaluate the impact of the rank-reduced WH basis introduced in Section 6.4 in terms of both smoothing accuracy and computational speed. For context, results are compared against those obtained using P-spline smoothing with the same number of basis functions.

To ensure a fair comparison, both approaches were implemented in the same computational framework, including the use of GLAM in the two-dimensional case – only the structure of the basis (and hence the model matrix) differs. The penalty structure and the unpenalized fixed effects (polynomials of degree $q - 1$) are identical.

In addition to the full basis of size 450 (30×15), three reduced basis of respective size 288 (24×12), 128 (16×8) and 32 (8×4) were considered.

As in Section 5.6, accuracy is assessed using the relative LAML error defined in Equation (5.2). Note, however, that since both reduced-rank and P-spline smoothers rely on different bases and penalization matrices, their LAML expressions are different from the one used for full-rank WH smoothing. Hence, a reduced model can exhibit a higher LAML than the full-rank version at its selected smoothing parameter.

Figure 6 summarises the average speed-up achieved by both the reduced-rank and P-spline smoothers compared to the full-rank WH smoothing. As the number of retained parameters decreases, computation time drops substantially. Compared to the full-rank WH smoothing (unoptimized):

- the 128-parameter basis achieves an $88\times$ speed-up;
- the 32-parameter basis achieves up to $256\times$ faster computation.

The alternate iteration and performance iteration strategies outperform the outer iteration in the reduced setting, primarily because model matrix construction becomes the new computational bottleneck – even with the use of the GLAM framework. In this context, the Newton algorithm combined with alternate iteration proves to be the most efficient, with the generalized Fellner–Schall update being nearly as competitive for smaller bases.

The gains in computational speed come with a moderate tradeoff in estimation accuracy. As shown in Figure 7, the relative LAML errors remain small:

- For the 128-parameter basis, the average error is just 0.82%.
- For the 32-parameter basis, it rises to 2.26%.

Across all sizes, the reduced-rank WH smoother slightly outperforms the P-spline smoother in terms of LAML error, confirming its effectiveness as a principled dimension reduction strategy.

7 How to extrapolate the smoothing?

Semi-parametric methods such as P-splines and WH smoothing naturally allow for extrapolation – that is, predicting values outside the range of the original data. Extrapolation is handled by solving an extended smoothing problem where extrapolated positions are associated with zero-weight observations.

However, in the two-dimensional case, extrapolation must be performed carefully: constraints are needed to ensure that the extrapolated solution remains consistent with the original smoothing result over the observed data. Following the approach introduced by Carballo *et al*. (2021) for P-splines, we now extend WH smoothing to support extrapolation while also enabling the construction of credibility intervals that capture uncertainty both inside and outside the original observation domain.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

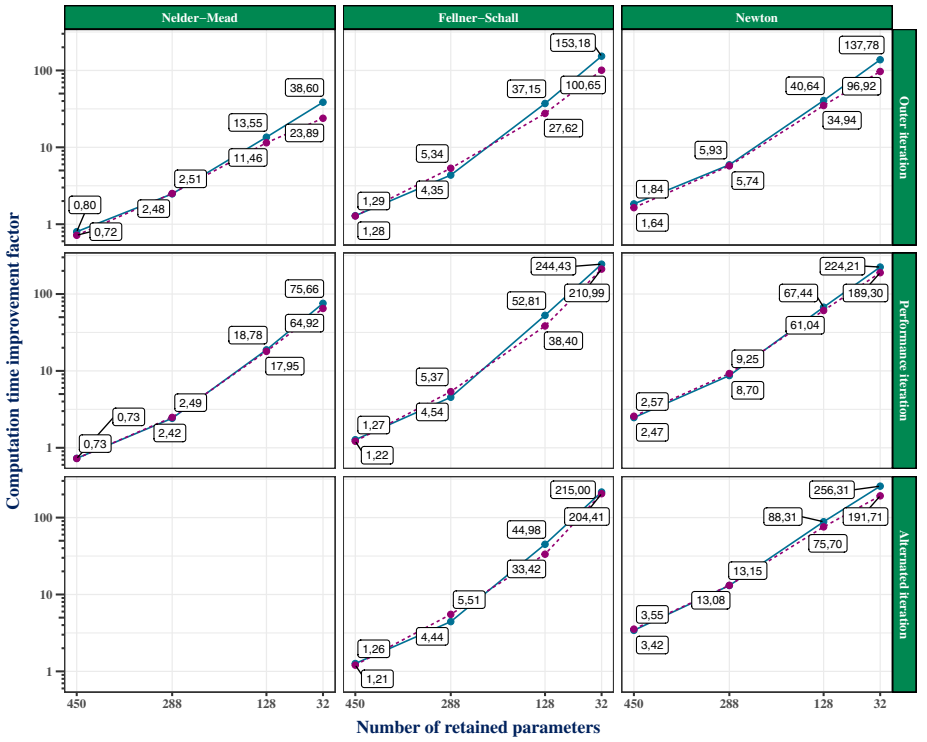


Figure 6. *Computation speed improvement from WH smoothing with a reduced-rank basis (solid lines) or P-spline basis (dotted lines), relative to unoptimized full-rank WH smoothing, as a function of basis size.*

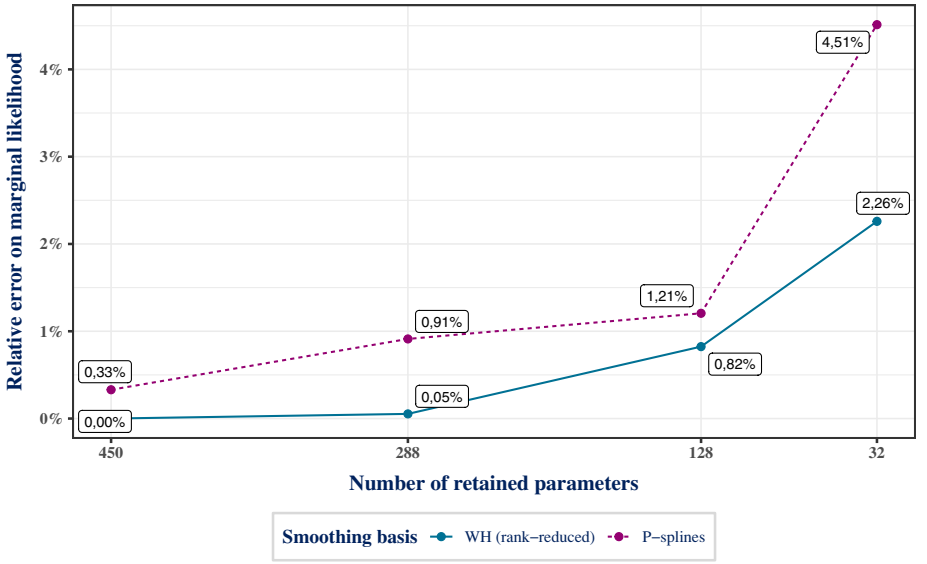


Figure 7. *Relative LAML error of WH smoothing with a reduced-rank basis (solid lines) or a P-spline basis (dotted lines), with respect to unoptimized full-rank WH smoothing, as a function of basis size.*

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

7.1. Defining the extrapolation of the smoothing

Let $\hat{\mathbf{y}}$ be the WH smoothing result obtained from an observation vector $\mathbf{y}$ defined over positions $\mathbf{x}$ (in 1D) or $(\mathbf{x}, \mathbf{z})$ (in 2D). We wish to extend predictions to a larger domain $\mathbf{x}+$ (or $(\mathbf{x}+, \mathbf{z}+)$), with $\mathbf{x} \subset \mathbf{x}+$ and similarly for $\mathbf{z}$.

To preserve WH smoothing's requirement for evenly spaced points, we assume that $\mathbf{x}+$ and $\mathbf{z}+$ are sequences of consecutive integers. Let n_+ be the length of $\mathbf{x}_+$ in the on-dimensional case. In the two-dimensional case, let n_{x+} and n_{z+} be the lengths of $\mathbf{x}+$ and $\mathbf{z}+$ and note $n_+ = n_{x+} \times n_{z+}$.

We define matrices C_x and C_z such that each extracts the indices of the original data from the larger domain. Specifically: $C_j = (O \mid I_{n_j} \mid O)$, where $j \in \{x, z\}$ and I_{n_j} is an identity matrix aligned with the observed positions. Define the matrix C as

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

Then C has the following useful properties:

- For any full-domain vector $\mathbf{y}+$, $C\mathbf{y}+$ returns the observed values only.
- $C^T\mathbf{y}$ embeds the observed values into a larger zero-padded vector.
- $CC^T = I_n$ and C^TC is a 2×2 block matrix with an identity matrix block and zeros everywhere else.

The extrapolated WH smoothing is defined as the solution to the following extended problem:

$$\hat{\mathbf{y}}_+ = \underset{\boldsymbol{\theta}_+}{\operatorname{argmin}} \left\{ (\mathbf{y}_+ - \boldsymbol{\theta}_+)^T W_+ (\mathbf{y}_+ - \boldsymbol{\theta}_+) + \boldsymbol{\theta}_+^T P_+ \boldsymbol{\theta}_+ \right\} \tag{7.1}$$

where:

- $\mathbf{y}_+ = C^T\mathbf{y}$ is the extended data vector (zeros for unobserved points),
- $W_+ = C^TWC$ is the extended weight matrix (zeros for unobserved points),
- P_+ is the penalization matrix over the extended grid, defined as

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

Importantly, the smoothing parameters λ, λ_x, and λ_z must remain fixed during extrapolation – they are inherited from the original fit and no new information is introduced.

The fidelity term in Equation (7.1) simplifies to:

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

where $\boldsymbol{\theta} = C\boldsymbol{\theta}_+$. This is the fidelity term from the original fit.

The smoothness criterion, on the other hand, now applies to the entire extended domain, constraining the extrapolated parts of $\hat{\mathbf{y}}_+$ to remain smooth and consistent with the trend learned from the data.

The same extrapolation approach applies directly to generalized WH smoothing, simply by replacing $\mathbf{y}$ by $\mathbf{z}_k$ and W by W_k, obtained at convergence of the PIRLS algorithm and setting $\sigma^2 = 1$ in the derived credible intervals.

7.2. Unconstrained solution for the 1D case

The solution to the extrapolation problem in Equation 7.1 can be obtained directly, as in Section 1.1, by taking derivatives with respect to $\boldsymbol{\theta}_+$ and setting them to zero. This yields the closed-form solution:

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

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

Assuming a Bayesian model where $\mathbf{y}+ \mid \boldsymbol{\theta}+ \sim \mathcal{N}(\boldsymbol{\theta}+, \sigma^2 W+^{-1})$ and $\boldsymbol{\theta}+ \sim \mathcal{N}(0, \sigma^2 P+^{-1})$, we obtain, as in Section 2, the following credible interval:

$$\mathbb{E}(\mathbf{y}_+) \mid \mathbf{y}_+ \in \left[(W_+ + P_+)^{-1} W_+ \mathbf{y}_+ \pm \Phi^{-1}(1-\alpha/2) \sqrt{\sigma^2 \mathbf{diag}\{(W_+ + P_+)^{-1}\}}\right].$$

To get a better understanding about how the variance-covariance matrix $V_+ = (W_+ + P_+)^{-1}$ for the unconstrained extrapolation problem of Equation (7.1) is related to the variance-covariance matrix $V = (W + P_\lambda)^{-1}$ of the original smoothing problem, introduce matrices $\overline{C}_j$ (for $j \in x, z$), which selects the rows in the extrapolated domain that are not part of the original data and define:

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

With this definition, Q is a permutation matrix moving observed positions to the top.

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

$$D_{n_+,q} = \begin{bmatrix} D_{2-} & D_{1-} & 0 \\ 0 & D_{n,q} & 0 \\ 0 & D_{1+} & D_{2+} \end{bmatrix} = Q^T \begin{bmatrix} D_{n,q} & 0 \\ D_1 & D_2 \end{bmatrix} Q \quad \text{where } D_1 = \begin{bmatrix} D_{1-} \\ D_{1+} \end{bmatrix} \text{ and } D_2 = \begin{bmatrix} D_{2-} & 0 \\ 0 & D_{2+} \end{bmatrix}.$$

The extended weight and penalization matrices may be rewritten:

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

where $P_+^{ij} = \lambda D_i^T D_j$, for $i, j \in \{1, 2\}$.

This block structure allows us to apply standard results for partitioned matrix inverses to derive:

$$V_+ = Q^T \begin{bmatrix} V_+^{11} & V_+^{12} \\ V_+^{21} & V_+^{22} \end{bmatrix} Q = Q^T \begin{bmatrix} V_+^{11} & -V_+^{11} P_+^{12} (P_+^{22})^{-1} \\ -(P_+^{22})^{-1} P_+^{21} V_+^{11} & (P_+^{22})^{-1} P_+^{21} V_+^{11} P_+^{12} (P_+^{22})^{-1} + (P_+^{22})^{-1} \end{bmatrix} Q$$

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

From the above, we retrieve:

$$C\hat{\mathbf{y}}_+ = CV_+ W_+ y_+ = CQ^T V_+ Q C^T W y = V_+^{11} W y.$$

This coincides with the original fit $\hat{\mathbf{y}}$ only if $V_+^{11} = V$. In general, this equality does not hold, since the extrapolation solution minimizes the total smoothness of the extended vector, not just of the observed part.

In V_+^{22}, we identify:

- a propagation term: $(P_+^{22})^{-1} P_+^{21} V_+^{11} P_+^{12} (P_+^{22})^{-1}$, capturing the uncertainty transferred from the known part to the extrapolated part;
- an innovation error term: $(P_+^{22})^{-1}$ associated with the prior on the extrapolated coefficients $\overline{C}\hat{\mathbf{y}}_+$.

In the one-dimensional case, D_2 is block-diagonal with invertible triangular blocks, so:

$$P_+^{11} - P_+^{12}(P_+^{22})^{-1} P_+^{21} = D_1^T D_1 - D_1^T D_2 (D_2^T D_2)^{-1} D 2^T D_1 = 0$$

which means that $V_+^{11} = (W + P_\lambda)^{-1} = V$. This confirms the result from Carballo *et al*. (2021), namely that with a difference-based penalty, a perfectly smooth extrapolation that leaves the original fit unchanged can always be constructed in the one-dimensional case.

This behaviour is illustrated in Figure 8, which shows the extrapolated fit (with $q = 2$) obtained from generalized WH smoothing applied to the annuity portfolio used previously. The extrapolation follows a straight line – the polynomial of degree $q - 1 = 1$ – and joins smoothly with the original curve.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

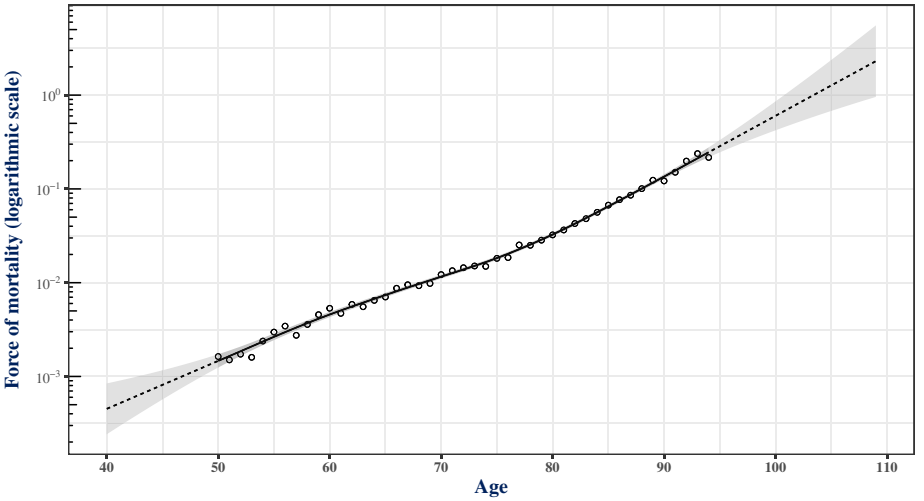


Figure 8. *Extrapolation of one-dimensional WH smoothing. The smoother is extrapolated on both sides of the initial observation range following a polynomial of degree $q-1$ (in this case a straight line as $q=2$).*

7.3. Constrained solution for the 2D case

In the two-dimensional case, while the extended penalization matrix P_+ still takes the same structure as previously described, the expressions of its block components P_+^{11}, P_+^{12}, P_+^{21}, and P_+^{22} are more complex. In particular, we no longer have the simplification $P_+^{11} - P_+^{12}(P_+^{22})^{-1}P_+^{21} = 0$, therefore $V_+^{11} \neq V$ and $C\hat{\mathbf{y}}_+ \neq \hat{\mathbf{y}}$. Solving the unconstrained extrapolation problem thus leads to a modification of the estimated coefficients for the observed data positions, as demonstrated by Carballo *et al*. (2021).

This difference arises because, unlike the one-dimensional case, the smoothness criterion in two dimensions penalizes both rows and columns simultaneously, making it impossible to extrapolate without increasing the penalization. Since no new data are introduced in the extrapolated region, the smoothness criterion weighs more heavily in the optimization, prompting adjustments to the originally fitted values in order to produce a globally smoother estimate.

To address this, we follow the approach proposed by Carballo *et al*. (2021) and formulate a constrained optimization problem that enforces preservation of the original fitted values in the smoothing region. This is done by introducing a Lagrange multiplier $\boldsymbol{\omega}$ and solving the following constrained problem:

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

This optimization admits a closed-form solution for the constrained extrapolated estimator $\hat{\mathbf{y}}_+^*$ as a linear transformation of $\hat{\mathbf{y}}$. The derivation details are provided in Section F of the Online Supplementary Materials. The final form is

$$\hat{\mathbf{y}}_+^* = Q^T \begin{bmatrix} I \\ -(P_+^{22})^{-1} P_+^{21} \end{bmatrix} \hat{\mathbf{y}}$$

and the associated variance-covariance matrix is

$$V_+^* = Q^T \begin{bmatrix} V & -VP_+^{12}(P_+^{22})^{-1} \\ -(P_+^{22})^{-1}P_+^{21}V & (P_+^{22})^{-1}P_+^{21}VP_+^{12}(P_+^{22})^{-1} + (P_+^{22})^{-1} \end{bmatrix} Q.$$

This formulation differs from the variance matrix of the unconstrained solution. Indeed, it enforces the constraint that the initial coefficients remain unchanged, as reflected by the presence of V (the original

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

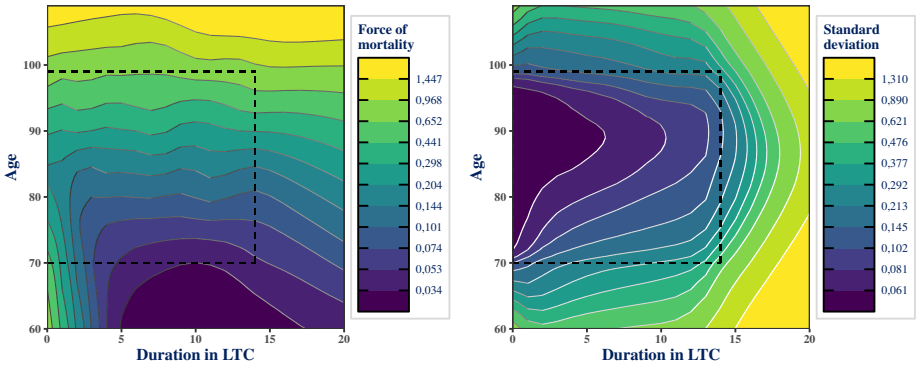


Figure 9. *Constrained extrapolation of 2D WH smoothing. The contour lines of mortality rates and the associated standard deviation are depicted. The dotted lines delimit the boundaries of the initial smoothing region.*

variance matrix) instead of V_+^{11}. The corresponding credible intervals are

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

The following figures illustrate the impact of the constrained extrapolation procedure discussed above, using the LTC portfolio of 100,000 policyholders as a case study.

- Figure 9, left (mortality rates): this panel shows the estimated mortality rates obtained after applying the constrained extrapolation procedure to the two-dimensional WH smoothing model. The dotted lines indicate the boundaries of the original smoothing region. Visually, the transition from the smoothing region to the extrapolated area is seamless – the extrapolated surface naturally extends the smoothed mortality rates while respecting the original fitted values within the data range.
- Figure 9, right (standard deviation): this panel displays the posterior standard deviation (or credible interval width) associated with the extrapolated estimates. It reflects both the uncertainty from the original smoothing and the innovation error introduced in the extrapolated region. As expected, the standard deviation increases as we move away from the observed region, illustrating growing uncertainty about farther values.
- Figure 10 (ratio of mortality rates): this heatmap shows the pointwise ratio between the unconstrained and constrained extrapolation of the mortality rates. A value above 1 indicates that the unconstrained version overshoots the constrained one at that location, while values below 1 indicate underestimation. We observe that discrepancies exist not only in the extrapolated region but also within the original data region – confirming that the unconstrained approach distorts the original estimates in order to achieve overall smoothness.
- Figure 11 (ratio of standard deviations): this final figure includes two panels comparing uncertainty estimates.
  - Left panel: ratio of standard deviation from the unconstrained extrapolation over that from the constrained extrapolation (including innovation error). The unconstrained version underestimate the actual uncertainty not only in the extrapolated region but also within the original data region, again reflecting the adjustments made to the original estimates in order to achieve overall smoothness.
  - Right panel: ratio of standard deviation from the constrained extrapolation without innovation error over the fully constrained version with innovation error. This illustrates the contribution of the innovation error to the total uncertainty – it is substantial and should not be neglected.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

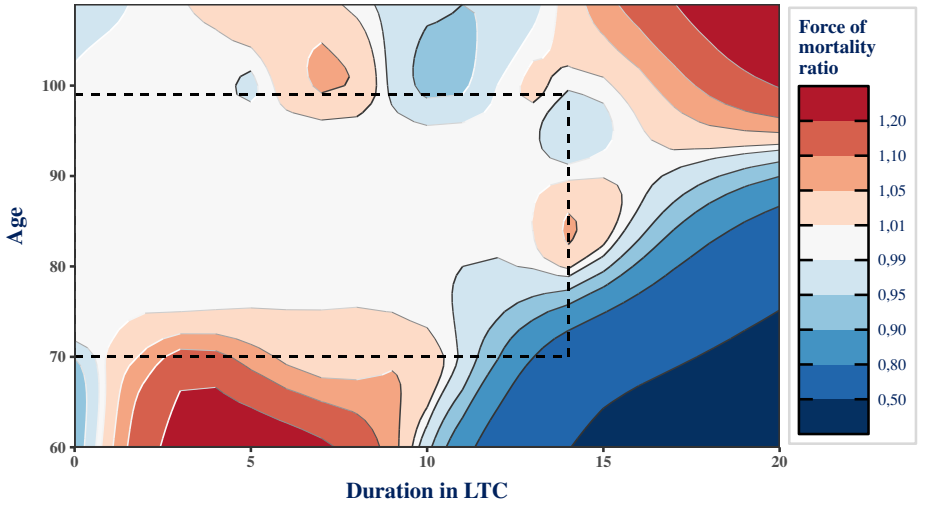


Figure 10. *Ratio of mortality rates resulting from the extrapolation of 2D WH smoothing. The numerator corresponds to the unconstrained extrapolation and the denominator to the constrained extrapolation presented in Figure 9.*

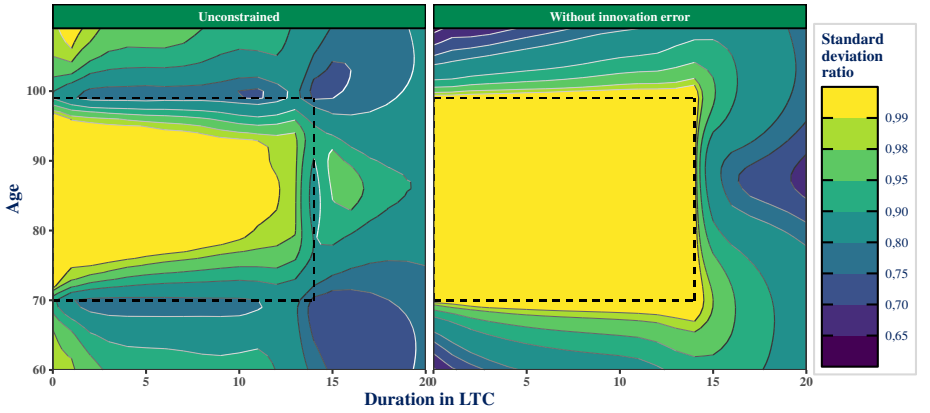


Figure 11. *Ratio of standard deviation of log-mortality rates from the three extrapolation methods. Left: unconstrained versus constrained with innovation error. Right: constrained without versus with innovation error. In both, the denominator is the fully constrained method of Figure 9.*

8. Discussion

Choosing the order of the penalization

Throughout this work, we have assumed second-order difference matrices for penalization. This choice is both standard and meaningful: from a Bayesian perspective, it corresponds to a prior belief that the log-transformed quantity of interest evolves linearly, which implies exponential behaviour on the original scale – consistent with actuarial models such as Gompertz.

The difference order directly shapes both the estimated trend and its extrapolation: higher-order penalties allow for more flexibility, but may induce unstable or erratic behaviour outside the data range. While Whittaker originally used third-order differences and higher orders can marginally improve model fit according to information criteria such as AIC, second-order penalties typically offer a robust compromise between smoothness, interpretability, and extrapolation stability. A detailed evaluation is provided in Section G of the Online Supplementary Materials.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

Summary of contributions

This paper revisits the classical WH smoothing approach through the lens of modern statistical modelling. Each section brought forward a key practical insight:

- Section 2 established that WH smoothing is more than an empirical method. It has a firm Bayesian foundation. Under Gaussian assumptions, credibility intervals may be derived and used as practical substitutes to confidence intervals.
- Section 3 clarified how to construct observation and weight vectors in survival analysis models: using log-crude rates as observations and event counts as weights yields a sound statistical formulation.
- Section 4 introduced generalized WH smoothing, in which the penalization is applied directly to the likelihood rather than a normal approximation. This refined method yields more accurate results, especially in situations where the available data volume is limited, but the number of combinations is high, such as in the two-dimensional case.
- Section 5 advocated for smoothing parameter selection via marginal likelihood (or its Laplace approximation, LAML), offering a principled and robust alternative to heuristic criteria like AIC or GCV.
- Section 6 presented two computational improvements: one exploits the banded structure of WH matrices to reduce runtime by up to a factor of 25; the other relies on reduced-rank smoothing basis, leading to even faster estimation (up to 250 $\times$ speed-up) with limited loss in accuracy, slightly outperforming P-splines.
- Section 7 addressed extrapolation: while WH smoothing naturally extends beyond the data range, constraints are needed in two dimensions to preserve the original fit. We proposed a method to extrapolate while accounting for both structural uncertainty and innovation error and provided credible intervals accordingly.

All these techniques are available in the `WH` package for the statistical software `R` (R Core Team, 2025), including automated smoothing parameter selection and constrained extrapolation with uncertainty quantification.

Limitations and outlook

Despite its strong practical appeal, WH smoothing has limitations that suggest several avenues for future work:

- Regular spacing requirement: WH smoothing assumes evenly spaced observations, which aligns well with standard life insurance grids (age and/or duration). However, this is less suitable when events are concentrated in a short period, such as in disability or long-term care claims. One solution is to combine finer discretization in early durations with methods like P-splines that accommodate irregular grids. Alternatively, and adaptive WH smoothing procedure (based on the ideas in Ruppert and Carroll, 2000; Krivobokova *et al*., 2008) could offer a way to retain regular spacing while varying the smoothness locally.
- Limited covariate handling: The basic WH framework does not accommodate additional explanatory variables (e.g., gender or policy features). However, WH smoothing can be extended using ideas from smoothing spline ANOVA and hierarchical models (Lee and Durban, 2011; Gu, 2013), allowing for structured random effects and flexible interactions. This opens the door to richer, more personalized experience modelling while preserving interpretability.

In sum, revisiting WH smoothing through a modern lens reinforces its theoretical foundations and offers practitioners fast, transparent, and adaptable tools for experience modelling. It remains a compelling alternative to more recent – yet often more opaque – techniques when working with evenly spaced discrete data.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/asb.2025.10061.

Acknowledgments. The author thanks the two anonymous referees for their valuable comments, which helped improve the quality and clarity of the paper.

Data availability statement. The synthetic data used to conduct the performance comparisons in this study are available from the author upon request.

Competing interests. The author declares no competing interests.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, *1973*.

Anderssen, R.S. and Bloomfield, P. (1974) A time series approach to numerical differentiation. *Technometrics*, **16**(1), 69–75.

Bohlmann, G. (1899) Ein Ausgleichungsproblem. *Nachrichten von Der Gesellschaft Der Wissenschaften Zu Göttingen, Mathematisch-Physikalische Klasse*, **1899**, 260–271.

Brent, R.P. (1973) *Algorithms for Minimization Without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall.

Brooks, R.J., Stone, M., Chan, F.Y. and Chan, L.K. (1988) Cross-validatory graduation. *Insurance: Mathematics and Economics*, **7**(1), 59–66.

Canadian Institute of Actuaries. (2017) Final Report: Canadian Mortality Improvement Model — MI-2017. Canadian Institute of Actuaries. https://www.cia-ica.ca/app/themes/wicket/custom/dl_file.php?fid=14926&p=35281.

Carballo, A., Durban, M., Kauermann, G. and Lee, D.-J. (2021) A general framework for prediction in penalized regression. *Statistical Modelling*, **21**(4), 293–312.

Carballo, A., Durban, M. and Lee, D.-J. (2021) Out-of-sample prediction in multidimensional p-Spline models. *Mathematics*, **9**(15), 1761.

Clifford, P. (1977) Nonidentifiability in stochastic models of illness and death. *Proceedings of the National Academy of Sciences*, 74(4), 1338–1340.

Cornea-Madeira, A. (2017) The explicit formula for the Hodrick-Prescott filter in a finite sample. *Review of Economics and Statistics*, **99**(2), 314–318.

Currie, I.D. (2013) Smoothing constrained generalized linear models with an application to the Lee-Carter model. *Statistical Modelling*, **13**(1), 69–93.

Currie, I.D., Durban, M. and Eilers, P.H.C. (2004) Smoothing and forecasting mortality rates. *Statistical Modelling*, **4**(4), 279–298.

Currie, I.D., Durban, M. and Eilers, P.H.C. (2006) Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(2), 259–280.

Delwarde, A., Denuit, M. and Eilers, P. (2007) Smoothing the Lee–Carter and Poisson log-bilinear models for mortality forecasting: A penalized log-likelihood approach. *Statistical Modelling*, **7**(1), 29–48.

Demmler, A. and Reinsch, C. (1975) Oscillation matrices with spline smoothing. *Numerische Mathematik*, **24**(5), 375–382.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22.

Eilers, P.H.C. and Marx, B.D. (1996) Flexible smoothing with *B*-splines and penalties. *Statistical Science*, **11**(2), 89–102.

Fellner, W.H. (1986) Robust estimation of variance components. *Technometrics*, **28**(1), 51–60.

Fix, E. and Neyman, J. (1951) A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biology*, **23**(3), 205–241.

Giesecke, L. and Defense Manpower Data Center. (1981) Use of the Chi-Square Statistic to Set Whittaker-Henderson Smoothing Coefficients. Smoothing.

Golub, G.H. and Van Loan, C.F. (2013) *Matrix Computations*. JHU Press.

Gschlössl, S., Schoenmaekers, P. and Denuit, M. (2011) Risk classification in life insurance: Methodology and case study. *European Actuarial Journal*, **1**, 23–41.

Gu, C. (1992) Cross-validating non-Gaussian data. *Journal of Computational and Graphical Statistics*, **1**(2), 169–179.

Gu, C. (2013) *Smoothing Spline ANOVA Models*, Vol. **297**. Springer.

Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models*, Vol. **43**. CRC Press.

Henderson, R. (1924) A new method of graduation. *Transactions of the Actuarial Society of America*, **25**, 29–40.

Hodrick, R.J. and Prescott, E.C. (1997) Postwar US business cycles: An empirical investigation. *Journal of Money, Credit, and Banking*, 1–16.

Hoem, J.M. (1971) Point estimation of forces of transition in demographic models. *Journal of the Royal Statistical Society: Series B (Methodological)*, **33**(2), 275–289.

Kauermann, G. (2005) A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planning and Inference*, **127**(1-2), 53–69.

Knorr, F.E. (1984) Multidimensional Whittaker-Henderson graduation. *Transactions of Society of Actuaries*, **36**, 213–255.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press

Krivobokova, T., Crainiceanu, C.M. and Kauermann, G. (2008) Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics*, **17**(1), 1–20.
Lee, D.-J. and Durban, M. (2011) P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, **11**(1), 49–69.
Marra, G. and Wood, S.N. (2012) Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, **39**(1), 53–74.
Nelder, J.A. and Mead, R. (1965) A simplex method for function minimization. *The Computer Journal*, **7**(4), 308–313.
Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, **135**(3), 370–384.
Patterson, H.D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.
R Core Team. (2025) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: *R Foundation for Statistical Computing*. https://www.R-project.org/.
Reinsch, C.H. (1967) Smoothing by spline functions. *Numerische Mathematik*, **10**(3), 177–183.
Reiss, P.T. and Todd Ogden, R. (2009) Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(2), 505–523.
Rodríguez-Álvarez, M.X., Lee, D.-J., Kneib, T., Durban, M. and Eilers, P. (2015) Fast smoothing parameter separation in multidimensional generalized p-splines: The SAP algorithm. *Statistics and Computing*, **25**(5), 941–957.
Rodríguez-Álvarez, M.X., Durban, M., Lee, D.-J. and Eilers, P.H.C. (2019) On the estimation of variance parameters in non-standard generalised linear mixed models: Application to penalised smoothing. *Statistics and Computing*, **29**, 483–500.
Ruppert, D. and Carroll, R.J. (2000) Theory & methods: Spatially-adaptive penalties for spline fitting. *Australian & New Zealand Journal of Statistics*, **42**(2), 205–223.
Schall, R. (1991) Estimation in generalized linear models with random effects. *Biometrika*, **78**(4), 719–727.
Society of Actuaries. (2018) A Practitioner's Guide to Statistical Mortality Graduation. https://www.soa.org/globalassets/assets/files/resources/tables-calcs-tools/2018-stat-mort-graduation.pdf.
Taylor, G. (1992) A Bayesian interpretation of Whittaker—Henderson graduation. *Insurance: Mathematics and Economics*, **11**(1), 7–16.
Verrall, R.J. (1993) A state space formulation of Whittaker graduation, with extensions. *Insurance: Mathematics and Economics*, **13**(1), 7–14.
Wahba, G. (1980) *Spline Bases, Regularization, and Generalized Cross Validation for Solving Approximation Problems with Large Quantities of Noisy Data*. University of Wisconsin.
Wahba, G. (1985) A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, 1378–1402.
Weinert, H.L. (2007) Efficient computation for Whittaker–Henderson smoothing. *Computational Statistics & Data Analysis*, **52**(2), 959–974.
Whittaker, E.T. (1923) On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41, 63–75.
Wood, S.N. (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.
Wood, S.N. (2017) *Generalized Additive Models: An Introduction with r*, 2nd ed. Chapman; Hall/CRC.
Wood, S.N. and Fasiolo, M. (2017) A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics*, **73**(4), 1071–1081.
Wood, S.N., Li, Z., Shaddick, G. and Augustin, N.H. (2017) Generalized additive models for Gigadata: Modeling the UK black smoke network daily data. *Journal of the American Statistical Association*, **112**(519), 1199–1210.

https://doi.org/10.1017/asb.2025.10061 Published online by Cambridge University Press