Revisiting Whittaker-Henderson Smoothing

Presentation at the IDS 2023 Conference

Guillaume Biessy

Head of R&D at LinkPact, Associate Professor at Sorbonne-Université

June 16, 2023

Introducing the paper

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

The paper proposes to reframe this smoothing technique within a modern statistical framework and addresses 6 questions of practical interest regarding its application :

1. How to measure uncertainty in smoothing results?
2. Which observation and weight vectors to use?
3. How to improve the accuracy of smoothing with limited data volume? (see the paper)
4. How to choose the smoothing parameter(s)?
5. How to improve numerical performance with a large number of data points? (see the paper)
6. How to extrapolate the smoothing results? (see the paper)

Whittaker-Henderson smoothing

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:

$$\hat{\mathbf{y}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \left\{ \underbrace{(\mathbf{y} - \boldsymbol{\theta})^T W (\mathbf{y} - \boldsymbol{\theta})}_{\text{fidelity criterion}} + \underbrace{\boldsymbol{\theta}^T P_\lambda \boldsymbol{\theta}}_{\text{smoothness criterion}} \right\}$$

$$\text{where} \quad W = \text{Diag}(\mathbf{w}) \quad \text{and} \quad P_\lambda = \begin{cases} \lambda D_{n,q}^T D_{n,q} & \text{in the one-dimensional case} \\ \lambda_x I_{n_z} \otimes D_{n_x,q_x}^T D_{n_x,q_x} + \lambda_z D_{n_z,q_z}^T D_{n_z,q_z} \otimes I_{n_x} & \text{in the two-dimensional case.} \end{cases}$$

$D_{n,q}$ is the order q difference matrix of dimensions $(n-q) \times n$, such that:

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

Whittaker-Henderson may easily be shown to have the explicit solution: $\hat{\mathbf{y}} = (W + P_\lambda)^{-1} W \mathbf{y}$.

Illustration of Whittaker-Henderson smoothing in the one-dimensional case

The effective degrees of freedom *edf* shown in this figure are calculated by summing the diagonal values of $H = (W + P_\lambda)^{-1} W$, the *hat matrix* of the model. They serve as a non-parametric equivalent of the number of independent parameters in parametric models but can take non-integer values.

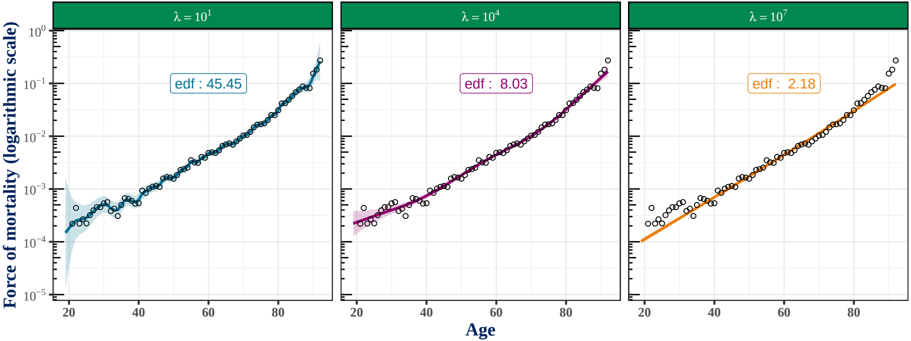

Figure 1: WH smoothing applied to a portfolio of synthetic mortality data for 3 choices of smoothing parameter.

How to measure uncertainty in smoothing results?

As $\mathbb{E}(\hat{\mathbf{y}}) = (W + P_\lambda)^{-1} W \mathbb{E}(\mathbf{y}) \neq \mathbb{E}(\mathbf{y})$ when $\lambda \neq 0$, the **frequentist approach** is biased and does not yield valid confidence intervals.

- The smoothness criterion may however be reframed as a $\boldsymbol{\theta} \sim \mathcal{N}(0, P_\lambda^-)$ **Bayesian prior**.
- Assuming $\mathbf{y}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, W^-)$ and using Bayes formula, it can be shown that :

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

- Therefore $\hat{\mathbf{y}}$ is also the **mode of the posterior distribution** of $\boldsymbol{\theta}|\mathbf{y}$.
- Using a second-order Taylor expansion at the mode, the posterior distribution can further be recognized as $\mathcal{N}(\hat{\mathbf{y}}, (W + P_\lambda)^{-1})$

Those assumptions yields **credibility intervals** for WH smoothing of the form:

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

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

Which observation and weight vectors to use?

- The previous credibility intervals requires that $\mathbf{y}$ be an vector of **independant observations** with **known variances** and that the weights $\mathbf{w}$ be chosen as the inverse of those variances.
- In the framework of left-truncated and right-censored longitudinal data, we assume independance between the insured lives' deaths and piecewise-constant force of mortality of the form: $\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{exp}(\boldsymbol{\theta})$ which is simply the crude rate estimator (the exp link ensure the rate positivity).
- The model log-likelihood takes the form: $\ell(\boldsymbol{\theta}) = \boldsymbol{\theta}^T\mathbf{d} - \mathbf{exp}(\boldsymbol{\theta})^T\mathbf{e_c}$ where $\mathbf{d}$ and $\mathbf{e_c}$ corresponds to the vectors of observed deaths and central exposure to risks respectively.
- The derivatives of the log-likelihood function for this model are given by:

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

This leads to the obvious solution $\hat{\boldsymbol{\theta}} = \mathbf{ln}(\mathbf{d}/\mathbf{e}_c)$. The properties of the maximum likelihood estimator imply that asymptotically $\mathbf{ln}(\mathbf{d}/\mathbf{e}_c) \sim \mathcal{N}(\mathbf{ln}\boldsymbol{\mu}, W^{-1})$, where W has elements $\mathbf{exp}(\hat{\boldsymbol{\theta}}) \odot \mathbf{e}_c = (\mathbf{d}/\mathbf{e}_c) \odot \mathbf{e}_c = \mathbf{d}$.

> This justifies applying WH smoothing to the observations vector $\mathbf{y} = \mathbf{ln}(\mathbf{d}/\mathbf{e}_c)$ and weights vector $\mathbf{w} = \mathbf{d}$ to estimate $\mathbf{ln}\,\boldsymbol{\mu}$ and then $\boldsymbol{\mu}$.

How to select the smoothing parameter(s)?

- To select the smoothing parameter, we adopt an (empirical) Bayes approach and try to maximize :

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

- Using the previous second-order Taylor expansion leads to the closed-form expression:

$$\ell^m_{\text{norm}}(\lambda) = -\frac{1}{2}\left[(\mathbf{y} - \hat{\mathbf{y}}_\lambda)^T W(\mathbf{y} - \hat{\mathbf{y}}_\lambda) + \hat{\mathbf{y}}_\lambda^T P_\lambda \hat{\mathbf{y}}_\lambda - \ln|W|_+ - \ln|P_\lambda|_+ + \ln|W + P_\lambda| + (n_* - q)\ln(2\pi)\right].$$

where $|A|_+$ denotes the product of non-zero eigenvalues of any square matrix A.

- $\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, \ell^m_{\text{norm}}(\lambda)$ does not have an explicit expression but may be obtained by numerical methods.

> For a given λ, all terms appearing in $\ell^m_{\text{norm}}(\lambda)$ are byproducts of the estimation of $\hat{\mathbf{y}}_\lambda$.

How to select the smoothing parameter(s)?

- The maximization of the marginal log-likelihood naturally $\ell^m_{\text{norm}}(\lambda)$ fits in the Bayesian interpretation of WH smoothing.
- While prediction error based criteria such as AIC or GCV have slightly better asymptotical properties, for finite-size samples they may lead to severe undersmoothing (see below).

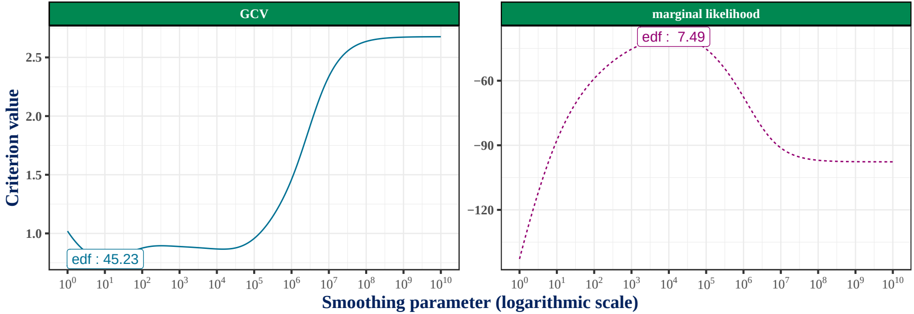

Figure 2: Comparison, in the context of one-dimensional WH smoothing parameter selection, of the Generalized Cross-Validation (GCV) criterion, which in this example leads to undersmoothing, and the marginal likelihood

Conclusion

Synthesis :

- Adopting a **Bayesian perspective** of Whittaker-Henderson smoothing provides a theoretical framework to obtain **credibility intervals** and select the **smoothing parameter(s)**
- This requires that $\mathbf{y}$ be a a vector of independant observations and the weights $\mathbf{w}$ be the inverse of the observations' variances. In the context of survival analysis, the maximum-likelihood estimator of **crude rates** asymptotically meets those requirements

What additional topics one may find in the paper :

- An intuitive representation of the smoothing based on **eigendecomposition of the penalization matrices**
- Further optimization of the smoothing **finite-size accuracy** and **large size** computation time and quantification of the associated gains in practical cases
- Natural **extrapolation** of the smoothing in the one-dimensional and two-dimensional cases

> The paper may be found here and the associated R package, named `WH` will soon be available on CRAN!