Statistical Modelling xxxx; **xx(x)**: 1–18

Sums of smooth exponentials to decompose complex series of counts

Carlo G Camarda[1], Paul HC Eilers[2] and Jutta Gampe[3]
[1]Institut National d'Etudes Démographiques, Paris, France
[2]Department of Biostatistics, Erasmus University Rotterdam, Rotterdam, the Netherlands
[3]Max Planck Institute for Demographic Research, Rostock, Germany

Abstract: Representing the conditional mean in Poisson regression directly as a sum of smooth components can provide a realistic model of the data generating process. Here, we present an approach that allows such an additive decomposition of the expected values of counts. The model can be formulated as a penalized composite link model and can, therefore, be estimated by a modified iteratively weighted least-squares algorithm. Further shape constraints on the smooth additive components can be enforced by additional penalties, and the model is extended to two dimensions. We present two applications that motivate the model and demonstrate the versatility of the approach.

Key words: additive components; composite link model; decomposition; penalty; shape constraints

Received March 2015; **revised** February 2016; **accepted** March 2016

1 Introduction

The generalized additive model can fit a sum of smooth components to observed counts; however, the sum is on the scale of the linear predictor. For count data, the usual model is a Poisson distribution with a log-link, that is, the logarithm of the expected values is expressed as a sum of smooth components. As a consequence, the model operates as a product of components on the scale of the count data. Often an additive decomposition of the expected values themselves is a more realistic description of the data generating process. Here, we introduce such a model that describes the expected values as a sum of components. We stay with the logarithmic link function and still require smoothness of the components, but the sum is taken after the exponential, instead of before. Hence, we call this a Sum of Smooth Exponentials (SSE) model. To further motivate the model, we give two examples where such an additive decomposition is a natural approach.

X-ray crystallography allows the exploration of the molecular and atomic structure of crystals. A physical sample is rotated while it is illuminated by a beam of

Address for correspondence: Carlo G. Camarda, Institut National d'Etudes Démographiques, 133, boulevard Davout, 75980 Paris Cédex 20, France.
E-mail: carlo-giovanni.camarda@ined.fr

 10.1177/1471082X16641796

X-rays. Depending on the angle, the number of diffracted photons varies and they are observed and counted by an optical detector. Figure 1 shows such an X-ray diffraction (XRD) scan of a thin layer of indium tin oxide. It was analyzed in detail by Davies *et al.* (2008), and the data are available in the R package diffractometry. The overall signal, the photon counts, is formed by a baseline and a number of peaks. The latter characterize the type of material and its physical state (like stress). The objective of such an XRD analysis is to decompose the signal into its components, to remove the baseline and to isolate the peaks, which can be analyzed further for their position, height, symmetry, and so forth.

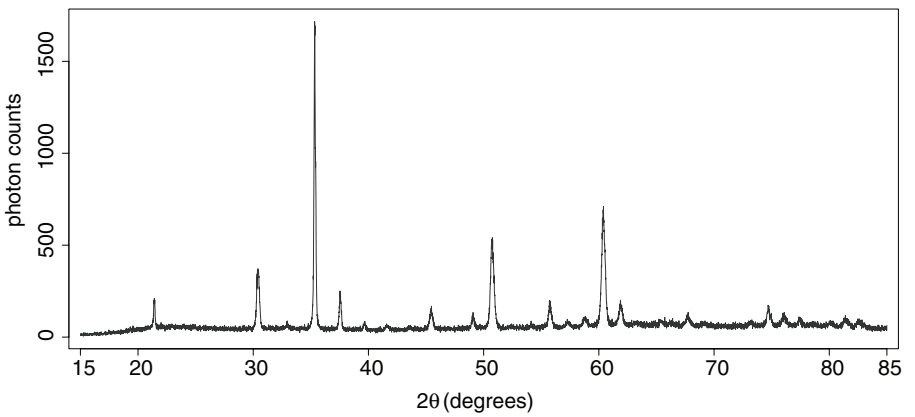

Figure 1 X-ray diffraction spectrum for indium oxide. The response y is the number of diffracted photons along twice the rotation angle (θ). The complete sequence contains measurements at 7 001 different angles between 15 and 85 degrees.

The second example deals with human mortality, and data are taken from the Human Mortality Database (2014). Death rates vary strongly across the age range, and so do causes of death. As a consequence, death at high ages is conceived as being different from the death of infants. Therefore, the decomposition of the mortality trajectory over age, which is shown in Figure 2, into several components has a long tradition. Heligman and Pollard (1980) suggested an additive three-component model, where the first component characterizes infant and child mortality, sharply decreasing after birth to very low values. The second component characterizes senescent mortality, which starts at low levels at middle-adult ages, around age 40, and increases exponentially. The central part, sometimes called middle-mortality, describes mortality after the onset of puberty and stretching into young adult ages. Although all causes of death contribute, this middle mortality is strongly related to risk-taking behaviour in young men and, at least historically, to mortality related to childbearing in women.

Siler (1983) modelled this central part as a constant hazard, and several modifications of this three-component additive model of human death rates were proposed to improve the fit to real data. However, the purely parametric forms of these models

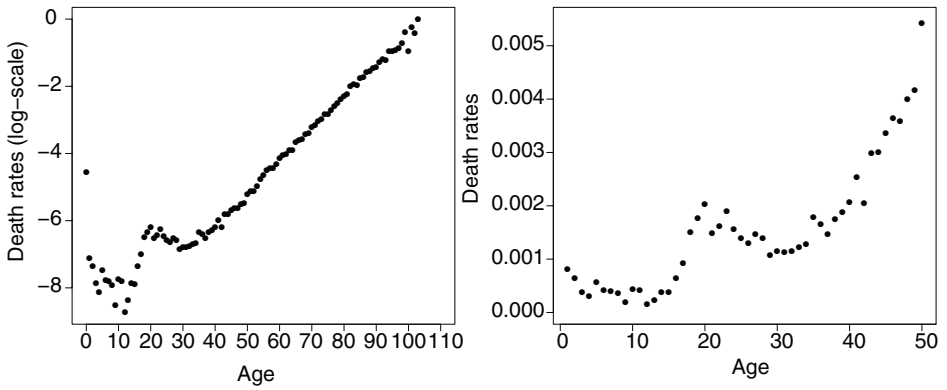

Figure 2 Age-specific death rates for Swiss males in 1980, ages 0 to 110 on log-scale (left) and ages 1 to 50 on original scale (right)

either do not fit well or they involve a rather high number of parameters. For instance, the model proposed by Heligman and Pollard (1980) attempts to describe mortality with eight or nine parameters.

In both examples, the overall mean trajectory, either of the photon counts or of the hazard of death, is seen as a sum of components, and the individual terms of the sum should be isolated in the analysis. In this article, we propose a novel approach to model and estimate such compound regression functions. While the individual components, which we assume to be smooth, are modelled on the log-scale, the final regression curve is an additive composition of these components on their original scale. Estimation of the SSE model can be achieved in a familiar framework as it can be formulated as a composite link model (CLM; Thompson and Baker, 1981) with a penalty added to guarantee smoothness (Eilers, 2007). Estimates are obtained by a modified version of the iteratively weighted least-squares (IWLS) algorithm.

The remainder of the article is organized as follows. In Section 2, we introduce the SSE model and demonstrate how it can be estimated as a penalized composite link model (PCML). We show its performance for the two examples presented in the introduction in Section 3 and extend the SSE model to two dimensions in Section 4. We conclude with a discussion.

2 Sums of smooth exponentials

2.1 The SSE model

We observe a series of counts $y_i, i = 1, \ldots, m$, at positions x_i, and the y_i are assumed to be Poisson distributed with expectation μ_i. The means μ_i are the sum of K components so that

$$\mu_i = \sum_{k=1}^{K} \gamma_{ik}. \tag{2.1}$$

Each component $\boldsymbol{\gamma}_k = (\gamma_{1k}, \ldots, \gamma_{mk})$ is assumed to be smooth across x and the logarithm of each component is expressed as a linear combination of B-splines:

$$\ln \gamma_{ik} = \sum_{j=1}^{J_k} B_{jk}(x_i)\alpha_{jk}. \tag{2.2}$$

The $B_{jk}(.)$ are the elements of the B-spline basis, and the α_{jk} are the corresponding coefficients. For each component, there is a separate set of basis functions $B_{jk}(.)$, which may also vary in size J_k, and a separate vector of coefficients $\boldsymbol{\alpha}_k$ with elements α_{jk}.

Hence, the mean (Equation 2.1) of the responses is

$$\mu_i = \sum_{k=1}^{K} \gamma_{ik} = \sum_{k=1}^{K} \exp\left(\sum_{j=1}^{J_k} b_{ijk}\alpha_{jk}\right) \tag{2.3}$$

where $b_{ijk} = B_{jk}(x_i)$. Smoothness of each component $\boldsymbol{\gamma}_k$ is enforced by a penalty on the coefficients in $\boldsymbol{\alpha}_k$. We choose a difference penalty of order d, see Eilers and Marx (1996). Since the mean in Equation (2.3) is expressed as the sum of exponentials of the smooth components, we call this a sum of smooth exponentials (SSE) model.

To estimate the parameters $\boldsymbol{\alpha}_k$, $k = 1, \ldots, K$, we minimize the penalized deviance

$$Q = \mathrm{DEV}(\boldsymbol{y}|\boldsymbol{\mu}) + \sum_{k=1}^{K} \lambda_k ||\boldsymbol{D}_k\boldsymbol{\alpha}_k||^2 = 2\sum_{i=1}^{m} y_i \ln(y_i/\mu_i) + \sum_{k=1}^{K} \lambda_k ||\boldsymbol{D}_k\boldsymbol{\alpha}_k||^2. \tag{2.4}$$

Here, the $\boldsymbol{D}_k$ are matrices that form d^{th}-order differences of $\boldsymbol{\alpha}_k$, and the λ_k are the smoothing parameters that tune the strength of the respective penalty. The penalty order d can vary across components.

Note that the penalties work on the level of the logarithm of the components, $\ln \boldsymbol{\gamma}_k$, while the mean is the sum of the $\boldsymbol{\gamma}_k$. If $\boldsymbol{\alpha}_k$ is a quadratic sequence then, for a third order penalty $\boldsymbol{D}_k$, all elements of $\boldsymbol{D}_k\boldsymbol{\alpha}_k$ will be zero, and so the penalty $||\boldsymbol{D}_k\boldsymbol{\alpha}_k||^2$ will be zero, too. That is, in terms of the third order penalty, such an $\boldsymbol{\alpha}_k$ is utmost smooth, and so is $\ln \boldsymbol{\gamma}_k = \boldsymbol{B}_k\boldsymbol{\alpha}_k$. But $\boldsymbol{\gamma}_k = \exp(\boldsymbol{B}_k\boldsymbol{\alpha}_k)$ has the shape of a normal density, which is relatively complicated and shows considerable variation in its curvature. This illustrates that adding such smooth exponentials not only guarantees a positive overall mean, but it can also can provide components of rather different curvatures.

The use of splines is not mandatory. For one or several of the K components, a parametric linear model, like a polynomial of low order, for $\ln \boldsymbol{\gamma}_k$ might be appropriate. Such parametric model components can be included easily, and they will have no associated penalty.

2.2 Fitting with the composite link model

The composite link model (CLM) was proposed by Thompson and Baker (1981) as a generalization of the generalized linear model (GLM; McCullagh and Nelder, 1989). Instead of modelling the expectations of observed data directly, it is written as a sum of GLM components. Eilers (2007) extended the CLM with a penalty for the estimation of latent smooth components from grouped or overdispersed data, the penalized composite link model (PCLM). The SSE model is a special case of the PCLM, as we will now show. To keep the description simple, we consider a model with two components, for example, the leftmost peak in the XRD scan with its neighbourhood, say diffraction angles between 20 and 23 degrees (see Figure 1).

Let $\boldsymbol{B}_1$ be a B-spline basis with a relatively large number of basis functions. It is used to model the peak, and so it should allow sufficient detail to follow the steep rise and decay. On the other hand, we do not need much detail for the baseline, so we can describe it with a relatively small number of B-splines in the basis $\boldsymbol{B}_2$. The model is

$$\boldsymbol{\mu} = \boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2 = \exp(\boldsymbol{B}_1\,\boldsymbol{\alpha}_1) + \exp(\boldsymbol{B}_2\,\boldsymbol{\alpha}_2). \tag{2.5}$$

Here, $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$, as well as $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, are vectors. If we define

$$\boldsymbol{\alpha} = \begin{pmatrix}\boldsymbol{\alpha}_1\\ \boldsymbol{\alpha}_2\end{pmatrix}; \quad \boldsymbol{\gamma} = \begin{pmatrix}\boldsymbol{\gamma}_1\\ \boldsymbol{\gamma}_2\end{pmatrix}; \quad \boldsymbol{C} = (\boldsymbol{I}\;\boldsymbol{I}); \quad \boldsymbol{B} = \begin{pmatrix}\boldsymbol{B}_1 & 0\\ 0 & \boldsymbol{B}_2\end{pmatrix}, \tag{2.6}$$

we can write

$$\boldsymbol{\mu} = \boldsymbol{C}\,\boldsymbol{\gamma} = \boldsymbol{C}\,\exp(\boldsymbol{B}\,\boldsymbol{\alpha}). \tag{2.7}$$

In the definition of the matrix $\boldsymbol{B}$, the zeros represent matrices of the proper size, filled with zeros. In the definition of $\boldsymbol{C}$, $\boldsymbol{I}$ is the identity matrix with m rows and columns, so that for each μ_i, the respective elements of $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are selected and added. By introducing

$$\boldsymbol{P} = \begin{pmatrix}\lambda_1 \boldsymbol{D}_1'\,\boldsymbol{D}_1 & 0\\ 0 & \lambda_2 \boldsymbol{D}_2'\,\boldsymbol{D}_2\end{pmatrix} \tag{2.8}$$

we can rewrite the sum of penalties as

$$\lambda_1 ||\boldsymbol{D}_1\,\boldsymbol{\alpha}_1||^2 + \lambda_2 |||\boldsymbol{D}_2\,\boldsymbol{\alpha}_2||^2 = \boldsymbol{\alpha}'\,\boldsymbol{P}\,\boldsymbol{\alpha}. \tag{2.9}$$

Again, in the definition of the matrix $\boldsymbol{P}$, the zeros represent matrices of the proper size, filled with zeros.

With the definitions in Equations (2.5) to (2.9), we now have a special case of the penalized CLM. Following Thompson and Baker (1981) and Eilers (2007), estimation of the model is achieved by repeatedly solving the system

$$(\check{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\check{\boldsymbol{X}} + \boldsymbol{P})\,\tilde{\boldsymbol{\alpha}} = \check{\boldsymbol{X}}'(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) + \check{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\check{\boldsymbol{X}}\tilde{\boldsymbol{\alpha}} \tag{2.10}$$

until convergence. Here,

$$\check{\boldsymbol{X}} = \tilde{\boldsymbol{W}}^{-1}\,\boldsymbol{C}\,\tilde{\boldsymbol{\Gamma}}\,\boldsymbol{B}; \qquad \tilde{\boldsymbol{W}} = \mathrm{diag}(\tilde{\boldsymbol{\mu}}); \qquad \tilde{\boldsymbol{\Gamma}} = \mathrm{diag}(\tilde{\boldsymbol{\gamma}}). \tag{2.11}$$

A tilde, as in $\tilde{\boldsymbol{\alpha}}$, indicates the current approximation to the solution.

Several practical issues need attention. The model is highly non-linear in the parameters, so reasonable starting values are important. A convenient way is to start with trial values for the $\boldsymbol{\gamma}$ components and take their logarithms to get a first $\tilde{\boldsymbol{\alpha}}$. When fitting peaks, like in the XRD example, it is wise to limit the domain of the B-splines to the neighbourhood of the peak. It reduces the number of parameters and thereby speeds up the calculations, but it also prevents numerical underflow. In the case of a peak, $\boldsymbol{\eta}_1 = \boldsymbol{B}_1\,\boldsymbol{\alpha}_1$ will resemble a concave parabola. If it is evaluated on a wide domain around the peak, it might happen that $\boldsymbol{\eta}$ takes on large negative values, which leads to a bad numerical condition of the equations. When using a smaller domain for $\boldsymbol{B}_2$, the composition matrix $\boldsymbol{C}$ has to be changed to $\boldsymbol{C} = (\boldsymbol{I} \quad \boldsymbol{E}_2)$, with $\boldsymbol{E}_2' = (\boldsymbol{O}_1 \quad \boldsymbol{I}_2 \quad \boldsymbol{O}_2)$. Here, $\boldsymbol{I}_2$ is an identity matrix with m' rows and columns, where m' is the number of rows of $\boldsymbol{B}_2$. The matrices $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$ are filled with zeros.

Additional features of the components, beyond smoothness, can be enforced by further penalties. In the mortality example, which we will discuss in detail in Section 3.2, we can require the senescent component to be strictly increasing with age, while the early adult component should be log-concave to obtain the hump shape. Such additional constraints, on monotonicity or on shape, can be implemented by a second penalty for the respective component (see Bollaerts *et al.*, 2006). Incorporating such additional knowledge about the components via an additional penalty can considerably facilitate, or is even required for, the identification of the single components in complex models.

Several smoothing parameters λ_k need to be determined; the specific number depends on the number of components in the model that are penalized for smoothness. We minimize the Bayesian information criterion (BIC) to determine the optimal values for the λ_k. This minimization is performed by a (multidimensional) grid search. The effective dimension (ED) of the model therein is calculated as the trace of the so called hat matrix $\boldsymbol{H}$:

$$\boldsymbol{H} = \check{\boldsymbol{X}}(\check{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\check{\boldsymbol{X}} + \boldsymbol{P})^{-1}\check{\boldsymbol{X}}'\tilde{\boldsymbol{W}} \quad \text{and} \quad \mathrm{ED} = \mathtt{trace}(\boldsymbol{H}).$$

In the following section, we illustrate the set-up of the SSE model in the two applications and show the estimation results. The R-code for all examples can be found in the supplementary material.

3 Applications

3.1 SSE model of X-ray diffraction spectrum

Figure 1 shows the XRD scan of indium oxide over 7 001 angles. To demonstrate the SSE approach, we focus on the first $m = 1\,750$ data points covering the angles

from 15 to 32.49, which contain the first two peaks of the spectrum. The signal will be decomposed in three additive components. The first component $\boldsymbol{\gamma}_1$ represents the baseline and is slowly varying so that we may capture it by a moderate number of *B*-splines. We placed 80 equidistant knots over the range of angles, leading to a total number of $J_1 = 83$ cubic *B*-splines (see Equation [2.2]).

The two further components $\boldsymbol{\gamma}_2$ and $\boldsymbol{\gamma}_3$, which represent the two peaks, will extend only in the neighbourhood of the peaks. To determine the position and neighbourhood of the peaks, we could either visually inspect the signal or choose a data-driven and less ad-hoc procedure as follows: We smooth the signal by deliberately oversmoothing and determine the first derivative of the resulting *P*-spline estimate, see Figure 3. The zero-crossings of the first derivative give the peak positions in good approximation, and the neighbourhood for the *B*-spline basis is determined from the relative maximum and minimum next to the zero-crossing, respectively.

This procedure selects 122 data points for the surrounding of the first peak (angles 20.71 to 21.92) and 137 data points for the second peak (29.72 to 31.08). In both cases, we put a knot at every fifth data point, resulting in $J_2 = 27$ and $J_3 = 30$ cubic *B*-splines for the two peak components $\boldsymbol{\gamma}_2$ and $\boldsymbol{\gamma}_3$, respectively. Wider surroundings could be chosen to be more conservative at the cost of higher computational time. Moreover, one should be aware that amount of smoothness in Figure 3 regulates the number of considered peaks: wiser choice is thus needed.

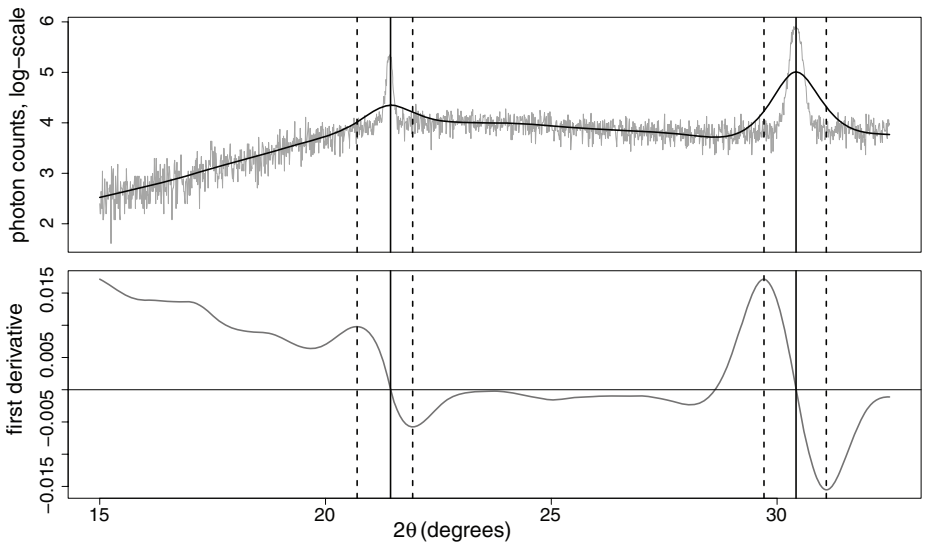

Figure 3 Range identification of spike components: The signal is smoothed by *P*-spline Poisson regression, deliberately oversmoothing the peaks (top panel). From the first derivative of the fit (bottom panel), the positions (zero-crossings) and relevant neighbourhood (relative maximum to minimum around the zero-crossing) are extracted for defining the domain of the corresponding SSE component.

All three components are assumed to be smooth, and the order of the difference penalty was chosen as $d = 2$ for the baseline and $d = 3$ for the two spikes. Additionally, the spikes are unimodal and, therefore, an additional shape-constraining penalty is added for each spike to ensure that the estimate is log-concave. The penalty matrix $\boldsymbol{P}_k^c$ that enforces the log-concave behaviour of component k is given by (Bollaerts *et al.*, 2006)

$$\boldsymbol{P}_k^c = \kappa \, \boldsymbol{D}_{k,2}' \, \boldsymbol{W}_k^c \, \boldsymbol{D}_{k,2} \,, \tag{3.1}$$

where $\boldsymbol{W}_k^c$ is a diagonal matrix that operates on the second-order differences of the coefficient vector $\boldsymbol{\alpha}_k = (\alpha_{jk})$ of the component (implemented via the matrix $\boldsymbol{D}_{k,2}$). Its diagonal elements are defined as

$$\boldsymbol{w}_{j,k}^c = \begin{cases} 1 \text{ if } \alpha_{j,k} \leq 2\alpha_{j-1,k} - \alpha_{j-2,k} \\ 0 \text{ otherwise.} \end{cases}$$

The penalty exerts influence only when the shape constraint is violated, and the weights in $\boldsymbol{W}_k^c$ are computed iteratively, that is, for each new value of $\boldsymbol{\alpha}_k$ during the iteration (2.10), the values of $\boldsymbol{W}_k^c$ are updated. The size of κ regulates how strictly the constraint is enforced. In the current example, we chose $\kappa = 10^5$. The overall penalty in this example hence is

$$\lambda_1 \, ||\boldsymbol{D}_1 \boldsymbol{\alpha}_1||^2 + \lambda_2 \, ||\boldsymbol{D}_2 \boldsymbol{\alpha}_2||^2 + \lambda_3 \, ||\boldsymbol{D}_3 \boldsymbol{\alpha}_3||^2 + \kappa \, \boldsymbol{P}_2^c \boldsymbol{\alpha}_2 + \kappa \, \boldsymbol{P}_3^c \boldsymbol{\alpha}_3 \,.$$

The matrix $\boldsymbol{D}_1$ implements second-order differences, while $\boldsymbol{D}_2$ and $\boldsymbol{D}_3$ implement third-order differences; $\boldsymbol{P}_2^c$ and $\boldsymbol{P}_3^c$ are defined as in Equation (3.1). Consequently, three smoothing parameters λ_k need to be determined. In this example, we chose $\lambda_2 = \lambda_3$ so that in the end, a two-dimensional grid search had to be performed to minimize the BIC

$$\text{BIC}(\lambda_1, \lambda_2 = \lambda_3) = \text{DEV}(\boldsymbol{y}|\boldsymbol{\mu}) + \ln(m) \cdot \text{ED} \,.$$

A different choice concerning the number of smoothing parameters is likely needed for analyzing the whole range of data as shown in Figure 1.

The grid extended over $16 \times 16 = 256$ λ-values (computing time 1.36 minutes on portable PC, Intel i5-3320M processor, 2.6 GHz, 4 GB RAM). Figure 4 shows a contour plot of the BIC profile.

The resulting estimates for the three components in the SSE model are shown in Figure 5. With this model, the $m = 1\,750$ data points are summarized by three components with effective dimensions 7.1 (baseline), 3.6 (first peak) and 7.9 (second peak), respectively, resulting in an overall effective dimension of ED = 18.6.

Once the signal is decomposed, the individual components can be analyzed further. For the XRD data, the position and shape of the peaks are of main interest, and relevant characteristics can be extracted from the estimated components $\hat{\boldsymbol{\gamma}}_2$ and $\hat{\boldsymbol{\gamma}}_3$, see central panels in Figure 5. Moreover, deviance residuals on the bottom panel

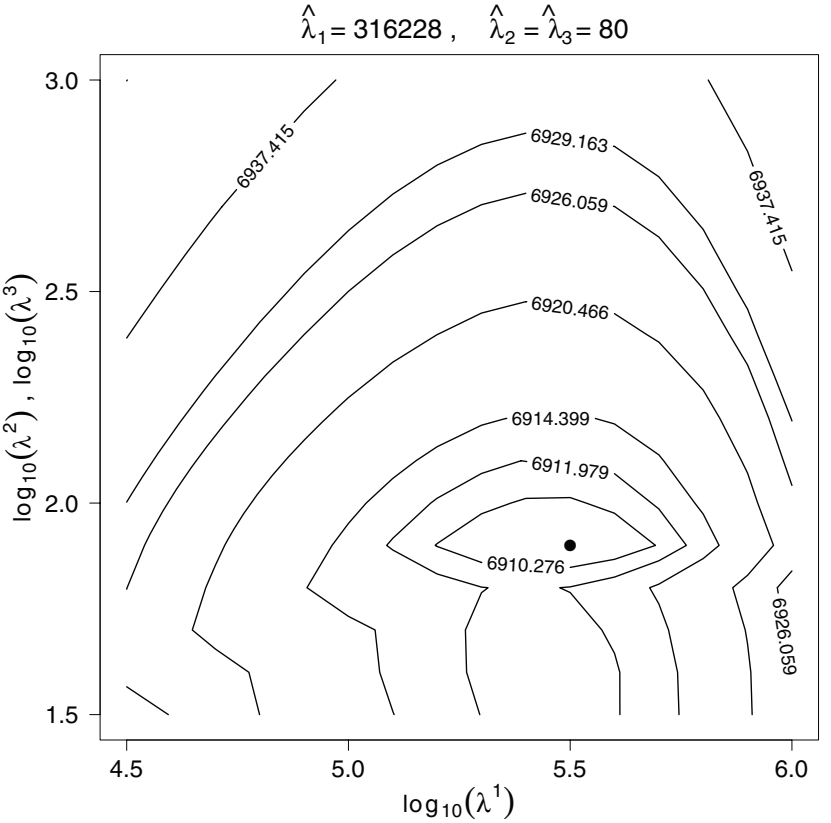

Figure 4 BIC contour plot over the two smoothing parameters for SSE model of the XRD spectrum example. The optimal values are at $\lambda_1 = 316228$ and at $\lambda_2 = \lambda_3 = 80$.

in Figure 5 suggest a good performance of the model: they are randomly scattered around zero with clear constant variance.

3.2 SSE model for the mortality trajectory

As death rates vary over several orders of magnitude, if the full age range is considered, they are often plotted on log-scale, as was done in the left panel of Figure 2. This also allows to recognize two features of human mortality that are typical and are generally found. First, mortality of newborns, that is, for age zero, is sharply higher than the death rates for later infant ages, with a majority of deaths occurring shortly after birth due to malformations, pre-term births, birth-related complications and the like. Therefore, the first year of life is usually dealt with separately, as in classic life table construction (Chiang, 1984), and constitutes a discontinuity in the mortality trajectory. Second, the exponential increase of death rates after about age

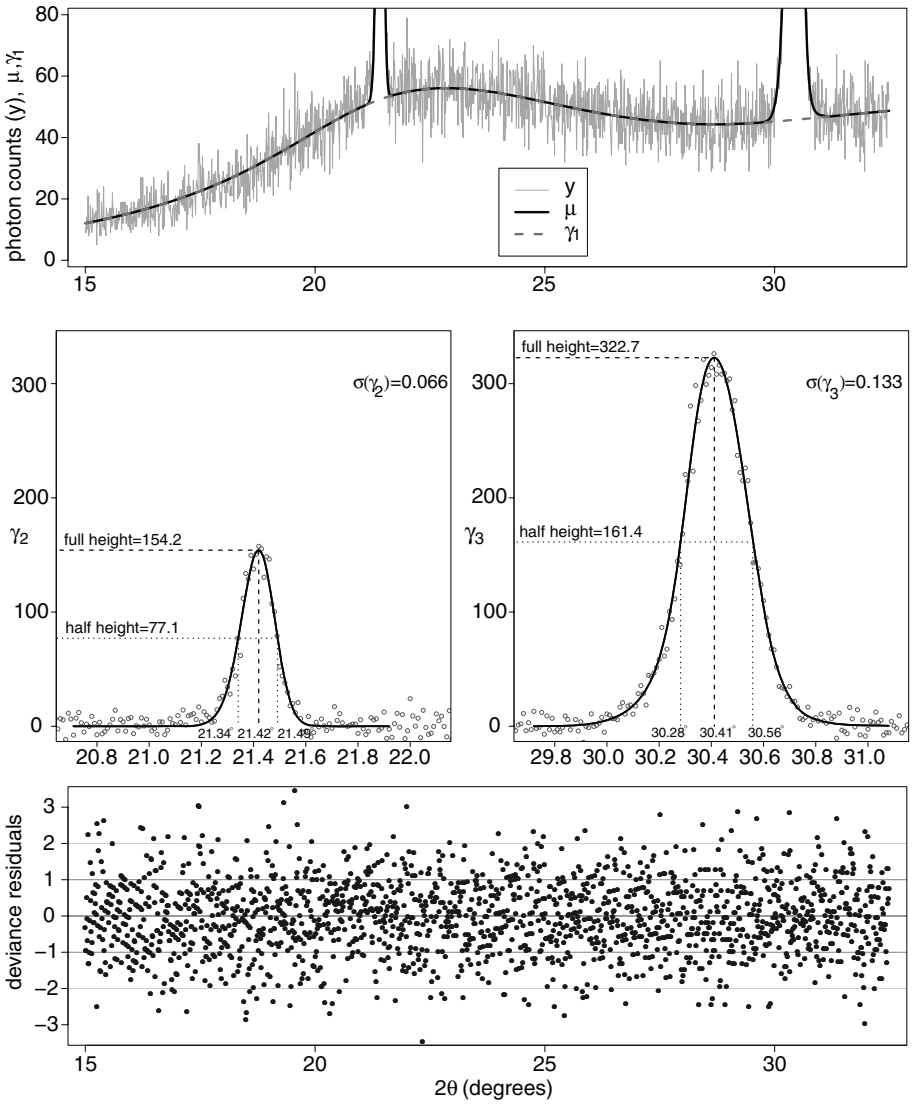

Figure 5 Estimates of the SSE model for the XRD spectrum data. Top panel: Observed counts, estimated baseline (dashed line) and sum of all three components (solid line). The vertical axis is clipped to better show the fit of the baseline signal. Central panel: Estimated components for the two spikes, along with the actual counts devoid of the estimated baseline. For each spike, the location of the mode as well as the angles where the component reaches its half-height are marked; the latter allows to characterize the (a)symmetry of the peak. The width of the photon distribution around the peak is described by the standard deviation (σ). Bottom panel: Deviance residuals for fitted SSE model.

40 is clearly seen as a linear pattern on the log-scale. This part is commonly referred to as senescence-related mortality.

Nevertheless, the additive decomposition that we have in mind is for the death rates themselves, portrayed in the right panel of Figure 2 for ages between 1 and 50 years. The falling mortality of children is followed by what is commonly called the 'accident hump', before the exponentially increasing senescent component takes over. The death rate at age zero is excluded for the reasons given above.

Hence, we intend to separate three components in an SSE model. The first component $\boldsymbol{\gamma}_1$ represents mortality for infants and children, while the second component $\boldsymbol{\gamma}_2$ captures senescence-related mortality. The third component $\boldsymbol{\gamma}_3$ describes the accident hump at early adult ages. It is realistic to assume that there is no sharp delimitation, where one component stops and another one continues, so that an additive composition adequately blends the transitions.

The data, taken from the Human Mortality Database, consist of the death counts y_i for ages $x_i = 1, 2, \ldots$, as well as the corresponding age-specific exposures e_i (person-years). The means of the counts $\boldsymbol{y}$ have to include the exposures $\boldsymbol{e}$ as a factor to the hazard of death so that Equation (2.3) becomes $\mu_i = \sum_k e_i \gamma_{ik}$. A simple change to the composition matrix $\mathbf{C}$, see Equation (2.6), allows us to incorporate the exposure information without changing the rest of the SSE model and its estimation. For a three-component model, we define

$$\mathbf{C} = \mathbf{1}_{1,3} \otimes \mathtt{diag}(\boldsymbol{e}) \,,$$

where $\mathbf{1}_{1,3}$ is a 1×3 matrix of ones and $\mathtt{diag}(\boldsymbol{e})$ is the diagonal matrix of the exposures.

All three components are modelled non-parametrically by cubic B-splines; however, besides the requirement of smoothness, additional but modest constraints are put on the components.

The first component $\boldsymbol{\gamma}_1$ is defined for ages 1 to 50 ($J_1 = 19$ cubic B-splines, difference penalty order $d = 2$) and is constrained to be monotone, decreasing by an additional penalty of the form (Bollaerts *et al.*, 2006)

$$\boldsymbol{P}_1^m = \kappa \boldsymbol{D}_{1,1}' \boldsymbol{W}_1^m \boldsymbol{D}_{1,1} \,, \tag{3.2}$$

where $\boldsymbol{D}_{1,1}$ forms first-order differences of the coefficients in $\boldsymbol{\alpha}_1$ and $\boldsymbol{W}_1^m$ is a diagonal matrix with diagonal elements equal to 1, if $\alpha_{j,1} \geq \alpha_{j-1,1}$ and 0 otherwise. Again, the value of κ regulates how strictly the monotonicity is enforced and we chose $\kappa = 10^5$.

The second aging-related component $\boldsymbol{\gamma}_2$ covers the full age range from 1 to 110 ($J_2 = 39$ B-splines, penalty order $d = 2$) and is required to be increasing. Therefore, it is constrained by a penalty similar to Equation (3.2), only the diagonal matrix $\boldsymbol{W}_2^m$ now filters differences for which $\alpha_{j,2} \leq \alpha_{j-1,2}$.

Finally, the third component $\boldsymbol{\gamma}_3$ is defined for ages between 1 and 80 years ($J_3 = 29$ B-splines, penalty order $d = 3$), and an additional penalty forces this component to be log-concave, see Equation (3.1). This guarantees the hump shape of early adult mortality.

In summary, three smoothing parameters, one for each component, need to be chosen. We perform a grid search over a total of $5 \times 7 \times 9 = 315$ combinations of λ-values to minimize the BIC, which was completed in 22 seconds (portable PC, Intel i5-3320M processor, 2.6 GHz, 4 GB RAM). The optimal values ($\lambda_1 = 10^4$, $\lambda_2 = 10^4$, $\lambda_3 = 10$) lead to three components with effective dimensions $ED_1 = 2$, $ED_2 = 4.7$ and $ED_3 = 3.7$. The results are shown in Figure 6.

The first component $\boldsymbol{\gamma}_1$ (infants and children) is estimated as being log-linear ($ED_1 = 2$), that is, the death rates are falling exponentially. This is in line with earlier suggestions (Siler, 1983); however, here it is the outcome of a more flexible model rather than an initial assumption. The second component $\boldsymbol{\gamma}_2$ (aging-related) is close to log-linear ($ED_2 = 4.7$); however, it is able to capture deviations from the Gompertz model (exponentially increasing death rates), which is expected to fit less well at higher ages (Horiuchi and Wilmoth, 1998). The accident hump component $\boldsymbol{\gamma}_3$ takes off in the middle of the teen years (onset of puberty) and reaches its maximum in the early twenties. By the end of the fifth decade, this component basically vanishes.

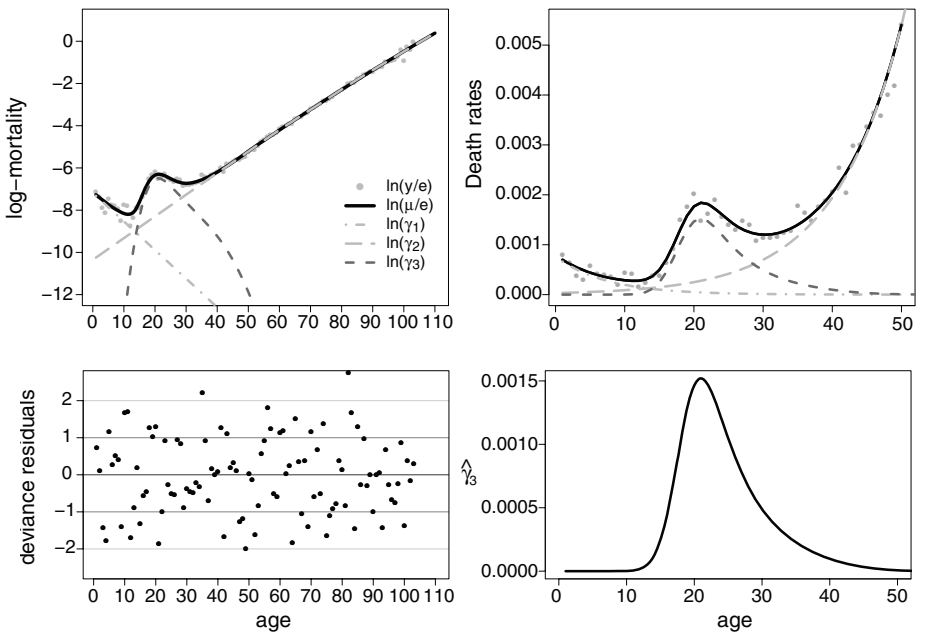

Figure 6 SSE model fitted to death rates, Swiss males in 1980. Top left: The three components and the overall fit, on log-scale for age range 1 to 110. Top right: SSE fit and the three components for ages 1 to 50. Bottom left: Deviance residuals for the SSE fit. Bottom right: Component $\boldsymbol{\gamma}_3$ ('accident hump') on original scale for ages 1 to 50.

4 Extending the SSE model to two dimensions

The model can be extended so that a sequence of SSE fits can be considered. As an illustration, we extend the mortality example of Section 3.2, which analyzed data in a single year, to the case where data in a series of calendar years $t_j, j = 1, \ldots, n$, are available. Changes in mortality over the years t_j possibly affect the single components differently, and we would like to see the evolution of the components separately.

We arrange the response data as a vector $\boldsymbol{y} = \mathtt{vec}(\boldsymbol{Y})$, where the matrix $\boldsymbol{Y} = (y_{ij})$ contains the observed numbers of deaths at age $x_i, i = 1, \ldots, m$ and in year $t_j, j = 1, \ldots, n$. The corresponding exposures are arranged in the same way as vector $\boldsymbol{e}$. Both $\boldsymbol{y}$ and $\boldsymbol{e}$ are of length $m \cdot n$.

The design matrices for the different components will depend on the assumptions that were made for the component. If a parametric specification is chosen for component k, then its design matrix will be

$$\boldsymbol{X}_k = \boldsymbol{I}_n \otimes \boldsymbol{X}_{k,x}.$$

A smoothness penalty enforces that the respective coefficients $\boldsymbol{\alpha}_k$ do not change abruptly between the measurement occasions (here: years):

$$\boldsymbol{P}_k = \lambda_k \boldsymbol{D}'_{k,d,n} \boldsymbol{D}_{k,d,n} \otimes \boldsymbol{I}_{p_k} . \tag{4.1}$$

The matrix $\boldsymbol{D}_{k,d,n}$ computes d-th order differences of the elements of $\boldsymbol{\alpha}_k$ across the years, and the amount of smoothness is tuned by λ_k. For instance, a linear structure in $\boldsymbol{X}_{k,x}$ with an additional penalty (Equation 4.1) implies log-linear $\boldsymbol{\gamma}_k$ over the ages with intercepts and slopes smoothly varying across years.

Non-parametric specifications again are modelled by B-splines. If a smooth two-dimensional surface is chosen for component k, the corresponding design matrix is given by

$$\boldsymbol{X}_k = \boldsymbol{B}_{k,t} \otimes \boldsymbol{B}_{k,x} .$$

The matrices $\boldsymbol{B}_{k,t} \in \mathbb{R}^{n \times q_k}$ and $\boldsymbol{B}_{k,x} \in \mathbb{R}^{m \times p_k}$ denote two univariate B-spline bases over years and ages, respectively. If different degrees of smoothness in the two directions are assumed (anisotropic smoothing), the penalty matrix is

$$\boldsymbol{P}_k = \lambda_{k,x} \boldsymbol{I}_{q_k} \otimes \boldsymbol{D}'_{k,d,x} \boldsymbol{D}_{k,d,x} + \lambda_{k,t} \boldsymbol{D}'_{k,d,t} \boldsymbol{D}_{k,d,t} \otimes \boldsymbol{I}_{p_k} , \tag{4.2}$$

where $\lambda_{k,x}$ and $\lambda_{k,t}$ are the smoothing parameters for the two dimensions of component k (Currie *et al.*, 2004). The difference matrices in Equation (4.2) act on the associated B-spline coefficients, either in the x- or in the t-direction.

Alternatively, an additive model can be considered for (the logarithm of) a given component. In this case, the design matrix is

$$\boldsymbol{X}_k = [\boldsymbol{1} : \boldsymbol{B}_{k,x} : \boldsymbol{B}_{k,t}]$$

with a block-diagonal matrix for the penalty term

$$\boldsymbol{P}_k = \lambda_k \, \mathtt{diag}(0, \, \boldsymbol{P}_{k,x} \, , \, \boldsymbol{P}_{k,t}) \, .$$

Both $\boldsymbol{P}_{k,x}$ and $\boldsymbol{P}_{k,t}$ are built as in the one-dimensional case. Since each of the columns of $\boldsymbol{B}_{k,x}$ and $\boldsymbol{B}_{k,t}$ sum to one, a small ridge penalty is needed for an additive component (Durbán *et al.*, 2002).

Again, shape constraints can be incorporated in the two-dimensional setting without changing the penalized IWLS algorithm. For example, if we assume a two-dimensional smooth surface for component k but we would like to enforce monotonicity or log-concaveness over ages for each single year, the additional penalty terms are given by

$$\begin{aligned} \boldsymbol{P}_k^m &= \kappa \, (\boldsymbol{I}_{q_k} \otimes \boldsymbol{D}_{k,1,x})' \, \boldsymbol{W}_k^m \, (\boldsymbol{I}_{q_k} \otimes \boldsymbol{D}_{k,1,x}) \qquad \text{and} \\ \boldsymbol{P}_k^c &= \kappa \, (\boldsymbol{I}_{q_k} \otimes \boldsymbol{D}_{k,2,x})' \, \boldsymbol{W}_k^c \, (\boldsymbol{I}_{q_k} \otimes \boldsymbol{D}_{k,2,x}) \, , \end{aligned} \tag{4.3}$$

respectively. The terms $\boldsymbol{W}_k^m$ and $\boldsymbol{W}_k^c$ are computed as in Equations (3.2) and (3.1). Alternatively, we can constrain monotonicity and log-concaveness over the years t_j for each single age by applying difference matrices over the t_j. A surface that is monotone in both directions can be estimated by adding monotonicity penalties for both dimensions.

4.1 Two-dimensional SSE model for mortality

We apply the two-dimensional SSE model to Swiss male mortality between 1980 and 2011. The three components—see Section 3.2—are each modelled as smooth surfaces over age and time. The time dimension (31 years) is covered by 13 cubic B-splines in all three components. The difference penalty in time direction is of order $d = 2$. The additional constraints—monotonicity on $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, log-concave shape of $\boldsymbol{\gamma}_3$—on the three components (in age-direction) are retained.

While smoothness in age-direction is determined separately for each component (has its own smoothing parameter), smoothness in the time-direction is assumed to be the same for all components so that only a single smoothing parameter is added. Hence, four smoothing parameters need to be chosen, and again we minimize the BIC over a grid of λ-values. In this application, we chose a grid of size $5^4 = 625$. The search was completed in 17.6 minutes (same hardware as before).

The estimated components for the optimal λ-values are shown for selected years in Figure 7. The evolution of the accident hump over the 31 years is shown in Figure 8.

This component shows a general decline of death rates—the peak level in 2011 is about 20 per cent of the value in 1980—but we can also see a change in the shape of this component during the 1990s, implying rising death rates during the fourth age-decade. This suggests further research on the causes-of-death structure and/or comparisons with other countries.

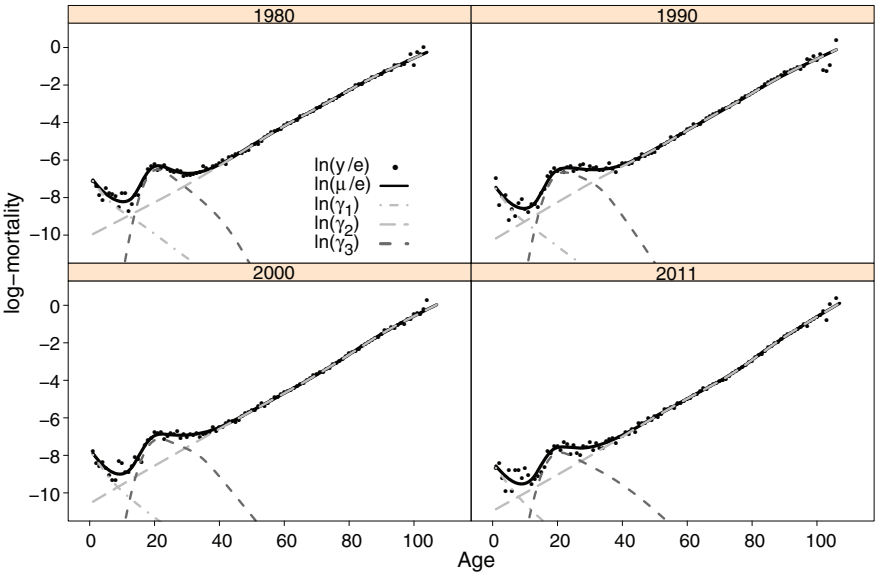

Figure 7 Components and overall fit resulting from a two-dimensional SSE model of Swiss male mortality 1980–2011. Results are shown for selected years and are plotted on log-scale. See also Figure 8.

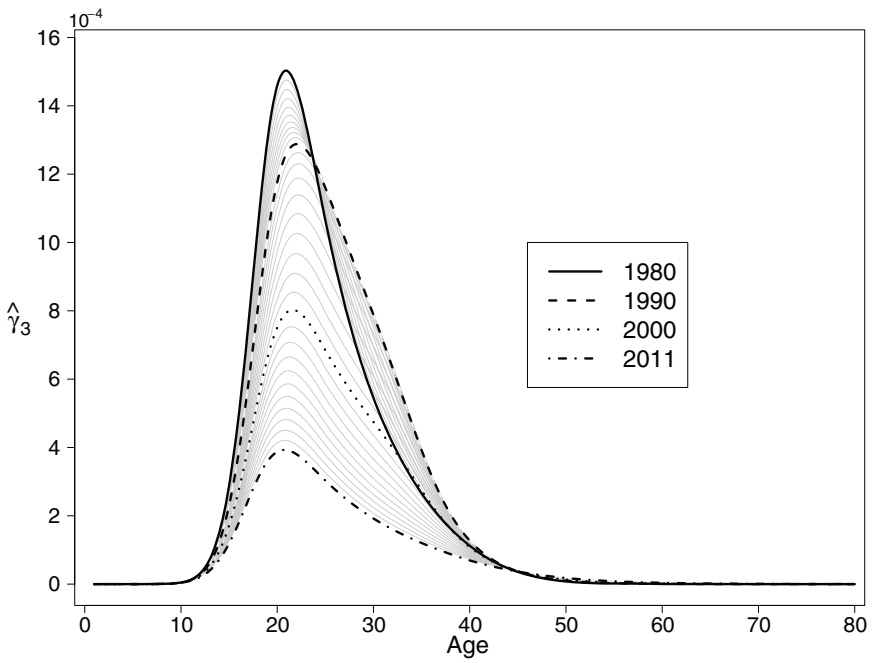

Figure 8 Component γ_3 (accident hump) estimated from a two-dimensional SSE model for Swiss male mortality 1980–2011. The curve for 1980 is very close to the estimate that was obtained from the univariate model and was shown in Figure 6.

5 Discussion

The SSE model offers a flexible but computationally feasible approach to directly model the mean trajectory of count data as a sum of smooth components. We view the methodology as a decomposition device to isolate the component(s) of interest. The two applications provide good examples and illustrate the usefulness of the approach.

With the decomposition idea in mind, the user will commonly have some knowledge of the phenomenon under study and the purpose of the analysis (e.g., removal of the baseline signal and analysis of the peaks). The number of components may be known (as in the mortality example) as well as some additional properties (such as the increase of aging-related mortality). The possibility of combining parametric with smooth components in the sum of predictors is an additional attractive feature of the SSE model in this context. Concerns about identifiability remain, particularly when all components are specified non-parametrically. Implementing known (or desired) properties via additional shape-constraining penalties can be essential. However, the examples provide evidence that this is a viable and successful endeavour. The individual components are readily available once the SSE model is fit, and their specific features can be analyzed immediately. Surely the sum of the exponentials gives a fit to the overall mean trajectory, and this can be used to smooth a highly non-linear regression function; however, this comes as a pleasant by-product. We do not recommend using the SSE approach as an alternative to adaptive smoothing, if there is no intention to identify components that are of interest in their own.

Finding good starting values can be important. In all applications, we provided starting components that were the results of some regression performed on subsets of the data. We refer to the supplementary material where these preliminary steps are detailed for each example. Once such starting values are determined, the computational effort is reasonable, even if the choice of several smoothing parameters is performed by a full grid search.

In both applications, there were plenty of data, and in the mortality example, we could have easily estimated a one-dimensional model for each calendar year separately. If data are sparser, however, the two-dimensional SSE model can be particularly useful for studying the change in some (or all) of the components.

To use SSE for the automatic analysis of XRD scans, reliable location of peaks is essential. Davies *et al.* (2008) used zero-crossing of the derivative to locate peaks, after initial smoothing with the taut string method. Apparently, their smoother introduces side lobes that result in a complicated pattern in the derivative at the peak centres. Our experience suggests that Poisson smoothing with *P*-splines (de Rooi *et al.*, 2014) and taking the derivative of the logarithm of the fit might work better.

de Rooi *et al.* (2014) consider the elimination of the $K\alpha_2$ artefact, the appearance of shadow peaks if the X-ray source emits at two different wavelengths. They use the PCLM. We plan to combine this idea with SSE.

Davies *et al.* (2008) model some peaks as a sum of two strongly overlapping peaks, using parametric models. It will be interesting to investigate the limits of SSE by trying to fit sums of overlapping components.

APPENDIX: COMPARISON TO ADAPTIVE SMOOTHERS

Our model is a structural smoother; it uses known characteristics of the data to decompose a signal as a sum of components with smooth logarithms. Our focus is on properties of these components, like locations and heights of peaks. However, the resulting fit to the data gives the impression as if a high quality adaptive smoother has been applied. This similarity has prompted the reviewers to ask for a comparison to adaptive smoothers. That is the subject of this appendix.

We have studied two packages, `AdaptFit` (Krivobokova, 2012) and `mgcv` (Wood, 2006), and challenged them with the indium oxide data. The Poisson family of distributions was specified with a logarithmic link. Unfortunately `AdaptFit` crashed on all occasions, reporting a fatal singularity. The other package was more robust and posed no problems. However, it is quite slow. On a portable PC, Intel i5-3320M processor, 2.6 GHz, 4 GB RAM, it can take more than 10 minutes to get a decent adaptive fit for a series of 1 750 observations.

Figure A1 shows data and fit, using cubic *B*-splines with two different settings. For the first estimation, we used 100 knots for the linear predictor and 50 knots for variable smoothness. Unwanted ripples are visible in the fit around the leftmost peak. Computation time was almost six minutes. The ripple gets smaller when 200 knots are used for the linear predictor. However, even with only 20 knots for adaptive smoothness, computation time is rather high (almost 11 minutes) and small ripples are still visible on the right tail of the first peak. We also increased the number of knots to 200 and 50 in linear predictor and variable smoothness, respectively. This

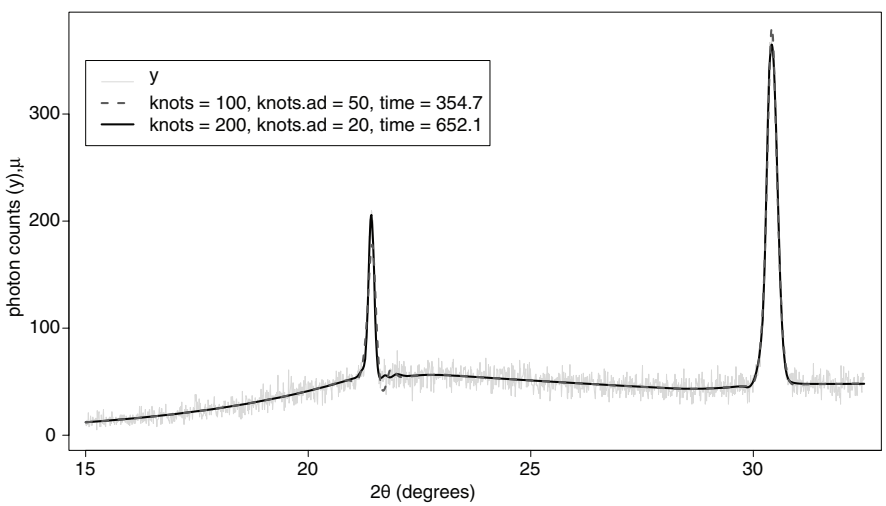

Figure A1 Adaptive smoothing for the XRD spectrum data using `mgcv`. For the linear predictor, different number of knots were used as well as different adaptive knots. The computation time reported was measured in seconds.

setting leads to 20 minutes of computation time and, though main peaks are well described, additional undesirable peaks are also captured (not shown here).

We also tested adaptive smoother from `mgcv` to our mortality examples (not shown here). Whereas in a one-dimensional setting this model provides a good estimation in few seconds, it is not capable to deal with our two-dimensional mortality example. Adaptive smoothing is thus not attractive.

Even if adaptive smoothing were fast and precise, it would not help us much for our applications, where the separation of components is the goal. We would have to estimate and subtract the baseline to get good estimates of the individual peaks. We do not want to position our model as an adaptive smoother either. It can only deal with sums of smooth positive components.

References

Bollaerts K, Eilers, PHC and van Mechelen I (2006) Simple and multiple P-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology*, **59**, 451–469.

Chiang CL (1984) *The Life Table and its Application*. Malabar, FL: Krieger.

Currie ID, Durbán M and Eilers PHC (2004) Smoothing and forecasting mortality rates. *Statistical Modelling*, **4**, 279–298.

Davies P, Gather U, Meise M, Mergel D and Mildenberger T (2008) Residual based localization and quantification of peaks in X-ray diffractograms. *Annals of Applied Statistics*, **2**, 861–886.

de Rooi JJ, van der Pers NM, Hendrikx RWA, Delhez R, Bottger AJ and Eilers PHC (2014) Smoothing of X-ray diffraction data and $K\alpha_2$ elimination using penalized likelihood and the composite link model. *Journal of Applied Crystallography*, **47**, 852–860.

Durbán M, Currie ID and Eilers PHC (2002) Using P-splines to smooth two-dimensional Poisson data. In *Proceedings of the 17th International Workshop of Statistical Modelling*, Chania, Greece.

Eilers PHC (2007) Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, **7**, 239–254.

Eilers PHC and Marx BD (1996) Flexible smoothing with *B*-splines and penalties (with discussion). *Statistical Science*, **11**, 89–102.

Heligman L and Pollard JH (1980) The age pattern of mortality. *Journal of the Institute of Actuaries*, **107**, 49–80.

Horiuchi S and Wilmoth JR (1998) Deceleration in the age pattern of mortality at older ages. *Demography*, **35**, 391–412.

Human Mortality Database (2014) University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Retrieved February 2014, from www.mortality.org

Krivobokova T (2012) *AdaptFit: Adaptive Semiparametic Regression*. Retrieved from https://CRAN.R-project.org/package=AdaptFit, R package version 0.2–2.

McCullagh P and Nelder JA (1989) *Generalized Linear Models*. Monographs on Statistics Applied Probability, 2nd edn. London: Chapman & Hall.

Siler W (1983) Parameters of mortality in human populations with widely varying life spans. *Statistics in Medicine*, **2**, 373–380.

Thompson R and Baker RJ (1981) Composite link functions in generalized linear models. *Applied Statistics*, **30**, 125–131.

Wood SN (2006) *Generalized Additive Models. An Introduction with R*. Boca Raton, FL: Chapman & Hall.