Joint Modelling of Male and Female Mortality Rates Using Adaptive P-splines

Kai Hon Tang *
Erengul Dodd †
Jonathan J. Forster ‡

Abstract

Raw mortality data often exhibit irregular patterns due to randomness. Graduation refers to the act of smoothing crude mortality rates. In this paper we propose a flexible and robust methodology for graduating mortality rates using adaptive P-splines. Since the observed data at high ages are often sparse and unreliable, we use an exponentially increasing penalty. We use mortality data of England and Wales and model male and female mortality rates jointly by means of penalties, achieving borrowing of information between the two sexes.

1 Introduction

Crude mortality rates often exhibits natural randomness and irregular patterns. Graduation or modelling of mortality rates refers to the act of smoothing crude mortality rates. Sometimes extrapolation to higher ages is also performed during the process since the usual assumed maximum lifespan of humans (e.g. 125) is often beyond the range of available data. Mortality rates are usually expressed in the form of life tables, which contains the number of expected lives and deaths at different ages. Life tables are widely used in the insurance sector as there are many insurance products related to mortality.

Several methods have been proposed in the past, however, due to the peculiar shape of the human mortality curve (e.g. accident hump in late teens, very high infant mortality rates), these methods often involve relatively complicated mathematical functions. It is only recently that non-parametric smoothing techniques

*Department of Infectious Disease Epidemiology, Imperial College London, UK, k.tang@imperial.ac.uk
†Mathematical Sciences, University of Southampton, UK
‡Statistics, University of Warwick, UK

have become popular and are employed in graduation of mortality rates. These non-parametric methods are more flexible and they do not assume any particular decomposition of the mortality schedule.

Mortality graduation models, sometimes also called 'laws of mortality', have become more sophisticated over time. One of the earliest models, Gompertz law of mortality (1825), states that the mortality rates is an exponential function of age, i.e. the log of mortality rates is linear in age. The law is then extended by Makeham in 1860 to the Gompertz-Makeham law where he proposed the addition of a constant, which captures mortality due to age-independent accidents. In spite of the simple and interpretable functions, these laws provide adequate fit to only limited age range, say 30 to 90, where the log-linearity assumption seems appropriate. The constant rate of increase (i.e. the log-linearity) in mortality rates might not be suitable at the very old ages. In fact, a decelerating rate of increase is often observed at these ages (Carriere, 1992). Perks (1932), Beard (1959) and Thatcher (1999) suggested to use logistic functions so that the mortality rates tend to an asymptote. At younger ages where mortality rates are lower, the logistic curve behaves similarly to the Gompertz or Gompertz-Makeham laws. Lindbergson (2001) suggested a piecewise model such that for ages less than a cut-off age the mortality rates follow the Gompertz-Makeham law, while after the cut-off age they are modelled by a linear function. Saikia & Borah (2014) did a comparative study on models for the oldest ages and showed that the logistic model is the most reasonable choice among the laws mentioned above. More recently, Pitacco (2016) provides a comprehensive review of old age models.

The models discussed so far are only applicable to the adult ages. Several more complicated mathematical functions have been proposed in an attempt to model mortality rates for the entire age range. Heligman & Pollard (1980) proposed the 8-parameters Heligman-Pollard Model. Carriere (1992) modelled the survival function and viewed it as a mixture of infant, young adult and adult survival functions. Despite the ability to model mortality rates of the whole age range with interpretable parameters, these models are often difficult to fit in practice, due to the high correlation in the estimated parameters. The high correlation also compromises interpretability.

Recently more flexible, non-parametric smoothing approaches are used to model mortality. These methods are used to produce the English Life Tables (ELTs), decennial life tables published by the Office for National Statistics (ONS) in every 10 years after every census. Since the 13-th English Life Table (ELT13), subsequent ELTs are all produced using spline-based methods. In ELT14 a variable knot cubic spline is adopted and the optimal number and location of knots are estimated. In ELT15 a weighted least squares smoothing spline is used, with a modification that the user specifies a set of weights based on their judgment in respect of the regions where the closest fit should be expected. In addition,

some data at the highest ages are discarded to maintain a monotonic (upward) progression in the graduated mortality rates. In ELT16 a Geometrically Designed variable knot regression spline (Kaishev *et al.*, 2006) is used. This is a 2-stage method where they first fit a variable knots linear spline, and then estimate the optimal control polygon of higher order splines, resulting in estimates of the optimal knot sequence, the order of the spline and the spline coefficients all together. Detailed information on methods used in the production of ELTs can be found in Gallop (2002). In the latest English Life Table, ELT17, Dodd *et al.* (2018) suggest to use a hybrid function for graduation. Specifically, mortality rates before some cut-off age are modelled using splines while mortality rates after the cut-off age are modelled using a parametric model (log-linear or logistic). At younger ages there is sufficiently dense data, therefore the flexibility of splines makes it a very effective tool in graduating mortality rates at these ages, whereas at the oldest ages a parametric function helps produce more robust estimates and extrapolation. The final graduation combines different threshold ages with weights determined by cross-validation error. This requires multiple fits and the transition from spline to the parametric model is not guaranteed to be smooth.

We extend the approach presented in Dodd *et al.* (2018) by using adaptive splines. Under this approach, instead of splitting the age into non-parametric and parametric regions, a P-spline is used for the whole age range with adaptive penalty. Contrary to traditional P-splines, instead of having a single smoothing parameter governing the overall smoothness, adaptive splines allow varying smoothness over the domain, hence enhancing flexibility and giving better fit to functions with changing smoothness. Ruppert & Carroll (2000) showed that even when the true underlying function is uniformly smooth, adaptive splines perform at least as well as traditional splines in terms of model fit. Different approaches have been suggested to estimate the varying smoothness penalty. Pintore *et al.* (2006) and Liu & Guo (2010) used piece-wise constant functions for the varying penalty. Ruppert & Carroll (2000) and Krivobokova *et al.* (2008) proposed to use a second layer of spline for the penalty. Storlie *et al.* (2010) suggested to estimate the changing smoothness from an initial fit with an ordinary spline while Yang & Hong (2017) recommended to weight the penalty inversely to the volatility of the data in proximity. Bayesian adaptive splines have also been investigated (see Baladandayuthapani *et al.* (2005), Crainiceanu *et al.* (2007), Jullion & Lambert (2007) and Scheipl & Kneib (2009)). In addition to smoothness, shape-constrained splines have also been proposed so that the resulting estimated splines satisfy some presumed functional form. Pya & Wood (2015) demonstrated how to achieve different shape constraints by re-parameterisation of the spline coefficients while Bollaerts *et al.* (2006) controlled the shape through iteratively adding asymmetric penalties. Camarda *et al.* (2016) proposed a sums of smooth exponentials model and demonstrated it for mortality graduation using shape-constrained

P-splines. Camarda (2019) applied shape constraints in the context of mortality forecasting.

The English Life Tables present smooth mortality rates for males and females. Although male and female mortality rates show similar patterns, male mortality rates are expected to be higher than female mortality rates at any age. When males and female mortality rates are modelled separately, problems may arise such as crossing-over rates at very high ages, where data is sparse, or a decreasing mortality trend in age. In the previous ELTs these problems are addressed using rather ad-hoc approaches. For example, in ELT14 an arbitrary value is chosen as the mortality rate at the closing age, while in ELT15 the data at the highest ages are discarded as it would produce downwards mortality profile otherwise. In ELT17, the male and female mortality rates are calculated as the weighted average of the respective graduated mortality rates starting at the age where they cross-over.

Various multi-population models have been proposed in the context of coherent mortality forecasting. Many of these concern the correlation between the period effects of different populations, for example, the common factor model by Li & Lee (2005) assumed that different populations share the same age-period effect, who further proposed the augmented common factor model where additional population specific age-period terms are included for a better fit, which are mean-reverting in the long run. Cairns *et al.* (2011) modelled mortality for two populations and assumed that the spread between the period effects (and cohort effects) of the two populations are mean-reverting. Dowd *et al.* (2011) also proposed a similar model where in addition to having mean-reverting spread between the time series, the period effect (and cohort effect) of a smaller population is actually being "pulled" towards that of the larger population, hence in the long run the projected mortality rates converge.

Another class of models is the relational models where usually a reference mortality schedule is first modelled and then the spread/ratio between another population of interest and the reference schedule. For example, Hyndman *et al.* (2013) proposed the Product-Ratio Model where the geometric mean of the populations are modelled as the reference schedule and the ratios are assumed to follow stationary processes so as to produce non-divergent forecasts. Villegas & Haberman (2014) modelled the mortality of different socio-economic sub-populations. Jarner & Kryger (2011) proposed the SAINT Model, where they model the ratio between the mortality rates of a sub-population to the reference using some orthogonal regressors. Biatat & Currie (2010) extended the 2-D spline model by Currie *et al.* (2004) to a joint model by adding another 2-D spline for the spread of a population to a reference population.

Limited focus has been placed on the convergence of the age-patterns of male and female mortality. Plat (2009) modelled the ratio between a sub-population and a main population and assumed the ratio to be one after a certain age, this is

because the difference between the sub-population mortality and main population mortality is expected to diminish at higher ages. Hilton *et al.* (2019) on the other hand used a logit link and assumed a common asymptote for males and females. In this paper we propose a coherent method for jointly smoothing male and female mortality rates. Our method allows us to borrow information, especially in areas where data is scarce and unreliable. The convergence of the male and female mortality curves is achieved by means of penalty.

In this paper we propose a model for mortality graduation using adaptive splines. This model is able to produce smooth mortality schedules for the entire age range and it is robust at the oldest ages where data is sparse. We also model male and female mortality rates together as modelling them independently causes inconsistencies such as divergent or intersecting trends, especially at the oldest ages. This is often addressed using ad-hoc methods (e.g. see previous ELTs). The joint model proposed in this paper does not require such interventions and is able to produce non-divergent and non-intersecting male and female mortality schedules. It also allows strength to be borrowed across sexes at the oldest ages which further increases robustness. More details can be found in Tang (2021).

The rest of the paper is structured as follows: in Section 2, the data is described and some notes on general mortality pattern are discussed. In Section 3, we introduce our methodology and present our results. In Section 4, the model is extended to model male and female mortality rates jointly. Finally, we provide a brief conclusion in Section 5.

2 Data

The data we use in this paper contains the number of deaths and mid-year population estimates by single year of age for males and females in England and Wales from 2010 to 2012, obtained from the Office for National Statistics (2019). Data by single age is available up to age 104 for both males and females, while data at age 105 and above are aggregated into one age group. Typically infant mortality rates are dealt with separately, therefore infant data are excluded, leaving us with data spanning from age 1 to 104. The central mortality rate at age x, m_x, is defined as

$$m_x = \frac{d_x}{E_x^C} \tag{1}$$

where d_x and E_x^C are the number of deaths and central exposure at age $[x, x+1)$ respectively. The crude central mortality rates $\tilde{m}_x$ are then obtained by approximating the central exposure with the mid-year population estimates. Furthermore,

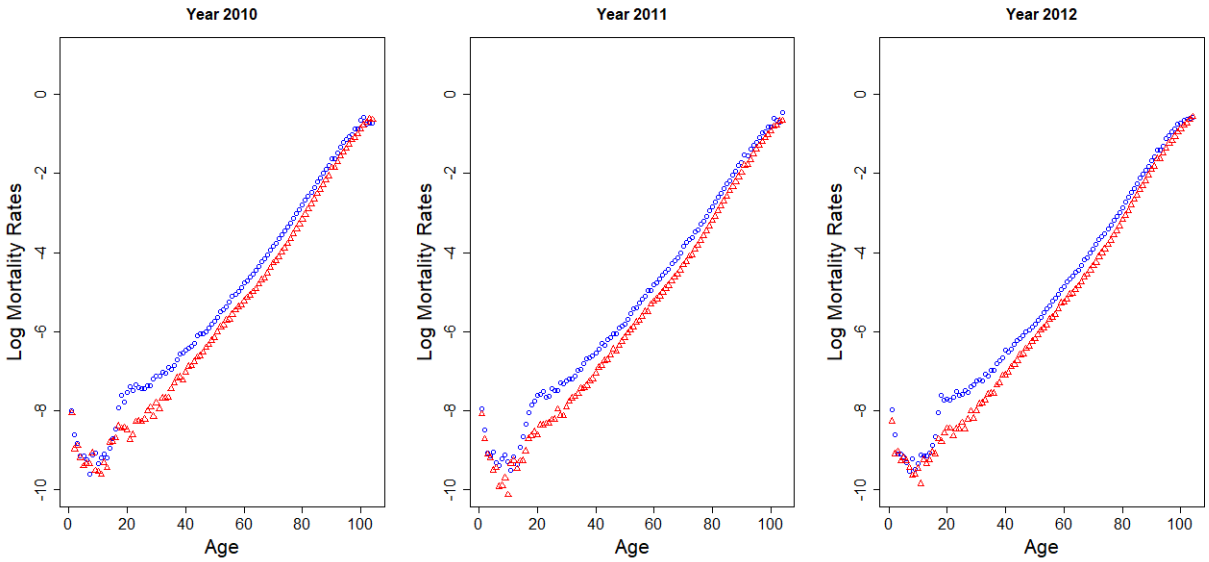


Figure 1: Crude Mortality Rates England and Wales. Blue and red points are the male and female crude mortality rates respectively.

it is assumed that the number of deaths at each age follow a Poisson distribution,

$$d_x \sim \text{Poisson}(E_x^C m_x) \tag{2}$$

Figure 1 shows the crude mortality rates for each year on a logarithmic scale. As expected, mortality is decreasing at child ages until around mid-teenage, followed by a sudden increase in mortality rates at late teenage, so-called "accident hump". Afterwards, the mortality increases steadily due to senescence. Male mortality rates lie above female mortality rates at almost all ages and that they are converging at the end. A decreasing rate of increase in mortality also seems to be taking place at the oldest ages. The crude mortality rates at the youngest and the oldest ages are more dispersed than those in the middle due to the sparsity of data at these ages.

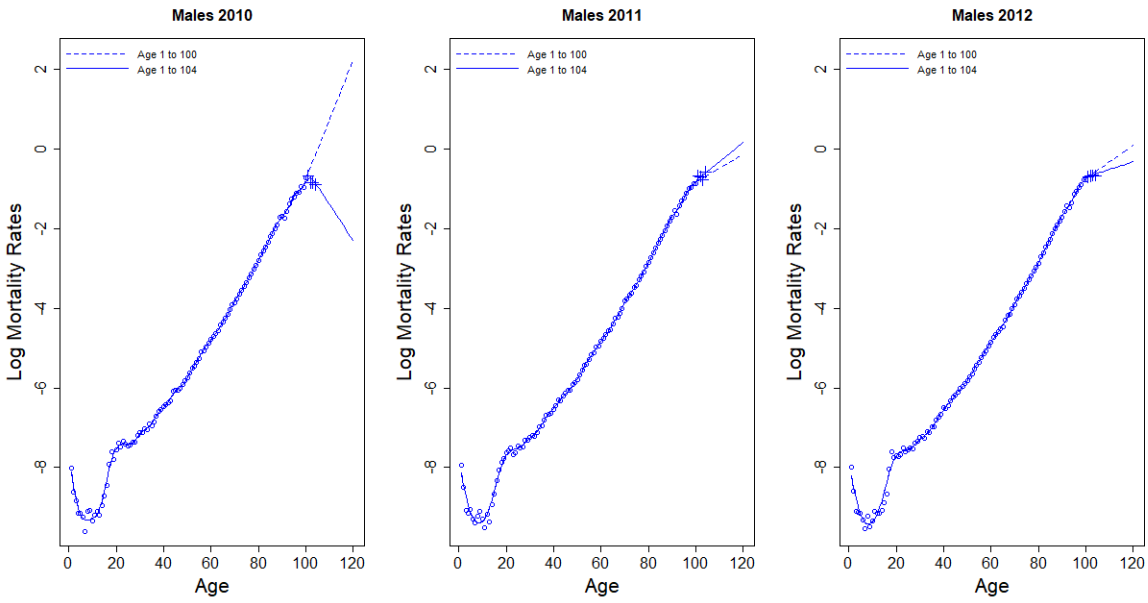
3 Methodology

In this Section we introduce our proposed model. Before moving onto the model specifics, first we shall have a look at the robustness problem when ordinary P-spline smoothing is used. Splines are piece-wise polynomials that are continuous up to certain order of derivatives and P-splines are one of the most commonly used splines having a B-spline basis with discrete penalties (Eilers & Marx, 1996). We use ordinary P-splines with 40 basis functions with equally spaced knots, and

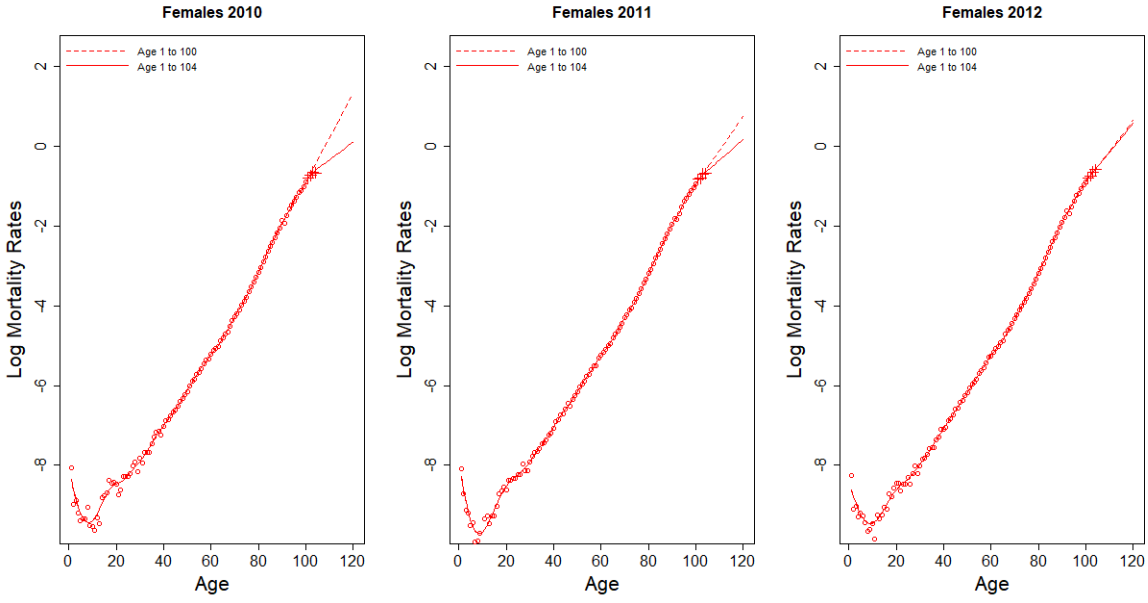
males and females have the same basis. The basis dimension is chosen according to preliminary analysis. The data is fitted with ordinary P-splines of dimension 20, 25, 30, 35, 40, 45, 50, 55, 60, 65 for each sex and the total effective degrees of freedom for males and females is examined. The total effective degrees of freedom starts to level-off at around 40 basis functions for each sex (80 total dimension), indicating that further increase in the dimension has limited benefits. Note that there is not a unified approach for choosing the dimension since the essence of penalised splines is that the user constructs a basis that is generous enough to capture the fluctuations of the underlying function and then penalises the roughness to prevent over-fitting.

Fitting the ordinary P-spline to the mortality data, several problems are revealed. Figure 2 plots the ordinary P-spline fits to the data. The circles and crosses are the crude mortality rates. The solid lines are the fits using data from age 1 to 104, while the dotted lines are the fits using data only up to age 100 (excluding the crosses). From Figure 2 it is clear that the fit to the oldest ages is not robust, the yearly variation in mortality pattern is irregular. Comparing the solid and dotted lines, the exclusion of the last 4 data points (crosses) changes the shapes quite drastically, again revealing the lack of robustness. Extrapolation based on these trends is very sensitive to the unreliable data at the oldest ages. Sometimes an unreasonable mortality schedule can also be obtained, for instance the estimated mortality rates for males in 2010 is decreasing at the oldest ages.

The root cause for the lack of robustness at high ages is that ordinary P-splines have only one smoothing parameter governing the overall smoothness, hence resulting in either over-smoothed young mortality rates (less likely due to the much bigger exposures than older ages) or under-smoothed old mortality rates. This motivates the use of P-splines with varying smoothness penalty. Contrary to a global smoothing parameter in ordinary P-spline tuning the overall smoothness, the smoothness penalty over the domain is allowed to vary. This is sometimes called "adaptive smoothing" or "adaptive spline". Ruppert & Carroll (2000) have shown that adaptive smoothing is more effective and gives better results than ordinary splines especially when the underlying function has varying smoothness. Here we suggest using an exponential function for the smoothness penalty. This is because the mortality pattern is expected to be increasingly smooth in age and having a heavier smoothness penalty at the oldest ages means that more strength is borrowed from neighbouring ages, therefore improving the robustness. Alternatively, linear or piece-wise constant functions could be used for the smoothness penalty. However, the linear function would not be as effective as the exponential function here as we require very low penalty at younger ages where there is plenty of data and significantly higher penalty at the oldest ages where data is sparse. On the other hand, a piece-wise constant function would require estimation of the levels and locations of the jumps. In our case, we do not believe that this additional



(a) England and Wales males



(b) England and Wales females

Figure 2: Ordinary P-spline fits extrapolated to age 120. The solid lines are the estimated mortality rates using data from age 1 to 104, while the dotted lines are the estimated mortality rates using data only from age 1 to 100.

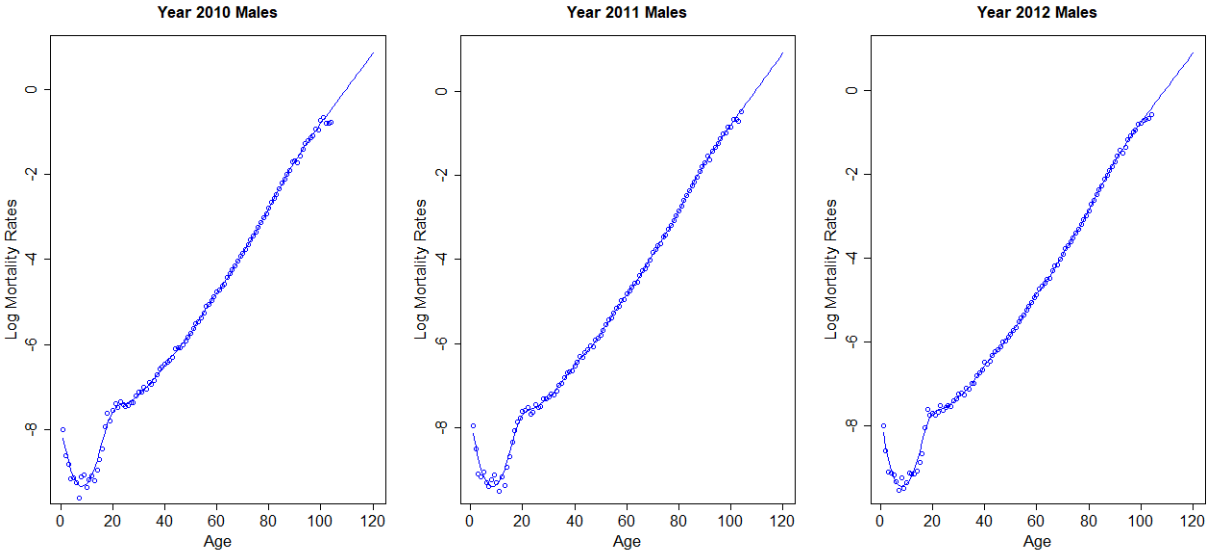
computational burden is justified. Using exponential function for the smoothness penalty we have

$$\begin{aligned} d_x &\sim \text{Poisson}(E_x^C m_x), \text{ where } \log(m_x) = \boldsymbol{b}_x' \boldsymbol{\beta} \\ \text{with penalty } &\sum_{i=3}^{k} \zeta(i)(\bigtriangledown^2(\beta_i))^2 \text{ and } \zeta(i) \;=\; \lambda_1 \, \exp(\, \lambda_2 \, i), \; \lambda_1 > 0 \end{aligned} \tag{3}$$

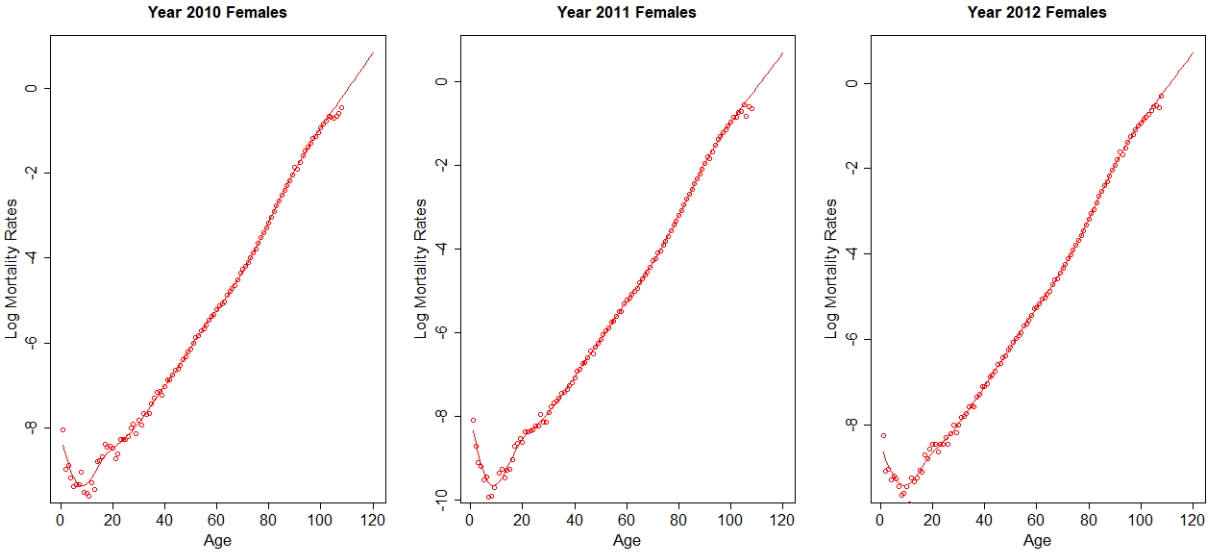
Here $\bigtriangledown^2$ is the second difference operator, $\boldsymbol{b}_x'$ is the row corresponding to age x of the design matrix $\boldsymbol{B}$ of the B-spline basis with the corresponding coefficient vector $\boldsymbol{\beta}$, λ_1 and λ_2 are smoothing parameters and k is the dimension of the basis. The penalty can be written more compactly as $\boldsymbol{\beta}'\boldsymbol{P}'\boldsymbol{\Lambda}\boldsymbol{P}\boldsymbol{\beta}$ where $\boldsymbol{P}$ is the second order difference matrix with appropriate dimensions and $\boldsymbol{\Lambda}$ is a diagonal matrix with $\boldsymbol{\Lambda}_{ii} = \lambda_1 e^{\lambda_2 \, i}$.

When the smoothing parameters λ_1 and λ_2 are known or specified, the fitting of the spline coefficients is straightforward by maximising the penalised likelihood. Otherwise they can be estimated together with the spline coefficients from the data. Here we estimate λ_1 and λ_2 by minimising the Bayesian Information Criterion (BIC). The Newton-Raphson method can be used to estimate the optimal smoothing parameters within each working model of the Penalised Iteratively Re-weighted Least Squares (P-IRLS) iteration, namely the performance iteration (Gu, 1992; Wood, 2006). The relevant gradient and Hessian for the minimisation of BIC can be found in the Appendix A. The BIC surface is relatively flat at regions with very low or very high penalty, hence these regions should be avoided as initial estimates. Alternatively, a two-dimensional grid search could be employed, the gradients and Hessian are prepared for the sake of the joint model in section 4, which contains six smoothing parameters and hence reduces the efficiency of grid search.

In Figure 3 we present the fitted mortality rates for our model. Compared to the ordinary P-spline fit in Figure 2, we can see that the fit is more robust at high ages under the exponentially increasing penalty. The yearly mortality pattern is more reasonable with less irregular variations. The spline is also capable of capturing the decreasing rate of increase in the oldest mortality rates. Figure 4 shows the P-spline fit for males in 2011 and the corresponding estimate of the exponential penalty. The increase in smoothness takes effect at around age 90, which is desirable as this is the region where the exposures and hence the reliability start to drop.



(a) England and Wales males



(b) England and Wales females

Figure 3: Adaptive exponentially increasing penalty P-spline fit to the period mortality schedule extrapolated to age 120.

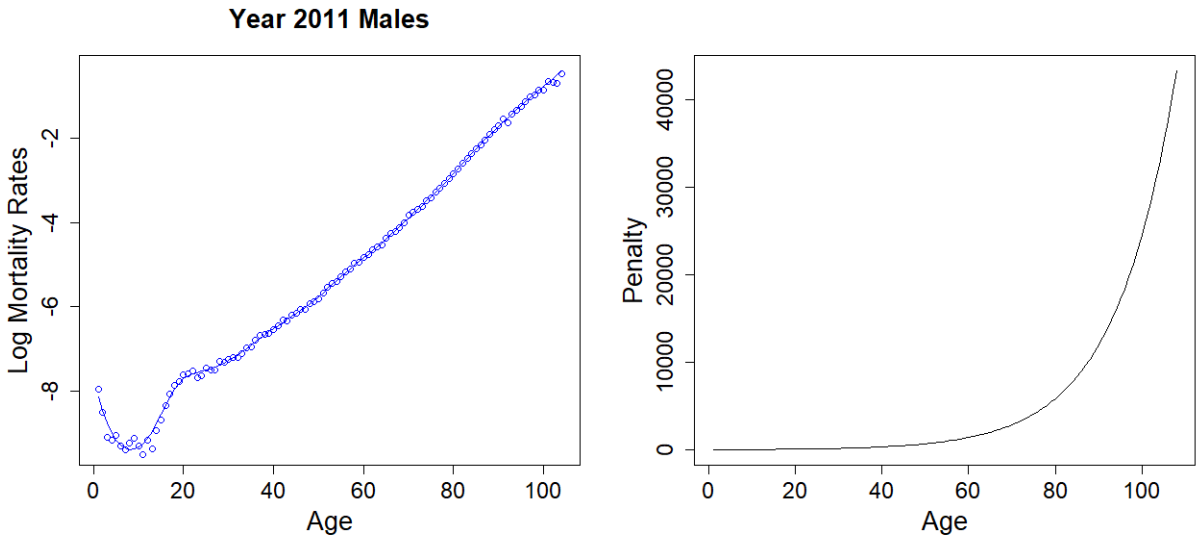


Figure 4: P-spline with exponential penalty fit for 2011 England and Wales males and the corresponding adaptive smoothness penalty.

4 Joint Model

We have been modelling male and female data separately. However jointly modelling male and female mortality rates has several advantages, such as avoiding cross-over at the highest ages and allowing borrowing of strength. Male and female mortality rates often display converging trends at high ages, information could be shared between the two sexes. This will produce coherent estimates especially at ages with low exposures as we are pooling the two data sets. On top of the exponential smoothness penalty for males and females, an additional penalty on the difference between the male and female P-spline coefficients is introduced. As stated before, males and females have the same P-spline basis.

At earlier ages, the splines shall enjoy more freedom as data at this region is more reliable. It is also believed that there is genuine difference in the levels and patterns of mortality between male and female mortality (especially for the accident hump). At adult ages the difference between male and female mortality rates seems to diminish gradually as age increases. Therefore an exponential penalty will be used again such that at younger ages the male and female coefficients are expected to be penalised less. In our applications it is found that estimation is easier when the differences between the first few male and female coefficients are left un-penalised. We suspect this is due to the apparent converging trend at the youngest ages as well. Therefore we only apply the difference penalty after a certain basis function. In our experiments, the selection of this coefficient is

immaterial as long as it excludes childhood mortality. Here the difference penalty is chosen to start at the 9-th coefficient (around age 24, which is just after the accident hump). The model is

$$\log \begin{pmatrix} \boldsymbol{m^M}(\boldsymbol{x}) \\ \boldsymbol{m^F}(\boldsymbol{x}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}(\boldsymbol{x}) & 0 \\ 0 & \boldsymbol{B}(\boldsymbol{x}) \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta^M} \\ \boldsymbol{\beta^F} \end{pmatrix}$$

$$\text{with the following three penalties} \quad \sum_{i=3}^{k} \zeta^M(i)(\bigtriangledown^2(\beta_i^M))^2$$

$$\sum_{i=3}^{k} \zeta^F(i)(\bigtriangledown^2(\beta_i^F))^2$$

$$\sum_{i=9}^{k} \zeta^D(i)(\beta_i^M - \beta_i^F)^2$$

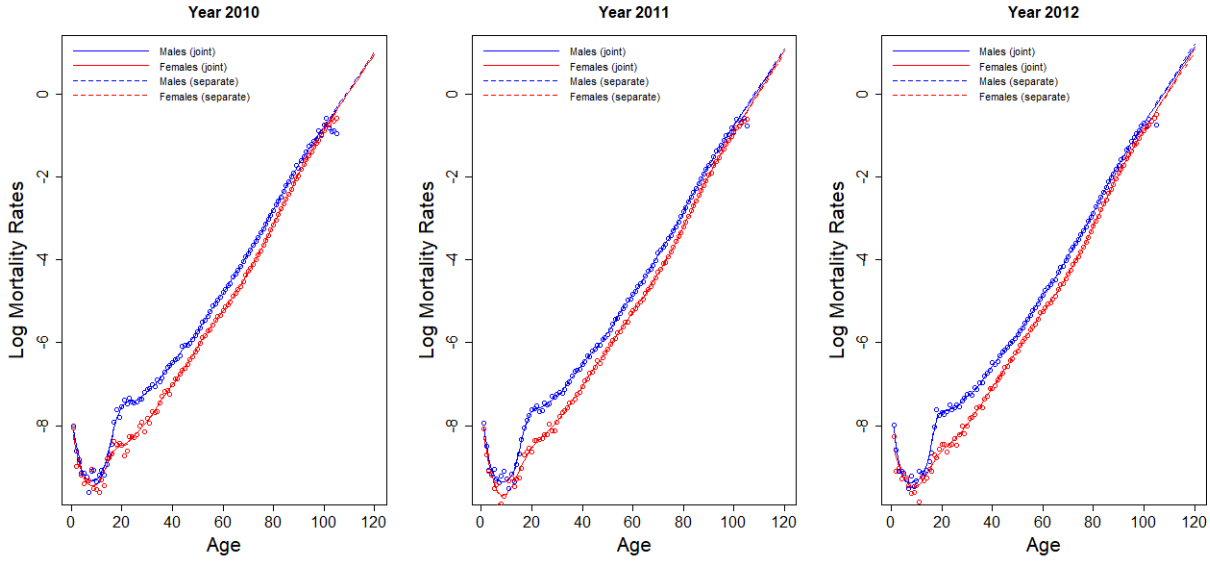
$$\text{where } \zeta^j(\cdot) = \lambda_1^j \exp(\lambda_2^j \cdot), \text{ with } \lambda_1^j > 0 \text{ for } j \in \{M, F, D\}. \tag{4}$$

The first two penalties relate to the smoothness of male and female mortality rates while the third penalty corresponds to the difference between male and female mortality rates. The penalties can be written more compactly as $\boldsymbol{\beta}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{\beta}$ where the overall coefficient vector is $\boldsymbol{\beta} = (\,\beta^M\ \beta^F\,)$ and the overall penalty matrix is $\boldsymbol{P} = \begin{pmatrix} \sqrt{\Lambda^M}\boldsymbol{P}^M \\ \sqrt{\Lambda^F}\boldsymbol{P}^F \\ \sqrt{\Lambda^D}\boldsymbol{P}^D \end{pmatrix}$, $\boldsymbol{P^M} = (\,\boldsymbol{\nabla_2}\ 0\,)$, $\boldsymbol{P^F} = (\,0\ \boldsymbol{\nabla_2}\,)$. Here $\boldsymbol{\nabla_2}$ is the second order difference matrix, $\boldsymbol{P^D} = (\,1\ -1\,) \otimes \boldsymbol{I}$ and $\mathbf{0}$ is simply a null matrix of appropriate dimension.

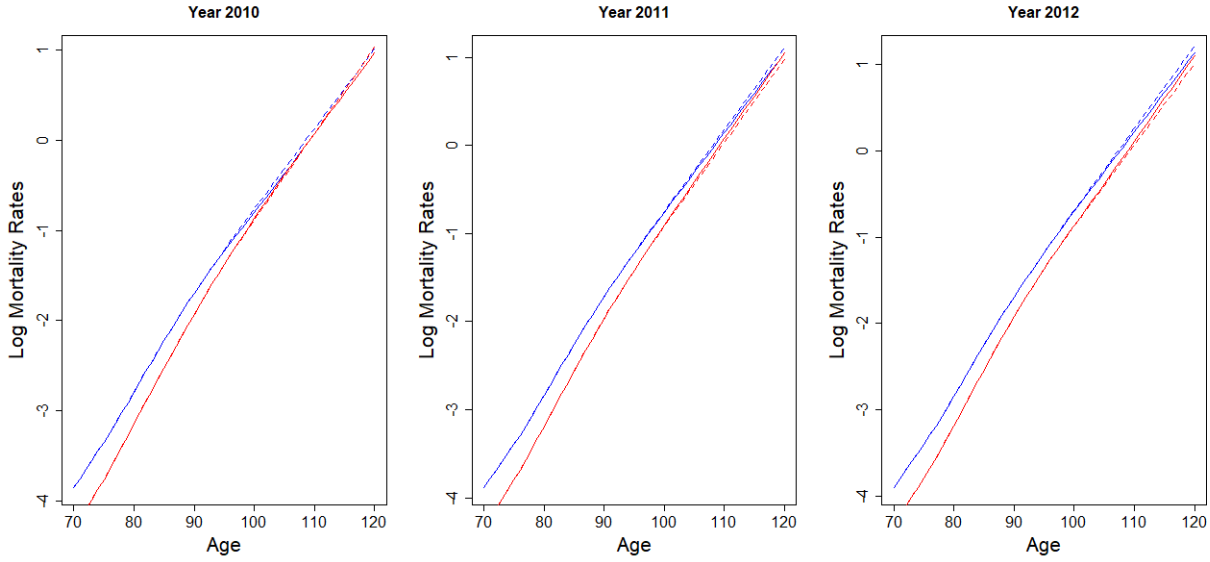
Figure 5 shows the results of the modelling. When males and females are modelled separately (dotted lines), the extrapolated trends sometimes diverge, which is not desirable. A clearer example would be year 2007 where the dotted lines drift apart (Figure 6). The difference penalty prevents divergence and further increases the robustness when extrapolating to higher ages even without data by learning from the existing data how fast and strongly the two trends should converge.

4.1 Preventing Cross-Over

One final improvement for the model is that, while the difference penalty dictates the convergence of male an female mortality curves, there is no guarantee that the graduated male mortality rates will always be higher than the female mortality rates. For instance, in Figure 5a, the male and female mortality estimates for 2012 cross-over at around age 10. Sometimes this could also happen at the oldest ages



(a) Full age range



(b) Age 70 to 120

Figure 5: P-spline fit with cross-sex penalty extrapolated to age 120. The dotted lines correspond to the fits without the cross-sex penalty (i.e. fitted separately) and the solid lines correspond to the fits with cross-sex penalty.

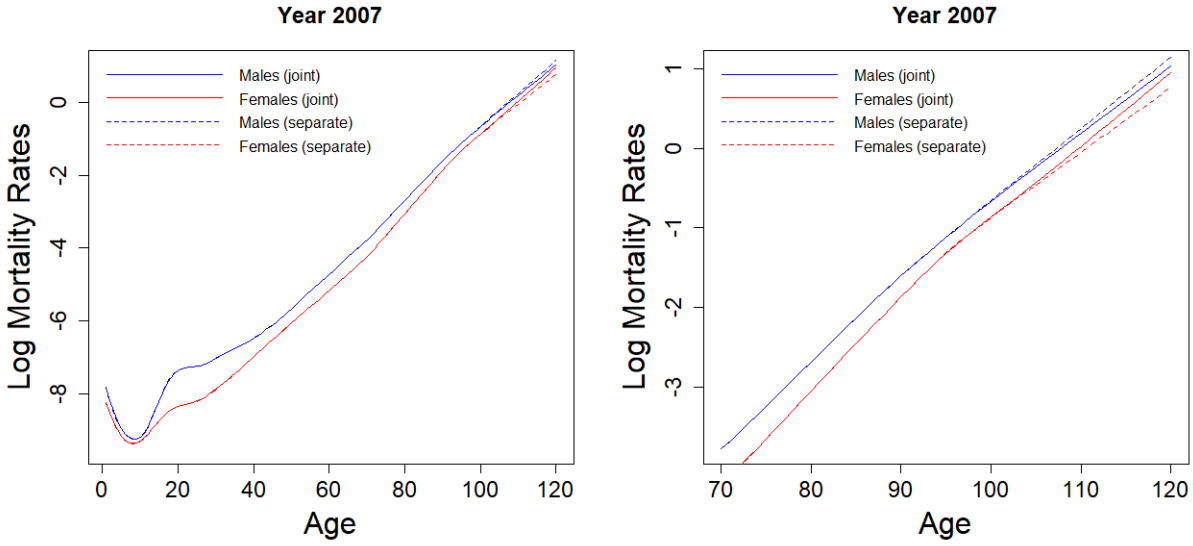


Figure 6: P-spline fit with cross-sex penalty extrapolated to age 120 for year 2007. The dotted lines correspond to the fits without the cross-sex penalty (i.e. fitted separately) and the solid lines correspond to the fits with cross-sex penalty.

or extrapolated range. To deny the possibility of male and female mortality rates crossing over, we impose a hard constraint on the coefficients such that the each of the B-spline coefficients for males are greater than or equal to that for females, i.e. $\boldsymbol{\beta^M} \geq \boldsymbol{\beta^F}$. Since it is assumed that males and females have the same knot sequence, this is a sufficient condition such that $m^F(x) \leq m^M(x) \quad \forall x$. In other words, the male mortality curve will always be above the female mortality curve. To do this, we first re-parameterise model 4 to

$$\boldsymbol{\beta^*} = \begin{pmatrix} \boldsymbol{\beta^F} \\ \boldsymbol{\beta^D} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta^F} \\ \boldsymbol{\beta^M - \beta^F} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{-I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta^M} & \boldsymbol{\beta^F} \end{pmatrix} = \boldsymbol{Z\beta} \tag{5}$$

Therefore, the design matrix and the penalty matrix in equation 4 have to be transformed accordingly, we have

$$\boldsymbol{B^* = BZ^{-1}} \quad \text{and} \quad \boldsymbol{P^* = PZ^{-1}} \tag{6}$$

and the non-negative constraint is applied only to $\boldsymbol{\beta^D}$. Non-negative least squares is performed within each P-IRLS iteration using the *lsei* package in R.

Since now this is not a linear smoother, i.e. we cannot write the general form $\hat{\boldsymbol{y}} = \boldsymbol{Ay}$, the gradient and Hessian for the BIC optimisation cannot be found analytically. In addition, to the best of our knowledge, the effective degrees of freedom of a non-negative least squares procedure is not known. We believe that

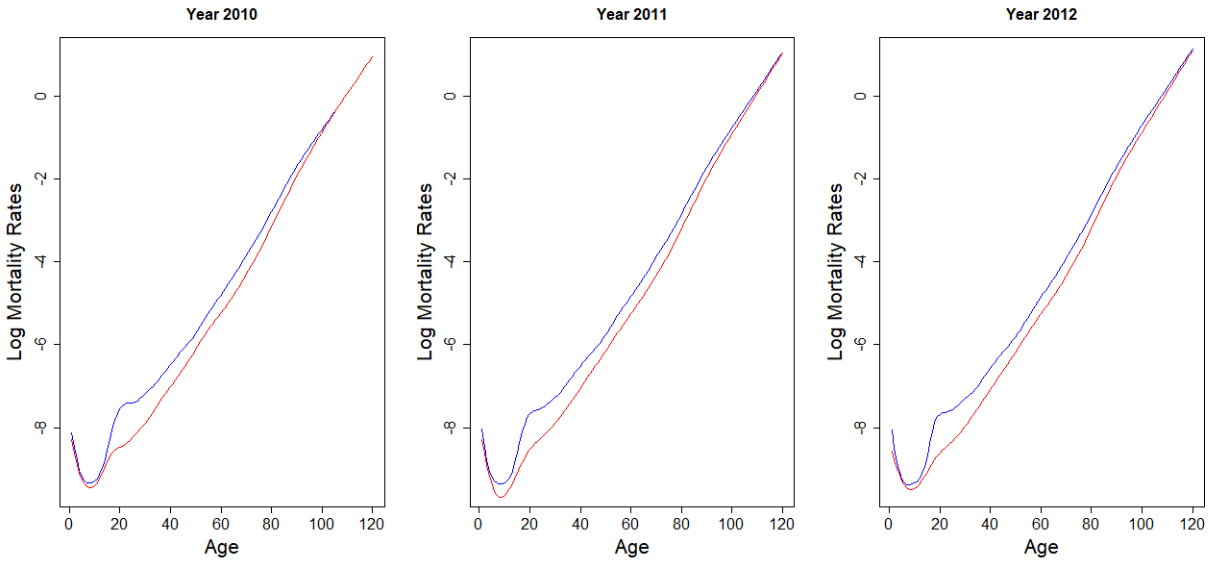


Figure 7: P-spline fit with difference penalty extrapolated to age 120 with non-negative constraints on the coefficients.

adding the non-negative constraints shall not change the overall optimal smoothness considerably, therefore we use the estimated smoothing parameters from the unconstrained model and treat them as known quantities in the constrained model. Figure 7 plots the graduated mortality rates under the non-negative constraints. As expected, the male mortality curve now always lies above female mortality curve.

5 Conclusion

Crude mortality rates often exhibit irregular and wiggly patterns, due to natural randomness. Therefore the crude rates have to be smoothed, or graduated, before they are used, for example in pricing of a life-related insurance product.

In this paper we propose a non-parametric approach of mortality graduation that is suitable for the whole age range using P-splines with adaptive penalties. It is assumed that the smoothness penalty is an exponential function, hence at younger ages where more reliable data is available, the model enjoys higher freedom (i.e. lower penalty); whereas at the oldest ages where data is sparse, the heavier penalty improves robustness. Under this penalty, the model benefits from the flexibility offered by P-splines and is robust at ages with low exposures, making extrapolation to higher ages more stable.

Modelling male and female mortality rates separately can cause some problems (e.g. crossing-over). These problems are addressed using rather ad-hoc

methods in previous ELTs. In addition, at very high ages the data is sparse and unreliable (for example, In ELT15 the data at the highest ages is discarded starting at the age that would produce an implausible trend), therefore a robust model is needed at these ages. Advantages can be gained from jointly modelling of male and female mortality rates. Information can be borrowed from each other, which is useful especially for males at the highest ages since there are usually more female data at these ages compared to males. In addition, cross-over of male and female mortality rates can be avoided. These are achieved by introducing an additional penalty for the difference between male and female spline coefficients and constraining the spline coefficients. The approach described in this paper provides a coherent way of mortality graduation across the whole age range even for the regions where data is sparse or non-existent.

One potential drawback of our model is that mortality rates at ages beyond available data is extrapolated almost linearly, while in the literature there is some evidence showing a decreasing increase in the log mortality rates at the highest ages. A remedy would be to use a logistic model instead of a log-linear model, so that an asymptotic limit is introduced. Another possible extension would be to use a fully Bayesian approach, as this would give us the opportunity to incorporate prior beliefs of the increasing smoothness and increasing similarity between male and female mortality rates at the highest ages into the model in a more natural framework. Over-dispersion is usually observed in mortality data and hence the Poisson assumption maybe too restrictive. A further way in which our model could be enhanced is to introduce a dispersion parameter to capture the over-dispersion. For example, a standard parameterisation of the dispersion model conveniently leads to a Negative Binomial, rather than Poisson, distribution for the counts of deaths, which is also often used to model mortality.

References

Baladandayuthapani, Veerabhadran, Mallick, Bani K, & Carroll, Raymond J. 2005. Spatially adaptive Bayesian penalized regression splines (P-splines). *Journal of Computational and Graphical Statistics*, **14**(2), 378–394.

Beard, Robert E. 1959. Note on some mathematical mortality models. *Pages 302–311 of: Ciba Foundation Symposium-The Lifespan of Animals (Colloquia on Ageing)*, vol. 5. Wiley Online Library.

Biatat, Viani D, & Currie, Iain D. 2010. Joint models for classification and comparison of mortality in different countries. *Pages 89–94 of: Proceedings of 25rd International Workshop on Statistical Modelling, Glasgow*.

Bollaerts, Kaatje, Eilers, Paul HC, & Van Mechelen, Iven. 2006. Simple and multiple P-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology*, **59**(2), 451–469.

Cairns, Andrew JG, Blake, David, Dowd, Kevin, Coughlan, Guy D, & Khalaf-Allah, Marwa. 2011. Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin: The Journal of the IAA*, **41**(1), 29–59.

Camarda, Carlo G. 2019. Smooth constrained mortality forecasting. *Demographic Research*, **41**, 1091–1130.

Camarda, Carlo G, Eilers, Paul HC, & Gampe, Jutta. 2016. Sums of smooth exponentials to decompose complex series of counts. *Statistical Modelling*, **16**(4), 279–296.

Carriere, Jacques F. 1992. Parametric models for life tables. *Transactions of the Society of Actuaries*, **44**, 77–99.

Crainiceanu, Ciprian M, Ruppert, David, Carroll, Raymond J, Joshi, Adarsh, & Goodner, Billy. 2007. Spatially adaptive Bayesian penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, **16**(2), 265–288.

Currie, Iain D, Durban, Maria, & Eilers, Paul HC. 2004. Smoothing and forecasting mortality rates. *Statistical modelling*, **4**(4), 279–298.

Dodd, Erengul, Forster, Jonathan J, Bijak, Jakub, & Smith, Peter WF. 2018. Smoothing mortality data: the English Life Tables, 2010–2012. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **181**(3), 717–735.

Dowd, Kevin, Cairns, Andrew JG, Blake, David, Coughlan, Guy D, & Khalaf-Allah, Marwa. 2011. A gravity model of mortality rates for two related populations. *North American Actuarial Journal*, **15**(2), 334–356.

Eilers, Paul HC, & Marx, Brian D. 1996. Flexible smoothing with B-splines and penalties. *Statistical science*, 89–102.

Gallop, Adrian. 2002. Mortality at advanced ages in the United Kingdom. *Living to 100 and Beyond: Survival at Advanced Ages*.

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

Heligman, Larry, & Pollard, John H. 1980. The age pattern of mortality. *Journal of the Institute of Actuaries*, **107**(01), 49–80.

Hilton, Jason, Dodd, Erengul, Forster, Jonathan J, & Smith, Peter WF. 2019. Projecting UK mortality by using Bayesian generalized additive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **68**(1), 29–49.

Hyndman, Rob J, Booth, Heather, & Yasmeen, Farah. 2013. Coherent mortality forecasting: the product-ratio method with functional time series models. *Demography*, **50**(1), 261–283.

Jarner, Søren Fiig, & Kryger, Esben Masotti. 2011. Modelling adult mortality in small populations: The SAINT model. *ASTIN Bulletin: The Journal of the IAA*, **41**(2), 377–418.

Jullion, Astrid, & Lambert, Philippe. 2007. Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational statistics & data analysis*, **51**(5), 2542–2558.

Kaishev, Vladimir K, Dimitrova, Dimitrina S, Haberman, Steven, & Verrall, RJ. 2006. Geometrically designed, variable know regression splines: asymptotics and inference.

Krivobokova, Tatyana, Crainiceanu, Ciprian M, & Kauermann, Göran. 2008. Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics*, **17**(1), 1–20.

Li, Nan, & Lee, Ronald. 2005. Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, **42**(3), 575–594.

Lindbergson, Maria. 2001. Mortality among the elderly in Sweden 1988–1997. *Scandinavian Actuarial Journal*, **2001**(1), 79–94.

Liu, Ziyue, & Guo, Wensheng. 2010. Data driven adaptive spline smoothing. *Statistica Sinica*, 1143–1163.

Office for National Statistics. 2019. *Population estimates and deaths by single year of age for England and Wales and the UK, 1961 to 2018*. `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/adhocs/10727populationestimatesanddeathsbysingleyearofageforenglandandwalesandtheuk1961to2018`.

Perks, Wilfred. 1932. On some experiments in the graduation of mortality statistics. *Journal of the Institute of Actuaries*, **63**(1), 12–57.

Pintore, Alexandre, Speckman, Paul, & Holmes, Chris C. 2006. Spatially adaptive smoothing splines. *Biometrika*, **93**(1), 113–125.

Pitacco, Ermanno. 2016. High age mortality and frailty. Some remarks and hints for actuarial modeling.

Plat, Richard. 2009. Stochastic portfolio specific mortality and the quantification of mortality basis risk. *Insurance: Mathematics and Economics*, **45**(1), 123–132.

Pya, Natalya, & Wood, Simon N. 2015. Shape constrained additive models. *Statistics and Computing*, **25**(3), 543–559.

Ruppert, David, & Carroll, Raymond J. 2000. Theory & Methods: Spatially-adaptive Penalties for Spline Fitting. *Australian & New Zealand Journal of Statistics*, **42**(2), 205–223.

Saikia, Pallabi, & Borah, Munindra. 2014. A comparative study of parametric models of old-age mortality. *International Journal of Science and Research*, **3**(5).

Scheipl, Fabian, & Kneib, Thomas. 2009. Locally adaptive Bayesian P-splines with a Normal-Exponential-Gamma prior. *Computational Statistics & Data Analysis*, **53**(10), 3533–3552.

Storlie, Curtis B, Bondell, Howard D, & Reich, Brian J. 2010. A locally adaptive penalty for estimation of functions with varying roughness. *Journal of Computational and Graphical Statistics*, **19**(3), 569–589.

Tang, Kai Hon. 2021. *Modelling and Projecting Mortality Rates Using Adaptive P-splines*. Ph.D. thesis, University of Southampton.

Thatcher, A Roger. 1999. The long-term pattern of adult mortality and the highest attained age. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **162**(1), 5–43.

Villegas, Andrés M, & Haberman, Steven. 2014. On the modeling and forecasting of socioeconomic mortality differentials: An application to deprivation and mortality in England. *North American Actuarial Journal*, **18**(1), 168–193.

Wood, Simon. 2006. *Generalized additive models: an introduction with R*. CRC press.

Yang, Lianqiang, & Hong, Yongmiao. 2017. Adaptive penalized splines for data smoothing. *Computational Statistics & Data Analysis*, **108**, 70–83.

A Gradient and Hessian for the Newton-Raphson for the estimation of smoothing parameters

Following the notations in Wood (2006), for linear spline models, let $\boldsymbol{X} = \boldsymbol{QR}$ be the QR-decomposition of the design matrix and $\left[\begin{smallmatrix}\boldsymbol{R}\\ \sqrt{\boldsymbol{\Lambda}}\boldsymbol{P}\end{smallmatrix}\right] = \boldsymbol{UDV'}$ be the SVD decomposition. Any rank deficiency is detected at this stage and removed by deleting the corresponding columns from the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$. Let $\boldsymbol{U_1}$ be the sub-matrix of $\boldsymbol{U}$ such that $\boldsymbol{R} = \boldsymbol{U_1 DV'}$. $\boldsymbol{X'X + P'\Lambda P} = \boldsymbol{VD^2V'}$ and hence the hat matrix
$\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X'X + P'\Lambda P})^{-1}\boldsymbol{X'} = \boldsymbol{QU_1 DV'}(\boldsymbol{VD^{-2}}V')\boldsymbol{VDU_1'Q'} = \boldsymbol{QU_1U_1'Q'}$.

To guarantee that λ_1 is positive, we parameterise $\lambda_1 = e^{\rho_1}$. Let $\boldsymbol{G} = \boldsymbol{X'X + P'\Lambda P}$, we have

$$\begin{aligned}
\frac{\partial \boldsymbol{G^{-1}}}{\partial \rho_1} &= -\boldsymbol{G^{-1}}\frac{\partial \boldsymbol{G}}{\partial \rho_1}\boldsymbol{G^{-1}} \\
&= -\boldsymbol{G^{-1}P'}\frac{\partial \boldsymbol{\Lambda}}{\partial \rho_1}\boldsymbol{PG^{-1}} \\
&= -\boldsymbol{G^{-1}P'\Lambda PG^{-1}} \\
&= -\boldsymbol{VD^{-2}V'P'\Lambda PVD^{-2}V'}
\end{aligned} \tag{7}$$

and

$$\begin{aligned}
\frac{\partial \boldsymbol{G^{-1}}}{\partial \lambda_2} &= -\boldsymbol{G^{-1}}\frac{\partial \boldsymbol{G}}{\partial_2}\boldsymbol{G^{-1}} \\
&= -\boldsymbol{G^{-1}P'}\frac{\partial \boldsymbol{\Lambda}}{\partial_2}\boldsymbol{PG^{-1}} \\
&= -\boldsymbol{G^{-1}P'N\Lambda PG^{-1}} \\
&= -\boldsymbol{VD^{-2}V'P'N\Lambda PVD^{-2}V'}
\end{aligned} \tag{8}$$

where $\boldsymbol{N}$ is a diagonal matrix with entries $\boldsymbol{N}_{ii} = i$. In practice we scale the entries $\boldsymbol{N}_{ii}$ to 0 to 1 to prevent overflow when λ_2 gets huge, as well as the corresponding entries $\boldsymbol{\Lambda}_{ii}$.

We then have

$$\begin{aligned}
\frac{\partial \boldsymbol{A}}{\partial \rho_1} &= \frac{\partial \boldsymbol{X}\boldsymbol{G}^{-\mathbf{1}}\boldsymbol{X}'}{\partial \rho_1} \\
&= -\boldsymbol{X}\frac{\partial \boldsymbol{G}^{-\mathbf{1}}}{\partial \rho_1}\boldsymbol{X}' \\
&= -\boldsymbol{X}\boldsymbol{V}\boldsymbol{D}^{-\mathbf{2}}\boldsymbol{V}'\boldsymbol{P}'\boldsymbol{\Lambda}\boldsymbol{P}\boldsymbol{V}\boldsymbol{D}^{-\mathbf{2}}\boldsymbol{V}'\boldsymbol{X}' \\
&= -\boldsymbol{Q}\boldsymbol{U}_1\boldsymbol{D}\boldsymbol{V}'\boldsymbol{V}\boldsymbol{D}^{-\mathbf{2}}\boldsymbol{V}'\boldsymbol{P}'\boldsymbol{\Lambda}\boldsymbol{P}\boldsymbol{V}\boldsymbol{D}^{-\mathbf{2}}\boldsymbol{V}'\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}_1'\boldsymbol{Q}' \\
&= -\boldsymbol{Q}\boldsymbol{U}_1\underbrace{\boldsymbol{D}^{-\mathbf{1}}\boldsymbol{V}'\boldsymbol{P}'\boldsymbol{\Lambda}\boldsymbol{P}\boldsymbol{V}\boldsymbol{D}^{-\mathbf{1}}}_{\boldsymbol{M}}\boldsymbol{U}_1'\boldsymbol{Q}' \\
&= -\boldsymbol{Q}\boldsymbol{U}_1\boldsymbol{M}\boldsymbol{U}_1'\boldsymbol{Q}'
\end{aligned} \tag{9}$$

and

$$\begin{aligned}
\frac{\partial \boldsymbol{A}}{\partial \lambda_2} &= \frac{\partial \boldsymbol{X}\boldsymbol{G}^{-\mathbf{1}}\boldsymbol{X}'}{\partial \lambda_2} \\
&= -\boldsymbol{Q}\boldsymbol{U}_1\underbrace{\boldsymbol{D}^{-\mathbf{1}}\boldsymbol{V}'\boldsymbol{P}'\boldsymbol{N}\boldsymbol{\Lambda}\boldsymbol{P}\boldsymbol{V}\boldsymbol{D}^{-\mathbf{1}}}_{\boldsymbol{M}^*}\boldsymbol{U}_1'\boldsymbol{Q}' \\
&= -\boldsymbol{Q}\boldsymbol{U}_1\boldsymbol{M}^*\boldsymbol{U}_1'\boldsymbol{Q}'
\end{aligned} \tag{10}$$

Moving on to the Hessian, first we find

$$\begin{aligned}
\frac{\partial^2 \boldsymbol{G}}{\partial {\rho_1}^2} &= \frac{\partial}{\partial \rho_1}\left(\frac{\partial \boldsymbol{G}}{\partial \rho_1}\right) \\
&= \frac{\partial \boldsymbol{P}'\boldsymbol{\Lambda}\boldsymbol{P}}{\partial \rho_1} \\
&= \boldsymbol{P}'\boldsymbol{\Lambda}\boldsymbol{P}
\end{aligned} \tag{11}$$

so

$$\frac{\partial^2 \boldsymbol{G^{-1}}}{\partial {\rho_1}^2} = 2\,\boldsymbol{G^{-1}}\frac{\partial \boldsymbol{G}}{\partial \rho_1}\boldsymbol{G^{-1}}\frac{\partial \boldsymbol{G}}{\partial \rho_1}\boldsymbol{G^{-1}} - \boldsymbol{G^{-1}}\frac{\partial^2 \boldsymbol{G}}{\partial {\rho_1}^2}\boldsymbol{G^{-1}}$$

$$= 2\,\boldsymbol{VD^{-2}V'P'\Lambda PVD^{-2}V'P'\Lambda PVD^{-2}V' - VD^{-2}V'P'\Lambda PVD^{-2}V'} \tag{12}$$

$$\frac{\partial^2 \boldsymbol{G}}{\partial {\lambda_2}^2} = \frac{\partial}{\partial \lambda_2}\left(\frac{\partial \boldsymbol{G}}{\partial \lambda_2}\right)$$

$$= \frac{\partial \boldsymbol{P'N\Lambda P}}{\partial \lambda_2} \tag{13}$$

$$= \boldsymbol{P'N\Lambda NP}$$

so

$$\frac{\partial^2 \boldsymbol{G^{-1}}}{\partial {\lambda_2}^2} = 2\,\boldsymbol{G^{-1}}\frac{\partial \boldsymbol{G}}{\partial \lambda_2}\boldsymbol{G^{-1}}\frac{\partial \boldsymbol{G}}{\partial \lambda_2}\boldsymbol{G^{-1}} - \boldsymbol{G^{-1}}\frac{\partial^2 \boldsymbol{G}}{\partial {\lambda_2}^2}\boldsymbol{G^{-1}}$$

$$= 2\,\boldsymbol{VD^{-2}V'P'N\Lambda PVD^{-2}V'P'N\Lambda PVD^{-2}V'} \tag{14}$$
$$\boldsymbol{- VD^{-2}V'P'N\Lambda NPVD^{-2}V'}$$

$$\frac{\partial^2 \boldsymbol{G}}{\partial \rho_1 \partial \lambda_2} = \frac{\partial}{\partial \lambda_2}\left(\frac{\partial \boldsymbol{G}}{\partial \rho_1}\right)$$

$$= \frac{\partial \boldsymbol{P'\Lambda P}}{\partial \lambda_2} \tag{15}$$

$$= \boldsymbol{P'N\Lambda P}$$

so

$$\frac{\partial^2 \boldsymbol{G^{-1}}}{\partial \rho_1 \partial \lambda_2} = \boldsymbol{G^{-1}}\frac{\partial \boldsymbol{G}}{\partial \rho_1}\boldsymbol{G^{-1}}\frac{\partial \boldsymbol{G}}{\partial \lambda_2}\boldsymbol{G^{-1}} + \boldsymbol{G^{-1}}\frac{\partial \boldsymbol{G}}{\partial \lambda_2}\boldsymbol{G^{-1}}\frac{\partial \boldsymbol{G}}{\partial \rho_1}\boldsymbol{G^{-1}} - \boldsymbol{G^{-1}}\frac{\partial^2 \boldsymbol{G}}{\partial \rho_1 \partial \lambda_2}\boldsymbol{G^{-1}}$$

$$= \boldsymbol{VD^{-2}V'P'\Lambda PVD^{-2}V'P'N\Lambda PVD^{-2}V'}$$
$$\boldsymbol{+VD^{-2}V'P'N\Lambda PVD^{-2}V'P'\Lambda PVD^{-2}V' - VD^{-2}V'P'N\Lambda PVD^{-2}V'} \tag{16}$$

Therefore

$$\frac{\partial^2 \boldsymbol{A}}{\partial {\rho_1}^2} = \boldsymbol{X} \frac{\partial^2 \boldsymbol{G^{-1}}}{\partial {\rho_1}^2} \boldsymbol{X'}$$

$$= \boldsymbol{Q U_1 D V'} (2\, \boldsymbol{V D^{-2} V' P' \Lambda P V D^{-2} V' P' \Lambda P V D^{-2} V'} - \boldsymbol{V D^{-2} V' P' \Lambda P V D^{-2} V'}) \boldsymbol{V D U_1' Q'}$$

$$= 2\, \boldsymbol{Q U_1 M M U_1' Q'} - \boldsymbol{Q U_1 M U_1' Q'}$$

$$= 2\, \boldsymbol{Q U_1 M M U_1' Q'} + \frac{\partial \boldsymbol{A}}{\partial \rho_1} \tag{17}$$

$$\frac{\partial^2 \boldsymbol{A}}{\partial {\lambda_2}^2} = \boldsymbol{X} \frac{\partial^2 \boldsymbol{G^{-1}}}{\partial {\lambda_2}^2} \boldsymbol{X'}$$

$$= \boldsymbol{Q U_1 D V'} (2\, \boldsymbol{V D^{-2} V' P' N \Lambda P V D^{-2} V' P' N \Lambda P V D^{-2} V'} - \boldsymbol{V D^{-2} V' P' N \Lambda N P V D^{-2} V'}) \boldsymbol{V D U_1' Q'}$$

$$= 2\, \boldsymbol{Q U_1 M^* M^* U_1' Q'} - \boldsymbol{Q U_1} \underbrace{\boldsymbol{D^{-1} V' P' N \Lambda N P V D^{-1}}}_{\boldsymbol{M^{**}}} \boldsymbol{U_1' Q'}$$

$$= 2\, \boldsymbol{Q U_1 M^* M^* U_1' Q'} - \boldsymbol{Q U_1 M^{**} U_1' Q'} \tag{18}$$

and

$$\frac{\partial^2 \boldsymbol{A}}{\partial \rho_1 \partial \lambda_2} = \boldsymbol{X} \frac{\partial^2 \boldsymbol{G^{-1}}}{\partial \rho_1 \partial \lambda_2} \boldsymbol{X'}$$

$$= \boldsymbol{Q U_1 M M^* U_1' Q'} + \boldsymbol{Q U_1 M^* M U_1' Q'} - \boldsymbol{Q U_1 M^* U_1' Q'} \tag{19}$$

$$= \boldsymbol{Q U_1 M M^* U_1' Q'} + \boldsymbol{Q U_1 M^* M U_1' Q'} + \frac{\partial \boldsymbol{A}}{\partial \lambda_2}$$

Finally, the BIC for is $-2\log(L) + \log(n)\ \mathrm{tr}(\boldsymbol{A})$, where L is the maximum

likelihood.

$$\frac{\partial \tilde{\sigma}^2}{\partial \rho_1} = -\frac{2}{n}\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{y}\right)' \frac{\partial \boldsymbol{A}}{\partial \rho_1}\boldsymbol{y} \qquad \text{and} \qquad \frac{\partial \tilde{\sigma}^2}{\partial \lambda_2} = -\frac{2}{n}\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{y}\right)' \frac{\partial \boldsymbol{A}}{\partial \lambda_2}\boldsymbol{y} \tag{20}$$

$$\begin{aligned}
\frac{\partial^2 \tilde{\sigma}^2}{\partial {\rho_1}^2} &= -\frac{2}{n}\Big[\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{y}\right)' \frac{\partial^2 \boldsymbol{A}}{\partial {\rho_1}^2}\boldsymbol{y} - \boldsymbol{y}'\left(\frac{\partial \boldsymbol{A}}{\partial \rho_1}\right)\left(\frac{\partial \boldsymbol{A}}{\partial \rho_1}\right)\boldsymbol{y}\Big] \\
\frac{\partial^2 \tilde{\sigma}^2}{\partial {\lambda_2}^2} &= -\frac{2}{n}\Big[\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{y}\right)' \frac{\partial^2 \boldsymbol{A}}{\partial {\lambda_2}^2}\boldsymbol{y} - \boldsymbol{y}'\left(\frac{\partial \boldsymbol{A}}{\partial \lambda_2}\right)\left(\frac{\partial \boldsymbol{A}}{\partial \lambda_2}\right)\boldsymbol{y}\Big] \qquad \text{and} \\
\frac{\partial^2 \tilde{\sigma}^2}{\partial \rho_1 \partial \lambda_2} &= -\frac{2}{n}\Big[\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{y}\right)' \frac{\partial^2 \boldsymbol{A}}{\partial \rho_1 \partial \lambda_2}\boldsymbol{y} - \boldsymbol{y}'\left(\frac{\partial \boldsymbol{A}}{\partial \rho_1}\right)\left(\frac{\partial \boldsymbol{A}}{\partial \lambda_2}\right)\boldsymbol{y}\Big]
\end{aligned} \tag{21}$$

$$\frac{\partial BIC}{\partial \rho_1} = \frac{n}{\tilde{\sigma}^2}\frac{\partial \tilde{\sigma}^2}{\partial \rho_1} + \log(n)\,\mathrm{tr}\left(\frac{\partial \boldsymbol{A}}{\partial \rho_1}\right) \tag{22}$$

Similarly,

$$\frac{\partial BIC}{\partial \lambda_2} = \frac{n}{\tilde{\sigma}^2}\frac{\partial \tilde{\sigma}^2}{\partial \lambda_2} + \log(n)\,\mathrm{tr}\left(\frac{\partial \boldsymbol{A}}{\partial \lambda_2}\right) \tag{23}$$

$$\frac{\partial^2 BIC}{\partial {\rho_1}^2} = n\left[\frac{1}{\tilde{\sigma}^2}\frac{\partial^2 \tilde{\sigma}^2}{\partial {\rho_1}^2} - \frac{1}{\tilde{\sigma}^4}\left(\frac{\partial \tilde{\sigma}^2}{\partial \rho_1}\right)^2\right] + \log(n)\ \mathrm{tr}\left(\frac{\partial^2 \boldsymbol{A}}{\partial {\rho_1}^2}\right) \tag{24}$$

$$\frac{\partial^2 BIC}{\partial {\lambda_2}^2} = n\left[\frac{1}{\tilde{\sigma}^2}\frac{\partial^2 \tilde{\sigma}^2}{\partial {\lambda_2}^2} - \frac{1}{\tilde{\sigma}^4}\left(\frac{\partial \tilde{\sigma}^2}{\partial \lambda_2}\right)^2\right] + \log(n)\ \mathrm{tr}\left(\frac{\partial^2 \boldsymbol{A}}{\partial {\lambda_2}^2}\right) \tag{25}$$

and

$$\frac{\partial^2 BIC}{\partial \rho_1 \partial \lambda_2} = n\left[\frac{1}{\tilde{\sigma}^2}\frac{\partial^2 \tilde{\sigma}^2}{\partial \rho_1 \partial \lambda_2} - \frac{1}{\tilde{\sigma}^4}\left(\frac{\partial \tilde{\sigma}^2}{\partial \rho_1}\right)\left(\frac{\partial \tilde{\sigma}^2}{\partial \lambda_2}\right)\right] + \log(n)\ \mathrm{tr}\left(\frac{\partial^2 \boldsymbol{A}}{\partial \rho_1 \partial \lambda_2}\right) \tag{26}$$

As mentioned, for non-gaussian data, optimisation can be performed within each P-IRLS iteration and is done simply by replacing the design matrix and data with $\sqrt{\boldsymbol{W}}\boldsymbol{X}$ and $\sqrt{\boldsymbol{W}}\boldsymbol{z}$ respectively, where $\boldsymbol{W}$ is a diagonal matrix with the current working weights and $\boldsymbol{z}$ is the vector of working pseudo-data.

B Gradient and Hessian for the Newton-Raphson of the Joint Model

Recall that the penalty matrix for the joint model is $\begin{pmatrix} \sqrt{\boldsymbol{\Lambda^M}}\boldsymbol{P^M} \\ \sqrt{\boldsymbol{\Lambda^F}}\boldsymbol{P^F} \\ \sqrt{\boldsymbol{\Lambda^D}}\boldsymbol{P^D} \end{pmatrix}$. Similarly to constraint λ_1^M, λ_1^F and λ_1^D to be strictly positive, they are re-parameterised to $\lambda_1^j = e^{\rho_1^j}$ and the gradients with respect to (ρ_1^j) are simply found by replacing $\boldsymbol{P}$ and $\boldsymbol{\Lambda}$ in derivatives equations 7 to 26 by the corresponding $\boldsymbol{P^j}$ and $\boldsymbol{\Lambda^j}$. For example,

$$\begin{aligned} \frac{\partial \boldsymbol{A}}{\partial \rho_1^M} = -\boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^M}\boldsymbol{U_1'}\boldsymbol{Q'} \quad \text{and} \quad \frac{\partial \boldsymbol{A}}{\partial \rho_1^D} = -\boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^D}\boldsymbol{U_1'}\boldsymbol{Q'} \\ \text{where } \boldsymbol{M^j} = \boldsymbol{D^{-1}}\boldsymbol{V'}\boldsymbol{P^{j\prime}}\boldsymbol{\Lambda^j}\boldsymbol{P^j}\boldsymbol{V}\boldsymbol{D^{-1}} \end{aligned} \tag{27}$$

The Hessian requires a little more attention and slight adjustments have to be made. We have,

$$\begin{aligned} \frac{\partial^2 \boldsymbol{A}}{\partial \rho_1^i \partial \rho_1^j} &= \boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^i}\boldsymbol{M^j}\boldsymbol{U_1'}\boldsymbol{Q'} + \boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^j}\boldsymbol{M^i}\boldsymbol{U_1'}\boldsymbol{Q'} - \delta_{ij}\,\boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^j}\boldsymbol{U_1'}\boldsymbol{Q'} \\ &= \boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^i}\boldsymbol{M^j}\boldsymbol{U_1'}\boldsymbol{Q'} + \boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^j}\boldsymbol{M^i}\boldsymbol{U_1'}\boldsymbol{Q'} + \delta_{ij}\,\frac{\partial \boldsymbol{A}}{\partial \rho_1^j} \end{aligned} \tag{28}$$

$$\frac{\partial^2 \boldsymbol{A}}{\partial \lambda_2^i \partial \lambda_2^j} = \boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^{*i}}\boldsymbol{M^{*j}}\boldsymbol{U_1'}\boldsymbol{Q'} + \boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^{*j}}\boldsymbol{M^{*i}}\boldsymbol{U_1'}\boldsymbol{Q'} - \delta_{ij}\,\boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^{**j}}\boldsymbol{U_1'}\boldsymbol{Q'} \tag{29}$$

$$\begin{aligned} \frac{\partial^2 \boldsymbol{A}}{\partial \rho_1^i \partial \lambda_2^j} &= \boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^i}\boldsymbol{M^{*j}}\boldsymbol{U_1'}\boldsymbol{Q'} + \boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^{*j}}\boldsymbol{M^i}\boldsymbol{U_1'}\boldsymbol{Q'} - \delta_{ij}\,\boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^{*j}}\boldsymbol{U_1'}\boldsymbol{Q'} \\ &= \boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^i}\boldsymbol{M^{*j}}\boldsymbol{U_1'}\boldsymbol{Q'} + \boldsymbol{Q}\boldsymbol{U_1}\boldsymbol{M^{*j}}\boldsymbol{M^i}\boldsymbol{U_1'}\boldsymbol{Q'} + \delta_{ij}\,\frac{\partial \boldsymbol{A}}{\partial \lambda_2^j} \end{aligned} \tag{30}$$

where δ_{ij} is the Kronecker delta function.