D-Splines: Estimating rate schedules using high-dimensional splines with empirical demographic penalties

Carl Schmertmann

September 19, 2018

1 Brief Abstract

I develop and test a new class of penalized spline models for estimating demographic rate schedules, with a particular interest in applying the methods to sparse data from small areas or small sub-populations. I propose adaptations of the P-spline approach (called *D-splines*) that regularize and smooth high-dimensional splines by using specific demographic knowledge rather than generic arithmetical rules. Preliminary tests of four alternative D-spline estimators on simulated small-area mortality data are promising: D-spline estimators appear to have low errors and to produce schedules that reliably reflect known properties of human mortality schedules.

2 Notation

2.1 High-dimensional spline function for a demographic schedule

Model a (generic) demographic rate schedule over A single-year ages $0 \ldots (A-1)$ as

$$g = \boldsymbol{S}\theta$$

where $\boldsymbol{S}$ is a $A \times K$ matrix in which each of the K columns is a B-spline basis function [4] over $0 \ldots (A-1)$ and knots are closely-spaced. For all of the analysis in this abstract, I will assume that there are $A = 100$ ages $0 \ldots 99$, and that $\boldsymbol{S}$ is a 100×36 matrix of cubic

B-spline basis functions constructed using knots at 32 ages $3, 6, \ldots, 96$.[1]

2.2 Data and likelihood

Suppose that there is observed data y for which we have a (log) likelihood model $L(\theta) = L[y|g(\theta)]$. In all of the analysis in this extended abstract I assume that y consists of age-specific exposure and death counts (N_x and D_x, respectively) for a small population, and that the log likelihood is Poisson. In that specific case the schedule g represents age-specific log mortality rates, and the log likelihood is

$$L(\theta) = c - \sum_x N_x exp(\theta' s_x) + \sum_x D_x(\theta' s_x)$$

where s_x is the $K \times 1$ vector that makes up the row of S that corresponds to age x.[2]

2.3 Penalties and regularization

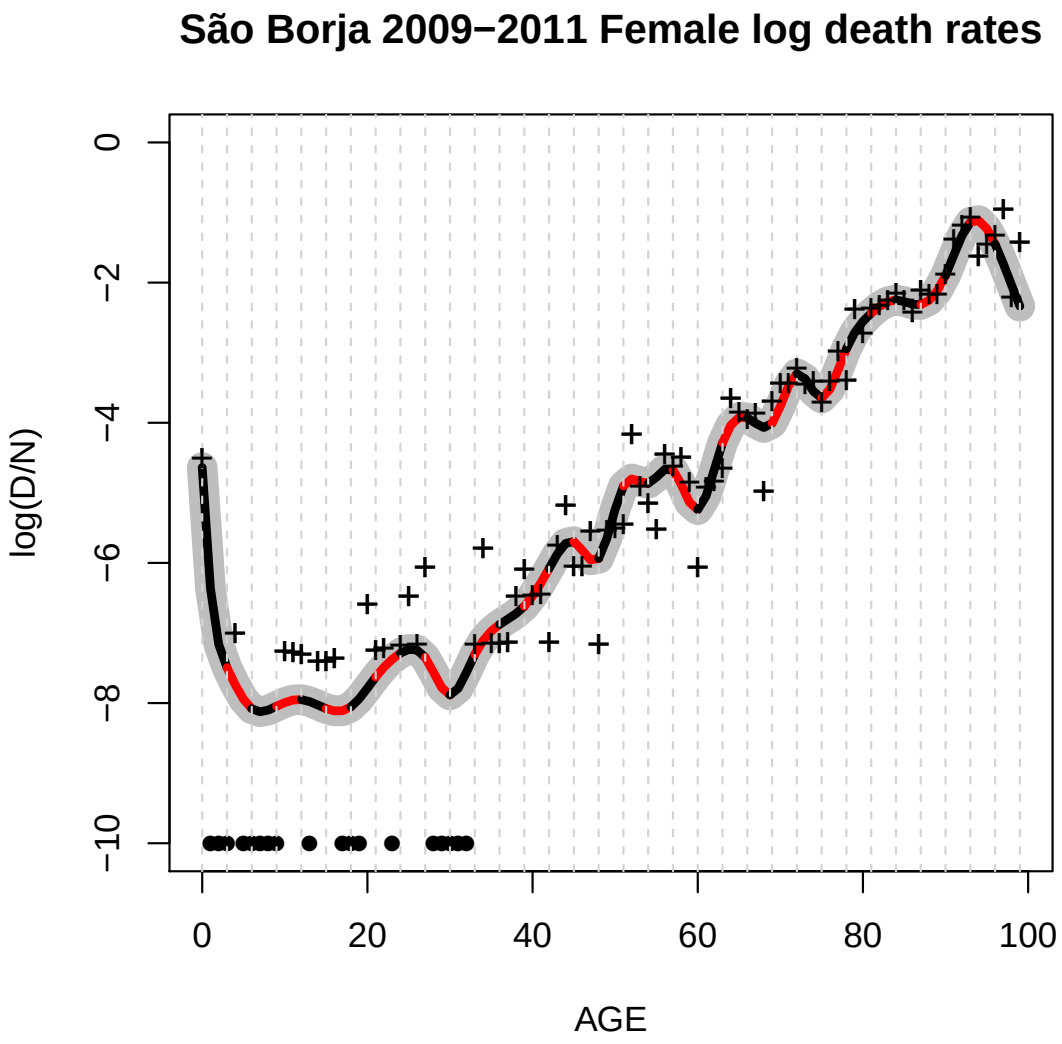
As an estimator, a high-dimensional spline function $g = \boldsymbol{S}\theta$ is often quite vulnerable to sampling noise. For example, Figure 1 displays the maximum likelihood fit for $\theta \in \mathbb{R}^{36}$ using the 100×36 cubic B-spline $\boldsymbol{S}$ matrix described above, with age-specific data for female exposure and deaths over 2009–2011 in the municipality of São Borja in the southern Brazilian state of Rio Grande do Sul. During this three-year period the municipality had approximately 94,000 woman-years of exposure and 688 female deaths over ages 0–99.

It is evident from Figure 1 that the "untamed" spline overfits mortality rates, in the sense that the true schedule of log rates is unlikely to have up-and-down fluctuations within small age ranges like those in the fitted model. This illustrates a classic bias-variance tradeoff: a high-dimentional spline function is flexible enough to represent small-scale features in the true rate schedule accurately (low bias), but that same flexibility means that it may overinterpret coincidental features of sample data (high variance). From many studies of large populations with reliable mortality statistics, we know that log mortality rates in human populations tend to increase fairly smoothly over ages 40–99. Thus, although a spline model correctly estimates the broad pattern of sharply decreasing mortality at young child ages followed by increases in adolescence and adulthood, the fitted spline function obviously exhibits too much curvature and non-monotonicity at adult ages.

[1] $\boldsymbol{S}$ is constructed in *R* with the command *splines::bs(0:99, knots=seq(3,96,3), degree=3,intercept=TRUE)*.

[2] More precisely, $s_x = S' e_x$, where $e_x = (0 \ldots 1 \ldots 0)'$ is an $A \times 1$ vector with 0s everywhere except for the position corresponding to age x.

Figure 1: 36-dimensional spline without penalties: Max Likelihood Fit



2.3.1 P-Spline penalties on parameters

A P-spline approach to estimation [5] "regularizes" or "tames" the spline by adding a penalty term to the log likelihood. For the purposes of this paper, the key feature of these penalties is that they apply to the vector of spline coefficients $\theta \in \mathbb{R}^K$ rather than to the estimated demographic schedule $g = \boldsymbol{S}\theta$. In a P-spline approach the researcher maximizes the penalized function

$$f(\theta) \;=\; L\,(\theta) - \lambda\,\theta'\boldsymbol{D'D}\theta$$

where $\boldsymbol{D}$ is a $(K-p)\times K$ matrix of constants such that $\boldsymbol{D}\theta$ is the vector of p^{th} differences in spline coefficients.

With equally-spaced knots for the spline basis functions in the columns of $\boldsymbol{S}$, the penalized log likelihood function $f(\theta)$ has higher values for for demographic functions that fit the data *and* which approximate p^{th}-order polynomials [2]. Maximizing the penalized function therefore requires a tradeoff between fitting the data and simplifying the fitted curve, where "simplifying" means choosing a smoother curve that looks more like a p^{th}-order polynomial.

P-splines represent an elegant approach to the bias-variance tradeoffs for spline models illustrated in Figure 1. Importantly, they are the foundation of the *MortalitySmooth* package in *R* [2]. As an example, Figure 2 shows the estimated log mortality rates for the São Borja data in Figure 1, using 3rd-order differencing penalties on θ, 33 internal knots (identical to those in Figure 1) and the *MortalitySmooth* defaults for other parameters, including the critical smoothing parameter λ.[3]

The P-spline fit in Figure 2 represents a significant improvement over the unpenalized maximum likelihood fit to the same data in Figure 1. The P-spline curve is a more plausible model for São Borja's mortality schedule: it is smoother, nearly linear over older adult ages, and (almost) monotonically increasing. These are common features in mortality schedules estimated from large populations, and the P-spline model finds a good tradeoff between fitting the small-sample data and smoothing the schedule.

2.3.2 An alternative: D-spline penalties on the fitted schedule

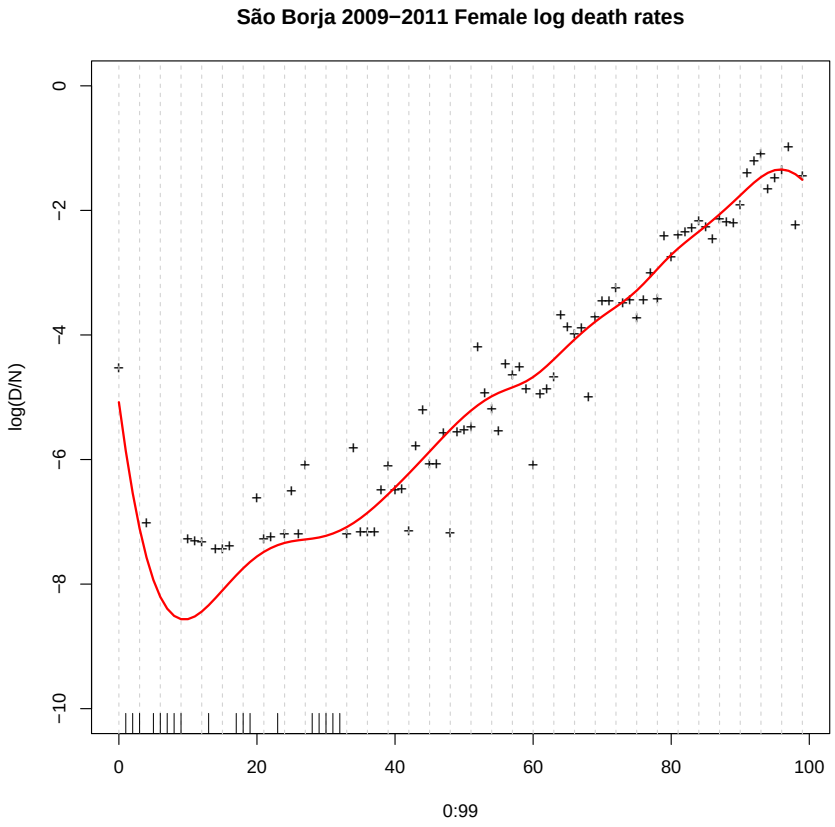
The P-spline approach to function smoothing and regularization has proven to be very valuable, but it is not specifically designed fitting demographic rate schedules. In particular, it implicitly relies on polynomial functions as a kind of gold standard for functional shapes. That reliance makes good sense for a generic curve-fitting tool, but it may not be optimal for fitting specific types of demographic curves (e.g. mortality or fertility age schedules) for which demographers already have specialized models and considerable prior knowledge.

In this paper I investigate an alternative approach that I call *D-splines*. D-splines also use penalized, high-dimensional spline functions, but penalties are based on deviations from demographic, rather than arithmetic, standards. The essential idea is to penalize features of the fitted schedule $g = \boldsymbol{S}\theta$ directly, using demographic prior knowledge. I propose several alternative penalties in Section 3, all constructed as follows

- define a *residual* vector $\varepsilon \in \mathbb{R}^G$ that should be close to zero for "good" schedules g
- calculate empirical residuals $\varepsilon_1 \dots \varepsilon_{487}$ across a set of 487 1x10 (single-year age by ten-year period) observed life tables in the Human Mortality Database (HMD) [10]

[3] With this dataset, the *Mort1Dsmooth* function does not converge when using the default 2nd-order penalties. Non-convergence was also major issue in many of the HMD experiments reported below.

Figure 2: 36-dimensional P-Spline fit to São Borja female data



- use the $G \times G$ empirical covariance matrix $\hat{\boldsymbol{V}}_{HMD} = \frac{1}{487} \sum_i (\varepsilon_i \varepsilon_i')$ as an estimate for the expected covariance of residuals
- replace the P-spline penalty on parameters θ with a "D-spline" penalty based on the residuals for fitted functions $S\theta$

The result is a penalized likelihood function

$$f(\theta) = L(\theta) - \frac{1}{2} \varepsilon'(\theta) \left[\hat{\boldsymbol{V}}_{HMD}^{-1} \right] \varepsilon(\theta)$$

for which the maximization tradeoff is "fit versus fidelity to patterns in demographic schedules", rather than "fit versus local smoothness".[4] The absence of the unknown smoothing constant λ is an additional benefit: in the D-spline approach λ does not have to be estimated. Instead the penalty term/tolerance for deviations is calibrated empirically to mimic

[4] The $\frac{1}{2}$ constant in the penalized log likelihood arises from an assumption that residuals would be approximately normally distributed over demographic schedules. I have not (yet) formally tested the normality of HMD residuals, but informal examination suggests that normality is quite reasonable: most residual distributions are close to symmetric, with thin tails.

residual patterns found in high-quality demographic data.[5]

3 Four experimental D-spline penalties for mortality schedules

3.1 Slope penalties

Log mortality schedules for human populations have characteristic shapes that can be described in terms of slopes (or equivalently, first-differences over single years of age). One possible approach to characterizing "good" spline schedules is therefore to measure the difference between the slopes in a proposed spline function $\boldsymbol{S}\theta$ and the average slopes in HMD schedules at the same ages.

For example, the average value of $(\ln \mu_1 - \ln \mu_0)$ across 487 1x10 HMD schedules is -2.005, with a standard deviation of 0.61. This suggests that steep negative slopes are likely between age 0 and age 1 in "good" schedules, but that we should be fairly tolerant about the exact slope value because of the large standard deviation. In contrast, between integer ages 75 and 76 the average value and standard deviation of $(\ln \mu_{76} - \ln \mu_{75})$ across HMD schedules are +0.109 and 0.056, respectively, which suggests that between these ages almost all "good" schedules have slopes within a narrow range of small positive values.

Applying this approach simultaneously to all 99 first-differences for intervals starting with ages 0...98, the HMD slopes have mean vector m_1 and covariance matrix $\boldsymbol{V}_1$. Defining $\boldsymbol{D}_1$ as the standard 99×100 first-differencing matrix, this produces a penalized log likelihood for the spline schedule $\boldsymbol{S}\theta$:

$$f_1(\theta) = L(\theta) - \frac{1}{2} (\boldsymbol{D}_1 \boldsymbol{S}\theta - m_1)' \, \mathrm{V}_1^{-1} \, (\boldsymbol{D}_1 \boldsymbol{S}\theta - m_1)$$

The value θ_1^* that maximizes this function produces a fitted "D-spline" schedule $\boldsymbol{S}\theta_1^*$ for mortality.

3.2 Curvature penalties

A slightly less demanding criterion than that in Section 3.1 might penalize curvature (second-differences) in the fitted spline schedule that failed to match HMD empirical patterns. Defining $\boldsymbol{D}_2$ as the standard 98×100 second-differencing matrix and defining m_2 and $\boldsymbol{V}_2$ as above

[5] The *calibrated spline* method for fitting fertility schedules that I developed in an earlier paper [8] uses the same fundamental logic.

(except for the 98 second-differences in HMD schedules) yields an analogous penalized log likelihood with different constants:

$$f_2(\theta) = L(\theta) - \frac{1}{2}\left(\boldsymbol{D}_2\boldsymbol{S}\theta - m_2\right)'\boldsymbol{V}_2^{-1}\left(\boldsymbol{D}_2\boldsymbol{S}\theta - m_2\right)$$

a different optimal value θ_2^* , and a different D-spline fit $\boldsymbol{S}\theta_2^*$.

3.3 Lee-Carter penalties

One can also define "good" spline schedules according to their fidelity to existing models. In this approach *residuals* might represent features of a schedule that cannot be represented within a specified model family. For example, in the most commonly used mortality modelling framework, Lee-Carter [7], schedules are modeled as

$$\ln \mu_x = a_x + k \cdot b_x$$

where $\{a_x\}$ and $\{b_x\}$ are $A \times 1$ vectors of pre-determined constants from that represent a baseline schedule and typical deviations from that schedule, estimated from a singular value decomposition on an empirical database.

In the Lee-Carter model the scalar parameter k determines the level of deviation from the baseline, so that (with $A = 100$ ages) the vector

$$\begin{bmatrix} \ln \mu_0 - a_0 \\ \ln \mu_1 - a_1 \\ \vdots \\ \ln \mu_{99} - a_{99} \end{bmatrix} = k \cdot \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{99} \end{bmatrix}$$

must lie in the column space of vector $b \in \mathbb{R}^{100}$. This suggest another way to define D-spline penalties using the HMD. Specifically, we can estimate the Lee-Carter a and b vectors from the HMD and then define the *residuals* for any schedule $\{\ln \mu\} \in \mathbb{R}^{100}$ as the part of $\{\ln \mu\} - a$ that lies outside of the column space of b. In matrix notation this vector of residuals is

$$\begin{aligned} \varepsilon &= \left[\boldsymbol{I} - b(b'b)^{-1}b'\right]\left[\{\ln \mu\} - a\right] \\ &= \boldsymbol{M}_b\left[\{\ln \mu\} - a\right] \end{aligned}$$

The mean of these residuals in the HMD equals zero by construction. After calculating the Lee-Carter residuals' covariance ($\boldsymbol{V}_{LC}$) across HMD schedules, the corresponding penalized log likelihood for D-spline estimation is

$$f_{LC}(\theta) = L(\theta) - \frac{1}{2}\left(\boldsymbol{M}_b\left(\boldsymbol{S}\theta - a\right)\right)' \boldsymbol{V}_{LC}^{-1}\left(\boldsymbol{M}_b\left(\boldsymbol{S}\theta - a\right)\right)$$

3.4 TOPALS penalties

TOPALS models [3, 6] are much less familiar than Lee-Carter, but share some properties that make them potentially useful for D-spline penalties. TOPALS is a relational approach, in which a demographic rate schedule is modeled as the sum of a fixed standard function and a variable offset function. In the case of mortality schedules a TOPALS model has the form

$$\{\ln \mu\} \;=\; \{\ln \mu^*\} + \boldsymbol{B}\alpha$$

where $\boldsymbol{B}$ is a $100 \times K$ matrix of *linear* B-spline functions over $0 \ldots 99$ constructed from knots at specified ages (here I will use $K = 7$ and linear spline knots at ages 0,1,10,20,40,70, and 100). Details about this model are in [6, 9]. The structure of a TOPALS mortality model is very similar to Lee-Carter: (fixed schedule) + (parameterized offset function). The principle difference is that the TOPALS offset function is a flexible linear spline rather than a scalar multiple of a fixed one-dimensional vector. The extra flexibility in the TOPALS approach relative to Lee-Carter brings the usual benefits and costs: lower bias with potentially higher variance in small samples.

For a D-spline approach based on TOPALS modeling, a schedule's *residuals* are the portion of $(\{\ln \mu\} - \{\ln \mu^*\}) \in \mathbb{R}^{100}$ that lies outside of the column space of $\boldsymbol{B}$. In matrix notation the vector of residuals is

$$\begin{aligned} \varepsilon \;&=\; \left[\boldsymbol{I} - B(B'B)^{-1}B'\right]\left[\{\ln \mu\} - \{\ln \mu^*\}\right] \\ &=\; \boldsymbol{M}_B\left[\{\ln \mu\} - \{\ln \mu^*\}\right] \end{aligned}$$

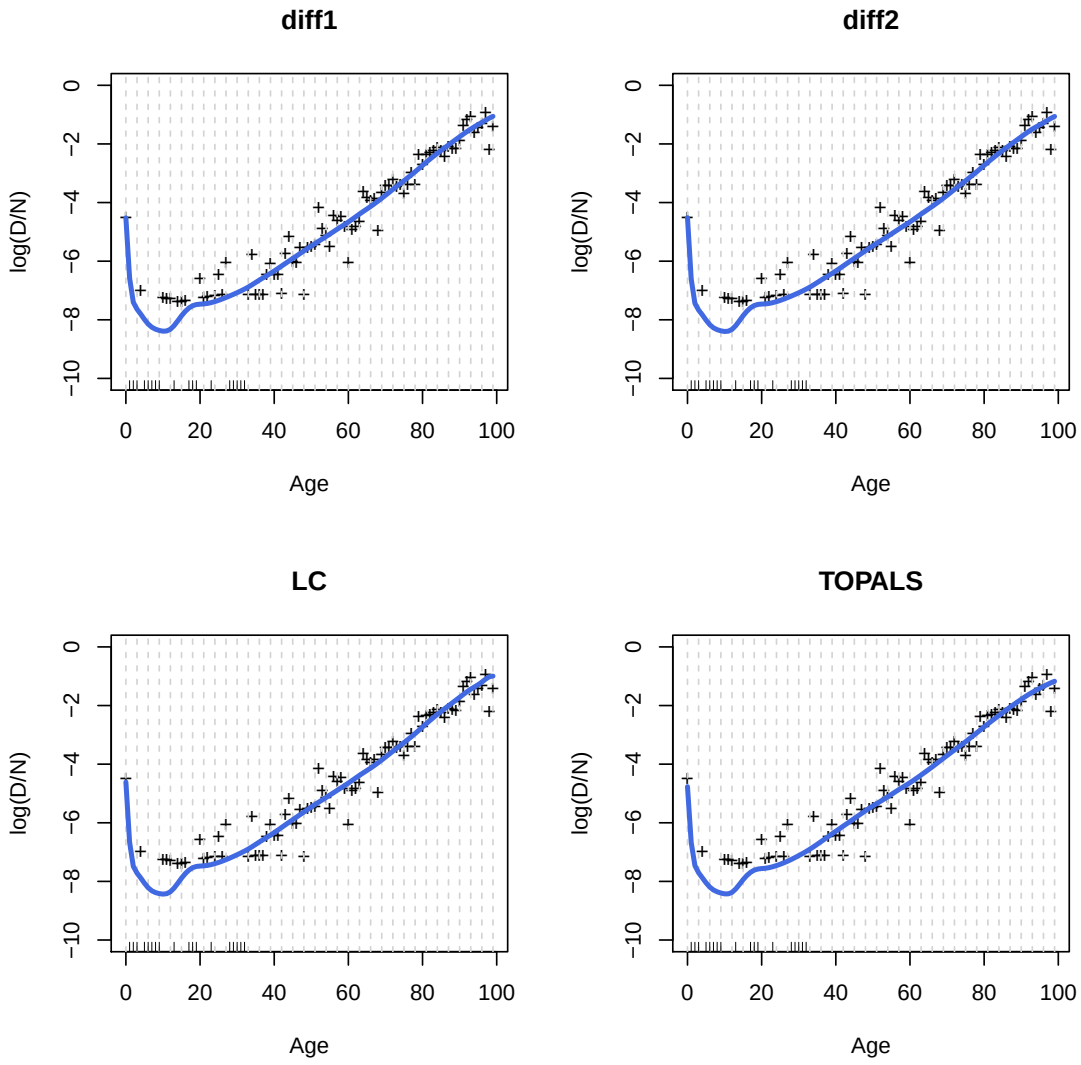
I use the mean HMD log mortality rate at each age as the standard $\{\ln \mu^*\} \in \mathbb{R}^{100}$, so that the average HMD residual is zero by construction.[6] After calculating the TOPALS residuals' covariance ($\boldsymbol{V}_{TO}$) across HMD schedules, the corresponding penalized log likelihood for D-spline estimation is

$$f_{TO}(\theta) = L(\theta) - \frac{1}{2}\left(\boldsymbol{M}_B\left(\boldsymbol{S}\theta - \{\ln \mu^*\}\right)\right)' \boldsymbol{V}_{TO}^{-1}\left(\boldsymbol{M}_b\left(\boldsymbol{S}\theta - \{\ln \mu^*\}\right)\right)$$

Figure 3 shows the spline functions that maximize the four alternative D-spline criteria when using the São Borja data used earlier in Figures 1 and 2. For this particular data set,

[6] This definition of the standard means that in all of the examples here, the TOPALS standard schedule $\{\ln \mu^*\}$ and the Lee-Carter a vector are identical.

Figure 3: Alternative D-spline fits to São Borja 2009-2011 female mortality data



all four D-spline methods produce extremely similar and extremely plausible fits.[7]

[7]The scale of the plots makes fine details difficult to see. The main differences between the four fitted schedules in Figure 3 are (1) slightly higher infant and child mortality with diff1 and diff2 penalties; (2) slightly higher mortality above age 80 with LC penalties; (3) slightly lower mortality above age 80 with TOPALS penalties.

4 Estimator comparison: preliminary results

4.1 Test data

I use 487 single-year age by ten-year period female mortality schedules in the HMD as the foundation for small-sample experiments. These schedules come from 49 different countries, over decades spanning the 1750s (for Sweden) to the 2010s.

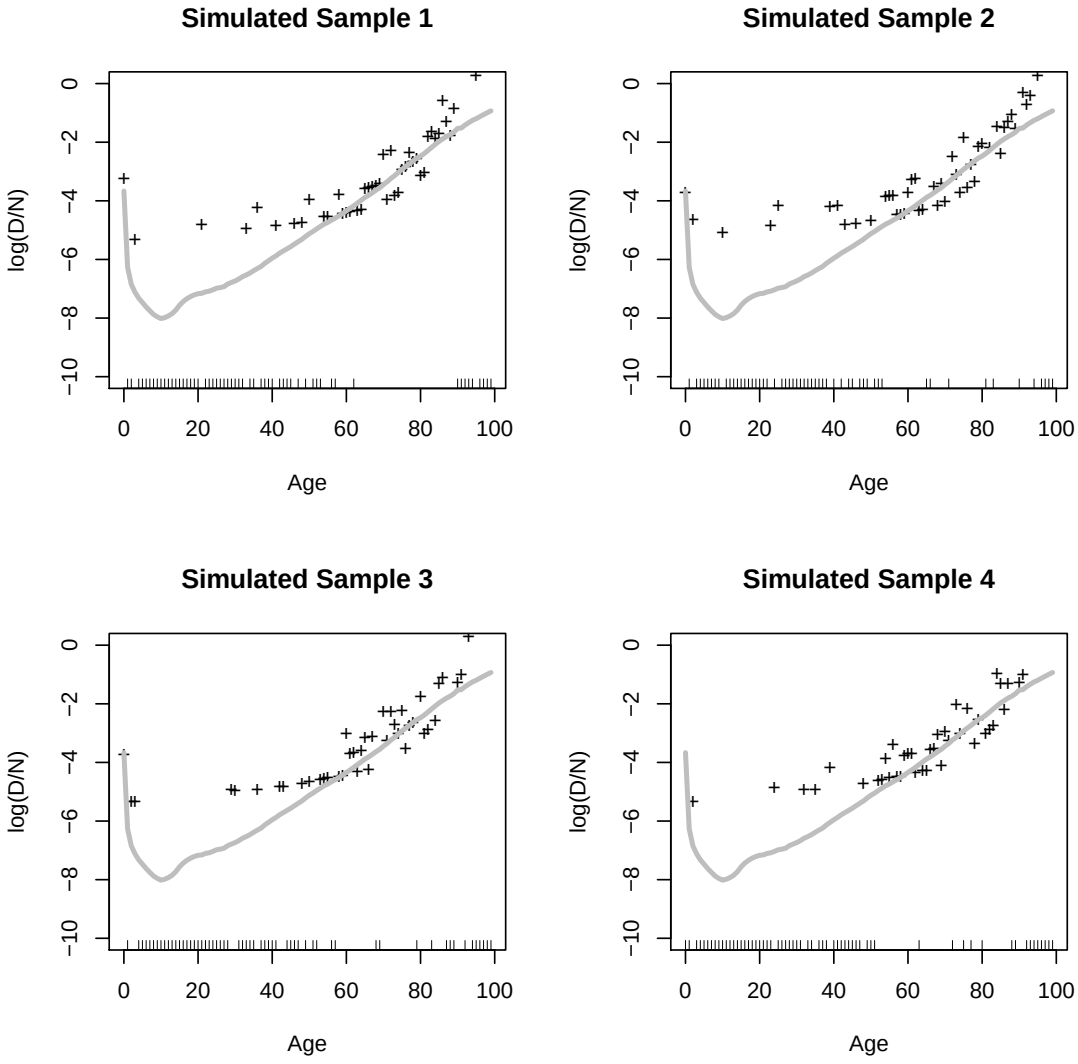
For each of the four proposed D-spline criteria, I calculated the 100×1 mean residual vector over the 487 schedules (equal to zero by construction for the Lee-Carter and TOPALS variants) and the 100×100 covariance matrix $\boldsymbol{V}$ of the residuals across the 487 schedules. These calculations provide the constants necessary for estimating the D-spline objective functions.

I then used the HMD 1x10 exposure data associated with each mortality schedule to create simulated (deaths,exposure) samples, as follows. For each schedule $s = 1 \ldots 487$ I rescaled observed age-specific exposures (N_{sx}) to represent a small population with the same age structure: $N^*_{sx} = P^* \cdot \frac{N_{sx}}{\sum_x N_{sx}}$, where the small population P^* is either 1000, 10000, or 100000. For each of the 487x3=1461 small populations, I drew 100 independent samples of simulated deaths at ages $x = 0 \ldots 99$ using the true log mortality rates from the corresponding HMD schedule:

$$D^{sim}_{sx} \sim Poisson(N^*_{sx} \cdot \mu_{sx}) \qquad \text{[repeated 100 times]}$$

Figure 4 illustrates the procedure, showing four simulated (death,exposure) samples for population size $P^* = 10000$ drawn from the 1950-1959 schedule for the USA. The true HFD mortality schedule is indicated with a grey line, and ages with zero simulated deaths are indicated with an extra tickmark along the horizontal axis.

Figure 4: Simulated Small-Population Data ($P^* = 10000$, USA 1950-1959 data) and true rate schedule



4.2 Experimental Design

For each of the 487x3x100=146,100 (HMD schedule, population size, simulation) combinations, I estimated the optimal schedule $\boldsymbol{S}\theta$ separately for each the four D-spline models penalty functions, where $\boldsymbol{S}$ is the 100×36 cubic B-spline matrix described earlier. D-spline calculations used a standard Newton-Raphson algorithm [1, p 137-139]. For each (HMD schedule, population size, simulation) combination I also estimated a P-spline function $\boldsymbol{S}\theta$ with the same basis $\boldsymbol{S}$, using the *Mort1Dsmooth* function from the *MortalitySmooth* package in *R* [2], with all settings other than the interior knots set at default values.

Of the 146,100x5=730,500 spline fitting problems to be solved in this design, algorithms

Table 1: Number of Convergence Failures for Spline Fits
(48,700 death and exposure samples in each cell)

		P^* =1000	10,000	100,000
D-spline	1st-Diff	27	0	0
	2nd-Diff	34	0	0
	Lee-Carter	0	0	0
	TOPALS	34	0	0
P-spline		7832	1388	20660

failed to converge to a solution in nearly 30 thousand cases, the vast majority of which were P-spline fits. Table 4.2 summarizes these problems.[8]

After removing non-convergent results, there were 700,525 estimated mortality schedules, each with a known true value. For each of the valid schedule estimates I produced three error measures:

1. mean absolute error $\frac{1}{100}\sum_x \{|\ln \mu_x^* - \text{true } \ln \mu_x|\}$.
2. $e_0^* - \text{true } e_0$
3. $e_{60}^* - \text{true } e_{60}$

The next section reports graphical summaries of these error measures over fitting methods and sample sizes.

4.3 Evaluation of Fitting Errors

4.3.1 Overall shape: mean absolute error

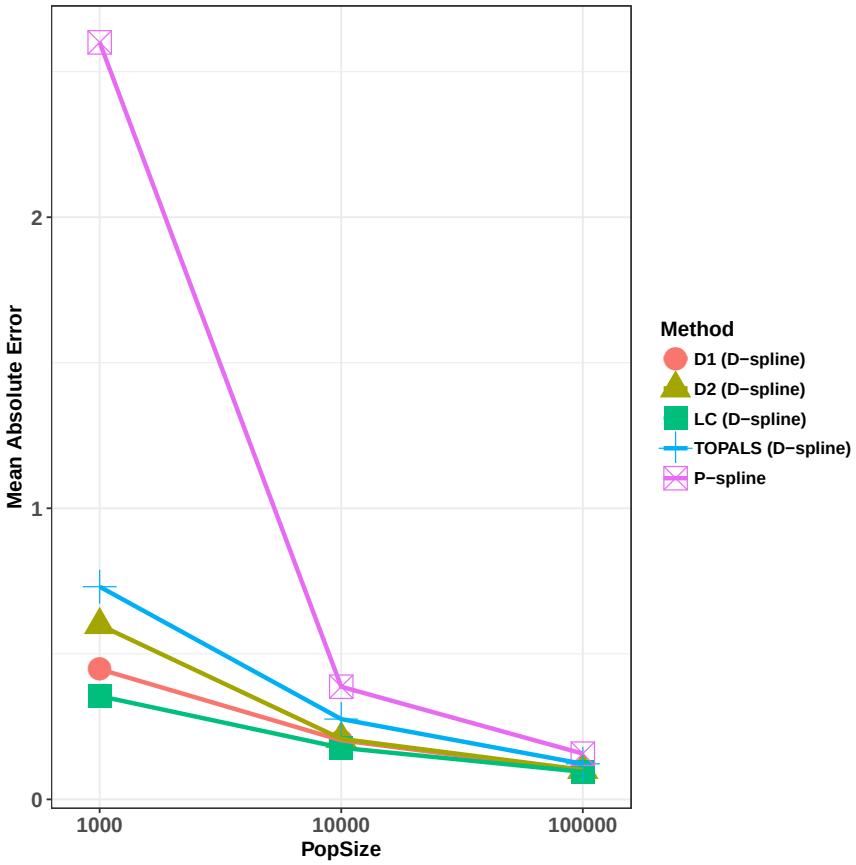
Figure 5 reports the mean absolute errors in log mortality rates for all estimated schedules, by simulated population size and spline fitting method. The information in the figure provides a measure of how well the alternative methods fit the overall shape of the corresponding true mortality schedules, with low values representing more accurate fits.

There is a clear ranking of estimators on this metric, with Lee-Carter D-splines performing best and standard P-splines worst. All methods produce good fits in larger populations with more deaths and exposure, but the P-spline approach does not do very well in small samples. This occurs because with sparse sample information the P-spline algorithm tends to oversimplify schedules and often produces nearly linear fits that increase monotonically

[8]Small changes to the *Mort1Dsmooth* default fitting parameters, for example from pord=2 to pord=3, had virtually no effect on convergence problems.

over the entire age range. In contrast, Lee-Carter D-splines do quite well at fitting small-sample schedules, probably because of the more rigid shape implied by low-ε schedules in the Lee-Carter version.

Figure 5: Mean Absolute Error by Population Size and Spline Method



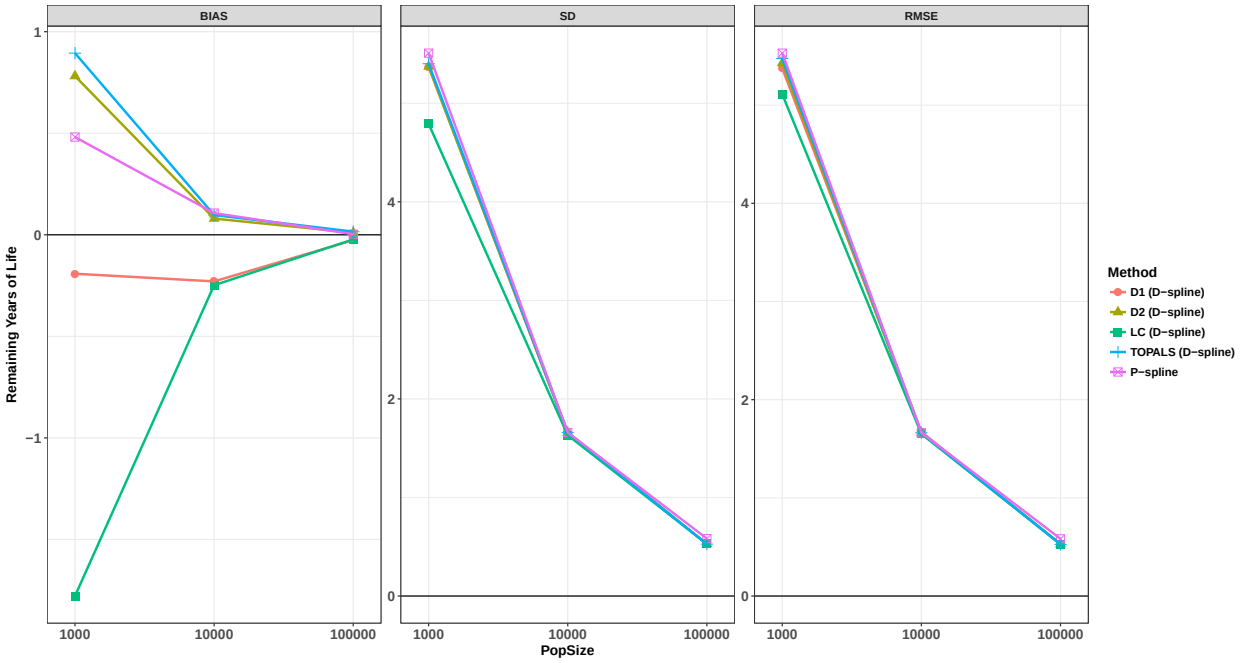
4.3.2 Overall level: e(0)

Perhaps the most important metric for success of a small-population method is accuracy in estimating life expectancy at birth. This is the most commonly reported mortality summary, so a good method should have low errors for e_0. Figure 6 reports summary measure of bias, standard deviation, and root mean squared error (RMSE) for all five spline estimation methods over the estimated schedules. An ideal method would have values of zero for all three measures at all population sizes.

Figure 6 displays information about the performance of the alternative estimators in terms of e_0. Note that the vertical scales differ in the three panels. Biases are comparatively small, and the main source of error at all three population sizes is sampling variance. Lee-Carter D-splines, which did best on overall shape in the previous section, once again appear to perform (slightly) better in terms of e_0. Although the bias of the Lee-Carter D-splines is

slightly higher than the other methods (Lee-Carter D-splines tend to overestimate mortality levels and thus underestimate life expectancy), their greater rigidity once again appears to be a net virtue.

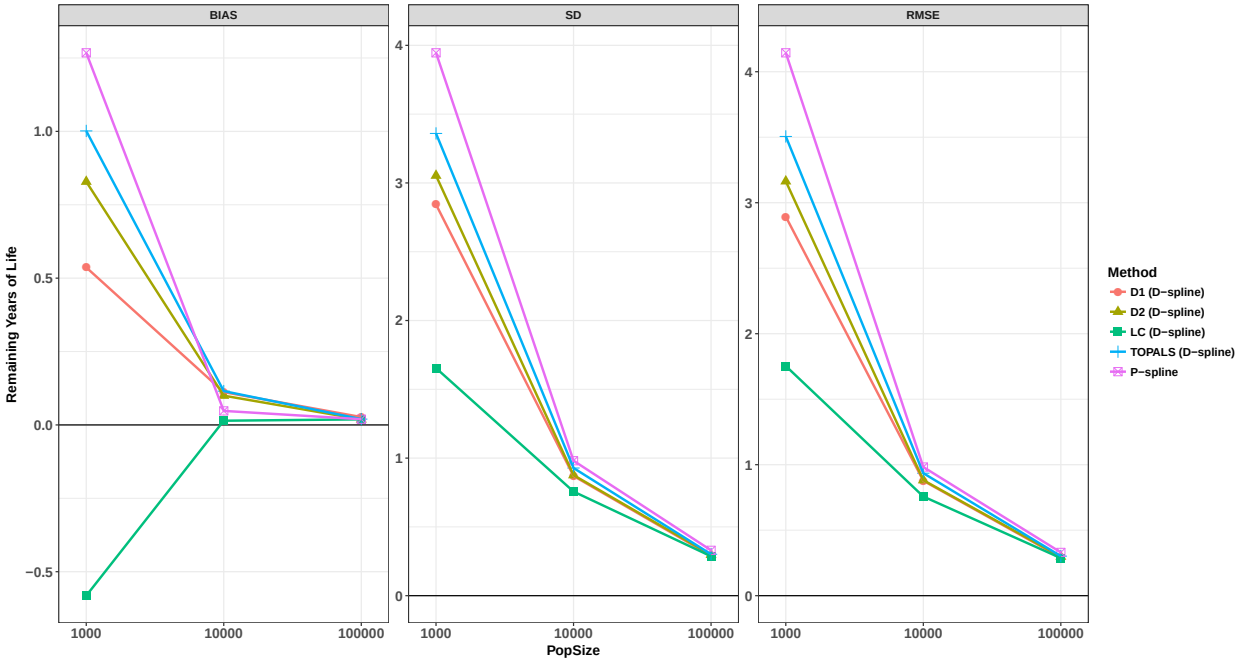
Figure 6: Estimation Errors for e_0



4.3.3 Level at older ages: e(60)

Finally, I summarize the performance of the estimators for estimating old-age mortality, using e_{60} as the predictand. Figure 7 shows errors for remaining life expectancy at age 60, in the same format as Figure 6. Again Lee-Carter D-splines perform best, with lower average errors (RMSE panel) at all sample sizes. It is also notable that for old-age mortality it seems that all four of the proposed D-spline methods outperform P-splines, with both lower bias and lower variance for e_{60}.

Figure 7: Estimation Errors for e_{60}



5 Further Work

This Extended Abstract reports on preliminary work and early results. The D-spline approach to penalized log likelihood appears to have promise, and by the time of PAA I expect to have a much more thorough analysis of the strengths and weaknesses of this framework.

Objectives for research between now and the PAA meeting in Austin include

- considering other D-spline residual definitions
- investigating the sensitivity of the experimental results to the choice of the schedules comprising the test bed. In the preliminary experiments here, I used the same set of schedules for (1) calibrating D-spline constants, and (2) testing fits. "Out-of-sample" performance could be different; that requires some kind of cross-validation approach.
- investigating more thoroughly the reasons for the good performance of the Lee-Carter D-spline method (what characteristics make it both low bias and low variance?)
- investigating the effective number of parameters for the various penalized models
- improving the convergence and performance of the P-spline estimators for better comparisons (although I suspect that eliminating the non-convergent cases probably improved, rather than worsened, the results for P-splines reported here)

References

[1] Takeshi Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

[2] Carlo Camarda. MortalitySmooth: An R Package for Smoothing Poisson Counts with P-Splines | Camarda | Journal of Statistical Software. 2012.

[3] Joop de Beer. Smoothing and projecting age-specific probabilities of death by TOPALS. *Demographic Research*, 27:543–592, October 2012.

[4] C. de Boor. *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer New York, 2001.

[5] Paul H. C. Eilers and Brian D. Marx. Flexible smoothing with *b*-splines and penalties. *Statistical Science*, 11(2):89–102, 1996.

[6] Marcos Roberto Gonzaga and Carl Paul Schmertmann. Estimating age-and sex-specific mortality rates for small areas with TOPALS regression: an application to Brazil in 2010. *Revista Brasileira de Estudos de PopulaÃ§Ã£o*, 33(3):629–652, 2016.

[7] Ronald D. Lee and Lawrence R. Carter. Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87(419):659–671, 1992.

[8] Carl P. Schmertmann. Calibrated spline estimation of detailed fertility schedules from abridged data. *Revista Brasileira de Estudos de População*, 31:291–307, 12 2014.

[9] Carl P. Schmertmann and Marcos R. Gonzaga. Bayesian Estimation of Age-Specific Mortality and Life Expectancy for Small Areas With Defective Vital Records. *Demography*, 55(4):1363–1388, August 2018.

[10] University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Human Mortality Database, 2014.