Parsimonious Parameterization of Age-Period-Cohort Models by Bayesian Shrinkage

Gary Venter[a,], Şule Şahin[b,]

[a] *School of Risk and Actuarial Studies, Business School, University of New South Wales, Sydney NSW 2052, Australia, email: gary.venter@gmail.com*
Columbia University, School of Professional Studies, Actuarial Sciences Program
[b] *Department of Actuarial Sciences, Hacettepe University, 06800, Ankara, Turkey, email: sule@hacettepe.edu.tr*
Institute for Financial and Actuarial Mathematics, University of Liverpool, Liverpool, L69 3BX, UK

Abstract

Age-period-cohort models used in life and general insurance can be over-parameterized, and actuaries have used several methods to avoid this, such as cubic splines. Regularization is a statistical approach for avoiding over-parameterization, and it can reduce estimation and predictive variances compared to MLE. In Markov Chain Monte Carlo (MCMC) estimation, regularization is accomplished by the use of mean-zero priors, and the degree of parsimony can be optimized by numerically efficient out-of-sample cross-validation. This provides a consistent framework for comparing a variety of regularized MCMC models, such as those built with cubic splines, linear splines (as ours is), and the limiting case of non-regularized estimation. We apply this to the multiple-trend model of Hunt and Blake [2014].

Keywords: Age-period-cohort models, mortality trends, MCMC, regularization, shrinkage priors.

Preprint submitted to ASTIN Bulletin: The Journal of the International Actuarial AssociationJune 15, 2017

1. Introduction

Age-period-cohort (APC) models model an array of data by assigning parameters to each row, column, and diagonal of the array. The values in each cell are then arithmetic combinations of the row, column and diagonal parameters. These models thus have parameters but no independent variables, although dummy variables can be constructed to mimic the arithmetic calculations. Actuaries use these models for mortality, and also for emergence of claims liabilities, by age and year. They can have many parameters, so it is easy to overfit to data noise. Regularization can reduce some of the response to noise and produce more parsimonious models.

Regularization is typically traced back to Hoerl and Kennard [1970]. It's most common forms are ridge regression and Lasso. These minimize the negative loglikelihood (NLL) plus the sum of the squares or absolute values, respectively, of the parameters. See Blei [2015]. The goal is to reduce estimation and predictive variance, even though shrinkage introduces bias in the estimate. The estimate is biased but closer to its true value.

Both ridge regression and Lasso actually start by scaling all explanatory variables to have mean zero and variance one, then minimize the NLL plus a selected shrinkage parameter λ times the sum of the squares or absolute values of the parameters. Part of the motivation for these approaches is that bigger parameters tend to be the source of much of the parameter variance. The constant term, which absorbs all the means of the explanatory variables, is not shrunk. The shrinkage is thus towards the overall mean. Lately Lasso is more popular because a number of the coefficients actually become zero, so it becomes a method of variable selection as well. Hastie et al. [2015] is a comprehensive source.

The Bayesian versions of ridge regression and Lasso impose shrinkage priors on the parameters. The ridge regression analogue is to use mean-zero normal priors with a small selected variance. Lasso is related to using low-variance Laplace (double exponential) priors, which push the parameters more strongly towards zero but also accept a few larger parameters. In the Bayesian version the Lasso parameters might get small without actually becoming zero, at least not in all samples. A major advantage of the MCMC implementation is the availability of a computationally efficient method of cross-validation – leave-one-out, or "loo" – that can be used to optimize the degree of shrinkage.

Actuarial credibility theory was shrinking rating class and territory parameters towards the mean before 1970, so can be considered to be the original form of regularization. The related James-Stein estimator was shown in a striking example by Efron and Morris [1975] to greatly reduce prediction variance by employing similar shrinkage.

APC models are reviewed in section 2 and MCMC optimization of them in section 3. We apply one such to US male mortality rates in section 4. Section 5 considers the results and possible extensions. Section 6 concludes.

2. The Model Class

Time-based data can come in any frequency, but for simplicity we will call our units "years". Modeled arrays have observations for each age in the column for that age, though ages are sometimes called lags in general insurance. Period refers to the observation time, i.e., the calendar year of the measurement. Cohort, or year of origin, is thought of as the year of birth in mortality and any sort of specifying year for liability, such year lawsuit filed or year contractually obliged. Year-of-birth cohorts are actually defined by year of death minus age at death. This will be the year of birth if death occurs after that year's birthday, otherwise it will be the year after birth.

We denote cohorts as $n = 1, ..., N$ and years as $u = 0, ..., U$. This produces periods of $n + u = 1, ...N + U$, although typically the observations run out before $N + U$. This puts a cohort in each row of the data, and periods are on NW to SE diagonals. Another popular notation system would make the periods the rows.

With a bit of loss of generality we denote $y[n, u]$ as the log of the incremental observation in the n, u cell, which could be mortality rates or incremental claims payments, for instance. More generally y could be some other transform of the data that nonetheless is modeled as below.

The oldest models in the actuarial literature use just two of the time dimensions. For instance the model of Lee and Carter [1992] models mortality by age with the q parameters and period, by r, but not cohort:

$$y[n, u] = q[u] + r[n + u]s[u] + \epsilon[n, u] \tag{1}$$

The idea of the $s[u]$ trend-weight-by-age parameters is that some ages benefit from the trend over periods more than others do. Verbeek [1972] used this AP model without the $s[u]$ factors for claims count emergence, and Taylor [1977] popularized that version for claim amounts. General insurance actuaries have been using AC models informally since at least the 1930s. Hachemeister and Stanard [1975] formalized these to show that a popular method is in fact MLE for a Poisson model.

Barnett and Zehnwirth [2000] discuss an APC model for claims, and recommend parameter reduction when using it, with cohort parameters $p[n]$:

$$y[n, u] = p[n] + q[u] + r[n + u] + \epsilon[n, u] \tag{2}$$

Renshaw and Haberman [2006] add age weights to both the period and cohort effects for mortality modeling:

$$y[n,u] = p[n]w[u] + q[u] + r[n+u]s[u] + \epsilon[n,u] \tag{3}$$

Cairns et al. [2009] however find problems with applying age weights to the cohorts. Our initial fitting suggested that with parameter reduction these age weights quickly collapse to a constant for our data. By using constrained parameterized curves, Xu et al. [2015] were able to fit age weights separately by cohort, and found different age sensitivity in different cohorts. That may explain why a single weighting function often does not work.

We ended up leaving out the age weights for cohorts, but used the multi-trend model described in Hunt and Blake [2014]:

$$y[n,u] = p[n] + q[u] + \sum_i r_i[n+u]s_i[u] + \epsilon[n,u] \tag{4}$$

The original model with a single trend assumes that the changes causing that trend, such as advances in medical treatments, always affect the various ages in the same way although to different degrees. But in some cases they differ more broadly. Modernization of more primitive societies is an example, where youth mortality rates decline substantially initially, then mortality trends at older ages become more significant. The additional trends allow different age take-up rates for the different trends at different times. They also allow for modeling events that may last several years but predominately affect specific age groups. The HIV epidemic is an example.

APC models have even a longer history in social sciences. See for example Ryder [1965]. Perhaps the original paper with a model having parameters in all three directions was the epidemiology paper of Greenberg et al. [1950], who in turn cite Frost [1939] for pioneering statistical analysis in the three directions. That literature early on addressed the issue of identifiability in such models. Fienberg and Mason [1978] summarizes that discussion. Basically you can get the same fit to every cell by using offsetting effects among the different parameters.

Actuaries use various constraints on the parameters to achieve identifiability. We adopt the following constraints:

- There is no long-term average trend across cohorts. We achieve this by making the cohort parameters the residuals of a regression over the years of birth. This forces the long-term trend to be represented by the period parameters only. The cohort parameters represent relative differences among the cohorts given that the overall trend is entirely taken up by the period parameters.

- The period trend is the trend for the age or ages with the greatest mortality change over the entire period. Such an age will get an age weight $s[u]$ of 1.0. The other ages will get lower age weights, which represent the degree to which they are affected by that trend. The assumption of the model is that every age u gets $s[u]$ times each year's trend, with the variability around that going into the residuals. There is a new trend parameter for each year, so an additional trend series is needed only when the age weights change.

Of course all these parameters are estimated simultaneously to find the combination of parameters that best fits the data under this model. The advantage of this particular choice of constraints is that each parameter group can be interpreted as having a precise meaning. The period parameters $r[n+u]$ are the by-period cumulative changes in mortality for the ages with the highest trend, given that there is no trend in the cohorts. The cohort parameters $p[n]$ are the relative cohort effects given that the overall trend is all in the period parameters, and the base mortality curve $q[u]$ represents the starting mortality rate for each age estimated across all the observations, given that the trends and weights $s[u]$ are as estimated.

Some recent papers apply MCMC to APC models. Although they do not aim at regularization, they find that MCMC provides a useful framework for estimation of more complex APC models. For instance, Antonio et al. [2015] uses MCMC to simultaneously estimate models for different populations. Chung Fung et al. [2016] uses it to incorporate stochastic volatility into APC models.

We estimate all the p, q, r and s parameters on linear splines – that is, line segments. There is a single-period line segment between any two adjacent parameters. The slope changes at each point are the underlying parameters of the model. These are second differences in the level parameters p, q, r, s. The level parameters are cumulative sums of the previous slopes, and the slopes are the cumulative sums of the previous slope changes.

The changes in slope at each period are the parameters that are shrunk towards zero. This produces little change in slope for most points, but larger changes occasionally. To model this, each slope change is given a Laplace (double exponential) prior with small variance. Smaller prior variances give more parsimonious models and smoother curves when graphing the parameters. Larger prior variances increase the likelihood, but allow more effective parameters, so do not necessarily increase the penalized log-likelihood. We use loo, a cross-validation method of penalizing the log-likelihood, to determine the optimal degree of shrinkage.

Since we are emphasizing the parameter shrinkage methodology, we just assume a normal distribution for y. However independent Bernoulli deaths would create a binomial distribution for number of deaths, and ages with a low number of deaths would then show greater proportional volatility of rates. Such distributional issues under this methodology would be a topic for further research.

3. MCMC estimation of age-period-cohort models

MCMC is a method to simulate (Monte Carlo) a sequence of samples from a probability distribution, where each sample is generated based on only the immediately previous sample (Markov Chain). It's main application is generating samples from the posterior distribution of a parameter vector θ when only the prior and conditional distributions are known. Thus it provides a way to do Bayesian estimation without being able to specify conjugate priors, or in fact any specific form of the posterior distribution. Good introductions are Ntzoufras [2010] and van Ravenzwaaij et al. [2016].

The key methodology to accomplish this is the Metropolis sampler. By Bayes' Theorem, the posterior distribution of θ given a sample X is:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \tag{5}$$

The denominator is usually intractable. Considering X to be constant, the posterior can be expressed as:

$$p(\theta|X) \propto f(\theta) \tag{6}$$

A sample from the posterior has to have more scenarios where the posterior probability is higher. Finding a new parameter set θ that increases $f(\theta)$ increases the posterior proportionally. To get from the latest θ to the next sample θ^*, the Metropolis sampler goes one parameter θ_j at a time, using a proposal distribution. For θ_j, this is a symmetric distribution, like normal, Laplace, or t, centered at θ_j. A parameter is drawn from the proposal distribution. If it makes $f(\theta^*) > f(\theta)$, it is accepted. If not another random number ρ is drawn, and if $\rho < f(\theta^*)/f(\theta)$, the parameter is accepted anyway. Otherwise the old θ_j is kept until the next round. The acceptance rule is designed to produce samples that conform in probability to the posterior distribution.

The Metropolis sampler is known to create a sample from the posterior under fairly general conditions, after a burn in period to get the parameters in the region of the maximum. The implementation after that is engineering, like the choice of the proposal distribution. It has been found that a narrow distribution produces too high an acceptance rate and can keep the sampler from getting to the maximum. But too low an acceptance rate prevents the sampler from moving much at all. Practice suggests that a rate around 40% is ideal, with anything in the range (0.25, 0.75) workable. The first major advance in methodology was the Metropolis-Hastings sampler, which no longer requires the proposal distribution to be symmetric, but has a bit more complicated acceptance rule.

Another major step was the use of the Gibbs sampler. See Casella and George [1992]. This started as a method of generating a sample from a bivariate distribution $p(x, y)$ by alternatively sampling from $p(x|y)$ and $p(y|x)$, using the last x to generate the next y and vice versa. The higher probability points are more likely to be drawn, so starting anywhere this eventually gets to be a sample of the bivariate distribution. For multivariate distributions, denote θ_{-j} as the vector without the j^{th} element. Then the sampling proceeds through the variables sampling from $p(\theta_j|\theta_{-j})$.

The application to Metropolis-Hastings is to take $f(\theta_j|\theta_{-j})$ as the proposal distribution and to set the acceptance rule to accept all samples. Here

θ_{-j} consists of the latest draws for all the other variables, so you can consider all of them as constants in $f(\theta_j|\theta_{-j})$, which is thus a univariate function. It is proportional to $f(\theta_{-j})$, so basically all you need is to be able to sample from a density proportional to $f(\theta_{-j})$, which might be done numerically.

Popular MCMC packages are JAGS, which uses the Gibbs sampler, and Stan, which is based on Hamilton mechanics, a method of dynamically tuning the proposal distribution of the Metropolis-Hastings sampler (Calderhead and Radde [2014]), which is itself controlled by the no U-turn sampler (NUTS). See Hoffman and Gelman [2014].

User-reported advantages of Stan include good error diagnostics, reliable convergence for a large class of models, and being able to work well with default settings. JAGS however may be considerably faster for some models.

We used Stan but found our model, with a large dataset and more than 400 parameters, to be about at the limit of what could be run on a personal computer. We drew samples of 3000 simulations, accepting the default of half of them as burn in. Convergence was verified by running several independent sampling chains, which can be done simultaneously with multiple processors, and doing a graphical comparison of the estimated parameters across the chains. A more formal test for each parameter is to take the ratio of its variance across all samples to its average variance inside the chains. The ratio is usually desired to be less than an admittedly subjective value of 1.1 for the model to be viewed as having converged. The runs took 4 days on the latest Mac laptops – which have a problem of erasing some needed temp files after 4 days of non-access. Mainframe systems or cloud computing would appear to be needed for any larger models. JAGS is certainly worth trying as well.

3.1. Priors

Priors can be used to constrain the parameters or to make some sets more likely. However it is typical to use fairly wide priors just to get the process going, except for the mean-zero priors for shrinking. For a parameter that can be positive or negative, just saying it has a prior that is uniform on the real line gives a wide prior. Typically MCMC packages will use double precision numbers, so this would give a uniform prior on $\pm 1.8 * 10^{308}$. A flat prior on the real line would have the vast majority of its probability outside of this interval. This would be easier to discuss with an extension of the real numbers to include infinitesimals, as in Keisler [2000]. Then the density of a uniform prior on the reals would just be an infinitesimal ϵ and the probability of being in or out of the interval $[-K, K]$ would be $2K\epsilon$ or $1 - 2K\epsilon$. In any case, there are not any truly improper priors in MCMC applications using double precision numbers, but that would not be noticeable in the posteriors.

Wide uniform priors generally produce estimates similar to classical unbiased estimates, with posteriors close to classical estimation-error distributions. This is not the case for parameters that have to be positive. A uniform prior on the positive line will tend to pull the estimate up a bit from the unbiased value. Heuristically this prior can be thought of as having an infinite pull upwards, compared to the uniform on all the reals which has balancing infinite pulls up and down. On the positive reals a prior proportional to $1/x$ has infinite weight at both ends and again tends to give classically unbiased estimates. This is sometimes easier to specify by making the log of the parameter uniform on the real line.

Mean-zero priors basically come in three varieties: light, medium, and heavy-tailed. The normal distribution is light-tailed. The double exponential, which is the exponential for positive values and its mirror image for negative, is medium-tailed. The Cauchy distribution (t-distribution with one degree of freedom) is heavy-tailed, as is the horseshoe distribution. The latter is a normal with σ^2 mixed with a Cauchy. It has a lot of probability near zero, so shrinks most parameters a good deal, but also has a heavy tail that allows occasional parameters to be large. It is regarded as being efficient in parameter reduction.

We use the double exponential shrinkage prior.

$$x > 0 : f(x) = e^{-x/b}/2b \tag{7}$$

$$x < 0 : f(x) = e^{x/b}/2b \tag{8}$$

The b parameter produces the standard deviation of the double exponential distribution. We are finding an optimal value around 0.04 for slope changes in the US male mortality model. The Cauchy is discussed in Appendix A.

3.2. Comparing Models

There is a growing view within statistics that models are approximations to more complex processes that are generating the sample data. This poses a challenge to traditional model testing, which is almost always based on the idea that the model generates the sample. The emerging consensus is that the most reliable model testing and comparison methods are based on testing the predictive power of the model on holdout samples.

MCMC does not maximize the likelihood – rather it generates the posterior distribution of the parameters. Also in non-linear models, and particularly for mean-zero priors, getting the effective number of parameters used, in order to adjust likelihoods for model comparison, is not clearcut.

But the likelihood of each holdout sample is a standard measurement it can do. This is called cross-validation.

A particularly appealing cross-validation test, popular lately because of a fast new algorithm for it, is leave-one-out cross validation, or loo. Each point is left out, one at a time, and the resulting model is tested on that point. Details may be found in Vehtari et al. [2016].

That part is not particularly new – see for instance Gelfand et al. [1992] or Gelfand [1996] – but estimates have tended to be unstable. What is new is a method called "Pareto-smoothed importance sampling," which addresses the instability. It is applicable to parameter sample sets generated by MCMC and is available in an R package `loo`. That takes the output of an MCMC run and estimates what the likelihood would be for a data point from the parameters fit by leaving it out.

Importance sampling is a numerical method often useful in Monte Carlo integration, used to compute the integral as the weighted average over a simulated sample of a more easily calculated integrand, using selected importance weights. It is used when, as with the hold-out sample here, the actual simulation of the process is too resource-intensive to be practical.

In the notation of Vehtari et al. [2016], y is the sample, an individual observation is y_i, and the sample leaving that point out is y_{-i}. A sample of S of the parameters is denoted as θ^S. Gelfand [1996] shows that

$$1/p(y_i|y_{-i}) = \int \frac{p(\theta|y)}{p(y_i|y_{-i},\theta)} d\theta \tag{9}$$

Using the MCMC-generated sample θ^S of simulated parameter sets, with importance weights w_i^s for y_i in the s^{th} parameter set, this can be estimated as:

$$1/p(y_i|y_{-i}) \approx \frac{\sum_s w_i^s}{\sum_s w_i^s p(y_i|\theta^s)} \tag{10}$$

or

$$p(y_i|y_{-i}) \approx \frac{\sum_s w_i^s p(y_i|\theta^s)}{\sum_s w_i^s} \tag{11}$$

Gelfand suggests trying weights of $w_i^s = p(y_i|\theta^s)^{-1}$. These give most weight to the samples that have poor fits at y_i, which can be anticipated to be more likely for the data excluding y_i. They result in estimating the holdout point's probability as its harmonic mean over the sample:

$$p(y_i|y_{-i}) \approx \frac{1}{average_s\left(p(y_i|\theta^s)^{-1}\right)} \tag{12}$$

Over time these weights have been found to be problematic in that some samples can give very low probabilities to the holdout point, giving a very high contribution to the average. As Wikipedia's page on the harmonic mean succinctly puts it, the harmonic mean tends to "mitigate the impact of large outliers and aggravate the impact of small ones."

Vehtari et al. [2016] address this basically using extreme value theory. They fit a Pareto to the probability reciprocals for each holdout point separately, and for each of these use the Pareto percentiles instead of the sample for the largest 20% of the $1/p$s. Their weights w_i^s are Gelfand's for 80% of the sample and the Pareto percentiles for the top 20% – with possible capping applied in some cases.

They test this and find that it performs reasonably well for many problems, especially for the sum over the dataset of the log of the holdout predictive probabilities, which is the loo cross-validation goodness-of-fit measure.

The symbol for this measure is $\widehat{elpd}_{loo}$, standing for "expected log pointwise predictive density." The true elpd is the expected value of the sum of the log of the probability densities for a new dataset not used in the fitting, where the expectation is over the actual distribution, not the fitted. This is not directly calculable and has to be estimated in some way.

The loglikelihood of the sample data is such an estimate, but it overstates the probability in that the parameters came from the same data. Fit testing measures like the AIC, etc. in fact can be viewed as attempts to correct for this bias. The priority of loo over other cross-validation metrics is that it is also a good estimate - some say the best available - of elpd, and so improves on AIC, etc. as well as being a cross-validation measure.

To summarize, then, $\widehat{elpd}_{loo}$ is a penalized log-likelihood measure that is arguably preferable to AIC, BIC, etc., particularly in shrinkage estimation where the parameter count is problematic. It can be used to determine the degree of shrinkage of a regularized model estimated by MCMC. That is a significant advantage to using Bayesian shrinkage instead of the classical versions. When the shrinkage is minimal, the standard MLE estimate results, so $\widehat{elpd}_{loo}$ includes that as a special case.

In seeking the degree of shrinkage, i.e., the b in the Laplace prior, that maximizes $\widehat{elpd}_{loo}$, for this model we found that increasing b from a low starting value gradually increases $\widehat{elpd}_{loo}$ up to a point, then increases both the loo penalty and the log-likelihood in step for considerably higher values of b, leaving $\widehat{elpd}_{loo}$ relatively stable around its maximum.

In such cases we prefer the lowest value of b that maximizes $\widehat{elpd}_{loo}$. This is the most parsimonious such model, which seems desirable in itself. But another point in its favor is that the derivation of the $\widehat{elpd}_{loo}$ estimate makes the usual statistical assumption that the data is generated by the model. Thus even though it is an out-of-sample test, $\widehat{elpd}_{loo}$ does not penalize for potential mis-specification of the model as being a somewhat simplified representation of a more complex process. It only penalizes for potential bias created by fitting to one particular sample of the process. This fact also tends to support using a more parsimonious estimator.

4. US Male Mortality

We model US male mortality rates from the Human Mortality Database (HMD) for ages 30–89, cohorts 1891–1975, and years 1970–2014. The data is considered error-prone prior to 1970.

Figure 1 graphs the cumulative changes in mortality by age decade (20s, 30s, 40s,...,80s) for 1970 – 2014. These are the changes in the log rates since 1970 averaged over each group. This does not look much like a single trend that the different ages participate in to various degrees.

The most dramatic exception to a single trend is the slowing or even reversal of the downward trend from 1985 – 1995 for ages under 50. Probably a good deal of this can be traced to HIV and the drug wars. There is also a flattening out of the downward trend starting in about 1997 for ages under 60, but less so for the 40s age group. The 80s ages have a much lower trend than the other ages up until about 2002, after which they appear to follow the general downward trend for the 70s, 60s and 40s age groups.

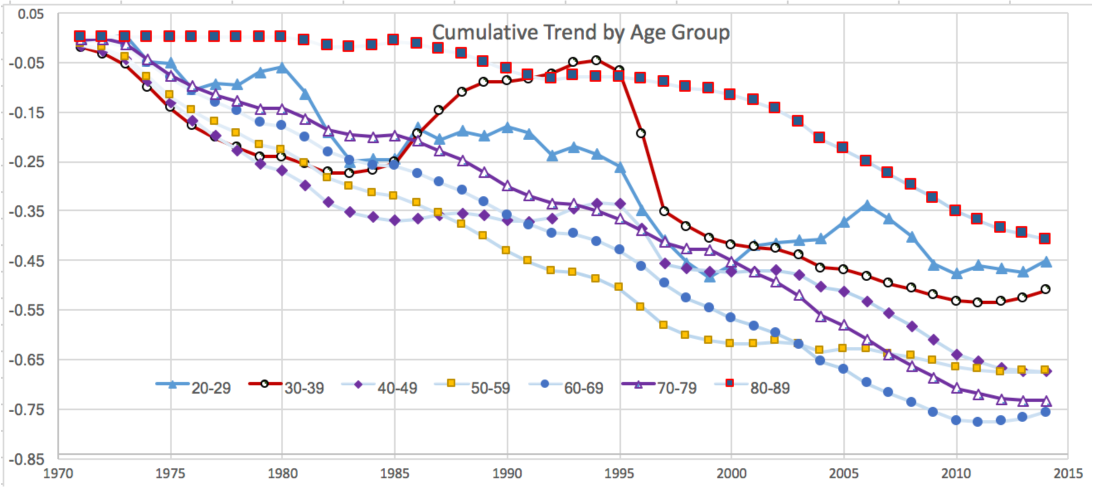

Figure 1: Cumulative trends in mortality rate by age group 1970-2014

These are challenging aspects for age / period / cohort models, which assign each age a constant percentage of the base cumulative trend for the entire period. However with several trends, each with their own weights, it may be possible to model much of the data. There are possibly a number of ways of defining multiple trends to model all this. One approach, which we will go through in detail, uses four key trends:

- A base trend that applies to all ages, which get their own weights.
- A trend from 1985 to 1995 forced to be non-negative, so towards higher mortality, for which ages get a possibly different set of weights.
- A trend, also forced to be positive, for ages in the 80s from 1975 until 2007, with all those ages getting the same weights.
- Another positive trend starting in 1997 which all ages under 80 participate in with their own weights.
- Because of volatility and differing trends, we did not include the 20s ages in the model.

There are thus 84 cohort parameters, for 1892 – 1975, with 1891 getting zero initially and 59 age parameters for 31 – 89, with age 30 getting zero. For trend there are 44 parameters for 1971 – 2014, with 1970 getting zero. Each age gets a weight, so there are 60 weight parameters for trend. These end up being constrained to have a maximum of 1.0. The HIV trend takes 11 parameters for 1985 – 1995. To these are applied age factors for ages 30 – 79, so 50 more parameters. There are 33 parameters for the additional trend for ages in the 80s from 1975 – 2007. The extra trend from 1997 takes 18 parameters, and it applies to ages 30 – 79, taking 50 more parameters.

All of these get double exponential priors with parameter b. The constant, consistent with Lasso practice, is not shrunk towards zero. Here it gets a uniform parameter on the reals, as does the log of the normal standard deviation. Thus there are 411 parameters that get priors. The shrinkage produces a smaller number of effective parameters.

For the cohort parameters a regression line to cohort year is fit to the p parameters. Then the residuals to this line become the detrended p parameters. For a regression of $y_1, ..., y_n$ on the integers 1,...,n, the slope and intercept are:

$$\begin{aligned} slope &= \frac{6\sum_j (2j-n-1)y_j}{n^3-n} && (13) \\ intercept &= average(y) - (slope)(n+1)/2 && (14) \end{aligned}$$

Making the age weights s positive was not so straightforward either. In Stan you can define parameters as being positive, but here the APC parameters are sums of slopes that can be negative. Just capping the final parameters from below at zero does not work well either, as that capping collapses derivatives that Stan uses to find posteriors. We ended up squaring the final parameters to make them positive, then dividing all those by the one with the highest value across the ages to make the maximum 1.0.

5. Fits

5.1. Optimizing Shrinkage

Besides being an out-of-sample test, $\widehat{elpd}_{loo}$ also is a correction of the loglikelihood for in-sample bias. The AIC is also derived to be a correction to the loglikelihood for in-sample bias. The small sample AIC, denoted by AICc, is an improved version of the AIC. With k parameters and sample size

n, the AICc correction subtracted from the loglikelihood is $p = kn/(n-k-1)$. Essentially a small sample here is any that makes this adjustment noticeable.

Like AICc, $\widehat{elpd}_{loo}$ is lower than the loglikelihood by a quantity a bit larger than the number of parameters. But it can be calculated in settings like ours where the effective number of parameters is not so apparent. In an attempt to count parameters in nonlinear models, Ye [1998] defines the generalized degrees of freedom used by a model as the sum of the derivatives of the fitted values with respect to the actual values. This agrees with the number of parameters in linear models (sum of diagonal of the hat matrix), and is a measure of how much the data can pull the fitted values towards it.

The generalized degrees of freedom itself is as quantitatively extensive to compute as the grind out loo elpd, which is prohibitive in many cases. Thus the R package `loo` appears to be the most effective way to penalize the loglikelihood for in-sample bias. The package also gives the elpd penalty p, denoted as $p-loo$. From this it is possible to back out how many parameters it would take for AICc to give the same penalty. This is $k = p(n-1)/(p+n)$. We will use this as an estimate of the parameter count, to get a sense of how much reduction from the original 411 parameters the shrinkage has produced.

Parameters can be fit for any selected value of the double exponential parameter b. Lower values of b restrict changes in slope of the final parameters more. A higher b can allow the model to fit more closely to the data. This can reduce the standard deviation σ of the normal distribution of the observations around their means, and can increase the loglikelihood. However it does not necessarily increase elpd. We will use the response of $\widehat{elpd}_{loo}$ to b to select b.

What we find is that higher values of b tend to give more parameters and

Multiple	σ	b	count	elpd
0.25	0.0196	0.005	175.1	6580.5
2.0	0.0160	0.032	282.0	6885.5
2.5	0.0158	0.040	285.2	6892.1
3.5	0.0159	0.056	290.1	6885.6
5.0	0.0155	0.078	304.1	6885.2
Cauchy	0.0161		239.4	6873.2

Table 1: The $\widehat{elpd}_{loo}$ measure and $p - loo$ count by Laplace b, itself modeled as selected multiple of residual standard deviation σ

a better in-sample fit. Up to a point this also increases $\widehat{elpd}_{loo}$. Eventually however, increasing b no longer improves $\widehat{elpd}_{loo}$. This is all measured with a bit of noise. The standard deviation of $\widehat{elpd}_{loo}$ in all these fits is around 45, and for $p - loo$ it is about 10.

In the fitting, b was not set explicitly. The model specifies b as an externally defined multiple of the residual standard deviation σ. When the multiple is made larger, σ decreases a bit but b still increases. Higher or lower values of b are obtained by making the multiple higher or lower. This has been recommended by some authors as a way to allow the model to more quickly get away from poor fitting local maxima of the posterior density, which would have higher σ. Table 1 shows various selected values of the multiple and the resulting values of σ, b, the parameter count calculated from $p - loo$ and $\widehat{elpd}_{loo}$.

Starting at about 2, this multiple gives models within a fairly narrow range of the maximum $\widehat{elpd}_{loo}$, even though higher values give more parameters and bit better σ. The parameter counts in this range are around 285, which is about a 30% reduction from the original 411. A much smaller multiple of 0.25 is shown for comparison. We also tried a Cauchy prior for reference, discussed in the Appendix.

For the range of multiples we used, we did not see clear over-fitting, where $\widehat{elpd}_{loo}$ significantly decreases with more parameters. This may not even be possible within this model, with 411 parameters compared to 2600 observations. On the basis of parsimony, as discussed above, we selected the multiple of 2.5, and those are the parameters discussed below. However there is little difference among the parameter sets in this range. The lower multiples tend to have just a bit smoother looking parameter graphs.

5.2. Parameters

Figure 2 graphs the estimated parameters with the exception of the base mortality, which in this log model looks very close to a straight line. The parameters interact in complex ways, so it is a bit risky to comment on them separately, but a few patterns seem to emerge.

For the cohorts, remember that these were forced to have no overall trend, so the trend would all be in the trend parameters. The pattern showing up is lowest mortality for the earliest and most recent births, with highest mortality rates for those born $1910 - 1935$ and again for those from the 1950s. There are some demographic changes that could explain much of this.

People born in 1900 or before did not get into this data unless they lived into their 70s. Thus the cohort parameters for them do not reflect the entire cohort, but just the quite select group who lived well past the life expectancy for that period. It is a reasonable possibility that this subgroup had lower mortality than did later-born populations at the same ages just because of this selection effect.

At the other end, those born in 1970 or later only show up here at ages

30 – 44. Thus the mortality rates are not necessarily representative of the entire cohort, and could well be interacting with other trends, so the cohort parameters in themselves might not be meaningful.

The dip for those born in the 1940s corresponds with the demographic research of Carlson [2008]. He is looking at a bit broader cohort – namely 1929-45 – which substantially overlaps. He shows that the smaller size of this generation, compared to those before and after, created unique economic opportunities, especially for males. This generation had the lowest unemployment rates and highest lifetime earnings, inflation adjusted, of any in US history, as well as lower mortality. They quickly ascended to management jobs managing the larger boomer generation that followed – and which they partially blocked from similar success. Other demographic trends like smoking rates probably interacted to make the lowest mortality group slightly different from the generation overall.

Looking at trends and trend weights, the main trend is pretty constant downward, but leveling off in the last few years. The ages most affected are 65 – 89, with a peak at 75. The HIV trend, from 1986 to 1996, most strongly affects men in their 30s, with some impacts at all ages. It may be due to a wider range of influences.

What we are calling the 30+ trend is generally upward from 2000 on, and has a similar age spectrum as the HIV trend. It is not clear what this trend is due to. It shows up as a leveling of mortality, except as an increase in the last 3 – 4 years for some ages.

The trend for the 80+ age group is more of a fine tuning. It was modeled with all ages in this range getting the same weight. Its small size means that the cohort impact of the earliest years largely accounts for the apparent slower declining mortality rates of the 80+ age group before 2000.

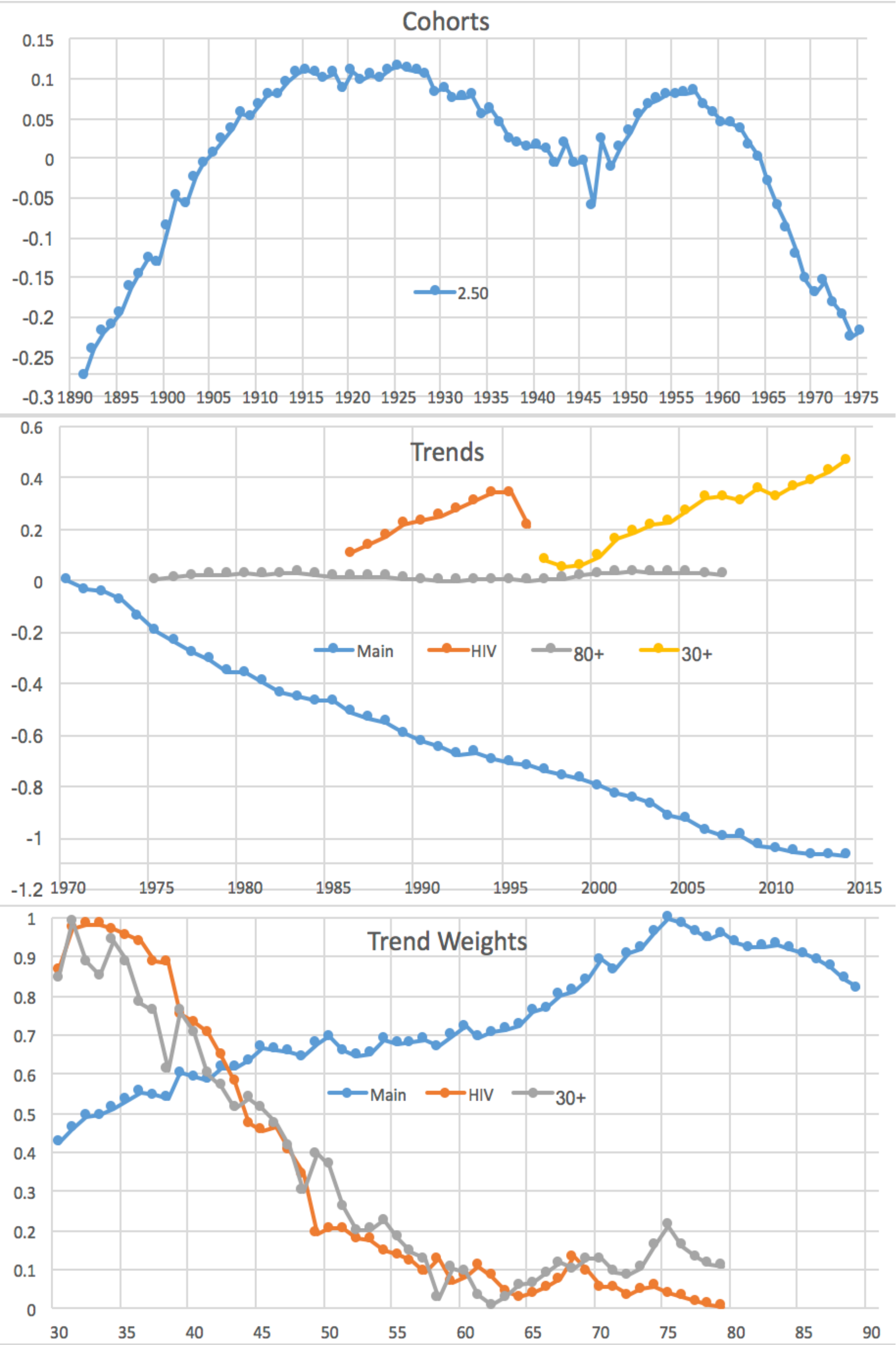

Figure 2: Final Level Parameters

Further parameter reduction may be possible in a few directions. The 80+ trend probably is not contributing much to the fit and could possibly be eliminated. However it is probably using only a very few effective parameters as well. Similarly the base mortality may be just as good with a parameterized curve, but again is probably not using very many effective parameters. Another possibility would be to have the HIV and 30+ trends use the same age weights, since those are fairly similar. That could save parameters. However a somewhat different pattern arises here with the Cauchy prior, discussed in Appendix A.

Future projections from this model would have to have a wide range of uncertainty. Two of the four trends continue to 2014, and these are what would need to be projected. The trouble is that the main trend has stopped, and it is hard to know how temporary that may be. The 30+ trend continues steadily upward, but it is not clear why or whether or not that may continue.

5.3. Goodness of Fit

Figure 3 graphs the age-group trends for the data, best fitting parameters, and the parameters from the multiple of 0.25. Both models show reasonably good fits by this measure.

The 0.25 and best models have residual standard deviations of 1.96% and 1.58%, respectively. Since these are on the log scale they are relative deviations from the means unlogged. Three times these give 99.73% probability of a point being within 5.9% or 4.7% of its mean, respectively.

The out-of-sample test shows the 0.25 model providing a worse fit as well. With 175 parameter equivalents compared to 285 for the best fit, it is considerably more parsimonious. Its trends are smoother and give an adequate intuitive feel for what the model trends are. The worst fit for both

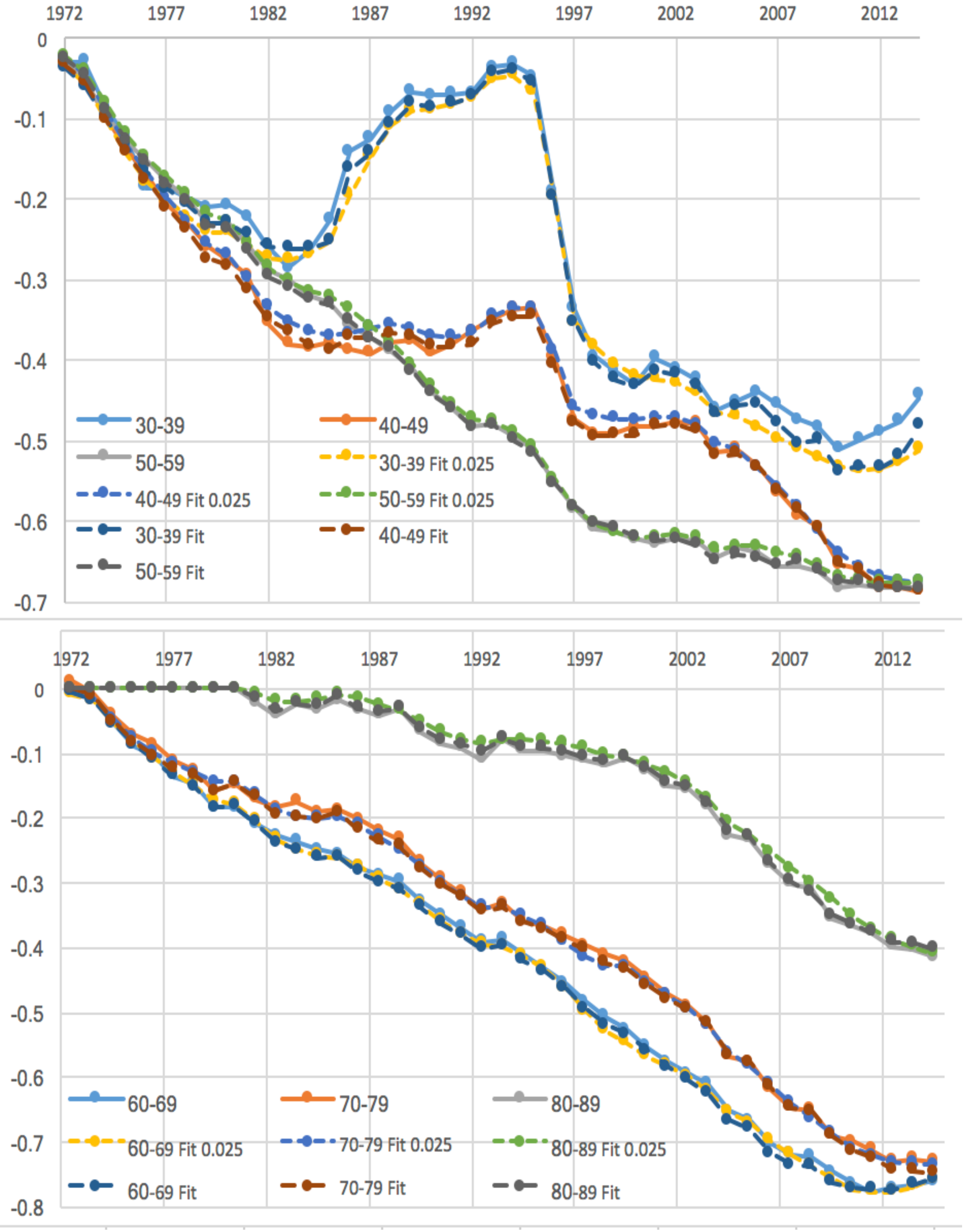

Figure 3: Cumulative Trend by Age Group. Top: 30s, 40s, 50s. Bottom: 60s, 70s, 80s. Solid: Actual; Dash: Best Fit; Dots: 0.25 Fit.

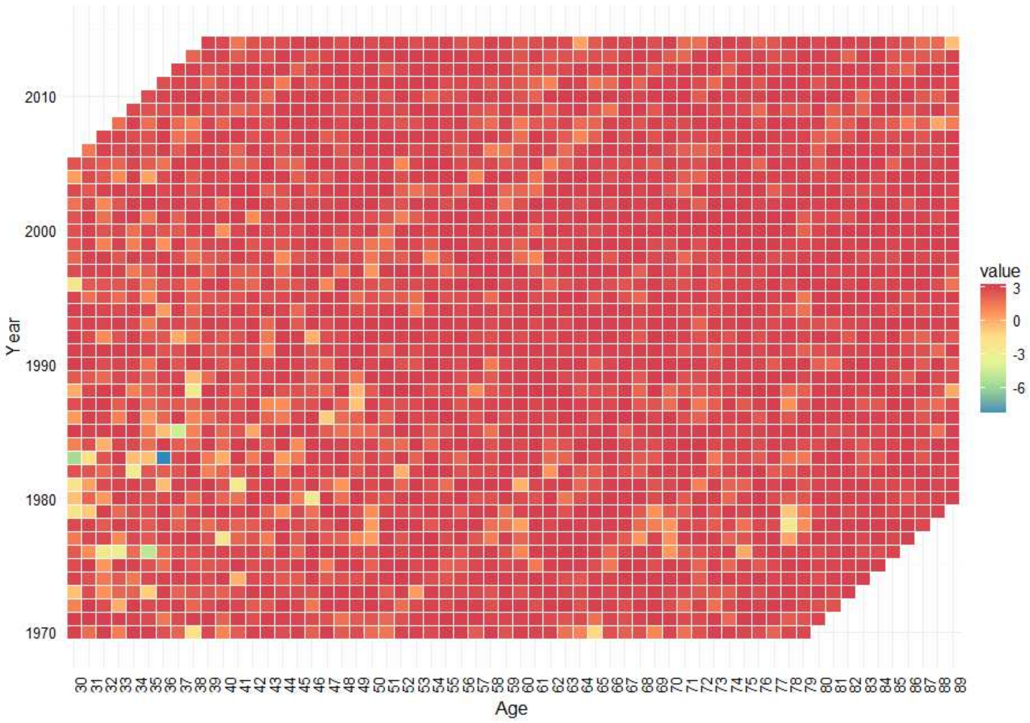

Figure 4: Loglikelihood by Data Point.

models is for ages in the 30s for the last ten years.

Figure 4 shows the loglikelihood at each data point. Here the rows are the periods of observation and the columns are the ages. The darker red points are the better fits, i.e., with the data closer to the fitted means. The poorer fitting points here are all lighter red or bluer colors regardless of the sign of the residual.

The worse fits tend to be at the younger ages and the earlier periods. This could be due to the distributional issues noted earlier – those ages have higher volatility.

6. Conclusions

Age-period-cohort modeling, either in mortality or general insurance loss reserving, appears to be an area where regularization can provide better fits than MLE, if measured by penalizing the log-likelihood for over-fitting by the use of cross-validation. Also, shrinkage facilitates simultaneous estimation of several trends in the Hunt and Blake model, which looks to be able to model complex trend patterns quite well.

For our model, Laplace b parameters greater than 0.04, and so by extension straight MLE, produce higher loglikelihoods, but were not better after penalizing for overfitting. We used linear splines for parameter shrinkage, but cubic splines could be formulated in MCMC and could be compared to linear splines by the $\widehat{elpd}_{loo}$ measure.

Although this model has four trends, the age 80+ trend appears to be handled well enough by cohorts, and the HIV and 30+ trends are non-overlapping in time and have somewhat similar weights by ages so could possibly be combined. Thus two trends, one fairly complex, could give a reasonable fit.

Parameter constraints are critical for getting convergence to a single model. The constraints discussed here, along with making the shrinkage standard deviation b a multiple of the residual standard deviation σ, appear adequate and allow for a clear interpretation of the parameters. The main estimation risk is that there may be local maxima for the posterior that are not good fits. These can usually be avoided by the choice of more specific priors.

The Cauchy prior seems worthy of future investigation.

References

Antonio, K., Bardoutsos, A., Ouburg, W., 2015. Bayesian Poisson log-biliner models for mortality projections with multiple populations. European Actuarial Journal 5, 245–281.

Barnett, G., Zehnwirth, B., 2000. Best estimates for reserves. PCAS 87, 245–303.

Blei, D. M., 2015. Regularized regression. Technometrics http://www.cs.columbia.edu/∼blei/fogm/2015F/notes/regularized-regression.pdf.

Cairns, A., Blake, D., Dowd, K., Coughlan, G., Epstein, D., Ong, A., Balevich, I., 2009. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. North American Actuarial Journal 13, 1–35.

Calderhead, B., Radde, N., 2014. Hamiltonian monte carlo methods for efficient parameter estimation in steady state dynamical systems. BMC Bioinformatics 15:253, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471–2105–15–253.

Carlson, E., 2008. The lucky few. Springer Netherlands.

Casella, G., George, E. I., 1992. Explaining the Gibbs sampler. The American Statistician 46:3, 167–174.

Chung Fung, M., Peters, G., Shevchenko, P., 2016. A unified approach to mortality modelling using state-space framework: characterisation, identification, estimation and forecasting. Working paper.

Efron, B., Morris, C., 1975. Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association 70:350, June, 311–319.

Fienberg, S. E., Mason, W. M., 1978. Identification and estimation of age-period-cohort models in the analysis of discrete archival data. Sociological Methodology 10, 1–67.

Frost, W. H., 1939. The age selection of mortality from tuberculosis in successive decades. American Journal of Hygiene 30:3A, 91–96.

Gelfand, A. E., 1996. Model determination using sampling-based methods. Markov Chain Monte Carlo in Practice, ed. W. R. Gilks, S. Richardson, D. J. Spiegelhalter London: Chapman and Hall, 145–162.

Gelfand, A. E., Dey, D. K., Chang, H., 1992. Model determination using predictive distributions with implementation via sampling-based methods. Technical Report No. 462 for the Office of Naval Research Department of Statistics – Stanford University.

Greenberg, B. G., Wright, J. J., Sheps, C. G., 1950. A technique for analyzing some factors affecting the incidence of syphilis. Journal of the American Statistical Association 45:251, 373–399.

Hachemeister, C. A., Stanard, J. N., 1975. IBNR claims count estimation with static lag functions. ASTIN Colloquium, Portimão, Portugal.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical learning with sparsity. CRC Press.

Hoerl, A., Kennard, R., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Hoffman, M. D., Gelman, A., 2014. Hamiltonian monte carlo methods for efficient parameter estimation in steady state dynamical systems. Journal of Machine Learning Research 15, 1351–1381.

Hunt, A., Blake, D., 2014. A general procedure for constructing mortality models. North American Actuarial Journal 18 (1), 116–138.

Keisler, 2000. Elementary calculus: An infinitesimal approach. http://www.math.wisc.edu/ keisler/calc.html.

Klugman, S. A., Panjer, H. H., Willmot, G. E., 2008. Loss models: From data to decisions. Wiley 3rd Edition, 669.

Lee, R. D., Carter, L. R., 1992. Modeling and forecasting U.S. mortality. Journal of the American Statistical Associationl 87, 659–675.

McDonald, J., 1984. Some generalized functions for the size distribution of income. Econometrica 52, 647–663.

Ntzoufras, I., 2010. Lesson 1 An introduction to MCMC sampling methods http://www.statistics.com/papers/LESSON1_Notes_MCMC.pdf.

Renshaw, A. E., Haberman, S., 2006. A cohort-based extension to the Lee-Carter model for mortality reduction factors. Insurance: Mathematics and Economics 38, 556–570.

Ryder, N. B., 1965. The cohort as a concept in the study of social change. Amecan sociological review, 843–861.

Taylor, G., 1977. Separation of inflation and other effects from the distribution of non-life insurance claims delays. Astin Bulletin 9, 217–230.

van Ravenzwaaij, D., Cassey, P., Brown, S., 2016. A simple introduction to Markov Chain Monte Carlo sampling. Psychonomic Bulletin & Review https://link.springer.com/article/10.3758%2Fs13423-016-1015-8.

Vehtari, A., Gelman, A., Gabry, J., 2016. Practical bayesian model evaluation using leave-one-out cross-validation and waic. arXiv preprint http://arxiv.org/abs/1507.04544.

Venter, G. G., 1983. Transformed beta and gamma distributions and aggregate losses. Proceedings of the Casualty Actuarial Society LXX, 156–193.

Verbeek, H. G., 1972. An approach to the analysis of claims experience in excess of loss reinsurance. Astin Bulletin 6, 195–202.

Wolfram, 2016. Gamma function. http://mathworld.wolfram.com/GammaFunction.html.

Xu, Y., Sherris, M., Ziveyi, J., 2015. The application of affine processes in multi-cohort mortality model. University of New South Wales Business School 2015ACTL13.

Ye, J., 1998. On measuring and correcting the effects of data mining and model selection. Journal of the American Statistical Association 93, 120–131.

Appendix A. Cauchy Prior

Appendix A.1. Statistical Properties

The Cauchy distribution centered at zero is a t-distribution with 1 degree of freedom. With parameter σ, its density and distribution function are:

$$\begin{aligned} f(x) &= \frac{1}{\pi}\frac{\sigma}{x^2+\sigma^2} && \text{(A.1)} \\ F(x) &= \frac{1}{2}+\frac{1}{\pi}arctan\left(\frac{x}{\sigma}\right) && \text{(A.2)} \end{aligned}$$

Its 75^{th} percentile is σ and its 25^{th} percentile is $-\sigma$. These could be used to estimate σ by matching percentiles, for instance.

If X is t-distributed, the distribution of $|X|$ is the folded t, whose density is twice the positive part of the t density. This is a power-transformed beta distribution. E.g., see McDonald [1984], Venter [1983] or Klugman et al. [2008]. The transformed beta density is:

$$f(x;\alpha,\beta,\tau,\theta) = \frac{\tau(x/\theta)^{\beta\tau}}{x(1+(x/\theta)^\tau)^{\alpha+\beta}}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{A.3}$$

with the k^{th} moment for $-\beta\tau < k < \alpha\tau$ being $\theta^k\frac{\Gamma(\beta+k/\tau)\Gamma(\alpha-k/\tau)}{\Gamma(\alpha)\Gamma(\beta)}$.

The folded t with ν degrees of freedom is the special case where $\alpha = \nu/2, \beta = \frac{1}{2}, \tau = 2, \theta = \sigma\sqrt{n}$. This does not require ν to be an integer. Having a t-distribution with non-integer degrees of freedom can facilitate estimation by MLE and MCMC. Its distribution function is an incomplete beta. With a half of a degree of freedom it becomes a heavier tailed version of the Cauchy. For $\nu = 2.1$ the variance is $21\sigma^2$, so finite.

With one degree of freedom, its k^{th} moment exists only for $-1 < k < 1$, and since $\Gamma(\frac{1}{2}) = \pi$, is given by:

$$E\left(X^k\right) = \frac{\sigma^k}{\pi}\Gamma\left(\frac{1+k}{2}\right)\Gamma\left(\frac{1-k}{2}\right) \tag{A.4}$$

Wolfram [2016] has the identity:

$$sin(\pi z) = \frac{\pi}{\Gamma(z)\Gamma(1-z)} \tag{A.5}$$

which with $z = (1+k)/2$ gives:

$$E\left(X^k\right) = \frac{\sigma^k}{sin\left(\frac{\pi}{2}(k+1)\right)} \tag{A.6}$$

The moments of the folded t are the moments for the absolute value of a Cauchy variable. Thus for a Cauchy with $|k| < 1$,

$$E\left(|X|^k\right) = \frac{\sigma^k}{sin\left(\frac{\pi}{2}(k+1)\right)} \tag{A.7}$$

For $k = \frac{1}{2}$, this gives $\sigma = \frac{1}{2}E(\sqrt{|X|})^2$, which can be used to estimate σ.

Appendix A.2. Parameters from Cauchy Prior

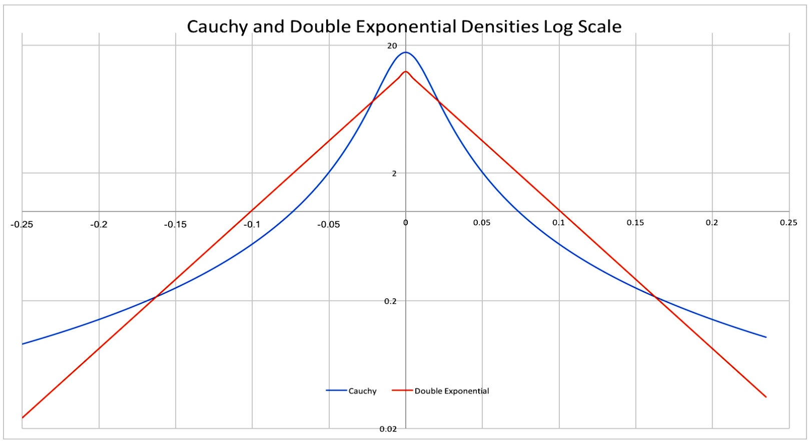

Figure A.5: Cauchy and Double Exponential Densities – Log Scale

In comparison, the double exponential distribution in σ has $\sigma = \frac{4}{\pi}E(\sqrt{|X|})^2$. Its 75^{th} percentile is $\sigma log(2)$. Thus its σ parameter is larger than the Cauchy's by a factor of $8/\pi = 2.55$ for the moment match, and by a factor of $1/\log(2) = 1.44$ for the percentile match.

The double exponential σ, called b in the text, was 0.04. In the end we looked at a Cauchy sigma around 0.018, which is more like matching the moment. This was actually done by setting σ to 1.1 times the residual standard deviation, which ended up at 0.0161 with this Cauchy prior.

This Cauchy has more weight near 0, but also more in the tails: there is 53% probability it is less than 0.02 in absolute value, compared to 39% for the double exponential. On the other hand, there is a 5.7% probability it is outside of [-0.2, 0.2], compared to 1.3% for the double exponential.

This can give it more parameter shrinkage than the double exponential in many cases, but also allows less shrinkage when needed.

The resulting parameters, graphed in Figure A.4, are usually a bit smoother than those from the double exponential fit, but the weights on the trends, graphed in Figure A.4, are an exception. The age weights for the main trend and the 30+ trend are much smoother than before, but the HIV age weights are more jagged. This combination may be why the Cauchy prior gave almost as good a fit as the double exponential with fewer effective parameters. The HIV and 30+ weights are fairly different from each other here, which may be a reason to keep them separate.

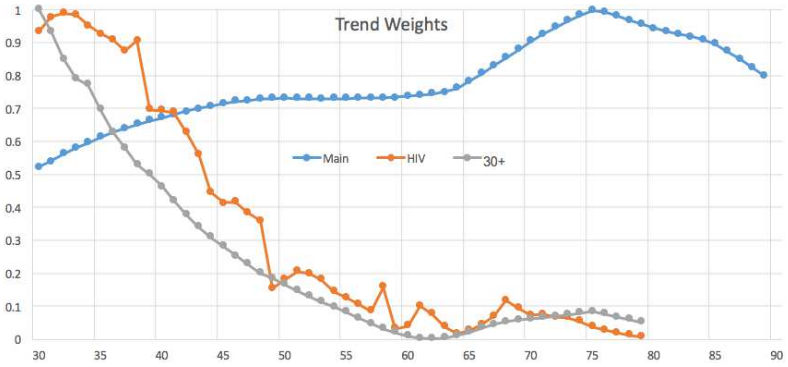

Figure A.6: Final Level Parameters Cauchy

We did not optimize the Cauchy fit, so another σ may give as good a fit as the double exponential, with fewer parameters. A major drawback to the Cauchy, however, is that computer run times for it are considerably longer – like by a factor of 100. This is in part because it requires smaller steps in the Stan fitting, according to error messages, and that can make the runs much longer. For a model this complex with a fairly large sample size, that puts it almost out of the range of feasible computation on a personal computer.

Perhaps model searching can be done with the double exponential prior, with final fits using the Cauchy. Matching the absolute half moment of the double exponential seems like a good starting point for the Cauchy.