Parameter Shrinkage for Joint Age-Period-Cohort Modeling of Related Datasets

Gary Venter
Columbia University Actuarial Science
gv2112@columbia.edu

Abstract: Actuaries use age-period-cohort (APC) models for mortality modeling and general insurance loss reserving. Several recent papers have addressed simultaneously modeling related datasets, such as loss triangles for subsets of a class of business or mortality data across regions. This paper does joint modeling by shrinking the differences among the same parameters for different datasets. This could loosely be described as credibility weighting for triangles, but it comes more directly from statistical approaches such as ridge regression and lasso. Like credibility, these seek to reduce estimation and prediction error by various forms of shrinkage.

The models discussed here already incorporate parameter reduction by smoothing linear spline slope changes. This is extended to also shrink the differences between the same slope changes for different datasets. Doing so can reduce prediction error, measured using penalized log-likelihood, by increasing model parsimony. Bayesian Markov Chain Monte Carlo (MCMC) estimation is used in an example to illustrate the method. A related classical approach based on random effects is introduced as an alternative. The example is a joint model of historical female mortality data for Spain and Japan – two of the world's longest lived populations.

Keywords: Joint estimation, Parameter shrinkage, MCMC, Age-period-cohort models

1 Background

APC models originated in epidemiology, notably in Greenberg, Wright, and Sheps (1950), but have been used extensively by actuaries, although not always called that. They are models with parameters in three directions of time. Cohort refers to year of origin, such as accident year in reserve data or year of birth in mortality data. Age is time since origin, such as payment lag or age at death. Period refers to the time at the observation, e.g., calendar year or year of death.

Sometimes dummy variables are introduced in APC model to use regression or GLM software. The parameters are multiplied by these dummy variables, particularly when used in models of the logs of the data. But this is not necessary. A cell mean could just as well be estimated directly as the product of a row parameter p_i with a column parameter q_j. Thus in some sense, APC models can be considered to be models with parameters but without any explanatory variables.

Reserving actuaries have been using age-cohort (AC) models since at least the 1930s. Verbeek (1972) introduced an age-period (AP) model for claim counts, and Taylor (1977) extended this to aggregate losses. The AP model of Lee and Carter (1992) has been widely applied to mortality data. (The Lee-Carter model is bilinear – that is there is an age effect, a period effect, and a multiplier on the period effect that varies by age. Thus there are two sets of age parameters. Bilinear models with age, period and cohort effects are considered to be APC models in this paper.) Full APC models were used informally by reserving actuaries for some time before Barnett and Zehnwirth (2000) finally published them. Renshaw and Haberman (2006) did this for mortality models. For mortality, cohorts were the later addition, but periods came later for reserving.

With a parameter for every row and column and often for every diagonal as well, APC models can easily become over-parameterized. Actuaries have developed methods to address this, including fitting parameterized curves across the parameters and smoothing by cubic splines. Linear splines are just line segments representing several consecutive parameters. These fits can be smoother or more jagged depending on how often the slopes change and by how much.

Statisticians have developed a variety of parameter shrinkage methods, broadly defined as "regularization." One direction, which is basically actuarial credibility theory with a Gaussian assumption, is that of Stein (1956). His finding that shrinkage always reduces predictive variance was unprecedented enough to get the name "Stein's Paradox." Other methods include ridge regression, lasso, random effects, and Bayesian shrinkage priors. The latter are fairly narrow mean-zero priors (sometimes with comparatively heavy tails) that favor small values for the parameters. That is the method used here, both for the APC parameters themselves and the differences between parameters for different datasets. The APC parameters are assumed to be on piece-wise linear curves (linear splines), and it is the slope changes that are given shrinkage priors.

Similar models have been used by actuaries. Barnett and Zehnwirth (2000) used shrinkage to space out or reduce piece-wise linear slope changes in APC reserve modeling. Gluck (1997) did something similar in the Cape Cod AC model. Venter and Şahin (2017) used Bayesian shrinkage of slope changes for an APC mortality model. G. Gao and Meng (2017) use a similar approach, but with cubic instead of linear splines, applied to a loss reserve triangle.

A number of papers have applied correlation methods to reserve modeling for related datasets, for example de Jong (2012) and Shi (2013). Shi, Basu, and Meyers (2012) do this in a Bayesian setting. Shi and Hartman (2014) use Bayesian parameter shrinkage, but on development factors, not in APC models. They do not use MCMC and do not have a methodology for penalized likelihood.

The appendix of Venter, Gutkovich, and Gao (2015) explores related triangle APC modeling with linear spline slope change shrinkage through classical random effects. They show, however, that the standard implementation of random effects contains a large number of hidden parameters and so is much less parsimonious than is commonly thought. Classical shrinkage models do not have an available penalized likelihood measure, because shrunk parameters are constrained and do not use a degree of freedom for every parameter, so penalization based on counting parameters, such as AIC and BIC, is misleading.

Li and Lee (2005) introduce the concept of using a joint stochastic period-trend process for mortality modeling of joint populations, with mean reverting variations from the joint process for each population. A similar approach by Jarner and Kryger (2009) models a large and a small population with the small population having a multi-factor mean-zero mean reverting spread in log mortality rates. Dowd et al. (2011) model two populations that can be of comparable or different sizes with mean reverting stochastic spreads for both period and cohort trends. A period trend spread in an APC model is estimated by Cairns et al. (2011), who use MCMC estimation. They assume fairly wide priors, not shrinkage priors. This uses the capability of MCMC to estimate difficult models. They could get a penalized likelihood, but the current method for this by Vehtari, Gelman, and Gabry (2016) was not yet available. Antonio, Bardoutsos, and Ouburg (2015) estimate the model of Li and Lee (2005) by MCMC.

2 Methodology

2a Linear APC Model

The notation is slightly easier if the data is assumed to be arranged in a rectangle with a row for each cohort and a column for each age. Assuming further that the ages are indexed to start at zero, the $[n, u]$ cell is then in the $n + u$ period. The linear form of the APC model is the simplest. Usually the modeling is done for the logs of the data, so the mean (or a parameter closely related to the mean, depending on the distributional assumptions) for the $[n, u]$ cell can be written as:

$$k + p[n] + q[u] + r[n + u]$$

Here $p[n]$ is the parameter for cohort n, $q[u]$ is the parameter for age n, $r[n + u]$ is the parameter for period $n + u$ and k is a constant term (which is often assumed to be absorbed by the other parameters, but is written separately here as it will have a different prior).

Linear spline smoothing posits slope change parameters a, b, and c for p, q and r, respectively. The slope at age u is then the cumulative sum of $b[0], ..., b[u]$, and $q[u]$ is the cumulative sum of the slopes from 0 to u. Thus the slope changes are in effect the second differences of the APC parameters p, q, r.

2b Bilinear APC Models

The bilinear models include additional parameters for interactions among the three directions. An age-period interaction is the most common assumption. The Lee-Carter plus cohorts model is the classical example, with mean of the logs of the mortality rates given by:

$$k + p[n] + q[u] + s[u]r[n + u]$$

The age modifier to trend $s[u]$ is intended to represent variation in period trend participation rates across the ages. Medical advances may affect mortality more for older ages, for example. The second differences in s will be denoted by d.

This version is less used in loss reserving, but has been sometimes applied to workers compensation incremental paid loss data. There there are two trend drivers – medical and wage inflation. When wage benefits are not indexed, the wage trend applies by cohort, i.e., by accident year, while medical inflation applies by period. Since wage benefits predominate in earlier claim ages, the period trend can then have less of an impact for those ages than it does later.

Age-cohort interactions are included in the model of Renshaw and Haberman (2006), but are not widely used. Some recent reserving models have found this useful, however, to capture a gradual acceleration in claim payments due to automation. In that case, the interaction is a cohort modifier to the age factor. See for example G. Meyers (2015), where the cohort modifier is a parameterized function of n.

The strongest assumption of the Lee-Carter model, and thus its weakest aspect, is that the period trend participation rate for a given age remains fixed over the periods. This can easily be violated. One example is modernizing societies in which youth mortality declines quickly initially, but then stabilizes while older ages get improvements. The model of Hunt and Blake (2014) addresses this by allowing multiple period trends, each with its own set of participation rates. It can be written as:

$$k + p[n] + q[u] + \sum_j s_j[u] r_j[n+u]$$

This model, with two trends j, is used in the example.

2c Constraints

Parameter constraints are needed for identifiablity and to reduce the chance of producing unreasonable parameter sets. The constraints from Venter and Şahin (2017) are used here:

- For ages, $a[0] = q[0] = 0$
- For periods, $c[1] = r[1] = 0$
- A trend line through the cohort parameters $p[n]$ has slope = intercept = 0
- The $s[u]$ trend weighting parameters are non-negative and achieve a maximum of 1.

Recall that ages are indexed to start at zero, but periods and cohorts start anywhere, here assumed to be at 1. Forcing the cohorts to have no overall trend may seem to arbitrarily rule out a decrease or increase across the cohorts over time, but such a trend could occur. However, APC parameters have no meaning in isolation. There is no true cohort trend, separately from the period trend. They are defined with reference to each other. Thus here the period trend is that trend given that there is no overall cohort trend, and the cohort effects are those effects under the assumption that all the overall trend is in the period parameters. That way at least they each have specific meaning.

The condition on the period trend participation by age factors s means that the period trend r is the trend for the age with the highest overall trend. This helps prevent overlap of the age parameters q and the s factors but also gives a more specific interpretation to the period

trend. It is the trend for the age with the highest trend given that there is no overall trend in the cohort parameters.

The easiest way to enforce the cohort condition is to get trial p parameters, then subtract a regression line through them to get the actual p parameters. For a regression of $y_1, ..., y_N$ on the integers 1,...,N, the slope and intercept are:

$$slope = \frac{6 \sum_j (2j - N - 1) y_j}{N^3 - N}$$

$$intercept = average(y) - \frac{(slope)(N+1)}{2}$$

2d Priors

The multi-dataset model is specified by providing priors for the parameters of a selected base dataset, and then setting the shrinkage priors for the parameters for the other datasets to have means equal to the corresponding base parameters. The same thing can be achieved by imposing zero-mean priors on the differences from the base.

MCMC estimation has methods to draw samples from a distribution that is proportional to the desired posterior distribution but with an unknown proportionality constant. Because of that, priors other than shrinkage priors can be specified as proportional to an unknown constant. For example, the constant k for each dataset is assumed to have prior 1.0 over the real numbers. Each dataset gets its own constant, which is not shrunk towards zero.

For the base dataset, all of the slope change parameters a, b, c, d are assumed to have a mean zero prior with a scale parameter *shrink* discussed below. For the other datasets, the slope change parameters are given prior means equal to the currently estimated base dataset slope change parameters, with each dataset given its own shrinkage scale parameter $shrink_j$. Priors are not specified for the p, q, r, s parameters, as these are computed from the a, b, c, d parameters.

If a slope change is zero, the slope continues as it was. Shrinking the slope changes towards zero thus tends to put the parameters on slowly changing line segments. However if a big change at some point is indicated by the data, that will overcome the push of the prior towards zero, and a larger slope change would appear. The parameters might thus appear to be on a smooth parameterized curve, but with more flexibility in movement than a single assumed curve shape would usually have.

Two datasets can end up with fairly different parameters with just a few differences in slope changes. Thus shrinking the other datasets' slope change parameters towards those of the base has the potential to produce different looking models without very many extra parameters for the other datasets. Again, when large differences in slope changes are needed, the estimation can overcome the prior's push towards zero. Otherwise small differences are favored by this choice of priors.

3 MCMC and Penalized Likelihood

MCMC incorporates a methodology to generate samples from a posterior distribution when only the likelihood and prior are known. Bayes Theorem gives:

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

The left side is the posterior distribution of the parameters given the data, and the numerator of the right side is the likelihood times the prior. The denominator is a constant for a given dataset, so maximizing the numerator maximizes the posterior. That is the key to the original Metropolis sampler. The idea of that is to have a proposal distribution to draw samples of the parameters. If a sample increases the numerator, keep it. If not there is a rejection rule to keep it or not, based on a [0,1] random draw. After a warmup period, the retained samples end up being representative of the posterior.

A refined version of that, the Hastings-Metropolis sampler, is more efficient. Further refinements include Hamiltonian mechanics and the no-u-turn sampler, which evolve the proposal distribution dynamically. This is the basis of the Stan MCMC package. The Gibbs sampler draws parameters one at a time from the posterior distribution of a single parameter given the data and the latest sample of all the other parameters. The JAGS package uses that.

A penalized likelihood measure for a parameter sample has been developed using the basic theory behind the AIC measure – to reduce the loglikelihood for the bias created by measuring it on the data set the model was fit to. Often the number of parameters is ambiguous because of shrinkage, which uses up fewer degrees of freedom than the parameter count, so a different method than AIC is needed.

For shrinkage models, including classical ones like lasso, a leave-one-out cross validation method is popular. That involves refitting the model many times, leaving out a different observation each time, measuring the likelihood of the omitted observation, and adding those up. This can be resource intensive, especially for the difficult estimation problems MCMC is particularly good at. To address this, Gelfand (1996) developed an approximation for a sample point's out-of-sample loglikelihood using a numerical integration technique called importance sampling. In his implementation, that probability is estimated as its weighted average over all the samples using weights proportional to the reciprocal of the point's likelihood under each sample. That gives greater weight to the samples that fit that point poorly, which would be more likely to occur if that point had been omitted. The estimate of the probability of the point comes out to be the reciprocal of the average over all the samples of the reciprocal of the point's probability in a sample.

That gave good but quite volatile estimates of the leave-one-out (loo) likelihood. Vehtari, Gelman, and Gabry (2016) solved that issue by something akin to extreme value theory – fitting a Pareto to the probability reciprocals and using the fitted Pareto instead of the actuals for the largest 20% of the sample. They call this "Pareto-smoothed importance sampling." It has been extensively tested and has become widely adopted. The penalized likelihood measure is labeled $\widehat{elpd}_{loo}$, standing for "expected log pointwise predictive density."

3a Selecting the Degree of Shrinkage

A starting point for selecting a mean-zero shrinkage prior is the Laplace distribution. This is basically an exponential distribution and its mirror image around zero. It is popular because it produces a Bayesian analogue to lasso shrinkage. (The normal prior gives an analogue to ridge regression.) It is controlled by a scale parameter, here called *shrink*. The smaller that parameter is, the stronger is the shrinkage towards zero, but the exponential tail will allow for an occasional large deviation from zero.

Selecting the *shrink* parameter thus requires a balancing of parsimony and goodness of fit. Taking the parameter that optimizes $\widehat{elpd}_{loo}$ is one way to proceed, and that was the approach taken in Venter and Şahin (2017). An alternative would be to give a sufficiently wide prior to *shrink* itself and let MCMC pick it. This is called a fully Bayesian method, whereas optimizing $\widehat{elpd}_{loo}$ is partially Bayesian and partially predictive – since loo is based on predictive accuracy. G. Gao and Meng (2017) is an actuarial paper using the fully Bayesian approach.

Expert opinion seems to be that for a properly specified model, the fully Bayesian and predictive approaches should get to about the same degree of shrinkage. The Bayesian one is a lot easier in that the predictive method involves rerunning the model numerous times. The Bayesian method is used here, but $\widehat{elpd}_{loo}$ is used to compare models.

The fully Bayesian approach can also be used to find a good form for the prior distribution. The Cauchy distribution is a heavier-tailed distribution growing in popularity as a shrinkage prior. A heavier-tailed distribution is also usually more concentrated near zero, except for the tails. As a prior, this can produce more shrinkage in general but also allow larger deviations from zero occasionally. It produces more parsimonious models that still have reasonably good fits, but which may or may not improve penalized likelihood.

The Cauchy is actually a t distribution with one degree of freedom. The normal is the limiting case of t distributions with ever larger degrees of freedom. Degrees of freedom ν does not have to be an integer, so a continuous prior on ν could be used to get a Bayesian solution to how heavy-tailed the shrinkage prior should be for a given data set. Even the Laplace distribution can be approximated by a t distribution. A t with 6 degrees of freedom matches a Laplace with scale parameter $\sqrt{3}/2$ in both variance and kurtosis (and so in fact matches all 5 moments that exist), so is a reasonable approximation. Figure 1 graphs their densities on a log scale.

A uniform prior for ν ranging from 0.4 to 15 seems to be a good starting point but oftentimes the range has to be tightened to get convergence. In the example, *shrink* was kept in an interval from 0.001 to 0.05, but for other models somewhat wider intervals have worked. It appears to be quite usual for ν to end up around 1, which gives a heavy-tailed prior. This tends to produce fairly smooth linear spline parameter curves, but with some jumps.

3b A Classical Alternative

The numerator of the posterior, that is $p(X|\theta)p(\theta)$, could be optimized numerically for θ and any parameters used in the prior. If the prior is normal and all but the constant term have

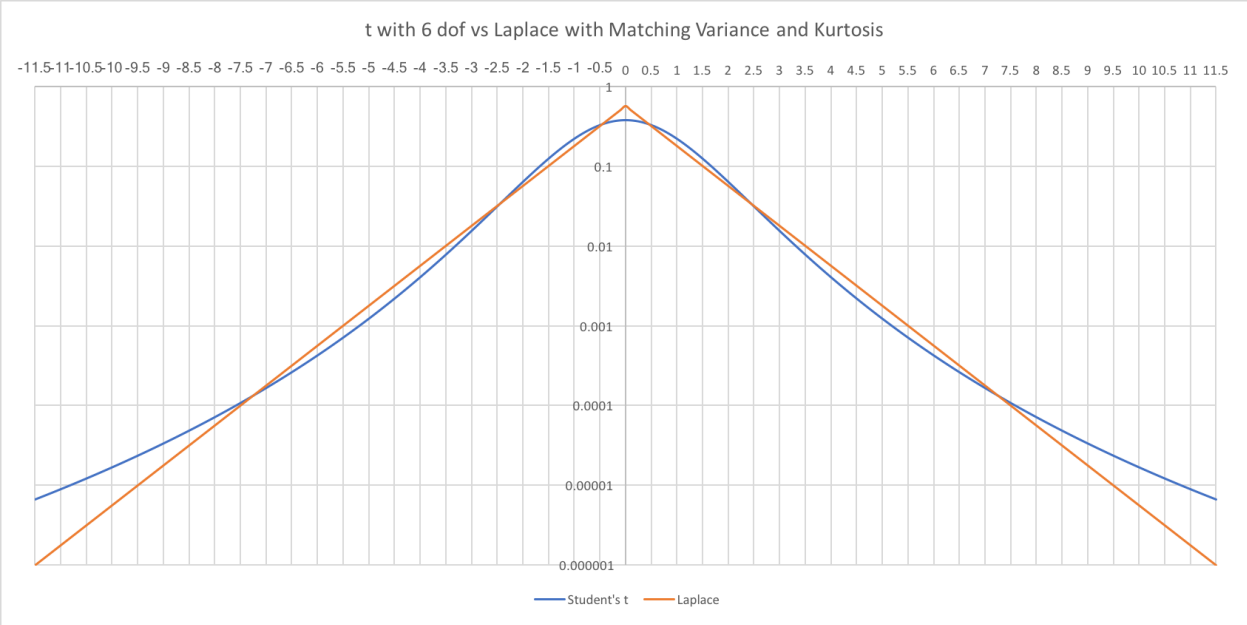

Figure 1: t with 6 Degrees of Freedom vs. Laplace

prior mean zero, that would be an implementation of classical random effects estimation.

The classical viewpoint may be difficult for a Bayesian to totally grasp, but it seems to go something like this: the θs to be shrunk towards zero are not parameters but predictions of the values of the effects they represent; each such effect has a known distributional form (typically normal), with mean zero and scale = a parameter to be estimated. Instead of a posterior distribution on the parameters, you get a distribution of the prediction errors. Regardless of the interpretation, a common way to implement random effects is to do a classical optimization of what Bayesians would think of as the product of the prior and the likelihood, maximizing which gives the posterior mode.

The default assumption in random effects is that each random effect has its own scale parameter *shrink*. But using the generalized degrees of freedom approach of Ye (1998), Venter, Gutkovich, and Gao (2015) found that having so many scale parameters can use up many degrees of freedom – that is, including them in the model makes the fitted values much more responsive to hypothetical small changes in the data points. Most random effects software allows users to specify having just one *shrink* for the whole model, which seems to give considerably more parsimonious models without sacrificing too much in goodness of fit.

Using a non-linear optimizer like Nelder-Mead is an alternative to random effects apps. Then there is no reason to stick to a normal prior. Assuming a t-distribution prior, with degrees of freedom and scale estimated as part of the optimization, would be parallel to the shrinkage model specified here. Thus classical optimization could theoretically get to the same estimates as the posterior mode of MCMC. Good starting parameters might be needed, however. It could be faster or slower than MCMC, and might be a good check. The classical parameters could also be used as starting values for MCMC (and vice versa, or iteratively) and then MCMC could be used to get the parameter distributions and the penalized likelihood.

4 Example Model

As an example of this methodology, Spanish and Japanese female mortality rates for 1961 – 2014, ages 40 – 100, are modeled. These are two of the longest-lived groups in the world, but have different histories and might not normally be modeled together. The example shows that joint modeling can improve model accuracy even in such cases. Cohorts were assigned to year of death minus age at death. Ages were truncated to age at last birthday, which makes the cohort the year of birth for those dying before their birthdays and the year after birth otherwise. Cohorts 1867 – 1968 are represented in the data.

Figure 2 graphs cumulative trends since 1947 for each population. The Japanese mortality rates have dropped faster, having started higher and ending a bit lower than the Spanish rates. In both countries, the cumulative trend gets steadily lower as the ages increase for most of the period. But for both, the younger ages included here show a flattening of trend in the more recent years.

Because the younger ages have the steepest trend in the beginning and the flatest at the end, a single period trend with constant age weights, as in the Lee-Carter model, seems unlikely to be adequate. Thus the model of Hunt and Blake (2014) was selected, with one trend for ages 40 – 62 and another for ages 63 – 100. The methodology as it stands requires all the modeled populations to use the same model structure, and 62 appeared to be a reasonable transition point for both.

A Poisson distribution for number of deaths was used. The models are then both:

$$deaths_{n,u} \sim Poisson\left[e^{k+p(n)+q(u)+s_1(u)r_1(n+u)+s_2(u)r_2(n+u)} exposures_{n,u}\right]$$

where $s_1(n) = 0$ for $n > 62$ and $s_2(n) = 0$ for $n < 63$.

Theoretically, deaths should be binomially distributed as the sum of Bernoulli processes, but with a small frequency the binomial is very close to the Poisson. For the oldest ages the frequency is not so small, but binomial did not provide a better fit and the Poisson converges better. The Poisson also might fit better if there is contagion in mortality, perhaps from weather, economic conditions or war. In such cases the negative binomial might fit even better, but here it did not.

This model is complex enough and the age range large enough that the constraints above were not enough to prevent nonsense parameter sets from arising. Additional constraints were imposed:

- The age parameters cannot decrease with older ages, so $q(u+1) \geq q(u)$
- The cumulative period trends have to be non-positive, so $r_j(t) \leq 0$

The two populations' parameters were first estimated individually. The Spanish data was selected as the base. Starting parameters for the combined case were the Spanish parameters and the difference between the Japanese and Spanish slope change parameters. Except for the constants k and the distribution parameters ν and $shrink$, shrinkage priors were used.

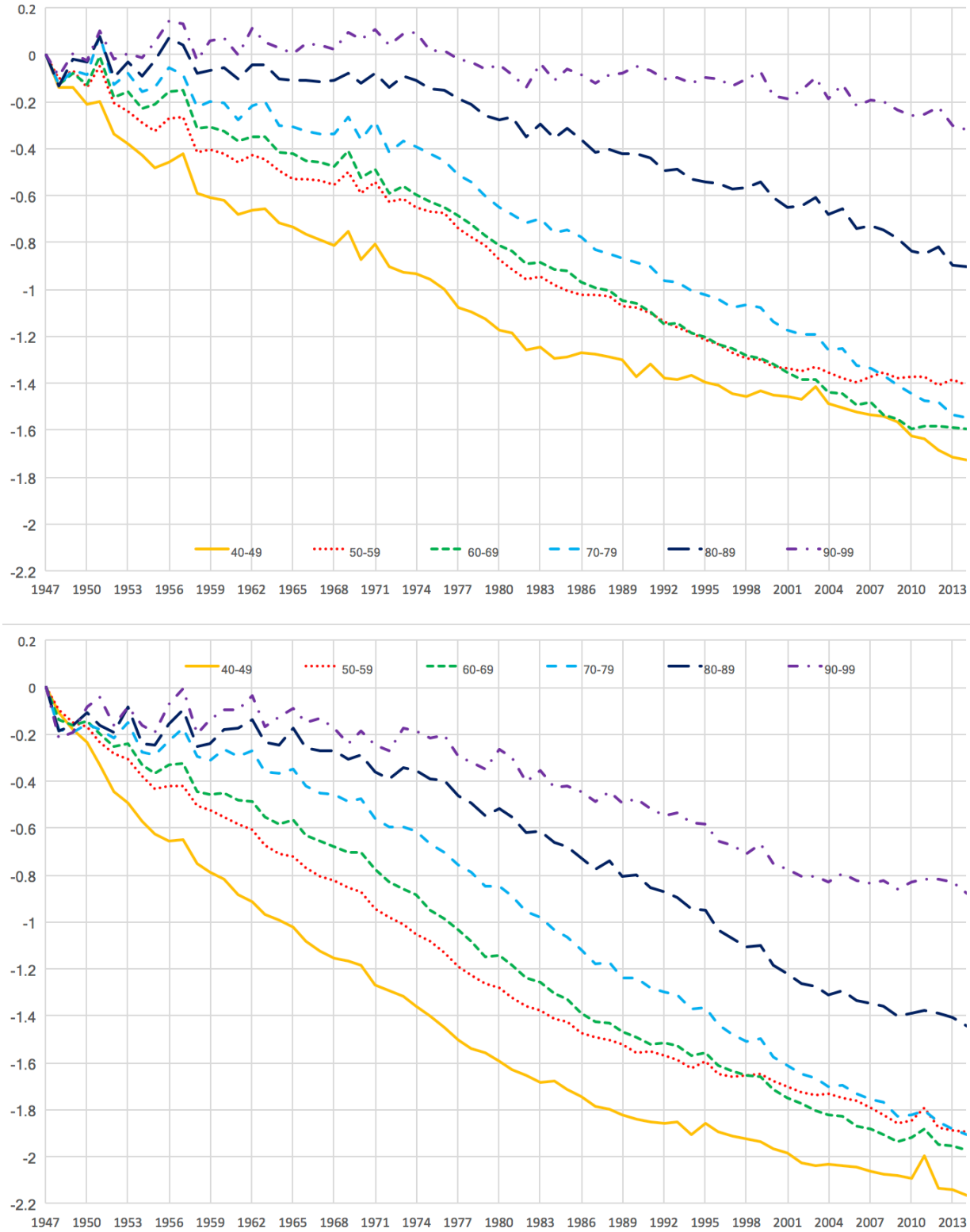

Figure 2: Female Mortality Cumulative Trend by Age Group 1947–2014 Spain (top) and Japan

Table 1: Negative Loglikelihoods Penalized and Unpenalized Individually and Combined

	Individual	Combined
NLL	37,805	37,826
$\widehat{elpd}_{loo}$	38,859	38,688

Table 2: Absolute Values of Differences Between Spanish and Japanese Parameters Individually and Combined

	Individual	Combined
a	3.39	2.30
b	6.47	5.78
c_1	22.72	20.35
c_2	6.92	5.41
d_1	16.08	15.40
d_2	7.79	7.26

The resulting negative log-likelihoods and loo-penalized NLLs are shown in Table 1 for the individual and combined estimations. The combined model was more parsimonious, and even though it had a higher NLL, it performed better on penalized log-likelihood.

The parameters of the combined model were only subtly different from the individual model's. Small changes in the slope-change parameters can produce more substantial changes in the fits, as the slope changes accumulate in the final parameters. Table 2 shows that the differences between the Spanish and Japanese slope changes were lower in the combined model for all the parameter sets. In this case, reducing the differences between the parameters produced a more parsimonious model, with an increase in the negative loglikelihood, but an improvement in the out-of-sample predictions of the omitted data points.

Actual vs. modeled expected numbers of deaths at the beginning, middle, and end of the period are shown for both populations in Figure 3. Some of the changes are due to population changes by age, but still the fits appear reasonable.

4.1 Estimated Parameters

The biggest differences between the two populations were in the period trends for the younger ages and their age weights. The c_1 and d_1 parameters are the slope changes for these. The cohort patterns, reflected in the a parameters, were similar in the two countries, as were the initial mortality patterns – the b's. The final model parameters are graphed for the two datasets in Figures 4 – 8.

Figure 4 shows that the starting mortality rates were higher in Japan, especially for ages above 80.

The period trend was much greater in Japan, especially for the first 20 – 25 years (Figure 5). For the last 20 years they have been pretty comparable, with Japan leveling off and even worsening slightly in the final 10 years. Even though the curves look fairly different, just a few slope change parameter differences were needed to get these results.

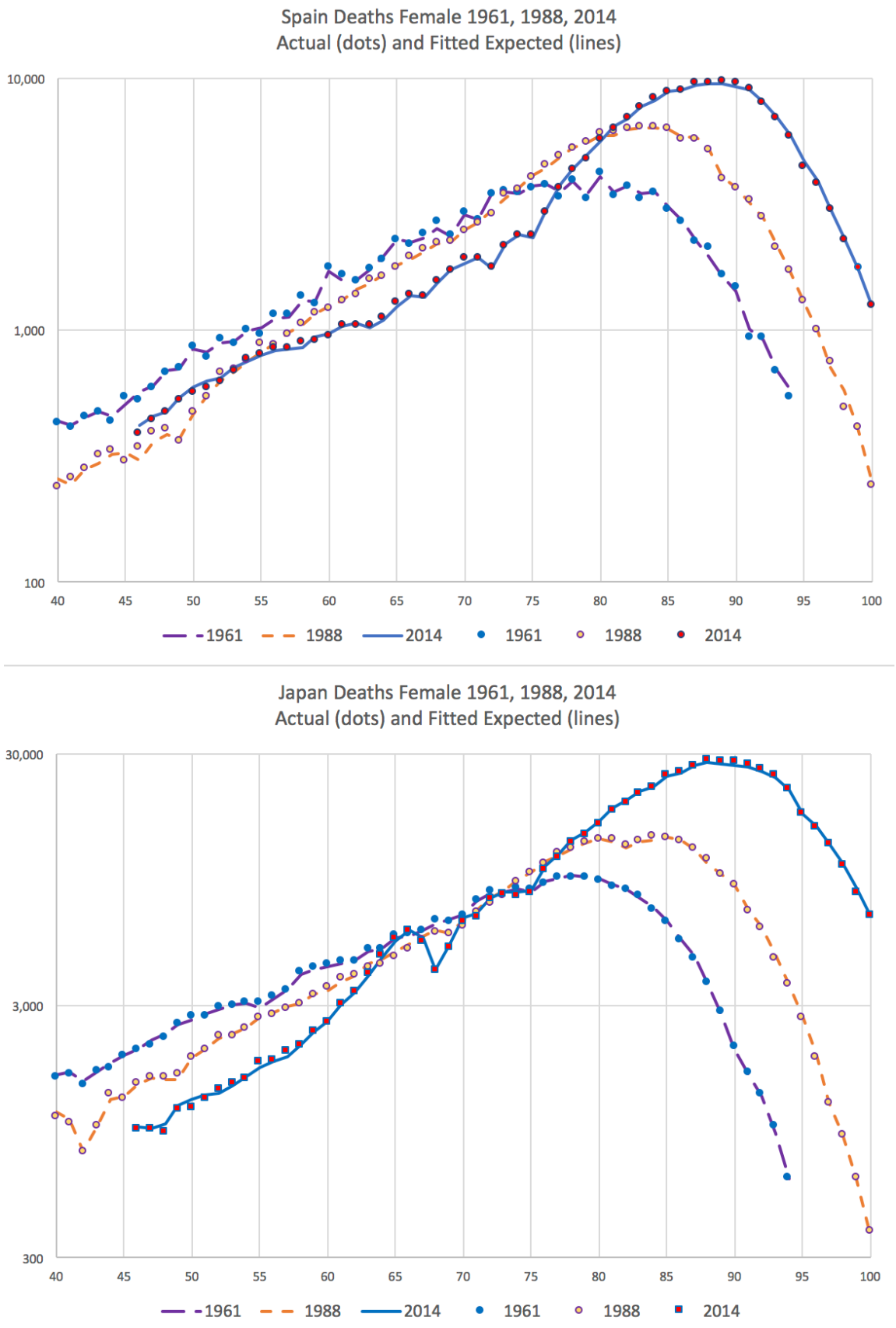

Figure 3: Actual and Fitted Deaths 1961, 1988 and 2014 Spain (top) and Japan log scale

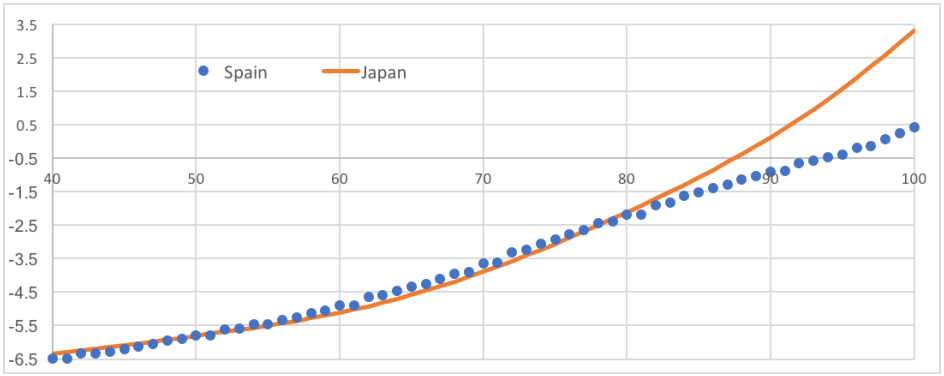

Figure 4: Initial Mortality by Age Including Constant, log scale

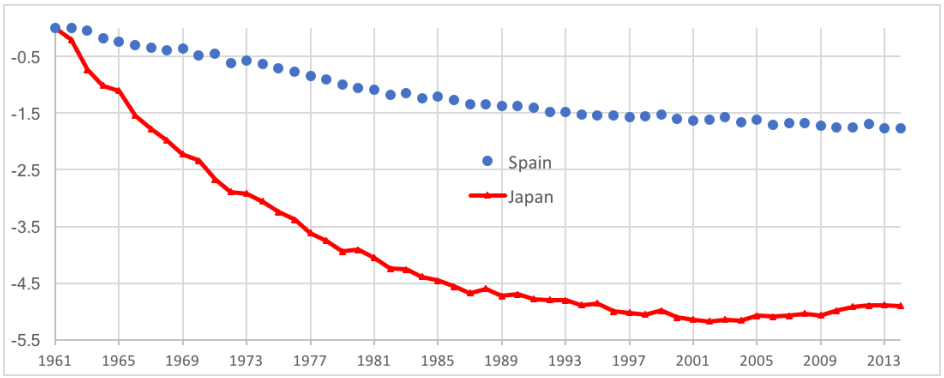

Figure 5: Period Levels by Year, Older Ages

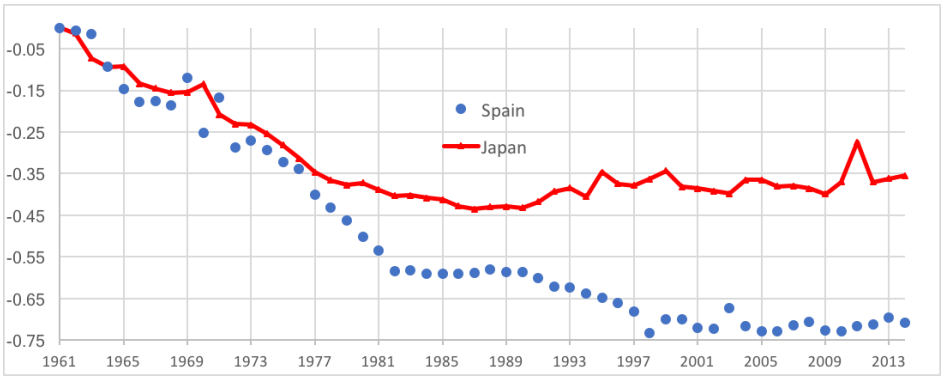

Figure 6: Period Levels by Year, Younger Ages

The trends for the younger ages seen in Figure 6 were more complex and differed more between the two populations. Japan shows a slight worsening in the more recent years for these ages too.

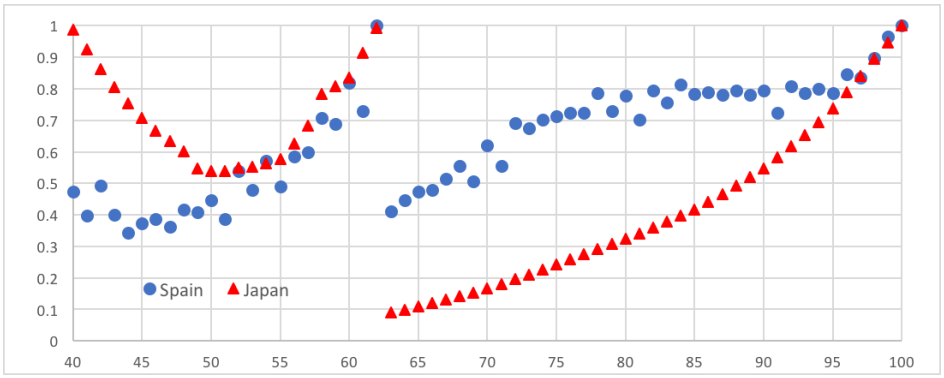

Figure 7: Age Weights for Period Trends

Figure 7 graphs the age weights s for both trends. The weights for the younger and older trends were fairly different, both across ages and countries. The high weights at age 62 were for generally lower trends than were the low weights at age 63, so there was not much discontinuity there.

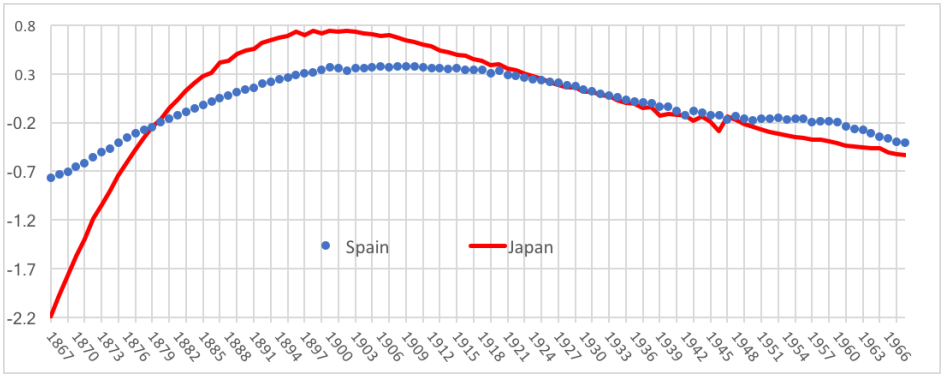

Figure 8: Cohort Adjustments

For both groups, the age weights were high at the oldest ages, which was unexpected as Figure 2 shows that the period trends were lowest for these ages, especially at the beginning of the data. The difference is accounted for by the cohort parameters in Figure 8, which show lower mortality for the earliest cohorts, with the highest mortality in cohorts from around the start of the 20th century.

The earliest cohorts were in the data only for the oldest ages, and in many countries it has been noted that these cohorts were particularly hardy once they had attained older ages. That does not mean they were longer-lived overall, however, just conditionally so. Both populations show an improvement in mortality for the latest cohorts as well, although this was flat in Spain for 1945 – 1960.

The Hunt-Blake model seems appropriate in that the trend for the younger ages is particularly complex, and probably would not be handled as well by Lee-Carter. Using linear splines with slope-change shrinkage produces generally smooth parameter curves, although with occasional jumps when the data overwhelms the prior push towards no change. The shrinkage priors ended up being close to Cauchy, with ν of 1.1 for Spain and 0.98 for Japan, with respective scaling parameters of 0.033 and 0.018.

5 Conclusion

Joint estimation of the same model for multiple datasets is a common actuarial problem, often with some of the datasets being sparse or volatile. The shrinkage approach here is to shrink the differences in the parameters across the datasets. This can be done in a Bayesian or classical setting, but is more natural in the Bayesian context, especially because there is a good, convenient penalized likelihood measure, $\widehat{elpd}_{loo}$, available in that case.

For APC models, parameter reduction is often appropriate anyway, and shrinkage of slope changes in linear spline models is one way of doing that. Then shrinking the differences in the slope changes across the fits for different datasets provides a convenient way to do the multivariate estimation. The example applied this to the Hunt-Blake model for mortality and got a more parsimonious joint model with improved out-of-sample predictions. Even though using this on two credible-enough datasets would not be the typical application, the improved prediction in that case is not surprising, especially when you consider the general predictive advantages of shrinkage such as those discussed in Stein (1956).

The use of hyper-priors on the shrinkage priors appears to be effective. In this case both the shape and scale of the shrinkage distributions were estimated. The resulting nearly-Cauchy prior is consistent with a good deal of recent discussions. It is considerably more heavy-tailed than the also popular Laplace prior, thus shrinking more towards zero but also allowing more large deviations from zero when called for.

References

Antonio, K., A. Bardoutsos, and W. Ouburg. 2015. "Bayesian Poisson Log-Bilinear Models for Mortality Projections with Multiple Populations." *European Actuarial Journal* 5: 245–81.

Barnett, Glen, and Ben Zehnwirth. 2000. "Best Estimates for Reserves." *PCAS* 87: 245–303.

Cairns, A. J. G., D. Blake, K. Dowd, G. D. Coughlan, and M. Khalaf-Allah. 2011. "Bayesian Stochastic Mortality Modelling for Two Populations." *Astin Bulletin* 41:1: 29–59.

de Jong, Piet. 2012. "Modeling Dependence Between Loss Triangles." *North American Actuarial Journal* 16:1: 74–86.

Dowd, K., A. J. G. Cairns, D. Blake, G. D. Coughlan, D. Epstein, and M. Khalaf-Allah. 2011. "A Gravity Model of Mortality Rates for Two Related Populations." *North American*

Actuarial Journal 15:2: 334–56.

Gao, Guangyuan, and S. Meng. 2017. "Stochastic Claims Reserving via a Bayesian Spline Model with Random Loss Ratio Effects." *ASTIN Bulletin*.

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: 145–62.

Gluck, Spencer M. 1997. "Balancing Development and Trend in Loss Reserve Analysis." *PCAS* 84: 482–532.

Greenberg, B. G., John J. Wright, and Cecil G. Sheps. 1950. "A Technique for Analyzing Some Factors Affecting the Incidence of Syphilis." *Journal of the American Statistical Association* 45:251: 373–99.

Hunt, Andrew, and David Blake. 2014. "A General Procedure for Constructing Mortality Models." *North American Actuarial Journal* 18 (1): 116–38.

Jarner, S. F., and E. M. Kryger. 2009. "Modelling Adult Mortality in Small Populations: The Saint Model." *Pensions Institute* Discussion Paper PI-0902.

Lee, R., and L. Carter. 1992. "Modeling and Forecasting U.S. Mortality." *Journal of the American Statistical Association* 87: 659–75.

Li, N., and R. Lee. 2005. "Coherent Mortality Forecasts for a Group of Populations: An Extension of the Lee-Carter Method." *Demography* 42:3: 575–94.

Meyers, Glenn. 2015. "Stochastic Loss Reserving Using Bayesian Mcmc Models." *CAS Monograph Series* 1: i–55.

Renshaw, A. E., and S. Haberman. 2006. "A Cohort-Based Extension to the Lee-Carter Model for Mortality Reduction Factors." *Insurance: Mathematics and Economics* 38: 556–70.

Shi, Peng. 2013. "A Copula Regression for Modeling Multivariate Loss Triangles and Quantifying Reserving Variability." *Astin Bulletin* 44:1: 85–102.

Shi, Peng, and Brian M. Hartman. 2014. "Credibility in Loss Reserving." *CAS E-Forum* Summer:2: 29–51.

Shi, Peng, Sanjib Basu, and Glenn G. Meyers. 2012. "A Bayesian Log-Normal Model for Multivariate Loss Reserving." *North American Actuarial Journal* 16:1: 29–51.

Stein, Charles. 1956. "Inadmissibility of the Usual Estimator of the Mean of a Multivariate Normal Distribution." *Proceedings of the Third Berkeley Symposium* 1: 197–206.

Taylor, Greg. 1977. "Separation of Inflation and Other Effects from the Distribution of Non-Life Insurance Claims Delays." *Astin Bulletin* 9: 217–30.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2016. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and Waic." *arXiv Preprint*

http://arxiv.org/abs/1507.04544.

Venter, Gary, and Şule Şahin. 2017. "Parsimonious Parameterization of Age-Period-Cohort Models by Bayesian Shrinkage." *Astin Bulletin*.

Venter, Gary, Roman Gutkovich, and Qian Gao. 2015. "Parameter Reduction in Actuarial Triangle Models" https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2992300.

Verbeek, H. G. 1972. "An Approach to the Analysis of Claims Experience in Excess of Loss Reinsurance." *Astin Bulletin* 6: 195–202.

Ye, J. 1998. "On Measuring and Correcting the Effects of Data Mining and Model Selection." *Journal of the American Statistical Association* 93: 120–31.