Modeling Mortality of Related Populations via Parameter Shrinkage

Gary Venter, School of Risk and Actuarial Studies, University of New South Wales, Sydney, and Columbia University, Actuarial Sciences Program, New York

Şule Şahin, Department of Actuarial Sciences, Hacettepe University, Ankara, and Institute for Financial and Actuarial Mathematics, University of Liverpool, Liverpool

Abstract: Parameter shrinkage is known to reduce fitting and prediction errors in linear models. When the variables are dummies for age, period, etc. shrinkage is more commonly applied to differences between adjacent parameters, perhaps by fitting cubic splines or piecewise-linear curves (linear splines) across the parameters. A common problem in mortality is modeling related populations where some commonality is desired. We do this by shrinking slope changes of linear splines for the largest population, then shrinking differences from those slope changes for the other populations.

There are frequentist and Bayesian approaches to shrinkage, and they have a good deal of similarity. Here we use a unified approach that compromises a bit with both of these paradigms. It uses Bayesian tools without some of the historical Bayesian concepts, and pushes the random-effects framework slightly to accommodate.

Modeling Swedish and Danish male mortality data is used as an illustration.

Keywords: Joint estimation, Parameter shrinkage, MCMC, Mortality

1 Introduction

Stein (1956) showed that shrinking multiple estimated means towards the overall mean to some degree always reduces estimation and prediction variances from those of maximum likelihood estimation (MLE). Similar methods have been part of actuarial credibility theory since Mowbray (1914), and in particular Bühlmann (1967). In regression, Hoerl and Kennard (1970) showed that some degree of shrinking coefficients likewise always produces lower fitting and prediction errors than MLE. This is quite related, as in their ridge-regression approach variables are typically scaled to have mean zero, variance one, and thus reducing coefficients shrinks the fitted means towards the constant term, which is the overall mean. Their initial purpose was to handle regression with correlated independent variables, and applied a bit more general procedure for ill-posed problems known as Tikhonov regularization, from Tikhonov (1943). As a result, shrinkage methods are also referred to as regularization.

Ridge regression produces parameter shrinkage by minimizing the negative loglikelihood (NLL) plus a selected shrinkage factor λ times the sum of squares of the regression coefficients. Some small enough λ will always reduce the estimation variance, but it usually takes extra steps using cross validation to select the shrinkage factor. A related method, lasso, uses the sum of the absolute values of the coefficients instead of the sum of their squares. It has become more popular because some coefficients usually shrink exactly to zero. Thus a modeler can start off with a long list of variables and use lasso to find combinations that work well together and exclude the rest.

Random effects modeling provides a more general way to produce shrinkage. Historically it postulated mean-zero normal distributions for the coefficients, with the variance parameters to be estimated, and optimized the product of those normal densities with the likelihood function. This also pushes the coefficients towards zero. Ridge regression turns out to be the special case where there is a single variance assumed for all the coefficients. Using the normal distribution for this shrinkage is common but it is not a conceptual limitation. If the double exponential (Laplace) distribution is assumed instead, lasso becomes a special case.

Bayesian shrinkage is quite similar to this, with normal or Laplace priors used for the parameters, which shrink them towards zero. Ridge regression and lasso estimates are produced as the posterior modes. This is the similarity we use to try to get a unified methodology that most frequentists and Bayesians could accept. They are already closer than in the past, as the Bayesian approach used for shrinkage departs from traditional Bayesian concepts in that the prior distributions are not necessarily related to prior beliefs. Rather they are some of the postulated distributions assumed in the model, just like the random effects are, and they can be rejected or modified based on the posteriors that are produced.

Mortality data, similarly to casualty loss reserve data, comes in arrays that are portions of a rectangle. Typical models assume there are factors for each row, column, and possibly diagonal of the rectangle. In the broader statistical literature, models that include parameters for all three directions are known as age-period-cohort (APC) models. Period refers to the time of observation for members of a cell, like year of death or year of payment, age is the age at death or lag, and cohort is the time of origin for the cell members, like accident year or year of birth. APC models trace back to Greenberg, Wright, and Sheps (1950).

Parameter shrinkage is a bit more complicated for these models, as you need to retain parameters for each age, period, and cohort. Actuaries have used a number of approaches to APC parameter reduction, including fitting curves (Gompertz, Vasicek etc.) across the parameters, and linear and cubic splines. Gluck (1997) uses a shrinkage method to allow some accident year variation in the Cape Cod loss reserving model. Barnett and Zehnwirth (2000) introduce shrinkage for piecewise linear slope changes in APC reserve modeling, using proprietary methodology. G. Gao and Meng (2018) use Bayesian shrinkage for fitting cubic splines to a reserve model. Venter and Şahin (2018) model US male mortality rates by Bayesian shrinkage with linear splines in the model of Hunt and Blake (2014), including multiple trends. Venter (2018) uses a similar model on US opioid mortality rates. Guszcz (2008) illustrates a different random-effects approach applied to the curves Clark (2003) fit to the loss triangle from Renshaw (1989), which was built using the data in the operational time reserving paper of Taylor and Ashe (1983). (Also modeled by Verrall (1990), Verrall (1991), Mack (1993), and Venter (2007).)

Quite a few papers have addressed joint APC modeling of related datasets in reserving and mortality, so this is a widely known issue. Venter, Gutkovich, and Gao (2017) explore related-triangle reserve modeling with linear spline slope-change shrinkage using random effects. 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 Bayesian Markov-Chain Monte Carlo (MCMC) estimation. They assume fairly wide priors, not shrinkage priors. This uses the capability of MCMC to estimate difficult models. Antonio, Bardoutsos, and Ouburg (2015) estimate the model of Li and Lee (2005) by MCMC. None of these use Bayesian shrinkage for joint populations, which is the emphasis here, and which takes advantage of shrinkage to reduce estimation and predictive variances.

We lay out a common framework for frequentist and Bayesian shrinkage in Section 2. Section 3 discusses the application to joint mortality modeling. Section 4 illustrates this with an example from Swedish and Danish mortality. That example explores using a combination of frequentist and Bayesian software packages to improve the efficiency of the estimation, with some code examples. Section 5 concludes.

2 Shrinkage Methodologies

For a model with parameters β_j , ridge regression minimizes $\text{NLL} + \lambda \sum_j \beta_j^2$ for a selected shrinkage constant λ . Hoerl and Kennard (1970) showed that for regression models there is always some $\lambda > 0$ for which ridge regression gives a lower estimation variance than does MLE. They did not have a formula to find an optimum value of λ , but could find reasonably good values by cross validation – that is, by testing prediction on subsamples omitted from the fitting. In the 1990s, lasso, which minimizes $\text{NLL} + \lambda \sum_j |\beta_j|$, became a popular alternative

in that it shrank some parameters exactly to zero. Thus it provides variable selection as well as shrinkage estimation.

Random effects modeling is a frequentist shrinkage tool that is in some ways similar to Bayesian analysis. For frequentists, parameters are constants and do not have distributions. (Although their estimates can have distributions.) A random effect is not exactly a parameter but is used similarly in models. One way to conceptualize these may be as model-postulated real-world, not mathematical, aspects of a system. For instance, a given point in the US may have an average annual temperature for 2019 that is different from that of the US as a whole. That difference could be modeled as a random effect, and then would have a distribution.

Such differences from averages are frequently modeled as random effects. Typically there would be many of them, each with its own observations. Each can be considered to have a distribution with mean zero, all with a single common variance. (However some popular random-effects software routines default to the assumption that each one has its own variance.)

We start by assuming that the random effects have mean-zero normal distributions with a common variance. Given a vector b of random effects and a vector β of parameters (fixed effects), with respective design matrices Z and X , and a vector of observations y , the model is:

$$y = X\beta + Zb + \epsilon$$

$$b \sim N(0, \theta\sigma^2), \epsilon \sim N(0, \sigma^2)$$

Sometimes fitting the model is described as estimating the parameters and projecting the random effects. A popular way to fit the model is to maximize the joint likelihood, which is $p(y|\beta, b)p(b)$. This is the likelihood times the probability of the random effects. The mean-zero distribution for the b s pushes their projections towards zero, so this is a form of shrinkage. Matrix estimation formulas have been worked out for this similar to those for regression.

Now we consider instead the double exponential, or Laplace, distribution for the random effects. This is an exponential distribution for positive values of the variable, and is exponential in $-b$ for negative values. Its density is

$$p(b|\lambda) = 0.5\lambda e^{-\lambda|b|}$$

Suppose all the random effects have the same λ . Then the negative joint loglikelihood is $\log(2) - \log(\lambda) + \sum \lambda|b_j| + \text{NLL}$. The constants do not affect the minimization, so for a fixed given value of λ , this is the lasso estimation. If λ is considered a parameter to estimate, then the $\log(\lambda)$ term would be included in the minimization, giving an estimate of λ as well. This same calculation starting with a normal prior gives ridge regression. Thus these are both special cases of random effects, and both provide estimates of the shrinkage parameter.

Lasso and ridge regression have problems with measures of penalized NLL like AIC and BIC, as these tend to rely on parameter counts, but shrunk parameters do not use as many degrees of freedom, because they have less ability to influence fitted values. Ridge regression can handle this using a regression measure called the diagonal of the hat matrix. The hat matrix

H gives the vector of mean values from the observations by $\mu = Hy$. The elements on its diagonal tell how much a fitted mean μ_j would change from a change in y_j , and so is the derivative of the fitted mean with respect to the observation. The sum of these derivatives is the degrees of freedom used by the model. Ye (1998) builds on this to define the generalized degrees of freedom used in a nonlinear model as that same sum of derivatives. With design matrix X , a standard regression result is $H = X(X^T X)^{-1} X^T$. With design matrix Z for ridge regression, H may be estimated by $H = Z(Z^T Z + \lambda I)^{-1} Z^T$, which gives a reduced parameter count for the random effects. See for instance Ginestet (2013). This is more problematic for lasso, however.

In any case, cross validation is more popular for model selection for frequentist shrinkage. There is usually some degree of subjectivity, or a random choice, involved, however. Bayesian shrinkage, as discussed below, has a fast automated cross-validation method for penalizing the NLL. It provides a good estimate of what the likelihood would be from a new sample from the same population, which is the goal of AIC as well. The weakness of this is that for financial modeling, often models are simplified approximations to more complex processes, so the next sample might not be quite from the model used. Other NLL penalties, like those for BIC and HQIC, relax this assumption, and lead to more parsimonious models. That might be accomplished in an ad hoc way by shrinking more than indicated when the model is regarded as this kind of approximation.

The Bayesian approach is to assume shrinkage priors for the parameters. For instance, a parameter might be assumed to be normally distributed with mean zero, or Laplace distributed. For a sample X and parameter vector β , with prior $p(\beta)$ and likelihood $p(X|\beta)$, the posterior is given by Bayes Theorem as $p(\beta|X) = p(X|\beta)p(\beta)/p(X)$. Here the probability of the data, $p(X)$, is not usually known but it is a constant. Since $p(\beta|X)$ must integrate to 1, the constant is not needed, so the posterior is determined by $p(X|\beta)p(\beta)$. Note that if β were random effects instead of parameters, this would be the joint likelihood. Thus maximizing the joint likelihood gives the same estimate as the mode of the posterior distribution. The main difference between Bayesians and frequentists now is that the former mostly use the posterior mean, and the latter use the posterior mode. Bayesians tend to suspect that the mode could be overly responsive to peculiarities of the sample. Frequentists come from a lifetime of optimizing goodness of fit of some sort, which the mode does.

Bayesians generate a sample of the posterior numerically using MCMC. This ranks samples using the joint likelihood, since this is proportional to the posterior. MCMC has a variety of sampling methods it can use. The original one was the Metropolis algorithm, which has a generator for the possible next sample using the latest accepted sample. If the new sample parameters have a higher joint probability, it is accepted. If not, a random draw is used to determine whether it is accepted or not. The first group of samples is eventually eliminated, and the remainder have been shown to follow the posterior distribution.

There is nothing particularly Bayesian about MCMC. It originated as a search procedure for numerical integration, and was first brought into probability theory by Bayesians. Over time, however, the methodology has become less Bayesian. Now the so-called prior distribution for the parameters is actually a postulated distribution for the model, and it is evaluated by the posterior it produces. If they are somehow at odds with each other, or there are convergence

problems, etc., the prior is modified to produce a better model.

We feel that random effects estimation can also make use of MCMC to calculate the posterior mean, but under a different framework. Now we will just consider the case where there are no fixed effects to estimate – all the effects in the model are random effects. A distribution is postulated for the effects. We know how to calculate the mode of the conditional distribution of the effects given the data, namely by maximizing the joint likelihood. By the definition of conditional distributions, the joint likelihood $= p(X|\beta)p(\beta) = p(X, \beta) = p(X)p(\beta|X)$. Thus the conditional distribution of the effects given the data is proportional to the joint likelihood, and so can be estimated by MCMC.

What the Bayesians would call the posterior mean and mode of the parameters can instead be viewed as the conditional mean and mode of the effects given the data. With MCMC, the effects can be projected by the conditional mean instead of the conditional mode that was produced by maximizing the joint likelihood. This does not require Bayes Theorem, subjective prior and posterior probabilities, or distributions of parameters, so removes many of the objections frequentists might have to using MCMC. Admittedly Bayes Theorem is obtained by simply dividing the definition of conditional distribution by $p(X)$, but you do not have to do that division to show that the conditional distribution of the effects given the data is proportional to the joint likelihood. To claim that this result effectively uses Bayes Theorem amounts to saying that Bayes Theorem is just a definition. In that case there is no need to invoke it.

2a Goodness of fit from MCMC

One advantage of using MCMC, whether the mean or mode is used, is that there is a convenient goodness-of-fit measure. The leave-one-out, or loo, cross validation method is to refit the model once for every sample point, leaving out that point in the estimation. Then the likelihood of that point is computed from the parameters fit without it. The sum of the loglikelihoods of the omitted points has been shown to be a good estimate of the loglikelihood for a completely new sample from the same population, so it eliminates the sample bias in the NLL calculation. This is what AIC, etc. aim to do as well.

But doing all those refits can be very resource intensive. Gelfand (1996) developed an approximation for an omitted point's likelihood from an MCMC sample generated from all the points. He used the numerical integration technique of importance sampling to estimate a left out point's likelihood by the weighted average likelihood across all the samples, with the weight for a sample proportional to the reciprocal of the point's likelihood under that sample. That gives greater weight to the samples that fit that point poorly, and turns out to be a good estimate of the likelihood that it would have if it had been left out of the estimation. This estimate for the likelihood of the point turns out to be the harmonic mean over the samples of the point's probability in each sample. With this, the MCMC sample of the posterior distribution is enough to calculate the loo goodness-of-fit measure.

This gives good but volatile estimates of the loo loglikelihood. Vehtari, Gelman, and Gabry (2017) addressed that by a method similar to extreme value theory – they fit a Pareto to the probability reciprocals and use the Pareto values instead of the actuals for the largest

20% of the reciprocals. They call this “Pareto-smoothed importance sampling.” It has been extensively tested and is becoming widely adopted. Their penalized likelihood measure is labeled $\overline{elpd}_{loo}$, standing for “expected log pointwise predictive density.” Here we call $-\overline{elpd}_{loo}$ simply loo. It is the NLL plus a penalty, so is a penalized likelihood measure.

The fact that this is a good estimate of the NLL without sample bias comes with a caveat. The derivation of that assumes that the sample comes from the modeled process. That is a standard assumption but in financial areas, models are often viewed as approximations of more complex processes. Thus a new sample might not come from the process as modeled. Practitioners sometimes respond to this by using slightly under-fit models – that is more parsimonious models with a bit worse fit than the measure finds optimal. BIC was designed for this kind of situations as well, and it also leads to more parsimonious models.

2b Postulated shrinkage distributions

With loo available it is possible to look for the value of λ that minimizes loo. However this runs the risk of being overly responsive to peculiarities of the sample, and also makes it more likely that the model has an under-estimated value of loo. An alternative is to provide a postulated distribution for λ itself, and use the conditional distribution of λ given the data. Our experience has been that the fitted values from this end up close to those for the λ with the highest loo, and often have a slightly better loo than any single λ produces. This can be represented as giving λ a prior, or as λ itself being a random effect with a postulated distribution. This is the approach used in the example. G. Gao and Meng (2018) do this as well.

The shape of the shrinkage distribution can also have an effect on the model fit. The Laplace distribution is more peaked at zero than the normal, and also has heavier tails, but is lighter at some intermediate values. An increasingly popular shrinkage distribution is the Cauchy, with $1/p(\beta) = \pi(1 + \lambda^2\beta^2)/\lambda$ and $-\log(p(\beta)) = -\log\lambda + \log\pi + \log(1 + \lambda^2\beta^2)$. It is even more concentrated near zero and has still heavier tails than the Laplace. For a fixed λ , the constants in the density drop out, so the conditional mode minimizes $\text{NLL} + \sum \log(1 + \lambda^2\beta_j^2)$. This is an alternative to both lasso and ridge regression. Cauchy shrinkage often produces more parsimonious models than the normal or Laplace do. It can have a bit better or bit worse penalized likelihood, but even if slightly worse, the greater parsimony makes it worth considering. It also seems to produce tighter ranges of parameters.

The Cauchy is the t-distribution with $\nu = 1$. If $\nu = 2$, the conditional mode minimizes $\text{NLL} + 1.5 \sum \log(2 + \lambda^2\beta_j^2)$, and if $\nu = 3$, this becomes $\text{NLL} + 2 \sum \log(3 + \lambda^2\beta_j^2)$. 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 distribution postulated for ν could be used to get an indication for how heavy-tailed the shrinkage distribution should be for a given data set. Actually, the Laplace distribution can be approximated by a t distribution. A t with $\nu = 6$ matches a Laplace with scale parameter $\sqrt{3}/2$ in both variance and kurtosis (and since the odd moments are zero, matches all 5 moments that exist for this t), so is a reasonable approximation. Figure 1 graphs their densities on a log scale.

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

We test this for the example in Section 4. The Laplace is assumed while building the model, but afterwards the shrinkage distribution is tested by assuming a t shrinkage distribution with an assumed distribution for ν , and finding the conditional distribution of ν given the data. For this case, the conditional mean is close to $\nu = 2$. This is heavier-tailed than the Laplace but less so than the Cauchy. With $\nu = 2$, the $f(\beta) = (2 + \beta^2)^{-1.5}$ and $F(\beta) = [\beta/\sqrt{2 + \beta^2} + 1]/2$. Still the Cauchy is a viable alternative to get to a more parsimonious model.

The versatility of the t for shrinkage also opens up the possibility of using different but correlated shrinkage for the various parameters, as the multi-variate t is an option in the software package we use. For $\nu = 2$, the covariances are infinite, but a correlation matrix can still be input. Two adjacent slope changes are probably quite negatively correlated, as making one high and one low can often be reversed with little effect on the outcome. Recognizing this correlation in the assumed shrinkage distribution might improve convergence. We leave this as a possibility for future research.

From here on we will refer to the model coefficients as estimates of parameters instead of projections of effects. Our model can be considered to have only random effects, with no fixed effects, as everything estimated has an assumed unconditional distribution, even the shrinkage variance. Thus a frequentist can translate the setup into the random-effects projection framework with no change in the formulas or results. We will not use the terms prior and posterior, as their traditional subjective probability interpretation is not being used. They are replaced with postulated distribution of the parameters and conditional distribution of the parameters given the data. Thus our approach appears to be completely consistent with frequentist theory. Ironically enough, Bayesians would view this approach as putting a prior on the shrinkage variance, which they call the fully Bayesian method.

3 Joint Mortality Modeling by Shrinkage

We are using data from the Human Mortality Database, which generally organizes tables by year of observation and age at death. The year of birth is approximately year of death minus age at death, depending on whether the death is before or after the birthday in the year. For convenience, we consider cohort to be the year given by year of death minus age at death. We denote cohorts as n , with parameters $p[n]$, age at death by u , with parameters $q[u]$, which makes the year of death $n + u$, with parameters $r[n + u]$. We also denote by c a constant term, that is not shrunk in the modeling. The parameters combine in various models to give the log of the mortality rate, and we model the number of deaths as Poisson in the mortality rate times the exposures, which are the number living at the beginning of the year in the n, u cell. Theoretically, deaths counts are the sum of Bernoulli processes, and so should be binomial, but for small probabilities, the binomial is very close to the Poisson, which is sometimes easier to model. Later we test the Poisson assumption against heavier-tailed distributions.

A basic model is the linear APC model, which does not include the Lee-Carter trend weights. This has become fairly popular with actuaries and it is our starting point, as it is a purely linear model that can be estimated with standard packages, including lasso packages. Its mortality rate m given by:

$$m[n, u] = c + p[n] + q[u] + r[n + u]$$

With exposures $e[n, u]$, the deaths $y[n, u]$ are Poisson in $e[n, u] \exp(m[n, u])$. The model of Hunt and Blake (2014), which is the model of Lee and Carter (1992) plus the cohort effect, allows the time trend $r[n + u]$ to vary by age, which reflects the fact that medical advances, etc. tend to improve the mortality rates for some ages more than others. Denoting the trend weight for age u by $s[u]$, the model we use here is:

$$m[n, u] = c + p[n] + q[u] + s[u]r[n + u]$$

Actually Hunt and Blake (2014) allow the possibility of different trends with different age weights either simultaneously or in succession. Venter and Şahin (2018) suggest an informal test for the need for multiple trends, and find that they indeed help in modeling the US male population for ages 30–89. In the example here we restrict the model to ages 50 and up, and do this test, which suggests that one trend is sufficient, so the model is presented in that form. In other populations we have noticed that modeling a wide range of ages can often require multiple trends.

The Hunt-Blake and Lee-Carter models are sometimes described as bilinear APC models, as they are linear if either the s or r parameters are taken as fixed constants. The age weights help model the real phenomenon of differential trends by age, but their biggest weakness is that they are fixed over the observation period, and the ages with the most trend can change over time as medical and public health trends change. The possibility of multiple trends in the Hunt-Blake model can reflect such changes.

To set the basic APC model up to use linear-modeling software packages, the entire array of death counts $y[n, u]$ is typically strung out into a single column. When doing this, it is

generally helpful to make parallel columns that record which age, period, and cohort a cell comes from. Then a design matrix can be created where all the variables $p[n]$, $q[u]$, $r[n + u]$ have columns parallel to the death counts, and are represented as 0,1 dummy variables that have a 1 only for the observations that variable applies to. Then a column vector β that has a value for the parameter for each variable can be multiplied on the left by the design matrix to give the fitted value of $m[n, u]$ for each cell.

This gets more complicated when the parameters to estimate are slope changes of linear splines fit to the p, q, r parameters. All the previous slope changes add up to the slope at a point, and the next point is the slope plus the previous point. That means that all the slope changes in the previous point go into the next point, and an additional copy of each of those slope changes is added because they are in the slope. For instance, the slope change parameter for the third cohort goes into $p[3]$. Then it is in the next slope, so it is added again in $p[4]$, and keeps on accumulating one more time at each step. More generally, the slope change for cohort j contributes $\max(0, 1 + i - j)$ times to $p[i]$. Thus an observation that is for cohort i , which would get a 1 for $p[i]$, gets a slope change dummy value of $\max(0, 1 + i - j)$ in the cohort j variable's slope change column. The same thing applies to the years, ages, etc.

For the two population case, assume that the dependent variable y column has all the observations for the first population followed by all the observations for the second population, and the design matrix has columns for all the first population variables followed by columns for all the second population variables. To create the overall design matrix, we put the first population's design matrix in the upper left quadrant, followed by a matrix of all zeros in the upper right quadrant. The lower left quadrant repeats the upper left, as the second population starts off with all the parameters of the first population. The same matrix is repeated in the lower right quadrant. But the right side variables represent the changes in the second population's slope changes from the first population's. Hopefully many of these will be small or zero. Then one more column was added for the change in the constant term for the second population. This is all zeros for the first population and all ones for the second population. (The software adds in an overall constant term, which is not shrunk, so a variable for it is not needed.)

MCMC incorporates a very good search for good parameters, but these variables make it difficult. Consecutive variables are very highly correlated – many 99.9% correlated, which makes them largely interchangeable, and their parameter values can end up negatively correlated as offsets. There could be hundreds of variables in APC models for reasonably large datasets as well. It could take days or weeks for a PC to converge to good parameters running MCMC software. As an alternative, we have started using lasso software, which is usually very fast, to identify variables that can be eliminated, and then use a reduced variable set for MCMC. Lasso software, like the R package `glmnet`, includes cross validation routines to help select λ . Typically, this would give a range of λ values that is worth considering. We use the lowest λ value this suggests, which leaves the largest number of variables, to generate the starting variables for MCMC.

When a slope-change variable is omitted, the slope of the fitted piecewise linear curve does not change at that point, which extends the previous line segment. If instead a small positive or negative value were to be estimated, the slope would change very slightly at that point,

probably with not much change to the fitted values. When lasso cross validation suggests that the slope change is not useful, that is saying it would not add to the predictive value of the model. It is possible that a few of the omitted points in the end could have improved the model, but probably they would not have made a lot of difference. So even with MCMC, there is no guarantee that all the possible models have been examined and the very best selected. Nonetheless this procedure is a reasonable way to get very good models.

Constraints are needed to get convergence of APC models. First of all, with the constant term included, the first age, period, and cohort parameters are set to zero. This is not enough to get a unique solution, however. With age, period, and cohort parameters all included, the design matrix is singular. Actuaries use various sets of constraints to deal with this. However the setup with slope change variables helps to some extent. If a straight regression is done for a simple age-period model, the slope change matrix gives exactly the same q, r parameter values that the 0,1 dummy matrix gives, and the APC matrix is still singular. But when enough slope changes are set to zero, the matrix is no longer singular. Our procedure, then, is to use lasso for an AP design matrix, take out the slope change variables that lasso says are not needed, then add in all the cohort variables, and run lasso again. The maximum variable set lasso finds for this is then the starting point for MCMC, without any other constraints on the variables.

The MCMC runs usually estimate a fair number of the remaining parameters as close to zero. We take out those variables as well, but look at how loo changes. As long as loo does not get worse from this, those variables are left out. Usually a fair number of the maximal variable set that lasso identified are eliminated in this process.

We then use the best APC model produced by this as the starting point for the bi-linear models. Those add age modifiers to the period trends. Some age will have the highest period trend by this procedure, and the models give every other age its own fixed percentage of the trend at every period. Some constraint is needed here, and we make the highest age factor 1.0, with all ages forced to be positive. This of course complicates the program. Typical lasso software does not provide capabilities for bi-linear modeling, but MCMC packages are generally more flexible. We end up making one design matrix for the age and cohort parameters, another for the period parameters, and a third for the age weights on trend. Then the weights are applied to the trend parameters, and added to the age-cohort parameters, and exponentiated to get the fitted mortality rates. A lasso approach could probably be devised by treating the period parameters from the APC model as fixed constants, but instead we just added all the age-weight variables to the model, and eliminated the parameters near zero as indicated by the MCMC outcomes.

4 Swedish-Danish Male Mortality Example

We model deaths from 1970-2016 for ages 50-99 and cohorts 1883-1953 from the Human Mortality Database. As a data investigation, Figure 2 graphs the cumulative mortality trends by decade age ranges for both populations to see if there are any apparent shifts in the ages with the highest trend. The trends are a bit different in the two countries, but they look a lot

like what the Lee-Carter model assumes, with the relative trends among the ages reasonably consistent over time.

Each population has 50 ages for 47 years coming from 71 cohorts. Leaving out the first age, period, and cohort of each population, and adding the dummy variable for Denmark, gives 331 variables for the design matrix. We started with a lasso run, using the R GLM lasso package `glmnet`, without the cohort variables. The `cv.glmnet` app does the lasso estimation as well as this package's default cross validation. One of the outputs is the smallest λ (thus the one with the most non-zero coefficients) that meets their cross-validation test. We ran the package twice. The first time was without the cohort variables, and then with the resulting non-zero variables from that plus all the cohort variables. Below is the key part of the code for that.

```
fit = cv.glmnet(x, y, standardize = FALSE, family = "poisson", offset=offs)
out <- as.matrix(coef(fit, s = fit$lambda.min))
write.csv(out, file="lasso_fit.csv")
```

Before running this, the mortality count vector for the two countries (4362 elements) was input into R and called `y`. The design matrix for the run is in `x`. The log of the exposure vector – also 4362 elements – was called `offs`. `Glmnet`'s Poisson app does a linear model with a log link then adds the offset, which here is the log of the exposures. This is like multiplying the estimated mortality rate by the exposures to get the Poisson mean for an observation. The output of cross validation is called `fit` in the first line. The next two lines output the coefficients with λ set at the indicated minimum value.

This resulted in 80 variables to use for the APC model. There were 22 age variables, 16 year variables, and 42 cohort variables that remained after lasso. We use the Stan MCMC application, specifically the R version, `rstan`. This requires specifying an assumed distribution for each parameter, and it simulates the conditional distribution of the parameters given the data. We used the double exponential distribution for all the slope-change variables. This is a fairly common choice, usually called Bayesian lasso, but we revisit it below. The double exponential distribution in Stan has a scale parameter $s = 1/\lambda > 0$. An assumed distribution is needed for that, as well as for the constant term.

For positive parameters, we prefer to start with a distribution proportional to $1/x$. This diverges both as x gets small and as it gets large, so it gives balancing strong pulls up and down, whereas a positive uniform distribution diverges upwardly only. As an example where the integral is known, consider a distribution for the mean β of a Poisson distribution with probability function proportional to $e^{-\beta}\beta^k$. With a uniform distribution for β , which is proportional to 1, the conditional distribution of β given an observation k is also proportional to $e^{-\beta}\beta^k$, which makes it a gamma in $k+1$ and 1, with mean $k+1$. But if the distribution of β is assumed proportional to $1/\beta$, the conditional is proportional to $e^{-\beta}\beta^{k-1}$, which makes it a gamma in $k, 1$, with mean k . Thus the $1/\beta$ unconditional distribution take the data at face value, whereas the uniform pushes it upwards. Numerical examples with other distributions find similar results.

The $1/\beta$ assumption is easier to implement in Stan by assuming $\log(\beta)$ is uniform. Stan

Figure 2: Cumulative Trends by Age Group

just assumes a uniform distribution is proportional to a constant, so can be ignored. If not specified, it takes the range $\pm 1.7977 \cdot 10^{308}$, which is the largest expressible in double-precision format. We usually start with a fairly wide range for uniform distributions, but if after working with the model for a while the conditional distributions tend to end up in some narrower range, we tend to aim for something a bit wider than that – mostly to save the program from searching among useless values. We ended up with assuming that the constant is uniform on $[-8, -3]$, and $\log(s)$ on $[-6, -3]$.

Initially we ran four sessions of Stan with two chains each for 10,000 iterations with 9,000 of those warmup. Each session had its own random number seed, which was repeated for every run. In each session at least one chain gave a reasonable fit. Then we eliminated variables with parameters less in absolute value than 0.001 – chosen by observation of ranges for this model – in each session, and used the parameters for the best chain in the session as starting values. With good starting values we ran only 500 iterations, with 200 warmup, again eliminating variables estimated below 0.001 in absolute value, but from then on used the average of the chains in a session for the next starting values. If eliminating the small parameters worsened the loo penalized NLL for a session, we put them back and ended that session. If no parameters were less than 0.001 in absolute value, we reran the session with 1500 iterations, 500 warmup, with the latest parameter values.

The best model had 54 parameters, a loo measure of 22,158.5, and NLL of 22,081.7. The sample-bias penalty was the difference, 76.8. This penalty is a bit high for that many parameters with shrinkage, a situation that some veteran R modelers tend to find suspicious in terms of predictive ability of the model. We know from the preliminary visual test that there is some variation in trend across the ages, and not reflecting this could be the problem. Slightly better fits could probably have been produced by eliminating some of the other small-slope-change variables, but already at a cutoff 0.001, there were cases where eliminating the variable made the fit slightly worse.

The next step was to put in the trend weights by age, which produces the Hunt-Blake model. This is no longer a simple linear model, which complicated the coding a bit, but not substantially so. We used the constraint that the highest weight for any age for a population would be 1.0, with all the weights positive. We enforced this by computing preliminary log weights as the accumulation of weight slope-change variables, then found the age with the highest weight for each population. The weight for that age was subtracted from all the weights for that population, and the results exponentiated, ending up positive with the highest value 1.0. In that way, the trend at any year is the trend for the age with the highest weight.

For Stan, we added in a separate slope-change design matrix for trend weights, with 100 rows and columns – one for each age in each population. When variables were eliminated, the columns were dropped but not the rows. Then the parameter vector times the design matrix would give the weights for each age. We followed the same procedure for eliminating variables and refining models as for the APC model. In the end, six more age, period, and cohort variables were eliminated, leaving 46 of those, and 72 slope-change variables were kept, for 118 variables all together. This produced a loo penalized NLL of 20,725.5, an NLL of 20,644.5, and a penalty of 81.0. With NLL and penalized NLL, since they are already in logs, it is

Figure 3: Trend Factors

the absolute difference, not the relative difference, than indicates how much improvement a better model produces. Here that is substantial. Also having 64 more variables only increased the penalty by 4.2. The age weights for trend apparently help the predictive value of this model quite a bit.

Next we tried an alternative shrinkage distribution, as discussed in Section 3, using the variables from the best model. Instead of double exponential, we assumed the slope-change variables were t-distributed. Recall that the double exponential is about equivalent to a t with $\nu = 6$, a Cauchy is $\nu = 1$, and a normal is the limit as ν gets large. We assumed $\log(\nu)$ was uniform on $[-1, 6]$. That gives a range for ν of about $[0.37, 403]$. The result was a very slightly worse loo, at 20,726. The mean of ν was 2.4, with a median of 2. The $[2\%, 98\%]$ range was about 1 to 6. This is heavier tailed than the double exponential, with larger parameters allowed and smaller ones pushed more towards zero. Two more parameters were < 0.001 .

We have seen before that loo does not change greatly with small changes in the shape of the shrinkage distribution, and it looked like estimating ν was using up degrees of freedom, so we tried a constant $\nu = 2$ and took out the two small trend-weight variables. The result was then 116 variables, with a loo measure of 20,723.4, NLL of 20,643.9, and a penalty of 79.5. A drop of 2 in the penalized NLL is usually considered a better model, and this is more parsimonious as well.

One assumption not checked so far is the Poisson distribution of deaths in each age-period cell. As an alternative, we tried the negative binomial distribution, which has an additional

Figure 4: Trend Age Weights

Figure 5: Age Factors

Figure 6: Cohort Factors

parameter ϕ , which does not affect the mean but which makes the variance $= \mu^2/\phi + \mu$. This actually gave a still lower NLL of 20,589.1, and loo of 20,648.1. The sample-bias penalty was much lower, at 59. There were also three fewer trend-weight variables, leaving 67 of those, as well as 46 age, period, and cohort variables as before. Now there is still a constant, the shrinkage parameter, and ϕ that are not shrunk, so 116 variables in total. The low penalty means that the loglikelihood of the omitted variables was not much worse than that of the full model, which is part of why the predictive measure loo was so much better.

The Poisson is an approximation to the binomial distribution that would come from a sum of independent Bernoulli processes. Having a more dispersed distribution suggests that there is some degree of correlation in the deaths. That could arise from high deaths in bad weather or disease outbreaks, for instance. However the dispersion is a measure of the departure of the actual deaths from the fitted means, and also includes model error and estimation error. Thus it is not entirely certain that deaths are correlated.

We just assumed that $\log(\phi)$ was distributed uniform on $\pm 1.7977 \cdot 10^{308}$. The conditional mean of ϕ given the data is 2839. The largest mean for any cell was slightly above 2000, and that would get a variance increased by 70%. This could be as little as a 10% increase for small cells. Compared to the Poisson, this makes it less critical to get a very close fit on the largest cells, which can give a better fit on the other cells. Of course, another reason for the better fit could be actual greater dispersion. The parameters graphed in Figures 3 – 6 are for the negative-binomial model as are the actual vs. fitted values in Figures 7 and 8.

Table 1: Summary of Goodness-of-Fit Measures

	Parameters	Loo	Penalty	NLL
APC (Poisson with Laplace shrinkage)	54	22,159	77	22,082
Hunt-Blake (Poisson with Laplace shrinkage)	118	20,726	81	20,645
Hunt-Blake (Poisson with $\nu = 2$ shrinkage)	116	20,723	79	20,644
Hunt-Blake (Negative Binomial with $\nu = 2$ shrinkage)	116	20,648	59	20,589

Table 1 summarizes the goodness-of-fit measures for all the models. Table 2 shows the number of parameters of each type by country and the sum of their absolute values in the negative-binomial model. Denmark has fewer parameters and lower impact as measured by parameter absolute values. This is consistent with its parameters being changes from the Sweden parameters, and that not so much change was needed. Trend is a bit of an exception, and as we saw, Sweden does not have much trend while Denmark has some. The age parameters are one area in particular where the populations are similar, and Denmark does not need much change from the Sweden values. It is surprising how dominant the age-weight parameters are in both size and count, and also how they add so little to the NLL penalty and so much to the goodness of fit. They are clearly important in this model.

Figure 3 shows the trend factors by population. Sweden has less trend and is a bit simpler. Denmark does not deviate from the Swedish trend until halfway through the period. Both show long stretches with few slope changes. These are a lot simpler than the empirical trend graph in Figure 1, which could be why the age weights are important.

Figure 4 shows the age weights. They are more complex than the trends, but still fairly smooth graphs. Denmark largely has the same overall pattern as Sweden, so is using its parameters to a fair degree.

Figure 5 graphs the age factors on a log scale, along with a rotation of the factors to horizontal, to better show their differences. There are some nuances of slope-change variation that shows up in the rotation. Figure 6 shows the cohort factors, along with fitted lines to each. Although the shapes are different, a lot of the slope changes are actually similar between the two populations. We like to take the trend out of the cohort parameters, partly to get uniqueness of parameters, and partly so they would not have to be projected. But eliminating some variables has solved the uniqueness issue, and even after removing the overall trend, both countries would show a downward trend in cohorts since the early 1920s.

Actual vs. fitted mortality rates are graphed for every cohort by age in three-year increments in Figure 7. Denmark’s actual rates are more volatile, but the fitted values look like reasonable representations of the trends in each country. For the most part, the rates are lower for later cohorts. The cohort parameters did not look exactly like this, but the fitted rates are a combination of age, period, cohort, and trend-weight parameters, which are all relative to each other. Figure 8 graphs this for every year of death instead of cohort, and the conclusions are similar.

Figure 7: Actual vs. Fitted Age by Cohort

Figure 8: Actual vs. Fitted Age by Year

Table 2: Parameter Count and Sum of Absolute Values by Type and Country

	Age	Cohort	Trend	Weight
Sweden Count	10	15	5	34
Denmark Count	3	9	4	33
Sweden Sum $ \beta $	0.126	0.062	0.054	0.363
Denmark Sum $ \beta $	0.022	0.041	0.093	0.189

5 Conclusion

Joint APC-modeling of related populations is a common need in actuarial work, including for mortality models and loss reserving, as shown for example by the quantity of papers on this topic. Parameter shrinkage is a fairly new approach for actuaries, and has advantages over MLE in producing lower estimation and prediction errors. We show how to apply that to APC models by shrinking slope changes of linear fits to the APC parameters, and then apply it to modeling related data sets by shrinking differences in slope changes between the data sets.

MCMC estimation can be applied to complex modeling problems. We show how it can be used here to produce samples of the conditional distribution of the parameters given the data. This is typically done in a nominally Bayesian setting, where the model starts with postulated distributions of the parameters, called priors, and yields the conditional distribution of parameters given the data, called the posterior. But frequentist random-effects modeling postulates distributions of the effects (which look like parameters to Bayesians) and there seems to be no reason it could not also use MCMC to sample from the conditional distribution of the effects given the data. Furthermore, when Bayesians take this approach, they often feel free to depart from the traditional view of priors as incoming beliefs and treat them instead more like random effects – that is postulated distributions. Thus the approach we take appears largely consistent with Bayesian and frequentist approaches, but different from the historical practices of each.

In the example, fewer and smaller parameters are used for the second population, suggesting that it is building from the model of the larger population. For this data, the Hunt-Blake model fits better than the linear APC model, the negative binomial fits better than the Poisson, and using the t-distribution with $\nu = 2$ works well as an assumed shrinkage distribution. A more parsimonious model, that looks not as good by loo, can be obtained by a Cauchy shrinkage distribution (the t-distribution with $\nu = 1$), which could be a good idea if the model is regarded as an approximation to a more complex process. These shrinkage distributions also provide regularization formulas similar to those of lasso and ridge regression. The t with $\nu = 2$ case, for instance, leads to minimization of $\text{NLL} + 1.5 \sum_j \log(2 + \lambda^2 \beta_j^2)$.

MCMC software can have problems with highly correlated variables, and the slope change variables are very highly correlated. Ridge regression and lasso were designed for this problem, and can be much more efficient in estimating conditional modes. We try using lasso, with minimal shrinkage, to eliminate the less necessary slope-change variables, and then use MCMC with the reduced parameter set. Common software requires linear models, so we start

off with a lasso Poisson regression for the linear APC model, with a log link, to reduce the variable set going into MCMC. This appears to work well, but may not find the absolutely best-fitting model.

Possibilities for Future Research

Other approaches to efficient estimation are probably worth exploring. One possibility would be to do a pure MLE estimation with no shrinkage, and use that as starting values for MCMC estimation. A number of ways to use MCMC software – number of iterations, number of warm ups, etc. – may be worth testing as well. Multivariate-t shrinkage distributions could be tried too. Highly correlated variables can lead to negatively correlated parameters, and this can be put into the correlation matrix. Even t-distributions with infinite or undefined variances can use correlation matrices. The rank correlations might be well defined, for example.

When extending this to more than two populations, there are choices as to how to set up the model. The third population can use changes from the second, or ignore the second and do separate changes from the first. Either one might work better in given cases, depending on the characteristics of the populations.

Even heavier-tailed distributions than the negative binomial could be tested also, such as the Poisson-Inverse Gaussian distribution. That cannot be implemented in Stan at present, but other MCMC packages, like the R package BayesianTools, present possibilities.

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.” *Proceedings of the Casualty Actuarial Society* 87: 245–303.
- Bühlmann, Hans. 1967. “Experience Rating and Credibility.” *Astin Bulletin* 4:3: 199–207.
- 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.
- Clark, David. 2003. “LDF Curve Fitting and Stochastic Loss Reserving: A Maximum Likelihood Approach.” *CAS E-Forum* Fall: 41–91.
- 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. 2018. “Stochastic Claims Reserving via a Bayesian Spline Model with Random Loss Ratio Effects.” *Astin Bulletin* 48:1: 55–88.
- 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.
- Ginestet, Cedrick E. 2013. “Regularization: Ridge Regression and Lasso.” *Boston University* http://math.bu.edu/people/cgineste/classes/ma575/p/w14_1.pdf.
- Gluck, Spencer M. 1997. “Balancing Development and Trend in Loss Reserve Analysis.” *Proceedings of the Casualty Actuarial Society* 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.
- Guszcza, James. 2008. “Hierarchical Growth Curve Models for Loss Reserving.” *CAS E-Forum* Fall: 146–73.
- Hoerl, A.E., and R. Kennard. 1970. “Ridge Regression: Biased Estimation for Nonorthogonal Problems.” *Technometrics* 12: 55–67.
- 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.
- Mack, Thomas. 1993. “Distribution-Free Calculation of the Standard Error of Chain Ladder Reserve Estimates.” *Astin Bulletin* 23:2: 213–25.
- Mowbray, Albert H. 1914. “How Extensive a Payroll Exposure Is Necessary to Give a Dependable Pure Premium.” *Proceedings of the Casualty Actuarial Society* 1: 24–30.
- Renshaw, A. 1989. “Chain Ladder and Interactive Modelling.” *Journal of the Institute of Actuaries* 116: 559–87.
- 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 C., and Frank R. Ashe. 1983. “Second Moments of Estimates of Outstanding Claims.” *Journal of Econometrics* 23: 37–61.
- Tikhonov, Andrey Nikolayevich. 1943. “On the Stability of Inverse Problems.” *Doklady Akademii Nauk SSSR* 39:5: 195–98.
- Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. “Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and Waic.” *Journal of Statistics and Computing* 27:5: 1413–32.
- Venter, Gary. 2007. “Refining Reserve Runoff Ranges.” *CAS E-Forum* <http://www.casact.org/pubs/forum/07sforum/07s-venter2.pdf>.
- . 2018. “Regularized Age-Period-Cohort Modeling of Opioid Mortality Rates.” *Applied Economics and Finance* 5:4: 12–23.
- Venter, Gary, and Şule Şahin. 2018. “Parsimonious Parameterization of Age-Period-Cohort Models by Bayesian Shrinkage.” *Astin Bulletin* 48:1: 89–110.
- Venter, Gary, Roman Gutkovich, and Qian Gao. 2017. “Parameter Reduction in Actuarial Triangle Models.” *Variance* <https://www.variancejournal.org/articlespress/articles/Parameter-Venter-Gutkovich-Gao.pdf>.
- Verrall, R.J. 1990. “Bayes and Empirical Bayes Estimation for the Chain Ladder Model.” *Astin Bulletin* 20: 217–43.
- . 1991. “On the Estimation of Reserves from Loglinear Models.” *Insurance: Mathematics and Economics* 10: 75–80.
- Ye, J. 1998. “On Measuring and Correcting the Effects of Data Mining and Model Selection.” *Journal of the American Statistical Association* 93: 120–31.