REPLACING MLE WITH BAYESIAN SHRINKAGE

CAS ANNUAL MEETING NOVEMBER 2018

GARY G. VENTER

ESTIMATION

- Problems with MLE known since Charles Stein 1956 paper
- He showed that when estimating 3 or more means, shrinking them all towards the grand mean reduces predictive variance
- James-Stein estimator same as Bühlmann's 1968 method – ratio of within and between variances determines degree of shrinkage
- Only difference is Stein assumed normal distribution MLE, Bühlmann assumed least squares – really the same thing

SOMETHING SIMILAR FOR REGRESSION

- Hoerl and Kennard 1970 paper minimized NLL plus selected λ times sum of squared parameters, excluding the constant term
- Produces shrinkage towards mean for fitted values, since they first standardize all variables to make them mean zero, variance one
- All fitted values are grand mean (= constant) + terms with mean zero
- Showed that there is always some value of λ that produces error variance less than that from MLE – but didn't have a good way to find it
- Application of a general method called regularization used for estimating difficult models, so sometimes is called regularization

NEXT

- That is called ridge regression based on their derivation
- Then in 1990s lasso minimized NLL + λ*sum of absolute values
- Modelers like that because some parameters go to exactly zero, so it is variable selection as well as error reduction
- Cross-validation used as way to select λ
- Divide sample into groups, estimate by leaving out a group, get NLL for omitted group, repeat for all groups. Find best λ.

ENTER BAYESIAN SHRINKAGE

- Giving priors mean & mode of 0 shrinks parameters towards 0
- Normal prior gives ridge regression as posterior mode
- Double-exponential = Laplace prior does this for lasso
- Has an extreme form of cross-validation, leave one out (loo), which makes every sample value an omitted group
- Loo=NLL of the omitted points – a good estimate of the NLL of a completely new sample – so is adjusted NLL like AIC, BIC, etc
- Can be computed very efficiently from the posterior estimates

IT'S NOT YOUR GRANDFATHER'S BAYESIAN ANALYSIS

- Simulation method for posterior (MCMC) does not need specification of the form of the posterior – just likelihood and priors. Good software available – Stan.
- Bayesian estimation not connected to beliefs – priors are part of the model and evaluated results they give
- Might change the priors after you see the posteriors
- Also can put prior on λ to get posterior estimate of it

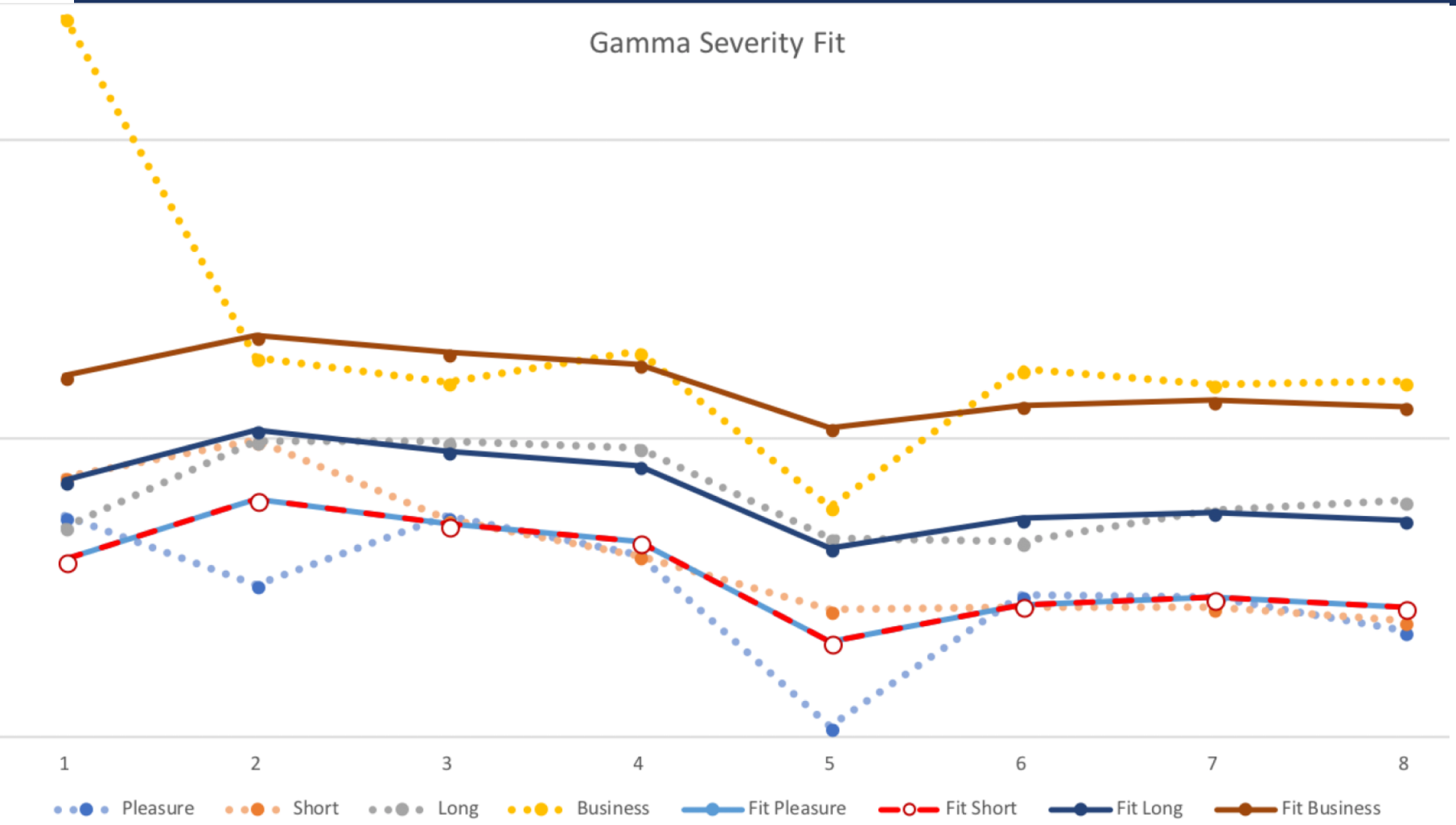
BAYESIAN SHRINKAGE REPLACES MLE AND LASSO, RIDGE REGRESSION TOO

- Reduces estimation and prediction variances over MLE
  - Also MCMC gives parameter distributions which lasso doesn't have
- MCMC usually fairly robust as to selection of priors
- Loo allows choice of λ as well as goodness of fit test; lasso lacking it
- Putting prior on λ usually similar to optimizing loo, and often having posterior for λ does slightly better than any one λ. So just run once.
- Good case that posterior mean better than mode that lasso gives
  - Mode can be overly responsive to features of the given sample

DIRECTLY APPLIES TO REGRESSIONS AND GLM USED FOR CLASS RATEMAKING

- Bayesian shrinkage better than MLE for any multivariate estimation
- Easiest for models with a vector of observations and a design matrix
- Shrinks parameters, maybe some to zero, eliminating some variables
- Can start with lot of variables and this chooses the best combination
- I tried it on data from Fu-Wu paper, Variance 01-02.
- Has loss severity by age and use, with claim count as volume measure
- Fitted multiplicative model with parameter for each age, use, log link

STAN KEPT ALL AGE VARIABLES BUT COMBINED SHORT DRIVE TO WORK AND PLEASURE USE CLASSES. GRAPH ACTUAL VS. FITTED



★ 8 age groups Business use, long drive to work, all other are fitted use groups

★ Smoothness of fitted values here optimized loo penalized NLL

USE IN HIERARCHICAL MODELS

- Generally considered to be models with data at various ”levels” – interpreted broadly
- E.g., levels could be by state, then county within state then municipality within county, with an additional variable of sourcing agent, or other things – distance to fire station, …
- A lot of states could start with expected zero difference from countrywide, then many counties with zero difference from state, similar for municipalities, and sourcing agent
- Could do the same with interaction terms between variables – a lot shrink away, some not
- Could have age group classes, then individual ages, but with many of those shrinking to the group average, etc. – can help to define age groups as well
- But levels could be layers of modeling assumptions too – make prior for λ, which is prior for β
- Also MCMC allows non-linear models – could have additive plus multiplicative model

USING ON LOSS TRIANGLES – ADDITIVE OR MULTIPLICATIVE MODELS

- Need all row, column factors so don't want to eliminate them
- One approach based on Barnett, Zehnwirth's 2000 CAS paper:
  - Fit piecewise linear curves to parameters in each direction, shrink slope changes of curves. Now can do it with Bayesian shrinkage
- Şahin & I do this for Bayesian shrinkage in mortality triangle model in 2018 Astin paper (like reserving but bigger triangles)
- Gao, Meng 2018 Astin paper similar for reserve model, but fits cubic splines instead of piecewise linear curves

DETAILS OF THIS FITTING

- Want to put in regression form, so string out the rectangle into a column, keeping track of row and column for each cell
- Regression would make a (0,1) dummy variable for each row, column, and diagonal, taking value = 1 at cells they affect, so coefficient * dummy gets to cells for right rows and columns
- Slope changes are 2nd differences of parameters and add up to the parameters – just need more complicated dummies
- The dummy for row u in a cell from row j takes value:
  - Max(0, 1+j-u). Same for columns, diagonals - numbered from 1

MODEL USING ROW, COLUMN, DIAGONAL PARAMETERS

- Mean for log of data with row, column, and diagonal parameters p_w , q_u , and r_{u+w} :
- $\mu_{w,u} = c + p_w + q_u + r_{u+w}$ – usually with a log link
- Used $e^{\mu_{w,u}}$ as the $a_{w,u}$ parameter of a gamma distribution with mean = $a_{w,u}b$ and Var = $a_{w,u}b^2$, with b constant across cells. Variance = b*mean, like in ODP (can't do this in GLM)
- Exponentiation of p_w, q_u gives the row and column factors
- Y = Xβ is the fitted $\mu_{w,u}$ vector.
- Same thing works when dummy variable is a slope change dummy max(0, 1+j-u).
- Still e^Y is the vector of gamma $a_{w,u}$ parameters
- With shrinkage, resulting row, column, diagonal factors are on piecewise linear curves

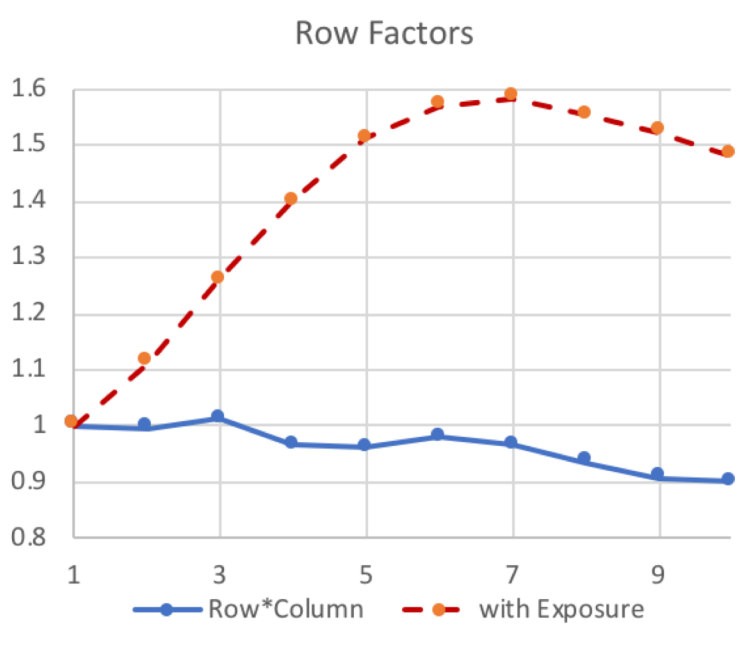
EXAMPLES

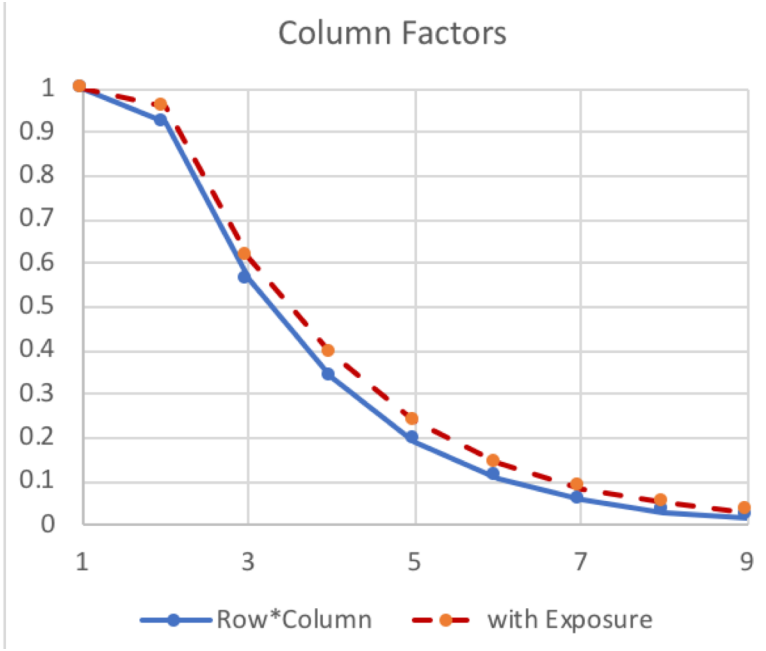
- Two 10x9 paid loss ratio triangles for US commercial auto
- Fit row-column (accident year, lag) and column-diagonal (lag, payment year) models first, then tried all 3 directions
- Took out any variables with parameters near zero with wide estimation ranges if doing so did not hurt loo penalized loglikelihood measure
- For State Farm, AY-lag model fit best by loo, for USAA lag-PY best
- Each model had two variables eliminated – so just continues existing piece-wise linear slope at those points
- Adding third direction didn't help either model

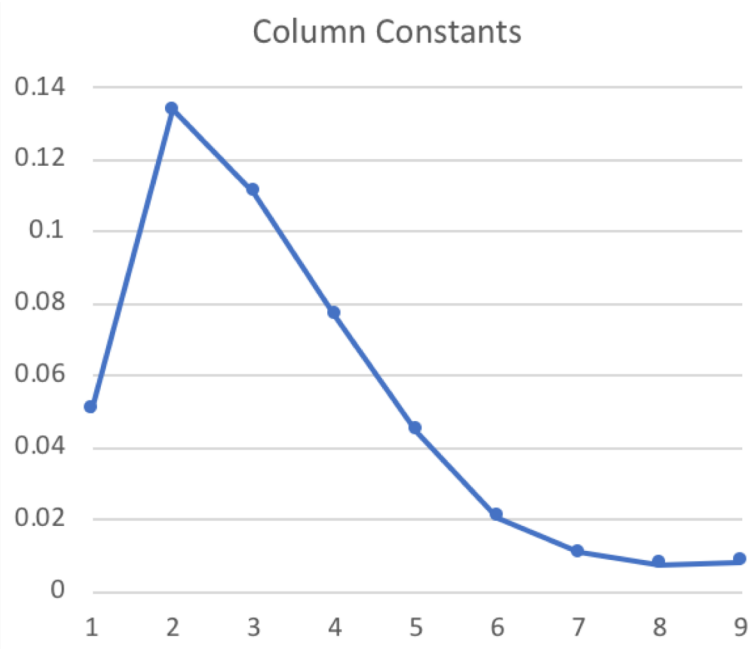
ADDITIVE ADJUSTMENT

- Muller's 2016 Variance paper suggested adding a factor for each column, which is multiplied by exposure by row and then added to the row*column mean
- Like adding in a Cape Cod model. The factor model $a_{w,u} = A_w B_u C_{w,u}$ goes to
- $a_{w,u} = A_w B_u C_{w,u} + D_u E_w$, with exposure E_w by AY, and lag factors D_u
- Again use 2nd difference dummies for the logs of the new factors
- Since triangle already divided by premium, made that the exposure and $E_w = 1$.
- This improved loo for both triangles, but for USAA none of the original lag parameters was then significant so became a purely additive model $a_{w,u} = C_{w,u} + D_u$

STATE FARM FACTORS, 2 MODELS





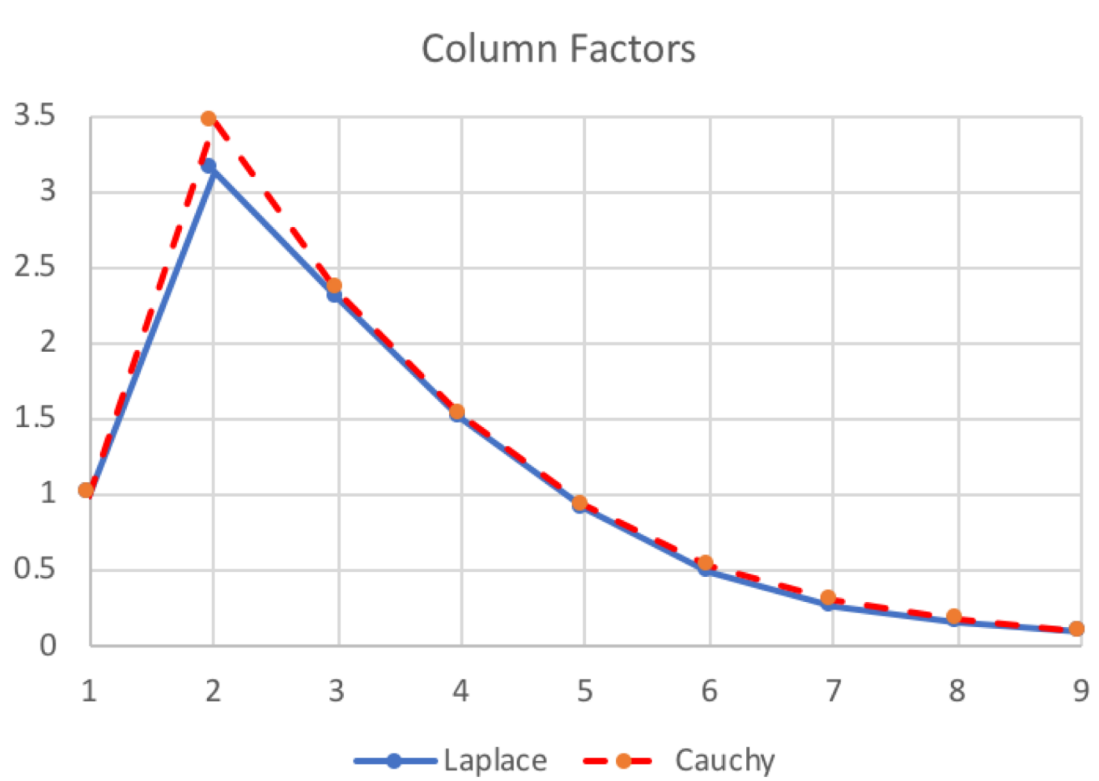


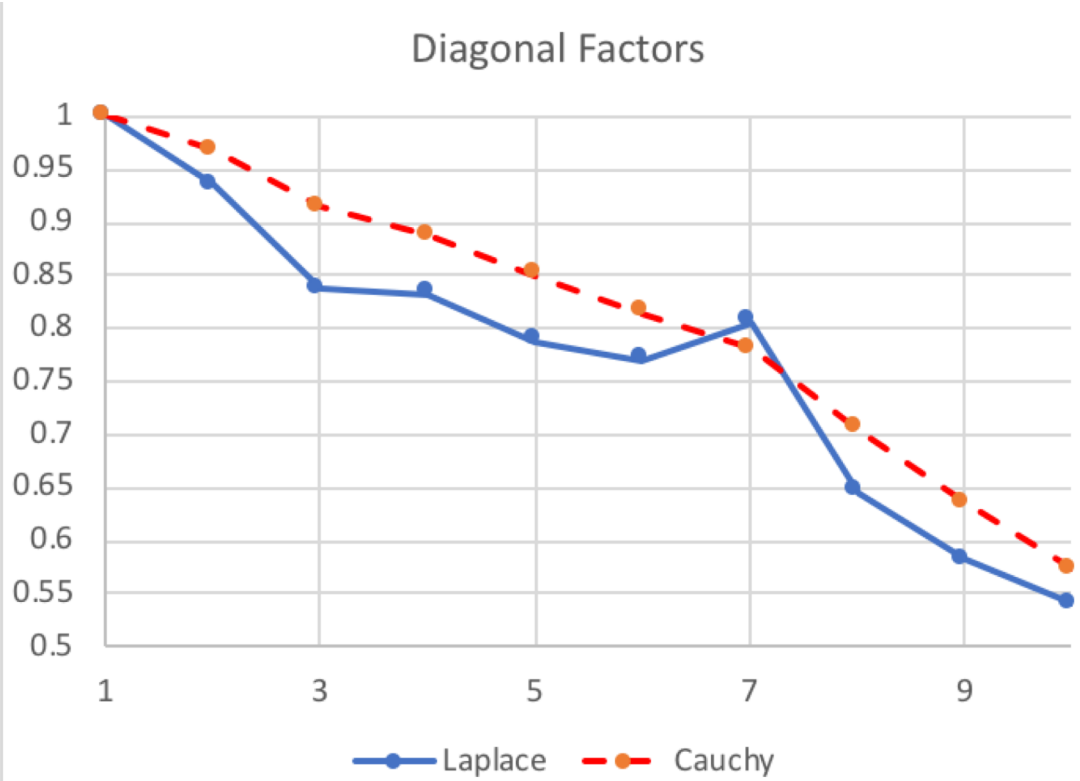
Exposure term makes each fitted value a linear model of the row factors, not just a multiple. Picked up acceleration of payments in more recent years.

WHICH SHRINKAGE PRIORS?

- Used double-exponential prior on all the 2nd difference parameters – like lasso
- But Student's-t with one dof, called Cauchy distribution, becoming popular too
- Heavier tailed but also stronger push towards zero – most parameters shrink more but some could be a lot bigger
- Tends to produce more parsimonious models but can have better fits by loo
- Tried this for USAA model before exposure adjustment – fit slightly worse but more parsimonious according to loo parameter penalty
- If process generating data is subject to change, this could be a better model
- Student's-t with two dof tried in other models, and seems to work very well
- Double exponential very similar to t with 6 dof – matches all 5 moments of that t

CAUCHY VS. DOUBLE EXPONENTIAL





CONCLUSIONS

- Bayesian shrinkage has lower predictive variance than MLE – can use instead of MLE to get better predictions in almost all models
- Recent advances include goodness of fit measure; direct fitting without a lot of shrinkage choices; no need to specify posteriors – so as easy as MLE
- Good R packages available
- Fitting process like for MLE – try models, compare fits
- Flexible choice of distributions and model forms like additive-multiplicative
- Fit curves to factors using 2nd differences for row-column models

REFERENCES

- 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.
- Hoerl, A.E., and R. Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." Technometrics 12: 55–67.
- 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, pp 373–99.
- Barnett, Glen, and Ben Zehnwirth. 2000. "Best Estimates for Reserves." Proceedings of the Casualty Actuarial Society 87: 245–303.
- Venter, Gary, and Şule Şahin. 2018. "Parsimonious Parameterization of Age-Period-Cohort Models by Bayesian Shrinkage." Astin Bulletin 48:1: 89–110.
- 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.
- Muller, Thomas. 2016. "Projection for Claims Triangles by Affine Age-to-Age Development." Variance 10:1:121–44.