REGULARIZED REGRESSION FOR RESERVING AND MORTALITY MODELS

GARY G.VENTER

TODAY

- Advances in model estimation methodology
- Application to data that comes in rectangles
- Examples

ESTIMATION

- Problems with MLE known since Charles Stein 1956 paper
- Showed that if estimating 3 or more means, shrinking them all towards the grand mean reduces predictive variance
- James-Stein estimator same as Bühlmann's 1968 method
- Only difference is they assumed normal distributions, he 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, except constant term
- A kind of mean shrinkage, especially since they first standardize all variables to make mean zero, variance one
- So all fitted mean values are constant plus term with mean zero
- Showed that for some λ, error variance is less than from MLE
- Application of already known general method called regularization used for estimating difficult models

NEXT

- That is called ridge regression based on their derivation
- From 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
- Cross-validation emerged as way to select λ
- Divide sample into groups, estimate all but one group, get NLL for omitted group, repeat for all groups. Find best λ.

ENTER BAYESIAN SHRINKAGE

- Making priors mean zero shrinks parameters towards zero
- Normal prior gives ridge regression as posterior mode
- Double-exponential = Laplace prior does this for lasso
- Extreme form of cross-validation, leave one out (loo) makes every sample value an omitted group
- The NLL of the omitted points a good estimate of NLL of a completely new sample – fitting the population
- 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.
- Priors no longer connected to previous beliefs – they are part of the model and evaluated on how they do
- 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

- Reduces estimation and prediction variances
- Usually fairly robust towards selection of priors
- Loo allows choice of λ as well as goodness of fit test
- Putting a prior on λ usually gives results similar to maximizing loo, and often a posterior distribution of λ slightly better than any single λ.
- Good case that posterior mean better than mode
  - Mode can be overly responsive to issues of the given sample

MODELING CONVENTIONS FORMING

- MLE has somewhat accepted model-building practices and similar ones are developing in the Bayesian shrinkage world
- Priors needed on non-shrunk parameters, like λ, the constant term and the distribution's shape parameters
- I like uniform priors, but on log of a positive parameter
- I start with fairly wide priors but reduce them towards the range of the posterior distribution or if run bombs
- If posterior concentrated near edge of range, I widen it
- Variables getting wide range around 0 taken out, loo checked

USING ON RECTANGLES OF DATA

- A lot of data comes in rectangular datasets
- Estimate parameters for rows, columns, and diagonals, and multiply or add them to estimate fitted means for cells where they meet
- Called age-period-cohort models in statistical literature since Greenberg etal. 1950 JASA paper
- Variables are dummies so you don't want to shrink or eliminate them
- One approach is in Barnett, Zehnwirth's 2000 CAS paper
  - Fit piecewise linear curves to parameters in each direction, shrink slope changes
- Şahin & I do this for Bayesian shrinkage in mortality model in 2018 Astin
- Gao, Meng 2018 Astin paper similar for reserve model, but fits cubic splines

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 goes to cells for right rows and columns
- Slope changes are 2nd differences of parameters so 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

- Mean for log of data with row, column, and diagonal parameters p_w , q_u , and r_{u+w} :
- $\mu_{w,u} = c + p_w + q_u + r_{u+w}$
- Actually 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
- Then exponentiation of the parameters becomes factors
- But since data strung out into a column Y, the variables are dummies for each row, column, and diagonal, all in design matrix X, with parameter vector β, so $Y = X\beta$ is the fitted $\mu_{w,u}$ vector.
- And each 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 smoothed curves

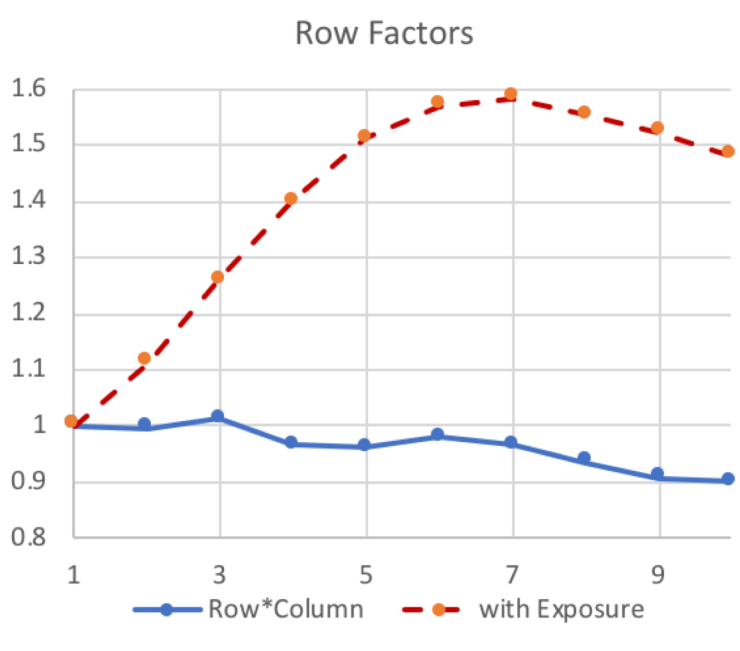
EXAMPLES

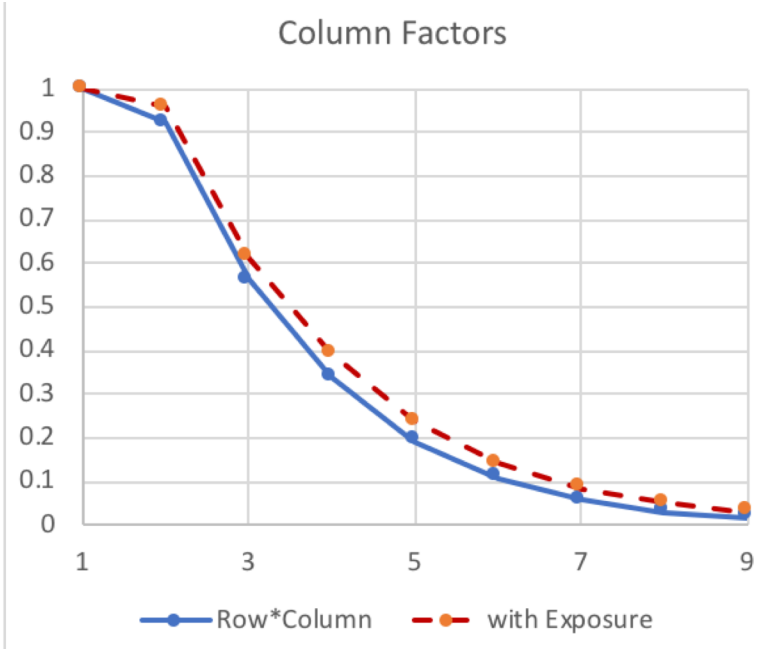
- Two 10x9 paid loss ratio triangles for US commercial auto
- Fit row-column (accident year, lag) and column-diagonal (lag, payment year) models first, then tried all 3 directions
- Eliminated variables with parameters near zero and wide estimation ranges positive and negative if doing so did not hurt 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 – just continues existing piece-wise linear slope at those points
- Adding third direction didn't help either one

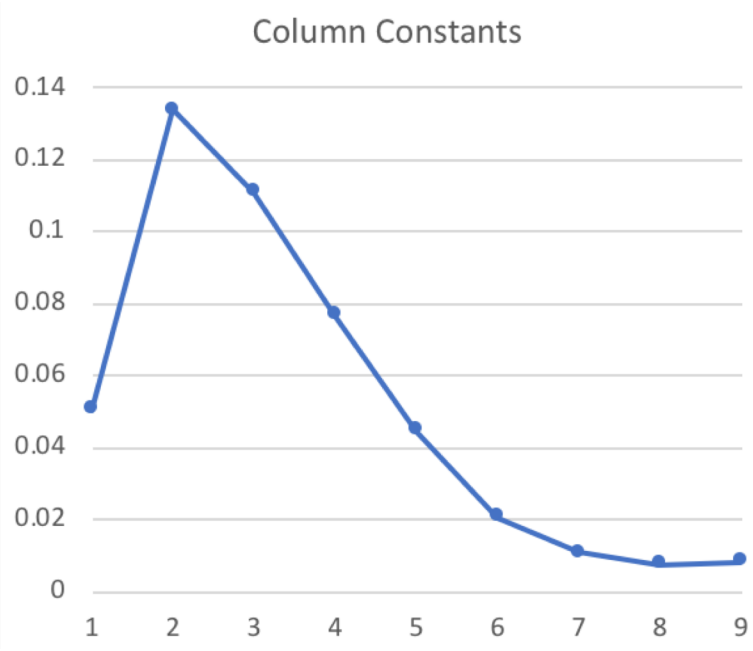
ADDITIVE ADJUSTMENT

- Muller's 2016 Variance paper suggested adding an exposure adjustment
- This needs a factor for each column, which is multiplied by exposure by row and then added to the product of the factors from the multiplicative model
- The factor model $a_{w,u} = A_w B_u C_{w,u}$ becomes
- $a_{w,u} = A_w B_u C_{w,u} + D_u E_w$, 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 fits by the loo measure for both triangles, but for USAA none of the original lag parameters was then significant so this became a purely additive model $a_{w,u} = C_{w,u} + D_u$

STATE FARM FACTORS, 2 MODELS





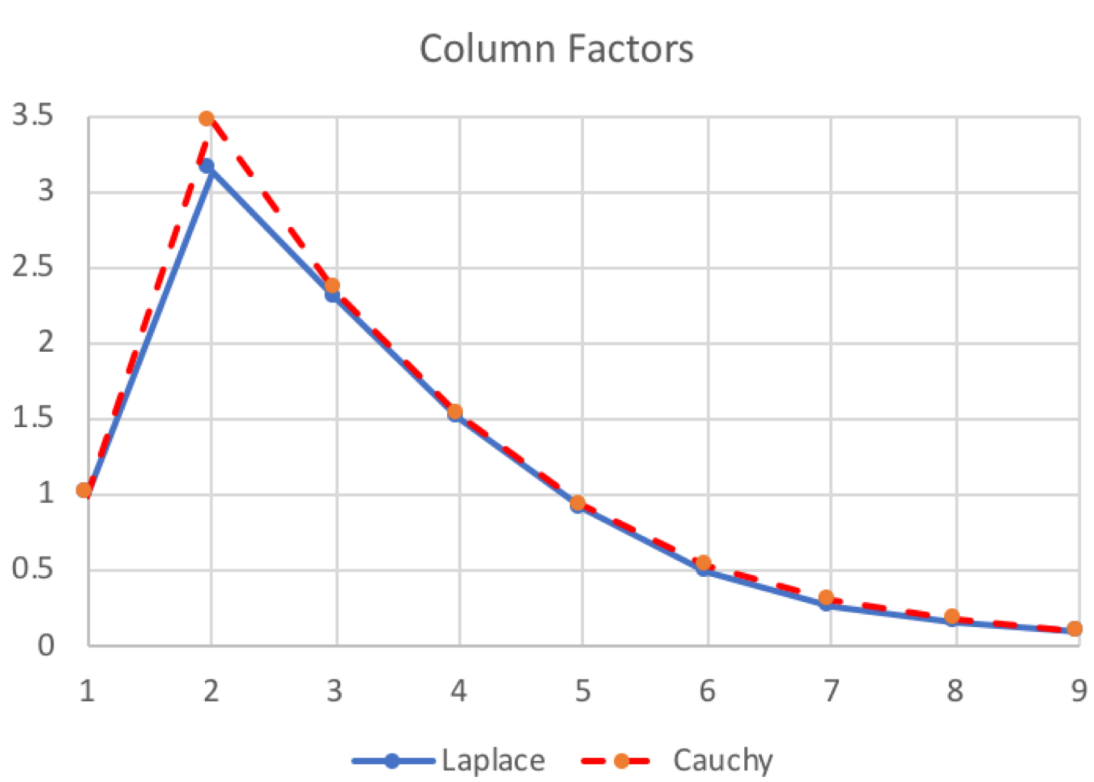


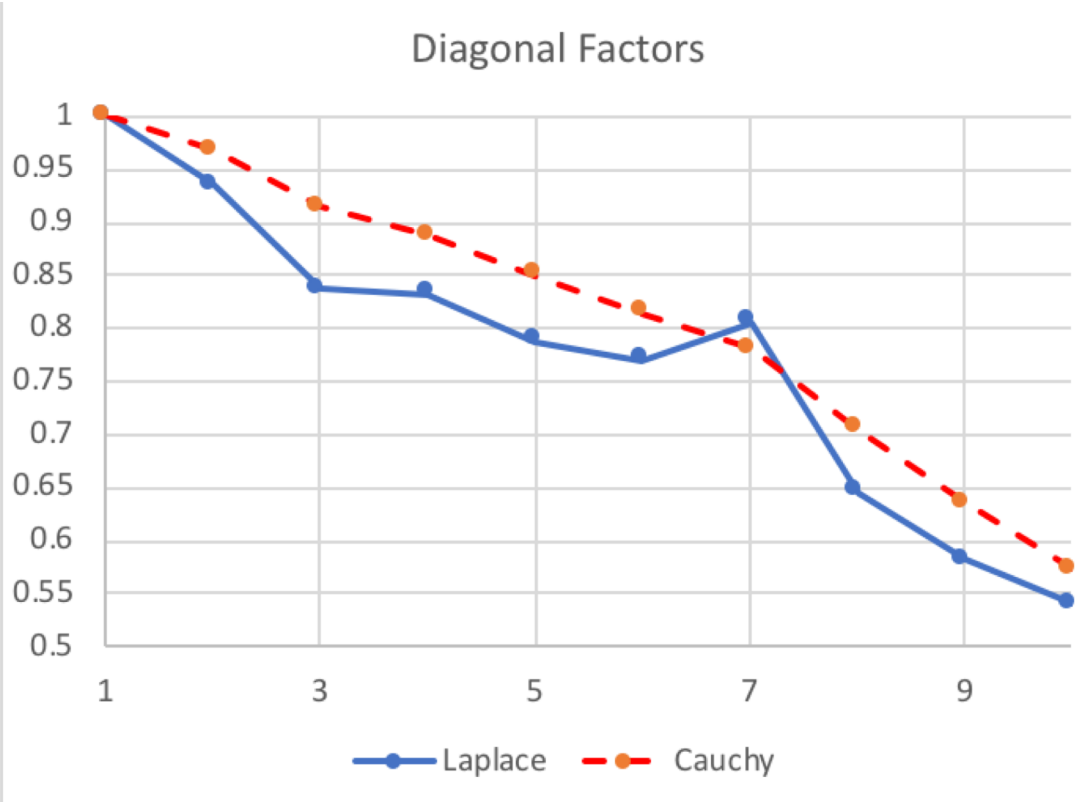
Exposure term makes each fitted value a linear model of the row factors, not just a multiple. Picked up acceleration of payments in more recent years.

SHRINKAGE PRIORS

- Usually used double-exponential prior on all the 2nd difference parameters
- Corresponds to lasso
- But Student's-t with one dof, called Cauchy distribution, becoming popular too
- Heavier tailed but also stronger push towards zero
- Most parameters shrunk more than with double-exponential, 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

CAUCHY VS. DOUBLE EXPONENTIAL





CONCLUSIONS

- Bayesian shrinkage has lower predictive variance than MLE
- Recent advances include goodness of fit measure; direct fitting without a lot of shrinkage choices; no need to specify posteriors
- Good R packages available
- Fitting process like for MLE – try models, compare fits
- Flexible choice of distributions and model forms like add-mult
- 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.