Simplified Smooth Splines for APC Models

Gary Venter, Columbia University, gv2112@columbia.edu

Abstract

Smoothing splines are splines fit including a roughness penalty. They can be used across groups of variables in regression models to produce more parsimonious models with improved accuracy. For APC (age-period-cohort) models, the variables in each direction can be numbered sequentially 1:N, which simplifies spline fitting. Further simplification is proposed using a different roughness penalty. Some key calculations then become closed-form, and numeric optimization for the degree of smoothing is simpler. Further, this allows the entire estimation to be done simply in MCMC for Bayesian and random-effects models, improving the estimation of the smoothing parameter and providing distributions of the parameters (or random effects) and the selection of the spline knots.

Keywords: Smoothing splines, Ridge regression, APC, MCMC

APC spline regressions

Regression parsimony usually removes less significant variables – those with small parameters that provide little improvement to the fit. Leaving those out is like making them zero. APC models estimate the effects of every age, period, and cohort, so these would have to be close to zero to leave out. Often, that is not possible. But what is feasible is reducing the differences between successive parameters, or maintaining a pattern of how parameters change – linear or cubic trends, for example. This is where splines come in. Another alternative is parametric curves across the parameters, but often these do not fit well.

Cubic and linear splines have basis design matrices – matrices of dummy variables for a regression to fit splines in each direction. There are variables for each APC parameter, except for enough left out to make the design matrix non-singular. If there are no further constraints, the spline regression fit would be the same as from regular regression. The spline curves could be very bumpy to fit through all the original parameters. Smoothing splines constrain the regression by minimizing the negative loglikelihood (NLL) plus a roughness penalty, giving smoother curves.

An early proposal for a roughness penalty on smoothing, from Whittaker (1922), is the sum of squares of the third differences of the smoothed values. Recently the integral over 1:N of the squared 2nd derivative of the spline curve has been a popular roughness measure. For cubic splines, this is a closed form but fairly complicated function of the parameters. For linear splines a comparable formula is the sum of the squared 2nd differences of the curve values at the points 1:N. With the right basis, this is the sum of the squared parameters. The 2nd derivative measure makes fitting each spline a numerical optimization, even when the regression can be done in closed form. The proposal here is to also penalize the sum of squared parameters for cubic splines. The smoothing is similar to using the 2nd derivative measure – fits with lower sums of squared parameters have lower 2nd derivative integrals, and vice versa.

Formally, let β_0, … β_N be the coefficients of the spline fits, with constant β_0, and let f(x) be a spline. Second-derivative fitting with smoothing constant λ minimizes:

$$NLL + \lambda \int_1^N f''(x)^2 \partial x$$

The proposed alternative is minimize:

$$NLL + \lambda \sum_{j=1}^{N} \beta_j^2$$

This is actually the formula for ridge regression with shrinkage λ, although sometimes the sum is taken starting at $j = 0$. Hoerl and Kennard (1970) proved that there is always some $\lambda > 0$ for which ridge regression has a lower predictive variance than MLE's. This is similar to Stein 1956, who showed that when estimating three or more means, shrinking the estimates towards the overall mean to some degree will always improve the predictive variance. Both bias the estimates, usually towards the grand mean, but that is less important than improving the accuracy. Improved predictive accuracy is what you want from splines.

Ridge regression has a closed-form estimation equation. For regression with design matrix x and observations y:

$$\beta = (x'x)^{-1}x'y$$

Let J be the diagonal matrix with upper right value = 0, for the constant. Then the ridge regression estimate is:

$$\beta = (x'x + \lambda J)^{-1}x'y$$

This can be done in a spreadsheet. Fitting is done excluding one group at a time. NLLs are measured on the omitted groups, and the sum of those is the cross-validation NLL, which is minimized over λ. This is a one-variable nonlinear minimization and can also be done in a spreadsheet. It is easier in popular programming languages like R and Python.

The cross-validation NLL is an estimate of the NLL excluding sample bias, which is understatement of the NLL coming from measuring it on the sample used to estimate the parameters. Penalized likelihood measures like AIC, BIC, HQIC, etc. also try to eliminate the sample bias, and cross validation can be considered a penalized likelihood measure for shrinkage estimation. Shrunk parameters do not use up as many degrees of freedom, so penalizing NLL on parameter counts does not work for shrinkage. Such measures are estimates of sample bias, not exact measurements of it. Statisticians do not like estimating parameters by optimizing penalized likelihood because that risks just finding the parameters with the greatest under-estimation of sample bias. They are still good measures for comparing a few models, however.

Markov chain Monte Carlo estimation (MCMC) gives a different way of estimating

λ. In Bayesian or random-effects estimation, the model would include distributions for the parameters or random effects $\beta_1, \ldots \beta_N$, e.g., each $\beta \sim \text{normal}(0,1/\lambda)$. Call their combination $p(\beta)$. Bayesians would also specify fairly non-constraining distributions for $\log(\beta_0)$ and $\log(\lambda)$. In random effects these are usually considered frequentist parameters, without distributions, but there appears to be little reason they could not be made into random effects themselves, with postulated distributions. In either case, the joint likelihood is $p(\beta)*p(y|\beta,\lambda) = p(\beta,\lambda) = p(\beta,\lambda|y)*p(y)$, by definition of conditional distribution. Here $p(y)$ is not known. MCMC is a way to generate samples from the product of two distributions known only up to an undetermined constant. Thus it can produce a sample of $p(\beta,\lambda|y)$. Frequentists usually estimate just the mode of this distribution, but it seems they could get the whole distribution MCMC by making β_0 and λ random effects, and estimate the mean of $\lambda|y$ without cross validation as Bayesians do.

Implementation example

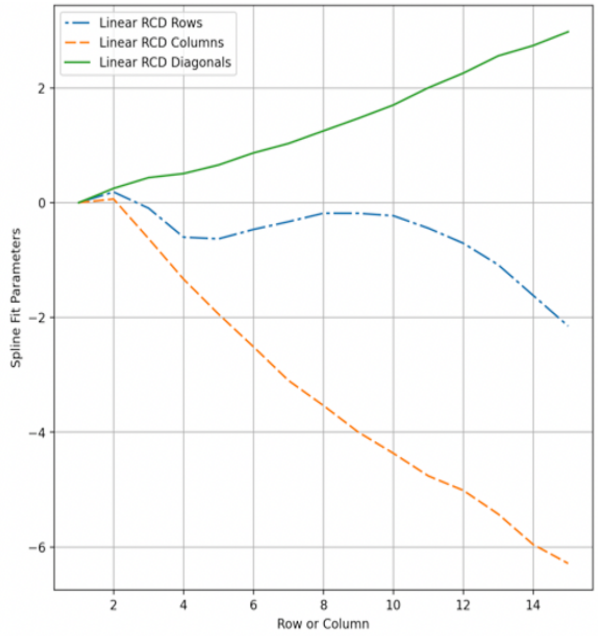
For a linear-spline design matrix, the variables estimated are the 2nd differences of the parameters in each direction. The 1st differences sum those up, and the A, P, or C parameters sum up the 1st differences. That ends up counting the 2nd difference parameters additional times for later observations. For an observation at age k, the age j variable ends up as $(1+k-j)_+$, and the same for periods and cohorts k and j. Usually the 1st A, P and C variables are left out to avoid a singular matrix when there is a constant term. One more variable needs to be omitted if all three directions are included, as discussed below.

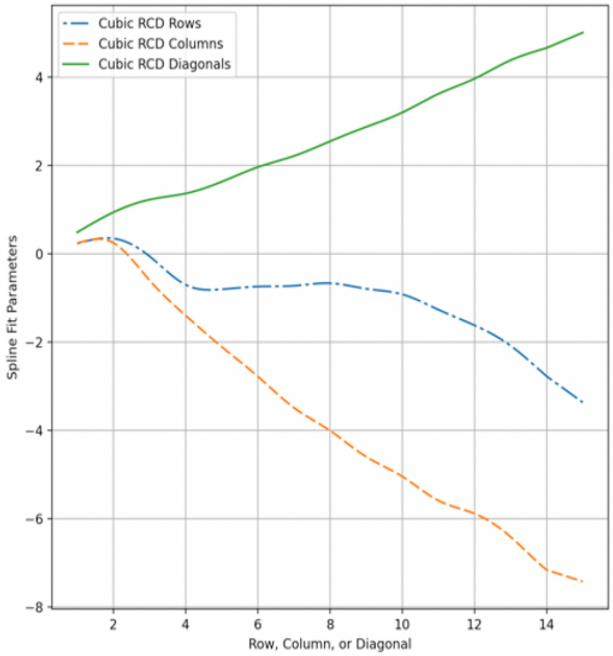
For cubic splines the design matrix here takes the first variable as the constant, the second in each direction as the A, P, or C number k for that direction for each observation, and the 3rd to Nth as $(2+k-j)_+^3$ for $j = 2, \ldots N$. Only one constant is used for the regression, so there are N–1 variables in each direction. This is a simplification of the basis given by Hastie, Tibshirani, and Friedman (2017), a derivation which is at https://stats.stackexchange.com/questions/172217/why-are-the-basis-functionsfor-natural-cubic-splines-expressed-as-they-are-es . For interpolating the cubic splines, use this basis for any real k in the interval. A disadvantage of the simplification is that the interpolation does not work in the last interval [N–1, N] and the spline does not work for extrapolation.

There is a linear relationship among the three directions. One more variable has to be eliminated to have a non-singular design matrix. This can be done by estimating the model using two directions only, then leaving out the variable with the lowest absolute value. This is at a point where the cubic or linear curve does not change shape much. Sometimes modelers leave out an entire direction because of the linear relationship, but that is like assuming there are no effects in that direction. This can be checked with residual analysis, and does not usually hold.

As an example, splines are fit to the logs of insurance loss payment data using cross validation. Payments are by year of accident (cohort=row) and years since accident (age=column), with year of payment (period) on SE-NW diagonals. Payments decrease with age, and increase from inflation by year of payment. The cohort size is volume increased by inflation. This data is an incomplete square with fifteen years in each direction. Linear and cubic splines were fit, first to AC data, then the smallest variable eliminated, then the period variables were included. For linear splines, the last cohort was eliminated, and for cubic splines it was the 2nd-to-last cohort. Leave-one-out cross validation was used, where every observation is taken as a left-out subsample. The linear splines gave a slightly better cross-validation NLL. The splines at the best λ's are graphed below, with the cubic splines interpolated to show the curves. The linear splines are fairly smooth themselves. Curve shapes barely change at several points. Those parameters are close to zero and their knots are effectively eliminated. Optimizing knot choice can be an awkward additional calculation in typical spline models.

The graphs look similar, but the scales are different, with the fits producing different estimated tradeoffs among the other directions. The parameters in each direction cannot be interpreted individually, which is not unusual.





References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2017. “The Elements of Statistical Learning.” Springer Corrected 12th Printing.

Hoerl, A. E., and R. Kennard. 1970. “Ridge Regression: Biased Estimation for Nonorthogonal Problems.” Technometrics 12: 55–67.

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.

Whittaker, E. W. “On a New Method of Graduation.” Proceedings of the Edinburgh Mathematical Society , Volume 41, February 1922 , pp. 63 – 75.