Ensembling Age-Period-Cohort Estimates by Averaging Multiple Window Constraint Models using Markov Chain Monte Carlo Methods

rounding makes equal probability --- NOT floor

# Abstract

Key words: Computational Sociology, Methods, Life Course, Bayesian, Culture

# Introduction

# Background

Both classic and modern studies have proposed different effects across three social dimensions of time: age, period, and cohort (Ryder 1965). These dimensions of time operate on individuals in different ways as they flow through social space. Age effects are driven either by the common experiences associated with the biological process of aging (Jackson, Weale, and Weale 2003), or persistent age-structuring institutions, like high school or retirement (Leisering 2003; Moen 2014). Period effects are responses of everyone to contemporaneous social experiences, like recessions or wars (Lam, Fan, and Moen 2014). Finally, cohort effects are socialization effects. They are broad-based historical events that stick to the populations experiencing them, even after the event has long since ended (Vaisey and Lizardo 2016). In contemporary research, cohorts are virtually always defined in terms of birth year.

These three dimensions of time sit at the center of some recurrent debates. For example, cultural sociologists are divided over whether the main driver of culture is a period process (cultural fragmentation) or a cohort process (acquired dispositions) (Vaisey and Lizardo 2016). Clinical reasearchers and biodemographers are looking for biological age—indicators of biological deterioration—but argue about confounding from cohort changes (Jackson et al. 2003). And yet others are concerned about separating long-term and short-term impacts of important events, like the Great Recession (Burgard and Kalousova 2015).

Unfortunately, it is not trivial to statistically estimate the unique effects of age, period, and cohort which would allow analysts to evaluate evidence that may settle these questions. The fundamental problem is one of statistical identification: these three dimensions are linearly dependent. Two of the dimensions define the third. If a researcher knows an individual’s age, and the year of the survey, cohort is also defined. Because these variables are exactly colinear, they are not estimable using classic statistical techniques. A number of solutions have been proposed to address this conundrum, these include traditional methods like block/window constraints (Anon. 2005), and new methods, including statistical transformation (the “intrinsic estimator”), and random effects models (Yang and Land 2013).

The past several years have seen a resurgence in debates surrounding on methodology (Bell and Jones 2017; Luo 2013; Luo and Hodges 2016). Perhaps the most common argument is that the “constraints”, *i.e.* assumptions used to break the APC identity, produce unknown biases into the models. These particular complaints represent the extreme case of multicollinearity, a well-known, though difficult problem in statistical theory (Wooldridge 2009, at p. 95-98), with numerous cautions in applied texts that separating highly colinear effects is difficult (Camm 2016, at p. 328). Viewing the APC identity problem as an issue of multicolinnearity in a Bayesian perspective provides a simple way to understand and adcress the problem. Ultimately, multicollinearity significantly reduce the ability of the data to supply information to produce the estimates, potentially introducing sensitivity to the prior (Gelman 2014, at p. 305-306), including assumtpions of linearity. To put it another way, the critique of APC models in general, and window constraints in particular (Luo and Hodges 2016), is that different sets of assumptions lead to different (and inconsistent) results. But, what if instead of focusing on finding a *best* fitting model, we use a simple nonlinear approach (block or window modelling), and hundreds of assumptions? Ensembling is the combination of estimates from different models subject to some weighting scheme. Bayesian Model Averaging (BMA) allows us to do just that with the APC problem: estimate hundreds of models with differeing assumptions, and then combine them, weighted by a measure of the model’s posterior probability.

# Bayesian Model Averaging

The theoretical backdrop of BMA applies to this curcumstance quite well. In principal, there is no one true (or best) model; instead estimates are conditional on models from the modeling space (), and have a posterior distribution, which is calculated as a weighted average of all models (Raftery 1995:144–45). It has been applied in diverse areas from weather forecasting to biology to social science (Fragoso and Neto 2015) . BMA operates under the simple fact that any particular estimate, including effect size and significance, have a posterior probability distribution which is calculated as “an average of the posterior distributions under each of the models considered, weighted by their posterior model probability.” (Hoeting et al. 1999). The major difficulties for BMA are (1) how to sample models to test, and (2) how to calculate the posterior model probability given the data (or where is the model, and is the data). For model selection, we use a Markov Chain Monte Carlo algorithm, similar in spirit and design to the reversible jump, MC3,and SVSS algorthims citations, but specific to modelling piecewise constant breaks in a set of highly collinear continuous variables. We call it the Stochastic Window Sampler (SWS). To implement SWS, we use the following steps:

1. Define a jumping distribution , so that is non-zero for all possible window constraints.
2. Specify a starting model, , and ellicit priors for models in .
3. Given that the chain is in state , draw , and accept it with probability

otherwise, retain . 4. After a sufficient number of iterations, use all accepted models, , and combine them using BMA to estimate APC effects.

## Step One: Defining a Window Constraint Sampler

In terms of window constraints, the target model, which is inestimable, is built by estimating each unique value of age, period, and cohort, as a dummy variable series:

is an estimated effect. , and index unique values for age, period, and cohort, and , , and stand for matricies of dummy variable series for age, period, and cohort. This model is unidentified, because, as with the continuous case, the dummy variables in any two of the matricies above fully condition the third matrix. In other words, the indicator variable in is a function of the indicators in and (in mathematical terms,the probability that any given cohort dummy variable is one or zero is exctly dependent on the values of A and P, so that ) is always exaclty zero, or exactly one, depending on the values of and ).

We can break this dependencey, however, by transofrming , , and —preferably without resulting to some prespecified arbitrary set of constraints (Gelman 2014, at p. 366). How do the window constraints break the linear dependency? By way of example, we can construct an age dummy variable series where is sliced into two groups based on some cut-point so that, for example, individuals who are older than 30 have a dummy variable of 1. This would identify a binary dummy variable for an older and a younger group. We can generalize this expression to an arbitrary vector of cut-points, with subscript so that the window constraints of in equation \_\_ are as follows:

If the vector and have the following properties: , , and , then can describe any posible sets of window restrictions for age. The first three requirement ensures that the dummy variable series is fully defined across the entire range of the continuous variable . In particular, the first restriction ensures that the dummy variable series with the oldest ages in contains the maximum value of . Similarly, the next two restrictins ensure that the smallest window constraint in contains the minimum value of . The final constraint requires that have at least three elements. Three elements in defines a dummy variable. Using the example above, if has a range of 5 to 50, then the dummy variable distinguishing older and younger respondents can be defined by .

Generalizing cross all dimensions of APC, permuting three similar vectors (say ) will describe any model for any possible window constraints detailed in equation 1. Accordingly, all permutations of G, as defined above, constitute the model space of window constraints (). For any given set of APC variables, is finite, but it can become large. For example, a set of continous ages, periods and cohorts weach with a range of 5 allows for 3,375 unique window models[[1]](#footnote-26). Bump them all to range of 6 and there are 38,304 unique window models. In any model space, only 1 model is inestimable because of perfect collinearity. Based on the statistical criticisms lobbied at APC models, critiques arge that this target model is (theoretially) the least biased, becasue it imposes the fewest assumptions on the models (Bell and Jones 2017; Luo and Hodges 2016). Although it is almost certainly the most parsimonious. The question is how to best use information from some subset of possible models in to estimate unbiased APC effects of the target model. Bayesian Model Averaging (BMA) provides a straightforward way to combine models. need transition

There are two features of that make an MCMC method an attractive way to sample window models. First, all of the models can be described by a set of three vectors, . Further, can be decomposed into two parts: (1) the number of windows, and (2) the break points for each of the windows. The probability that a break-point for a window constraint occurs between two numbers can be represented by a Simplex (a vector of numbers between 0 and 1 which sum to 1). If we multiply a Simplex vector by the number of window breaks, we have a value that represents the probablity a window break occurs between two of the continuos vairables. A scalar for the window breaks, and a a Simplex distribution are easy to sample from, and it is easy to generate a jumping distribution to explore a set of scalars and across the Simplex.

*Sampling the Window Groups (G).* As noted above, there are two basic features of vector . First, is the number of window breaks, or the rank of , and second is the location of the window breaks, or the values of . We use a uniform distribution over the range of A,P, or C to sample the number of window breaks. This is simple, straightforward, and is a noninformative prior distribution. We use the Dirichelet distributon to sample the location of window breaks. The Dirichelet is well-suited to this task, and is commonly used in classification tasks citation. In additon, Taddy et al. (2015) show that the Dirichelet is a natural prior distribution for variable selection and value-splitting in classification and regression trees (CART) algortithms, common for machine learning. Our approach is a similar, as developing is fundamentally a classificaiton task which aggregates similar ages, periods, and cohorts. Finally, a classical description of the Dirichelet distribution is one of “string-cutting.” This is preciesely the task in sampling —to cut a continuous string of variables into some subset of shorter lengths, where the estimates are assumed to be equal.

To implement our sampling scheme, we use two sets of auxiliary variables for each dimension () of APC to match the decompostion of described above. The fist is a Simplex () for each unique value of A, P, or C. We draw a unique simplex for each dimension , with the same length of unique elements in () using the Diricelet distribution. The second auxiliary variable is a scalar for each dimension, . We use a uniform distribution to sample . is simply the product of times . More formally, we consruct as follows.

this is a confusing paragraph

Where is the index number for the APC effects (the , , or of equaiton 1) is the continuous vector of values (the ,, or ). We take the floor rounded value of the product of the cumulative sum of and , which provides a stirng of integers from 1 to . These integers identify continuous groupings of windows. Back to the binary example above, the floor-rounded cumulative sum, would be a vector of 26 ones, followed by 20 twoes. The starting value is the minimum value of the , , or , but each subsequent value of the vecotr is the value of age where the vector of indexes in times change. is simply defined as the unique elements of , begins before the minumum value, and ends after the maximum value. In other words , sample d in this manner, describes a unique window model of the form laid out in Equation \_\_.

There are a number of possibilites for the jumping distribution of . Specifically, we use a multivariate normal distribution with a specified variance to update from one set of models to the next. As such, this produces a random walk on the auxiliary parameter which sequentially updates the probability of window breaks, by adjusting the probability of a window break at each value of APC. Based on the acceptance probability described below, the chain will converge to better fitting sets of window models as the chain approaches infinity.

## Step 2: Specify a Starting Model

This constuction makes the ellicitation of more and less informative priors straightforward. In particular, the for the Dirichelet distribution are for each of the unique values in the dimension () for APC. The Dirchelet distribution draws a random projection on the standard simplex, which represents a vector of weights to indicate athe probability of a window break. An equal to 1 is a uniform (or noninformative) distribution across the simplex (Gelman 2014:69). Values less than 1 indicate less weight, *i.e.* it is less likely to be a window break; wherease, values more than 1 indicate more weight, or it is more likely to be a window break. The meaning of to the Dirichelet distriution is pictured in Figure 1 below. Each panel depicts 1,000 random draws from the Dirichelet distributions with 10 dimensions. In the top panel, each is equal to 1. The mean probability that a window break occurs across each dimension. For the toop panel, the mean probability is 10% for all of the windows. In the bottom panel, $= [0.2,0.2,0.2,1,1,1,1,2,2,2]. The mean probability of a window break is low for the first three (2%), roughly even for the middle 4 (9%), and high for the final 3 (19%).

[Figure 1 About Here]

The hihger the value on the simplex, the higher the probability that the value will be a window break. Accordingly, the Dirichelet also makes it easy to specify more and less informative priors. A dirichelet with values of 1 for constitute equal probability of a window break. A researcher can ellicit a more informative prior by placing larger values of on particular values of a dimension to make it more probable that this will constitute a window break. Multiplying the vector of weights () by a scalar provides a set of numbers which sum to the scalar value .

## Step 3: Accept or Reject Proposed Models in the Markov Chain

## Step 4: Ensemble Accepted Models Using BMA

# A Simulation

# An Empirical Example from the GSS

# Discussion

# Conclusion

Anon. 2005. “Strategies for Estimating Age, Period, and Cohort Effects.” Pp. 12–35 in *Cohort analysis*. 2455 Teller Road,?Thousand Oaks?California?91320?United States of America? SAGE Publications, Inc.

Bell, Andrew and Kelvyn Jones. 2017. “The Hierarchical Age-Period-Cohort Model: Why Does It Find the Results That It Finds?” *Quality & Quantity*.

Burgard, Sarah A. and Lucie Kalousova. 2015. “Effects of the Great Recession: Health and Well-Being.” *Annual Review of Sociology* 41(1):181–201.

Camm, Jeffrey D. 2016. *Essentials of Business Analytics*. Mason, OH: Cengage South western.

Fragoso, Tiago M. and Francisco Louzada Neto. 2015. “Bayesian Model Averaging: A Systematic Review and Conceptual Classification.” *arXiv preprint arXiv:1509.08864*.

Gelman, Andrew. 2014. *Bayesian Data Analysis*. Third edition. Boca Raton: CRC Press.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. “Bayesian Model Averaging: A Tutorial.” *Statistical science* 382–401.

Jackson, Stephen HD, Martin R. Weale, and Robert A. Weale. 2003. “Biological Age—what Is It and Can It Be Measured?” *Archives of gerontology and geriatrics* 36(2):103–15.

Lam, Jack, Wen Fan, and Phyllis Moen. 2014. “Is Insecurity Worse for Well-Being in Turbulent Times? Mental Health in Context.” *Society and Mental Health* 4(1):55–73.

Leisering, Lutz. 2003. “Government and the Life Course.” Pp. 205–25 in *Handbook of the life course*, edited by Jeylan T. Mortimer and Michael J. Shanahan. New York: Kluwer Academic/Plenum Publishers,

Luo, Liying. 2013. “Assessing Validity and Application Scope of the Intrinsic Estimator Approach to the Age-Period-Cohort Problem.” *Demography* 50(6):1945–67.

Luo, Liying and James S. Hodges. 2016. “Block Constraints in Age–Period–Cohort Models with Unequal-Width Intervals.” *Sociological Methods & Research* 45(4):700–726.

Moen, Phyllis. 2014. “Constrained Choices: The Shifting Institutional Contexts of Aging and the Life Course.” Pp. 175–216 in *New directions in the sociology of aging*, edited by Linda J. Waite, Thomas J. Plewes, and National Research Council (U.S.). Washington, D.C: National Academies Press.

Raftery, Adrian E. 1995. “Bayesian Model Selection in Social Research.” *Sociological Methodology* 25:111–63.

Ryder, Norman B. 1965. “The Cohort as a Concept in the Study of Social Change.” *American Sociological Review* 30(6):843–61.

Taddy, Matt, Chun-Sheng Chen, Jun Yu, and Mitch Wyle. 2015. “Bayesian and Empirical Bayesian Forests.” *arXiv preprint arXiv:1502.02312*.

Vaisey, Stephen and Omar Lizardo. 2016. “Cultural Fragmentation or Acquired Dispositions? A New Approach to Accounting for Patterns of Cultural Change.” *Socius: Sociological Research for a Dynamic World* 2:2378023116669726.

Wooldridge, Jeffrey M. 2009. *Introductory Econometrics: A Modern Approach*. 4th ed. Mason, OH: South Western, Cengage Learning.

Yang, Yang and Kenneth C. Land. 2013. *Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications*. Boca Raton, FL: CRC Press.

1. A continuous variable of 5 integers can be sliced into continous window constraints in 15 ways (where | indicates a window break for dummy variables):

   2 windows, 4 combinations: 1|2345, 12|345, 123|45, 1234|5

   3 windows, 6 combinations: 1|2|345, 1|23|45, 1|234|5, 12|3|45, 12|34|5, 123|4|5

   4 windows, 4 combinations: 1|2|3|45, 1|2|34|5, 1|23|4|5, 12|3|4|5

   5 windows, 1 combination: 1|2|3|4|5

   Since the hypothetical assumes 3 variables (A,P,and C) of 15 combinations each, total combinations are . Similar calculations over an integer of 6 leads to 34 window combinations over each dimension for a total of 39,304 possible models. [↑](#footnote-ref-26)