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

# Abstract

Stochastic Window Sampler (SWS). incorporate this

Key words: Computational Sociology, Methods, Life Course, Bayesian Methods, Data Science

# 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 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—perhaps even thousands—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 (MCMC) 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 and Jumping Distribution

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

In Equation 1, 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 resorting 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 1 are as follows:

If the vector and have the following properties: , , and , then permutatios of descirbe all 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 single 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 by 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. MCMC methods from Bayesian Model Averaging (BMA) provides a straightforward way to combine models. Generally speaking, these methods search across some predefined neighborhood of models, and then use a weighted average of the posterior esitmates for inference.

Because as defined in equation \_\_ describes all possible window models for equation 1, any process that samples from by definition samples from all potential models of window breaks. Accordingly, sampling window breaks is simply a matter of smapling appropriate values for . In Bayesian statistics and data-driven inference like machine learning, thsese sorts of clustering problems are often solved using a Dirchelet distribution.

*Sampling the Window Groups (G) using the Dirchelet Distribution.* 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 probablility that particular values of the continuous variables will occur in . We use the Dirichelet distributon to sample both the location and number 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 effects are equal.

To implement our sampling scheme, we use a set of auxiliary variables for each dimension () of APC. These are the intensity, or weight values for the Dirchelet distribution (), which are numbers between 0 and infinity. Various combinations of sample from the simplex (represented as below) for each unique value of A, P, or C. The simplex is a vector of numbers between 0 and 1 which sum to 1. is the rounded, cumulative sum of the product times the maximum possible number of window breaks (*i.e.* the number of unique values). More formally, we consruct in three steps as follows:

First, we draw a random simplex from a Dircelet distribution across each dimension () of . The dimension of the simplex () is as large as the index () for each of the APC effects (the , , or of equation 1.

Second, We take the cumulative sum of from step 1 and multiply it by the the maximum index value of A, P, or C (). This provides a string of integers from 1 to , covering the entire continuous range of each dimension. These integers identify a vector of window groupings by index numbers for each dimension (. In terms of the dummy variable example from above, a continuous vector of ages between 5 and 50, recoded into a dummy variable with an younger (less than or equal to 30) and older group as wculd be described in terms of as a vector where 26 is repeated 26 times, (because ), and where 45 is repeated 24 times (because .

Third and finally, we use the vector of index numbers to construct as described in equation \_\_. To do this, we begin the vector a value of 1 less than the minimum observed value (for the reasons described above). For each subsequent value, of , we take use the value of described by each unique index number contained in , as follows:

Following the steps above, is simply 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 \_\_. Moving from one realization of to another, is a matter of changing the from one set of models to the next. We visualize the impact makes in determining the shape of below, where we discuss prior elicitation and the starting model, .

As commonly used in MCMC methods, we use a normal proposal distribution with a specified variance to update from one model () to the next () (**???**). 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 with each additional draw of .

## Step 2: Specify a Starting Model

Sampling window breaks as outlined in step 1 makes the ellicitation of more and less informative priors straightforward. In particular, the identifies the probability that a break occurs between two contiguous groupings of age, period or cohort. An larger value represents a more diffuse simplex (Gelman 2014:69). Smaller values indicate less weight, *i.e.* it is less likely to be a window break, and larger values indicate more weight, or it is more likely to be a window break. The meaning of parameter in the Dirichelet distriution for equations \_\_ to \_\_ is pictured in Figure 1 below.

Each panel of Figure 1 depicts 1,000 random draws from Equations \_\_ and \_\_ across a continuous variable of 8 dimensions. In the top panel, each is equal to 1. The bar graph reports the proportion of times that each dimension represents a new windows break. The first dimension is equal to 100% (as required to span the entire range, the first dummy variable must begin with the minimum value). Beyond that, the rest begin a window break nearly all of the time, with an average number of breaks equal to 7.2. By contrast, the bottom panel has a non-uniform array of values spanning 0.3 to 17.8. In response, the probability of window breaks is different between each value, with the highest probablities corresponding to the highest values in , and the lowest values corresponding to the lowest values of . The average number of window breaks in for this distribution is about one-third less than the top panel at 4.4.

[Figure 1 About Here]

Figure 1 illustrates that the higher the value , the higher the probability that the corresponding observed value will be a window break. Accordingly, the Dirichelet also makes it easy to specify more and less informative priors. 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. The final requirement of this step is to choose a starting model . We choose such that it starts close to the target model, *i.e.*, there are a large number of window breaks. To accomplish this, we set the initial value of each to the dimensionality of the A, P, or C. As with the example outlined in Figure 1, the mean number of window breaks between the maximum value of each A, P, or C. This is consistent with the interpretation of $\alphas$ as “prior sample size” cite gelman. This prior sample size is approximately the size of the window breaks less one (as the mean value in the xample above). Alternative starting values are possible, as are different starting values for multiple MCMC chains.

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

Having identified the sampling procedure, the proposal density, and the starting vaues, the next step is sample proposal models and accept or reject them. As outlined above, each chain takes the current model, , and samples by permuting the vector of $\alphas$ from a normal distribution. These values are used to draw a model of window constraints . Whether to accpet in favor of becomes a problem of model comparison. We adopt the jumping kernel described in Hoeting, et al. (1999), for the MC3 algortihm, which accepts model with a probability based upon the minimum of 1 or the posterior model probabilities. We calculate the ratio of model probabiliteis using the BIC approximation to the Bayes factor (Raftery 1995).

The chain is accepted over if is greater than a value drawn from a uniform distribution between 0 and 1 (). Otherwise, is retained for the next draw in the chain.

## Step 4: Ensemble Accepted Models in the MCMC Chain

Finally, after a specified number of iterations across the Markov Chain, any desired estimate, can be calculated using the weighted average of the effects times the posterior model probabilities.

As before, the probabilites can be difficult to calculate. However, decades of work have developed robust estimators, including an extensive classification laid out by Raftery (1995) more than 20 years ago in *Sociological Methodology*. We use the BIC for the linear model outlined by Raftery in order to calculate the weighting scheme, so that add the equation here for pi.

# Exploratory Simulations

To evaluate the effectiveness of our ensembling strategy, we simulate 6 different data sets with different quadratic age, period, and cohort effects. We draw three different sets of APC linear effects from a uniform distribution between -1 and 1; and three different sets of quadritic effects from a uniform distribution between -0.1 and 0.1. The smaller scale for the quadratic effects produces a more reasonable distribution of effects. We report the values used to simulate effects in the Appendix, S1. In order to compare cases with true nonlinear APC effects, we duplicate these three sets of effects, but delete a random A, P, or C.

Given these effects, we simulate 1,000 observations as described in Luo and Hodges (2016). First, we generate 400 unique combinations of 20 ages (1 through 20) and 20 periods. We repeat this collection 25 times for a total of 1,000 total observations. We generate cohort as a linear combination of of age and period (period - age), and produce a squared term for each A, P, and C. We then square the observations and apply the effects in Appendix S1 to generate a predicted outcome value. To scale the , we use a normally distributed error with mean 0 and standard deviation of 5.26. We store the average, predicted effects for comparison after we apply the BMA algorithm outlined above.

To estimate effects on the simulated data, we apply the algorithm described above. We run 4 parallel MCMC chains of 250 iterations each for a total sample size of 1,000. In other words, we run separate 1,000 window models on the data. Nearly all of these models are different, as the window constraints are drawn given the of the Dirchelet distribution (as depicted in Figure 1). Accordingly, even when Model is rejected, a new model is drawn based on the prior values for the Dirchele distribution. For each of the window breaks, we use a Gibbs sampler to draw 1,000 samples of each model, for a full posterior sample size of 1 million draws. For the Gibbs sampler, we use a uniform prior for the in equation 1, and an Inverse Gamma Distribution. These priors exactly coincide exactly with frequentist estimates and standard errors, but allow provide more information to evaluate model fit (Gelman 2014). We draw posterior samples using the composition method described in Lynch (**???**).

Because is is computationally intensive to sample 1,000,000 models, we use the simulation as an exploratory tool as opposed to a demonstrative tool. That is, we do not recalibrate the normal jumping distribution to change the accpetance rate, nor do we calibrate with other statistics. Instead, we simply run the models using identical, but arbitrarily determined characteristics, and compare the results to the true effects. We also use this to assess the effectiveness of unique tools for evaluating model fit from Bayesian estimators, such as test statistics from posterior predictive distributions.

A unique tool for assessing model fit with Bayesian estimators is through posterior predictive data (PPD). To generate the PPD, we draw a sample of 1,000 sets of parameters from the entire posterior of 1 million samples using a weighted sampling scheme equal to the posterior model weight from the BMA analysis. We use these the PPD parameters to simulate predictive data of the observed (though simulated) outcome variable, . The PPD distribution of is , and we calculate the Bayesian p-value of the mean, or, the (**???**). Better-fitting models have Bayesian p-values near 0.5 (meaning that the mean of the replicated data is centered on the mean of the observed data) (Gelman 2014:151), so we also compute the absolute deviance of this p-value (). Though we do not report them here, we also examined alternative test-statistics. The overall p-value we report here was highly correlated with these other test statistics, so for simplicity we retain the deviance measure only.

We generate information for each set of models as reported in Table 1, and plot the best-fitted model, which is also the most concordant with true effects (Simulation 3) in Figure 2. Figure 2 shows extremeely close tracking of estimated values (dashed line) with actual values across all dimensions. In the supplement, Figure S-1 plots similar effects across all simulated effects.

[Figure 2 about here]

Table 1 reports the simulation number, and a number of variables to evaluate the model fit, MCMC convergence, and model concordence with the true effect. This table shows (1) the absolute deviance of the Bayesian p-value is an indicator of model fit. That is to say–generally–the smaller this deviance, the better the model matches the simulation values. (2) The absolute deviance value is sensitive to model . (3) The SWS sampler is more effective at estimating nonlinear effects across all three dimensions.

[Table 1 about here]

The first two columns of Table 1 report the simulation number and whether the simulation includes all three effects. The thrid column colculates the proportion of effects (80 total) where the true effect is contained within the confidence intervals of the BMA-calculated main effects. A value of 1.0 would indicate exact concordance, and a value of 0 would indicate no overlap between estimates and true effects. A value of 0.05 would indicate 5% of the values fall outside of the 95% confidence intervals of the estimates. Note that this is an overly-conservative ratio, because it does not take into account expected uncertainty in estimating effects.

For the best-fitting model (pictured in Figure 1), only 1 of the 80 effects fall outside of the predicted values. The other models do not fare as well, although the the SWS performs better when all three dimensions have actual effects (simulations 1,3,and 5). This is not surprising, however, given that the construction of age-period-and cohort implies extremely high multicollinearity. Under these conditions, adding a variable which is highly correlated to a covariate, but not the outcome, should produce collider bias I think this is correct -- add citation. One particularly ill-fitting model (Simulation 2) misses almost 90% of the actual effects. It is important to note, however, that the situation, even for this simulation, is not as dire as the tabel might imply. The depictions of the model fits in Table S-1 show that the general trends are by and large reproduced. That is–appart from the null effect–trends generally track trends: when the actual effect is increasing, the estimated effect is increasing and when the actual effect is decreasing, the estimated effect is also decreasing. In addition, as suggested by Figure S-1, general groupings of ages, periods, and cohorts as similar would be generally accurate.

The fourth column of Table 1 shows the absolute deviance of the Bayesian p-value of the mean as described above. Generally speaking, the smaller this value, the better the estimates track the true values. Simulation 3 has both the smallist discordance between estimates and true values, and the smallest absolute deviance. It is not always the case, however, and smaller model fits provide less predictable relationships (as identified by the exception to this rule for Simulation 6, which has a relatively large deviance, lower discordance, but also low model r-squared).

While the simulation results do not suggest that the SWS method is a silver-bullet for the APC identity conuncrum, these exploratory simulations show significant promise for addressing the identity issue. In particular, the SWS provides a way to test multiple model constraints, to aggregate APC classifications into groups, and to identify large-scale trends. These simulations also show that Bayesian tests, as generated from a PPD, correlate with concordance of true effects. In this way, they provide a measure of the fit of the APC model. We take this knowledge and apply it to an empirical example using trends of egalitarian attitudes as measured in the GSS.

# An Empirical Example Using the GSS

Using the knowledge we

# Discussion

# Conclusion

**References**

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)