The Dyad Ratios Algorithm for Estimating Latent Public Opinion: Estimation, Testing, and Comparison to Other Approaches

James A. Stimson
University of North Carolina at Chapel Hill

October 25, 2017

Contents

Abstract

The central logic of the dyad ratios algorithm is explained. Central focus is on the use of ratios as a starting point, the recursive estimation procedure, validity estimation by iterative procedure and the bootstrapping of standard errors. The ability of the algorithm to estimate a known longitudinal path is tested with artificial data. Then dyad ratios is compared to principal components analysis for a particular real data problem where both are possible. A final section makes a limited comparison between dyad ratios and item response theory.

The dyad ratios algorithm is a method for the extraction of a common dimension in data such as survey marginal responses over time. The problem it solves is that such data are massively incomplete: most variables (survey questions) do not exist for most cases (time points), and therefore the technology of principal components analysis is undefined.

The algorithm is far from new. It has existed in code and in usage since 1988, almost three decades. Hundreds of scholarly articles and books have exploited it. But the algorithm itself has not been the subject of a scholarly article.[1] It has been described in appendices, in software user guides, and the like. But there is no scholarly publication that develops the methodology, validates the result, and compares it to alternatives. That is what this article is to be.

Public Policy Mood is now a staple of American politics. The literature is now too extensive for citation. It has found application in Britain (Bartle, Dellepiani & Stimson 2010, Green & Jennings 2012, McGann 2013), France (Brouard & Guinaudeau 2015, Stimson, Tiberj & Thiébaut 2010, Stimson, Tiberj & Thiébaut 2012), Europe (Guinaudeau & Schnatterer 2017), Mexico (Baker et al. 2015), with efforts underway or completed in Spain (Bartle, Bosch & Orriols 2014), Portugal, Germany, and Japan. In the United States and Europe it has found application for explaining public policy (Erikson, MacKuen & Stimson 2002, Bartle, Dellepiani & Stimson 2010). And there is a related literature on its causes in the U.S. (Enns & Kellstedt 2008, Owen & Quinn 2016, Ellis & Faricy 2011).

Here is the plan of this article. In Section 1 below I undertake a detailed description of the logic of the dyad ratios algorithm, drawing comparison at various points to its similarities to principal components analysis. In Section 2 I examine an artificial data test of the performance of the algorithm. In Section 3 I develop a controlled comparison between the algorithm and principal components analysis for a rare case where both estimation strategies are possible. And then in a final section, 4, I draw a more distant comparison to the Bayesian IRT model for similar problems, a product of recent developments.

[1] The world of scholarly publication is much friendlier to methodological developments today than it was thirty years ago. And *Political Analysis*, the one journal that might have published such work was unavailable to me because I was its editor.

1 The Dyad Ratios Model

What does it mean when we compare the result of a policy preference question from survey research to the same question posed at another time? The starting point of the dyad ratios conception is that the ratio of the two is empirical evidence of change in time for the underlying latent variable. This is true of course only to the degree to which item i is a valid indicator of latent concept C.

Further, if we have several such items, i, j, k, l, m, and n, the degree to which ratios at particular times covary across items is evidence of shared variance, both between the items and between the items and the latent concept.

Item validity in this conception is a variable property of items, a matter to be determined empirically by observing how much of the item's variance is shared with the latent concept. Note the contrast with the now popular Item Response Theory alternative. IRT requires assumed validity of items. It is based on its founding analogy to test theory where, of course the test designer selects items for their assumed validity. An item about the meaning of a square root might be selected for a test of mathematical ability or knowledge, for example. An item of knowledge of Shakespeare's plays would not.[2] The difference between the two is assumed validity.

A similar difference arises in interpretation of cross sectional item marginal totals. In the IRT framework, with assumed validity, a difference between two items in marginal percents—percent left in this case—is interpreted to mean that the item with the smaller score is more difficult to agree to than is the larger one (Caughey & Warshaw 2015). Shared variance conceptions, such as PC and DR make no such interpretations because they do not assume the relatedness of items. That opens the possibility that differences in marginals might occur because two items measure something different, not more or less of the same thing.

Now I proceed to nuts and bolts, my goal to explain how starting with an

[2]This difference at the theoretical heart of DR and IRT leads also to a difference of procedure in the two approaches. DR can be applied to any set of items, because it selects and weights those that share variance. IRT analysts must use theoretical judgment to preselect valid items (McGann 2013).

assumption of meaningful ratios can lead to a full dimensional analysis, with all the complexity that entails.

The algorithm doesn't dictate how that single value is to be obtained. But for illustration I use:

$$x_{it} = 100 \times \frac{Left_{it}}{(Left_{it} + Right_{it})} \tag{1}$$

where the Left and Right totals are summed over multiple responses, if present and the subscript *it* references item *i* at time *t*.

After aggregation into T regular time periods, we have a matrix of N items for T periods, x_{it} where i indicates variables and t indicates period. Because no survey item is ever posed at every consecutive time point in the sample, most of the matrix is missing data, represented as zero. We assume that items are positive numbers scored to represent the concept in question. I will refer to the concept as C_t and its estimated value as $\hat{C}_t$. It eases exposition to assume that all items are scored in the same direction, that higher scores indicate more of the concept and lower scores less of it (regardless of the polarity of the underlying survey question).

Our definition of a variable or item is that it is the same question, the same response options, the same sampling design and typically posed by the same organization (although that requirement can sometimes be waived). Thus a common variable name implies that all of the cases of that variable may be meaningfully compared.

1.1 Ratios

Define a dyad as values for the same variable, x_{ik} and x_{il} for any two time points, k and l, where $k \neq l$. Then as a starting point we can say that the ratio

$$r_{ikl} = \frac{x_{ik}}{x_{il}} \tag{2}$$

is a meaningful indicator of the concept C to the degree that item i is a valid indicator of C. (The validity issue is taken up below.) This assumption of meaningful dyad ratios is the foundation of the algorithm and the source of its name.

Thinking of data as ratios rather than variable scores has one major advantage. Whereas in general scores are not comparable across items, ratios are. Two variables will, in general, produce different scores. Because there is no science of question wording, we do not know what level of support or opposition each item should draw. If we had a full set of cases for each variable (as in principal components analysis) we could estimate the variable means and use that knowledge to compare across items. But we do not. Because of the missing values issue, we have neither a full set of cases nor a representative sample of them. Thus we cannot know item expectations. But ratios, r, have a known expected value across cases, 1.0. That common expectation justifies comparing across items.

But how to combine items? We have a multitude of dyad ratios, with typically large numbers for each time point in the series to be estimated. There are multiple ratios for each item. Time k, that is, has a ratio for every other time point that is available for item i. And then there a multiple variables as well. So information exists in abundance and the problem becomes simply how to combine it sensibly.

The problematic aspect of ratios is that they are relative information. What we want to know is the absolute level of our concept for some time t. But what we have tells us instead how t is related to t+k. *If* all variables were available for one or more common times, then the problem is easy. We make the ratios relative to those times and then we have absolute information. But in general our real world data do not provide the convenience of a time point available for all variables. The easy option is precluded by the data we have.

1.2 Recursive Estimation

There is no global analytic solution for combining information across items. We could just ignore comparability issues and average across all the different ratio estimates for each case. That would probably produce a decent approximation of a good measure. But it depends upon an assumption, that we have a representative sample of time points for each item, that is known to be false. Recursion is a second best approach. Begin with the final point of the series, T. Having lost our metric information in the computation of ratios, we can assign an arbitrary value to this one time, say 100.

Now there exists a subset of items which include an available value for time T. For those items (only) we can estimate absolute values for each item and each t by simply projecting the known value at T (100) onto the ratio of T to other time points. If, for example, item i has an average T to T-1 ratio of .92, then the estimated score for that item at t-1 is 92. Then we can average across all existing cases that have non-missing values for T and T-1 to get an estimate of t-1. (We are still assuming perfect validity for items here. That issue will be dealt with below.) All earlier times are also projected for later use.

For the latent concept C denote our estimated value for case t-1 as $\hat{C}_{t-1}$. So now we know two values, $C_T = 100$ and a data determined value for $\hat{C}_{T-1}$. The data determined value reflects the true ratio of T and T-1 estimated from all of the data which exists. No missing values enter the computation and no existing values are ignored.

Now we can repeat the process for comparing T-1 to T-2, but this time using the estimate for T-1, $\hat{C}_{T-1}$, rather than an arbitrary number, for the value at T-1. Following this procedure we eventually work back to time 1, the beginning of the series to be estimated.

When using backward recursion later periods tend to dominate the solution. Each later value has influence on the values of early items, but not the reverse. Also the estimates are not unique. Reversing the order and starting with time 1 and working forward—forward recursion—produces a similar, but not identical, set of estimates. Forward recursion has the reverse weighting of backward, early items contribute more to the solution than do later ones.

So we end up with two estimates of the latent concept C, C forward ($\hat{C}_F$) and backward($\hat{C}_B$), which are equally valid time series. Averaging the two (for each time point) accomplishes two things, (1) it uses all available information for the solution rather than using one and ignoring one, and (2) it corrects the weighting effects to produce a summary score in which all items weight equally in the solution.

Smoothing Now we face a choice. We could just average the computed values of $\hat{C}_F$ and $\hat{C}_B$. But sampling theory suggests that we could do better. Sampling theory tells us that if nature produced smooth outcomes—i.e., if opinion change were gradual and regular rather than abrupt and jumpy—then our observed estimates of it would be noisy. Because the data points are the result of survey samples, they will capture the true level of the phenomenon while adding or subtracting a small error in each due to sampling. So if nature were smooth, we would still not observe smoothness due to sampling errors.

Thus the choice: in estimating C_F and C_B do we prefer those which are strictly data determined (and therefore also sampling error determined) or a smooth approximation based upon the prior knowledge that nature is smoother than empirical estimates of it? In my view, which the reader need not share, the smoothed approximation is superior to the data-driven estimates.

The particular smoothing model chosen is exponential smoothing. It has the virtue that it is sensitive to how much noise in present in the data series (and will not alter an already-smooth series). The exponential smoothing model is:

$$y_t = \alpha x_t + (1 - \alpha)x_{t-1} \tag{3}$$

where y is the smoothed version of x. The intuition is that if the past, x_{t-1}, provides any useful information for predicting y_t, then some portion of the variation in x_t is a deviation from the smooth path of x. This is seen in zigzag behavior, where the series tends to return to normal levels after extreme movements away from them.

The α parameter is estimated (iteratively) by minimizing within sample forecast error. Thus it is fully determined by the data. Exponential smoothing has the desirable property that it will not oversmooth. If the data are already smooth, a situation that often occurs with annual aggregation levels, then α

converges on 1.0 and y_t converges on x_t. Smoothing occurs in both forward and backward directions in time, the result of which is that the raw data series become exponentially weighted moving averages of past and future values.

Smoothing operates on the raw values of C_F and C_B during estimation. That means, in effect that the smoothed value is presumed to be a better measure of the true level of the series than is the original, and that it is the smoothed values of the series that drive the ultimate measure.

The impact of smoothing varies in direct proportion to the apparent randomness of the series. Where the original series are highly patterned, the impact of smoothing is rarely discernible. Where, in contrast, they exhibit a good deal of period-to-period zig-zag fluctuation, the effect of smoothing is larger.

What consequence? Smoothing has a big (and helpful) effect on periods in which data are relatively thin and usually modest effects when data are rich. This is to be expected because having multiple estimates of a quantity averaged together (when data are rich) produces natural smoothing, the Central Limit Theorem in action.

1.3 Validity Estimation

The issues that arise in validity estimation in the dyad ratios algorithm are essentially the same as the validity issues in principal components analysis. In principal components analysis there are three standard approaches for validity estimation, (1) assuming perfect validity—essentially ignoring the issue—(2) estimating from the R^2 of multiple regressions of item i as dependent on all other items, and (3) iterative estimation. These amount in PC to treatment of the main diagonal of the input matrix, that it is (1) 1.0 for all items, (2) R_i^2, or (3) a convergence result where μ_i^2 (validity assumed for item i) becomes equal to $\hat{\mu}_i^2$ (validity estimated from the squared loading of $\hat{C}$ on x_i.) The first approach violates our understanding of measurement theory, albeit usually with small consequences. The second is impossible due to missing data issues. The third is implemented in software producing the dyad ratios estimates.

In the theory of vector decomposition mathematics if you could somehow "know" the right values of the validity estimates for each item, μ_i^2 (the proportion of all variance in the item i that is shared with the concept C) then the estimate produced after estimation, the squared correlation between the latent factor $\hat{C}$ and the raw item would be the same value, $\mu_i^2 = \hat{\mu}_i^2$. That theory provides a solution criterion. Iterative solution is reached with μ_i^2 estimated from the previous iteration differs by less than a trivial amount (.001) from μ_i^2 estimated from the current iteration for all i. There are N such estimates and the solution requires that all N be less than .001 different for solution.

Where validity comes into play in dyad ratios is that the estimate for each time point is, instead of a simple average of ratios, r_i,

$$\hat{C}_t = \frac{\sum_{i=1}^{N} r_i}{N} \tag{4}$$

a weighted average of ratios weighted by item validity.

$$\hat{C}_t = \frac{\sum_{i=1}^{N} \hat{\mu}_i^2 r_i}{\sum_{i=1}^{N} \hat{\mu}_i^2} \tag{5}$$

1.4 Bootstrapping Standard Errors

Users of estimated time series like to have cross-sectional—that is, period by period—standard errors around such estimates. Journal editors seem to like them even more. Never having seen such information actually employed for inference, I am agnostic on the matter. But it is worth considering how variability estimates might be constructed.[3]

[3] I write in a hypothetical vein here even though I have implemented the procedure and reported the results. The reason is that the computer code for doing so is in a language (Visual Basic) no longer supported and is compiled on an obsolete compiler for Windows versions no longer supported. Time permitting, a modern and sharable version will be developed.

When it is possible to know standard errors from an estimator, those are the definitive, proven, values and obviously the best way to proceed. But there are many situations in which no such derivation exists. In these cases bootstrapping presents a robust second best option. The fundamental idea of bootstrapping is that we can subject the estimator to variations of known magnitude in data input and then observe its behavior. With a sufficient number of such observations we get a distribution of values that are produced for each particular case and that distribution becomes empirical evidence for the properties of the distribution, the one in question being the standard deviation. The standard deviation of the distribution of observed estimates is the best (empirical) estimate of the standard error of the estimator.

The data input to the dyad ratios algorithm is survey based estimates of proportion "left" response. In one particular aggregation period—day, month, quarter, year, or multiple year period, that is—a particular value, say 65, is observed and reported. Sampling theory tells us that that value, call it $\bar{x}_{it}$ for item i and time t, is a best estimate of the true value, μ_{it}, but since it is the observed outcome of survey sampling with a sample of a particular size, it is merely an estimate of μ_{it}, subject to some fluctuation due to the fact of having randomly drawn a particular sample and not another equally valid draw.

That tells us how to proceed to varying the input in order to observe the behavior of the output. For each k repetitions of the algorithm[4] we alter that input value from the fixed original value, 65 in the example at hand, to a random draw from a distribution with mean x_{it} (e.g., 65) for item i at time t, and standard deviation σ_{it}, which is readily computed from sample size for the binary indicator.

When the period by period standard errors are estimated, it is easy to put confidence intervals around the estimated time series outputs. Because such estimates are empirically derived, they have no provable properties. That being the case, some scientific skepticism is appropriate.

With the technical details settled, I turn to three sections on performance. The first of these asks the question, given a known pattern—which is known because it is artificially generated—how well do we do recovering the pattern?

[4] I usually employ k = 1,000.

2 Testing the Algorithm with Artificial Data

Artificial data offers the advantage of generating a perfectly known pattern in the latent concept to be estimated. That allows assessment of the accuracy of estimation given real world factors in the data.

The Data Generating Process For an artificial latent concept a sine function is generated over 1,000 cases. This is chosen to partially mimic the irregular cycles often found in public opinion data.

$$y = sin(\frac{d\pi}{180}) \tag{6}$$

where d is 1 to 1,000, representing degrees.

With a constructed mean of 50 and a standard deviation of 10 the sine function, except for its error-free smoothness, looks much like the result of estimating a latent concept from survey data.

The simplest test would be the create the latent concept and the artificial "items" without error. We will not do that because the result can be known, the estimates would converge perfectly—r = 1.000—on the latent concept.

In the real world items always come from survey research and survey research always has sampling error. Thus we generate items that perfectly capture the known function and then add sampling error in controlled amounts. For the test case we have five artificial items, x1—x5, with means of 40—60 in 5 point intervals.

Sampling error is emulated by normal draws from N($\bar{x}_i$,σ^2) where σ^2 is varied over the 9 tests from 0.5 to 5.0 in intervals of 0.5.

In real data we have empirical estimates of the longitudinal standard deviation of the items. But that does not tell us sampling error because that observed standard deviation is composed of three pieces

$$\sigma^2_{total} = \sigma^2_{valid} + \sigma^2_{unique} + \sigma^2_{error} \tag{7}$$

(where total is observed variance, valid variance is estimated from r^2 of the loading of variable i on the latent dimension, unique variance is systematic variance in item i that is not shared with the concept, and error is the sampling error). It is only the total that we know from the empirical standard deviation.[5]

The dyad ratios algorithm is mainly an exercise in taking vary large numbers of empirically observed ratios within and between items and then averaging over all of them. Such elementary arithmetic operations ought to converge closely on the real underlying signal. Figure 2 shows that in fact it does. Correlations range from a low of .983 for the maximum sampling error of 5.0 to .999 for the minimum tried, 0.5.

Our real world data are imperfect due to sampling. By aggregating across items—even here for only five items—that imperfection is largely eliminated. With larger numbers of items[6] the performance is even stronger. The moral of this story, often observed, is that more data is better than less. The worst case, standard practice until recently, is choosing one item to stand for the concept.

A typical result is illustrated in Figure 2. The figure shows the underlying sine wave, a perfect function of time, in bold. The thinner line is actual estimates of the underlying concept produced using items with a sampling error of 2.5 added. The figure is a little hard to make out because the estimates converge on the reality of the artificial sine wave. But that convergence is the point of the figure, estimates converge on reality.

I turn in Section 3 to a comparison of the behavior of the algorithm to that of its near relative, principal components analysis.

[5]In real data unique variance—systematic variation due to the item and not shared with the concept—is usually present. In the simulation it is not.

[6]I have used up to 250 items in estimating Mood in the United States.

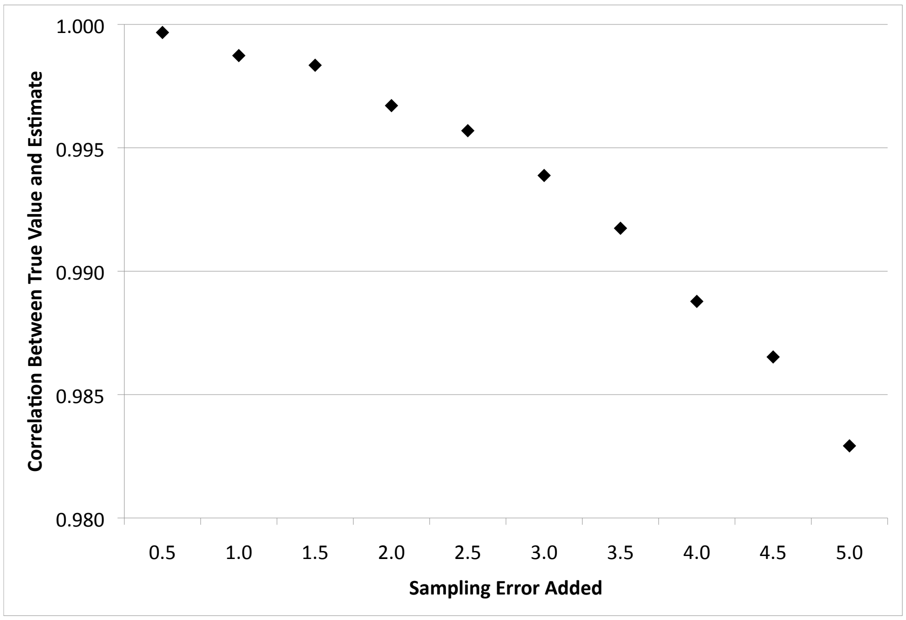

Figure 1: Algorithm Performance Over a Range of Assumed Sampling Error Magnitudes: Correlations Between True Values and Estimates

3 Dyad Ratios Compared to Principal Components

While the foundational idea of dyad ratios differs from the variance apportioning scheme at the heart of principal components analysis, much is also the same. This is no coincidence. They share a great deal because dyad ratios explicitly borrows a great deal from the established technology of the principal components model.

Both take as a starting point variance associated with variables and transform it to associations with latent dimensions. Both derive interpretation

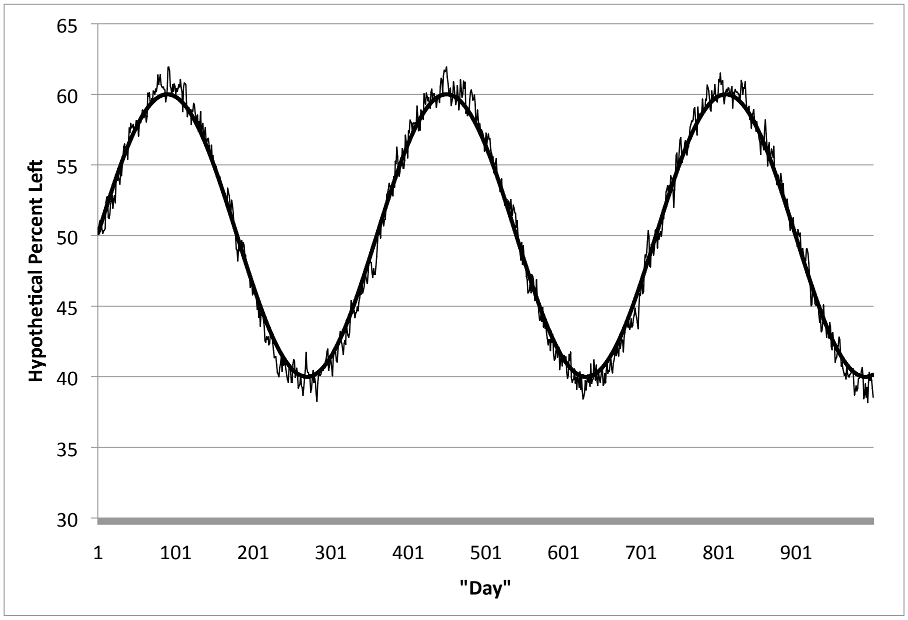

Figure 2: Estimating a Sine Wave in Artificial Data: Sine Wave and Dyad Ratios Estimates With Sampling Error of 2.5. The bold perfect curve is the generated Sine Wave. The jagged thin line is the Empirical Estimates from Items with Sampling Error.

from the “loading” of those starting variables on the derived dimensions. Both estimate variable validities as the core problem of latent dimension estimation.

The big and obvious difference is that principal components requires a complete set of cases, a near impossibility for the public opinion data for which the dyad ratios algorithm was developed. Since principal components is undefined for the data which dyad ratios operates on, direct comparison is usually not possible.

But it is possible to compare a somewhat artificial special case, a rare public opinion collection with no missing cases. Such a test bed is provided by the General Social Survey's items on spending priorities in the United States. Six items pose priorities questions (spending is too much, about right, or not enough) for cities, education, environment, welfare, healthcare, and race. They have been posed continuously in all GSS studies, 1973-2016.[7]

Even with identical cases there is no expectation of identical findings. The two models do differ. But the question of similarity is itself interesting. Do we reach similar or differing conclusions about the structure underlying a set of observed relationships?

Table 1: First Dimension Loadings for Six GSS Items Estimated by Principal Components and Dyad Ratios Algorithm

Variable	Dyad Ratios Loading	Principal Components Loading
Spend for Cities	0.75	0.70
Spend for Education	0.84	0.75
Spend for Environment	0.87	0.88
Spend for Welfare	0.78	0.73
Spend for Healthcare	0.52	0.48
Spend for Race	0.83	0.76
Percent Variance Explained	60	53
Correlation between loadings = .97		

The answer becomes clear in Table 1, where the dimension loadings on the first (and only common) dimension are displayed. It displays in column two the estimated loadings from dyad ratios and in column three the comparable loadings from principal components.[8] The loadings do differ, but the two sets show a similar rank order and correlate at .97.[9]

[7]The studies themselves are not continuous, sometimes at annual intervals and sometimes biennial. There are 30 studies over the 46 year span. I simply create a series that consists of the 30 available years.

[8]The PC results are from Stata iterated principal factor, the closest analogy to the DR iterative method.

[9]The Stata method for comparison is iterated principal factor with regression scoring.

Interpreting the result of a dimensional analysis is a mix or art and science, imprecise under the best of circumstances. Here the interpretation of the two results is the same. The latent concept is domestic left-right ideology in the United States. None of the differences in loading estimates is large enough to challenge that. All of the items tap that concept and the loadings are much more similar across them than different.

The purpose of a dimensional extraction is usually to create a measure of the latent concept. The question then is how similar are the measures derived from dyad ratios and principal components? The answer is that they are very similar. The two measures correlate at .986. (See Figure 3 for a visual picture of the two series.) For comparative reference, with completely identical dimensional solutions the two Stata scoring methods, regression and Bartlett, correlate at .99.

So, for identical data the two methods produce highly similar results. Dyad ratios may then be regarded as an extension of the capabilities of principal components to a realm where its data requisites are not met.

4 Dyad Ratios Compared to IRT

Dyad Ratios, like the larger family of principal components dimension extraction techniques, has an exploratory flair. Staring with a group of items of unknown relatedness and unknown dimensionality, it solves for a low dimensional (maximum two) solution and observes relatedness. Item Response Theory, IRT for short, begins with the attitude that the researcher knows the true dimensionality, typically one, and imposes it on a set of items prescreened for face validity. Its modern manifestation is tied to Bayesian estimation techniques (McGann 2013, Caughey & Warshaw 2015).

Is a meaningful comparison of the two possible? What argues for the possibility is that the goal of both models is the same, observing the latent dimension that underlies a group of items. As in the previous comparison with principal components, it is rare that we could employ both on the same set of items. If IRT analysts prescreen for face validity and Dyad Ratio analysts choose by other criteria—such as all available items on a topic—they

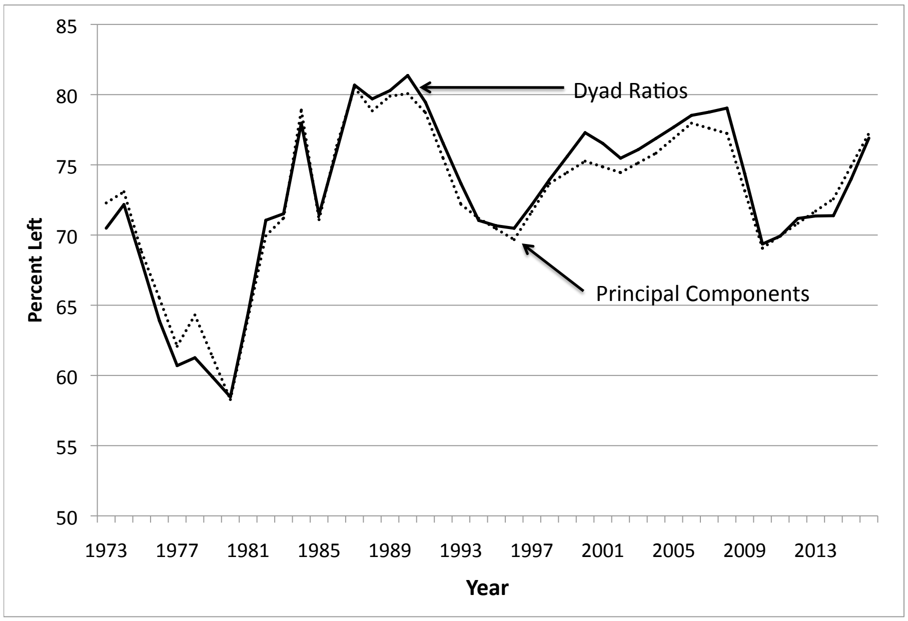

Figure 3: Six GSS Series Latent Structure Estimated by Principal Components and Dyad Ratios Algorithm

will not in general be the same.

The case at hand is the Caughy-Warsaw (Caughey & Warshaw 2015) estimates of Public Policy Mood in the United States. For this case, not only is the model of dimensionality different, the input data differ quite substantially. In particular Caughey and Warshaw screen not only for validity, but they also screen out questions with a relative frame. These are policy preference questions that involve a comparison to the status quo, matters of more or less rather than absolute preferences. Typical language is of the sort, "Should the government do more, less, or about the same as it is doing now?" These relative frame questions make well behaved time series and form an important and stable component of estimated mood.

In fact the degree of overlap between input data is unknown. Because of the exclusion of relative frame questions we know for certain that the data are different. How different, a matter of degree, is hard to tell. But at the outset we are warned that some similarity of estimates is the best that can be expected. This is not a comparison where only the mathematical model differs. The model and the data differ.

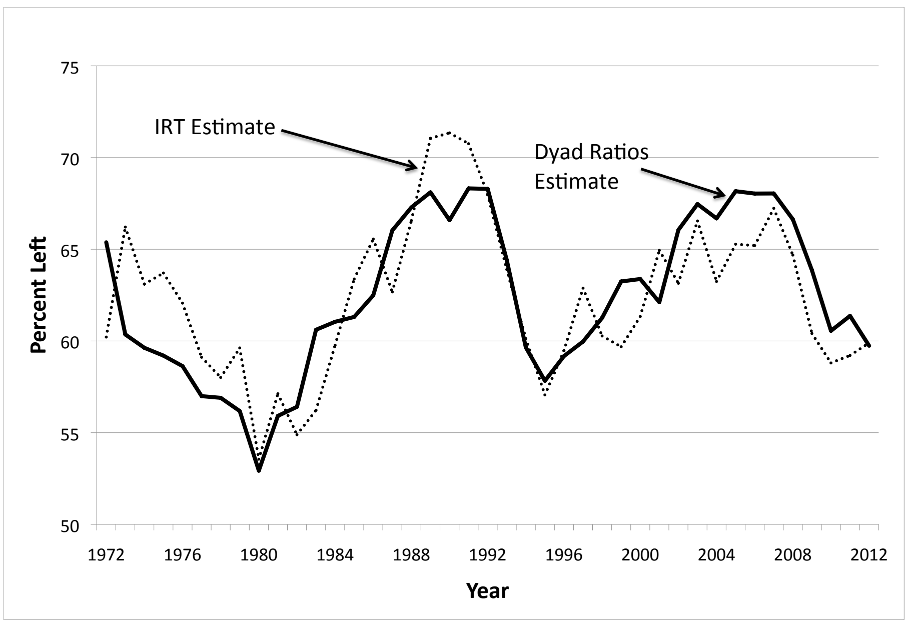

Figure 4: US Policy Mood Estimated by Dyad Ratios Compared to Caughey-Warshaw IRT Estimates, 1972-2012

The Caughey-Warshaw data (provided by Caughey and Warshaw) estimate Mood for the period 1992 to 2012. So that defines the possible overlap. For that period the similarities of the two estimates are strong (r = .783). That is more than suggestive evidence that the same concept is tapped by both approaches, even though model and data both differ.

If the two estimates widely diverged, it would set up an argument about which is right and which is wrong. Since the convergence is instead reasonably strong, the evidence would seem to validate both. We could reach a stronger conclusion if it were possible to observe both for identical items. But as Figure 4 clearly shows, there is much more similarity than difference in the two estimates. An analyst describing the ideological flow of politics in the United States would tell the same story with either Dyad Ratio or IRT estimates of the concept.

References

Baker, Andy, In Jorge I Dominguez, Kenneth F Greene, Chappell Lawson & Alejandro Moreno. 2015. "Public Mood and Presidential Election Outcomes in Mexico." Mexicos Evolving Democracy: A Comparative Study of the 2012 Elections pp. 107–27.

Bartle, John, Agustí Bosch & Lluís Orriols. 2014. The Spanish policy mood, 1978-2012. In 8th ECPR General Conference. University of Glasgow. pp. 3–6.

Bartle, John, Sebastian Dellepiani & James A. Stimson. 2010. "The Moving Centre: Policy Preferences in Britain, 1950-2005." British Journal of Political Science 41:259–285.

Brouard, Sylvain & Isabelle Guinaudeau. 2015. "Policy beyond politics? Public opinion, party politics and the French pro-nuclear energy policy." Journal of Public Policy 35(1):137170.

Caughey, Devin & Christopher Warshaw. 2015. "Dynamic Estimation of Latent Opinion Using a Hierarchical Group-Level IRT Model." Political Analysis 23(2):197–211.

Ellis, Christopher & Christopher Faricy. 2011. "Social Policy and Public Opinion: How the Ideological Direction of Spending Influences Public Mood." The Journal of Politics 73(04):1095–1110.

Enns, Peter K. & Paul M. Kellstedt. 2008. "Policy Mood and Political Sophistication: Why Everybody Moves Mood." British Journal of Political Science 38:433–454.

Erikson, Robert S., Michael B. MacKuen & James A. Stimson. 2002. The Macro Polity. New York: Cambridge University Press.

Green, Jane & Will Jennings. 2012. "Valence as macro-competence: An analysis of mood in party competence evaluations in Great Britain." British Journal of Political Science 42(2):311–343.

Guinaudeau, Isabelle & Tinette Schnatterer. 2017. "Measuring Public Support for European Integration across Time and Countries: The European Mood Indicator." British Journal of Political Science pp. 1–11.

McGann, Anthony J. 2013. “Estimating the Political Center from Aggregate Data: An Item Response Theory Alternative to the Stimson Dyad Ratios Algorithm.” Political Analysis 22(1):115–129.

Owen, Erica & Dennis P Quinn. 2016. “Does economic globalization influence the US policy mood?: A study of US public sentiment, 1956–2011.” British Journal of Political Science 46(1):95–125.

Stimson, James A., Vincent Tiberj & Cyrille Thiébaut. 2010. “Au service de l’analyse dynamique des opinions.” La Revue Française de Science Politique 60:901.

Stimson, James A., Vincent Tiberj & Cyrille Thiébaut. 2012. “The Evolution of Policy Attitudes in France.” European Union Politics .