Dissertation Outline

Essays on Political Science Applications of the Mixture Index of Fit

Juraj Medzihorsky
Department of Political Science
Degree: Doctor of Philosophy
Supervisor: Professor Tamás Rudas

August 2015

Abstract

This dissertation proposes new applications of the Rudas-Clogg–Lindsay mixture index of fit and log-linear models that improve inferences in several areas of substantive research in political science. These include problems from electoral research–detection of electoral fraud from digit distributions, allocations of seats according to votes, territorial variability of electoral support and competition, and analysis of voter transitions with aggregate data–as well as analysis of roll call data in the study of legislative politics, and statistical analysis of political text. The improvements are due to the fact that the methods allow to abandon conventional assumptions known to be difficult or false. Most importantly, the mixture index abandons the assumption that the whole population is described by the model. Furthermore, the index also allows to abandon the assumption that the data was stochastically sampled. Log-linear models allow to represent associations in multivariate categorical data without assuming continuity or requiring transformations that produce it. The thesis is accompanied by an R package named `pistar` that implements procedures for the application of the mixture index of fit in a variety of settings.

1 The Mixture Index of Fit and the pistar Package

This chapter discusses the Rudas–Clogg–Lindsay π^* mixture index of fit. The existing procedures for the computation of the index in a variety of contexts are presented. The presentation is accompanied by the introduction of `pistar`, an R package that implements the index for a variety of statistical models.

2 Latent Class and Log-Linear Election Forensics

Digit-based election forensics typically relies on null hypothesis significance testing, with undesirable effects on substantive conclusions. This chapter proposes an alternative free of this problem. It rests on decomposing the observed numeral distribution into the 'no fraud' and 'fraud' latent classes, by finding the smallest fraction of numerals that either needs to be removed or reallocated to achieve a perfect fit of the 'no fraud' model. The size of this fraction can be interpreted as a measure of fraudulence. Both alternatives are special cases of measures of model fit–the mixture index of fit and the dissimilarity index, respectively. Furthermore, independently of the latent class framework, the distributional assumptions of digit-based election forensics can be relaxed in some contexts. Independently or jointly, the framework and the relaxed assumptions allow to dissect the observed distributions using models more flexible than those of existing digit-based election forensics. Reanalysis of Beber and Scacco's (2012) data shows that the approach can lead to new substantive conclusions.

Note: An earlier version of this chapter has been accepted to *Political Analysis* as Medzihorsky, J. (2015) 'Election Fraud: A Latent Class Framework for Digit-Based Tests.'

3 The Generalized D'Hondt Index

The D'Hondt method is a widely used seat apportionment procedure. The method minimizes the maximum ratio of seats over votes, also known as the D'Hondt index, and sometimes used to evaluate seat distributions produced by other methods. Despite the method's widespread use, some of its properties are not well understood. This chapter shows that the method divides the votes into two classes, one of which is represented proportionally, and the other not at all, while minimizing the size of the unrepresented class, and can therefore be understood in the context of the mixture index of fit. This allows to generalize the D'Hondt index and method to situations with partially observed vote distributions. Moreover, a new kind of residual analysis becomes available that rests on inspecting the unrepresented votes. The residual analysis is illustrated with 16 British general elections from 1950 to 2010, and the generalization to partially observed votes with the Brazilian federal lower house election of 1982.

4 Analysis of Electoral Support with the Dissimilarity Index

This chapter presents a general framework for the analysis of electoral support based on the dissimilarity index. It rests on inspecting the fraction of votes that would need to be cast differently for the reality to conform to ideal-typical models, such as stable support or territorially homogeneous competition. Measures are formulated within the framework, which relate to those of electoral volatility and party nationalization and regionalization, are easy to interpret and use, allow comparisons across observations and theories, and account for party and territorial electorate sizes. The measures include the index of residential segregation of votes, which is compared with other measures from the literature on a set of 1495 elections in 119 countries from 1789 to 2013. Two families of models are used with the framework–log-linear models to capture territorial and spatial variability of electoral support and latent class models to inspect territorial heterogeneity of electoral competition. Their use is demonstrated on general elections in Canada (2006–2011), UK (2015), and Belgium (1946–1995).

5 Minimum Mixture Models of Voter Transitions in Aggregate Data

The rate of voters who switched party from one election to another is often measured from aggregate data. The existing methods involve trade-offs between how substantively informative the generated quantities and how strong and testable the underlying assumptions are. This chapter proposes a new latent class approach to the analysis of aggregate electoral data based on the Rudas–Clogg–Lindsay mixture index of fit. It is light on assumptions, but highly flexible, and provides a measure of transitions applicable to any number of elections and parties. The measure is easy to compute and interpret. The approach is extended to conditionally constant voting, and proportional and uniform swing. The use is demonstrated on data from the 2004 and 2008 elections in Montana and the 1966 and 1970 UK general elections.

6 Roll Call Analysis with the Mixture Index of Fit

This chapter introduces the mixture index of fit to roll call analysis. The index provides a general framework applicable to a variety of problems from this domain and instead of replacing the existing methods enhances them. In this chapter, it is applied to measure partisan and other kinds of group voting, and evaluate and substantively interpret ideal point models. The applications are illustrated with congressional roll calls related to the Civil Rights Act of 1964.

7 The Mixture Index of Fit in Text Analysis

Text is analysed statistically across a variety of fields to a variety of aims. However, most model evaluation metrics in use focus on predictive performance, and there are no metrics for description and exploration currently in wide use. This chapter introduces the Rudas–Clogg–Lindsay mixture index of fit to statistical text analysis. The index is applied to classification as well as scaling problems. The models include the unigram model, the mixture of unigrams model, latent class analysis known in this context as probabilistic latent semantic analysis/indexing, correspondence analysis, and log-linear models for multivariate categorical associations. The demonstrations use a dataset built from ten platforms of five U.S. parties from 1996 and 2000.