Advance Access publication October 19, 2011

Political Analysis (2012) 20:92–112
doi:10.1093/pan/mpr039

Model-based Clustering and Typologies in the Social Sciences

John S. Ahlquist

Department of Political Science, University of Wisconsin, Madison, 1050 Bascom Mall,
Madison, WI 53706, and United States Studies Centre at the University of Sydney, Australia
e-mail: jahlquist@wisc.edu (corresponding author)

Christian Breunig

Department of Political Science, University of Toronto, 100 St. George Street, Toronto, Ontario
M5S 3G3, Canada
e-mail: c.breunig@utoronto.ca

Edited by Jonathan N. Katz

Social scientists spend considerable energy constructing typologies and discussing their roles in measurement. Less discussed is the role of typologies in evaluating and revising theoretical arguments. We argue that unsupervised machine learning tools can be profitably applied to the development and testing of theory-based typologies. We review recent advances in mixture models as applied to cluster analysis and argue that these tools are particularly important in the social sciences where it is common to claim that high-dimensional objects group together in meaningful clusters. Model-based clustering (MBC) grounds analysis in probability theory, permitting the evaluation of uncertainty and application of information-based model selection tools. We show that the MBC approach forces analysts to consider dimensionality problems that more traditional clustering tools obscure. We apply MBC to the "varieties of capitalism," a typology receiving significant attention in political science and economic sociology. We find weak and conflicting evidence for the theory's expected grouping. We therefore caution against the current practice of including typology-derived dummy variables in regression and case-comparison research designs.

1 Introduction

Classifying complex objects into some smaller number of categories is fundamental to the scientific enterprise. The natural and behavioral sciences are replete with taxonomies, categorization schemes, and typologies. Canonically, categorization is closely related to definition and measurement: objects of interest are measured along several dimensions and then said to represent one or another "type" depending on the values taken on these measurements. Put another way, the researcher defines a category as one covering some set of measured attributes.

But taxonomy is not always the primary objective in trying to classify objects. We believe that it is just as common for researchers to have either developed theoretical expectations that observations hang together in some way or taken "supervised" or "unsupervised" strolls through the data looking for patterns. All three activities—definition, theory testing, and inductive data work—have generated typologies (Collier, Laporte, and Seawright 2008). The latter two certainly involve statistical inference, but the methodological literature in political science, typically dominated by qualitative researchers, has ignored inference in favor of discussions of concept validity and the like.

Authors' note: Previous versions of this paper were presented at the 2008 meetings of the Society for Political Methodology, the 2009 10th Anniversary Conference for the University of Washington Center for Statistics and the Social Sciences, and colloquia at Duke, Purdue, and MPIfG. We thank Justin Grimmer, Martin Hoepner, Ryan Moore, Kevin Quinn, Adrian Raftery, and Mike Ward for comments and helpful conversation. Carlisle Rainey provided research assistance. Replication materials are available at the *Political Analysis* dataverse, Ahlquist's dataverse (http://dvn.iq.harvard.edu/dvn/dv/ahlquist), and Breunig's Web site http://individual.utoronto.ca/cbreunig/.

In this paper, we build on the distinction between supervised and unsupervised (machine) learning to better understand the different roles typologies play in the social sciences. When the goal is inference and the exact parameters of a cluster are unknown or, even better, implications of social theory, then the appropriate statistical tools for unsupervised learning can help us better evaluate theoretical claims. Recent advances in the theory of mixture models with applications to cluster analysis and model selection (Banfield and Raftery 1993; Fraley and Raftery 1998; Biernacki, Celeux, and Govaert 2000; Fraley and Raftery 2002; Bensmail and Meulman 2003; Zhong and Ghosh 2003; Fraley and Raftery 2005; Raftery and Dean 2006; Baudry et al. 2010) provide a principled way to discuss and evaluate proposed typologies as well as easy-to-use $\mathcal{R}$ and MATLAB software for applied researchers. We consolidate these recent advances for applied audiences, arguing that among the various unsupervised machine learning algorithms, model-based clustering (MBC) provides a particularly useful way to discuss and evaluate typologies and guard against their reification. Traditional cluster analysis tools such as k-means and hierarchical cluster analysis provide little help in this regard. In using a resolutely model-based approach, we clarify how a typology's origins as either a definition or a theoretical implication is relevant to its application. We revise methodological advice about category formation and "concept stretching" (Sartori 1970; Collier and Mahoney 1993) accordingly while revisiting the appropriate use of "explanatory typologies" (Elman 2005). Specifically, from a multivariate inference perspective, the geometry of cluster analysis implies that increasing the number of cases permits us to simultaneously consider a greater number of attributes and yet still make informed model-based inference.

Typologies derived from theoretical arguments have, in some instances, morphed into explanatory "variables" in subsequent analysis, usually without sufficient empirical justification. It is only appropriate to use a theoretically derived typology as an explanatory construct if the existence of meaningful structure consistent with theoretical expectations is found in the data. This in turn requires that the number of cases exceeds the number of attributes believed to define the categories. An example of one such reified typology is the widely discussed "varieties of capitalism" (VoC) (Hall and Soskice 2001). Some scholars have taken this theoretical categorization scheme to the point of including it as a covariate in regression models. We use MBC to replicate two important studies in the literature. Given the seriousness with which scholars have appropriated the VoC, we find weak evidence that the VoC-posited categories are discernable in the data.[1] The clustering we identify in the data does not coincide with the VoC's theoretical arguments nor is it consistent across different sets of variables (and combinations thereof) that should, according to theory, exhibit VoC clustering. This mismatch has direct implications for substantive researchers: VoC categories are neither measurement exercises nor empirical realities and they should not be treated as either in the data analysis.

We anticipate that machine learning and classification techniques will see rapid adoption in the social sciences in coming years. Currently, the most aggressive applications have been in areas of textual analysis,[2] where supervised learning techniques predominate (but are by no means exclusive). Do certain word stems appear together more often in some media outlets rather than others (McCombs 2004)? Does the clustering of words used by politicians in public speech reveal something about their intentions and priorities (Sulkin 2005; Klebanov, Diermeier, and Beigman 2008)? Mixture models and computational tools are already being developed to address this exact problem (Hillard, Purpura, and Wilkerson 2008; Monroe, Colaresi, and Quinn 2008; Quinn et al. 2010; Grimmer 2010). But these tools need not be limited to one set of issues or one subfield. For example, a long-standing and ongoing debate in comparative politics surrounds the classification of political systems (Bueno de Mesquita et al. 2003; Vanhanen 2003; Marshall, Jaggers, and Gurr 2004; Treier and Jackman 2008; Pemstein, Meserve, and Melton 2010). Within countries generally considered to be democracies, there is further active research in identifying various types (LeDuc, Niemi, and Norris 1996; Lijphart 1999). Classifying nondemocracies has proven even more difficult; Geddes (2003) only recently introduced a comprehensive data set classifying authoritarian regimes. Do these different classification schemes readily map into one another? Even more broadly, a movement is afoot within the American Political Science Association to take a systematic audit of the various indicators of democracy and governance in an effort to systematize the use and evaluation of

[1] See Ahlquist and Breunig (2009) for an extended MBC analysis of the empirical applications of the VoC.

[2] Indeed, the Autumn 2008 issue of *Political Analysis* was dedicated to textual analysis methods.

scholar-generated measurement in political science.[3] In short, we see several areas of applied political science likely to benefit from increased awareness of MBC tools.

The paper proceeds in four parts. The next section briefly reviews typologies and concept formation in the social and behavioral sciences. Section 3 discusses mixture models and MBC in more detail, with special attention to their relationship to other clustering and data reduction techniques commonly used in the social sciences. Section 4 presents our empirical application. We conclude by suggesting other possible MBC applications in political science.

2 Categorizing Typologies: Denotive and Inferential

For our purposes, we will define typologies as any rule that takes a high-dimensional object and returns one and only one value from a finite set. Typically, the cardinality of this set is small. Formally, let x be an object measured on k attributes, let X be the set of n such objects, and let $Z = \{a, b, c, \ldots\}$ be the set of categories such that $|Z| \leqslant k$. A typology t is any single-valued mapping $t\colon X \mapsto Z^n$. Note that the objects X and the attributes k are taken as given.

We identify two broad ways of linking theory and categorization. The most familiar method of categorization involves the formation of a concept, the promulgation of a definition, measurement, and the assignment of labels (Collier, Laporte, and Seawright 2008). Much of the literature on concept formation and measurement concentrates on the links between precise definition, measurement (or "scoring"), and validity. Second, there is also a long tradition, dating back to at least Max Weber, in which categorization is primarily an exercise in theory building. Less methodological attention has been paid to the role of categorization in theory building and testing.[4]

Most social scientists think of typologies as related to measurement. The steps go more or less as follows: first, based on theory and/or empirical observation, a concept is born; second, a precise definition of the concept is given; third, some set of objects of interest is identified; fourth, the measurements *required by the definition* are taken; fifth, measurements are compared against the definition and a label or category is assigned to each observation accordingly.

At the risk of adding yet another "typology of typologies" (Elman 2005), we will refer to this sort of categorization exercise as *denotive* as its primary purpose is to identify and label empirical objects. The assignment of labels can be simple naming or it can take a more nuanced form by assigning ordinal or interval-scaled values to particular sets of observations. For denotive categorization, theory defines the *domain*, or universe of objects relevant to our categorization; the *attributes* or dimensions we care about; and the possible *combinations* of these attributes. Formally, preexisting theory defines X, $|Z|$, and possibly t.

Typologies and categorization can also form the basis of theory construction and testing. Weberian "ideal type" reasoning clearly fits in this tradition. Ideal types are theoretical concepts meant to help the theorist lay out her reasoning in its most stark form and to refer to an "example" where one does not, in fact, exist in the observable world (Weber 1949, 90). By extension, the analyst has some reason to expect observations or traits to agglomerate in certain ways. In other words, the posited clusters are (testable) implications of her theory.

Alternatively, the analyst might inductively search for a recognizable pattern in the data rather than relying on deductive theorizing. But both looking for patterns and examining a theoretical implication involve evaluating the mapping t and/or the set Z, given X. The analyst either is looking for or expects to find cases taking similar values on particular attributes clustering together. Whether these expected clusters are discernable in the data is a problem of statistical inference. We therefore call this sort of activity *inferential* categorization. In inferential categorization, any pattern identified is conditional on the sample employed and thus cannot be used to then "explain" the same observations used to infer the pattern in the first place.

[3] http://sites.google.com/site/democracyaudit/.

[4] But see Elman (2005) for a recent treatment on the use of typologies as explanatory constructs in political science as well as an extensive review of the enormous literature on typologies. See Collier, Laporte, and Seawright (2008) for an extensive list of political science typologies.

This distinction between denotive and inferential categorization has direct analogues in the machine learning literature. Machine learning algorithms, of which MBC is one family, are typically divided into "supervised" and "unsupervised" variants. Supervised algorithms, much like denotive clustering, take predefined categories of objects and then attempt to categorize new observations. Unsupervised learning algorithms, much like inferential categorization, seek to discover or confirm structure in the data. This difference in both the derivations and the roles of categorization schemes has been underappreciated in the social science literature on typologies.

To see the stark difference in denotative and inferential categorizations, consider the so-called "curse of dimensionality." In denotative categorization, the ability of the analyst to assign an observation a label is not restricted by the dimensionality of the definition. Put another way, the number of attributes (dimensions) that compose the definition does not affect our ability to assign any single observation a label. For example, there is an extensive literature on the measurement of "democracy." Given a definition of democracy and adequate measurement on the required attributes, any country can be labeled a "democracy" or not. With an adequate measurement instrument, one country's regime status is independent of that of all other countries. The elements of this set of democracies have no immediate bearing on the falsification of any theory of democracy or on our ability to apply the definition to other countries that may emerge in the future. Therefore, to the extent the attributes composing the denotive categorization are well measured and t is well defined, we can use denotative categorizations as "explanatory" in a meaningful sense.

But consider the situation in which we have a collection of objects and we are attempting to *infer* the existence of some sort of pattern or clustering that simplifies our understanding of the relationships between observations. As with any statistical inference problem, our ability to discern clustering is affected by *both* dimensionality and the presence or absence of other observations. More precisely, we can say that the number of observations relative to the number of dimensions in which these objects are purported to exist affects our ability to infer the existence and shape of clusters.

This last point is important as it clarifies a long-standing issue in political science, via the extent to which a concept "travels" outside the realm of cases for which it was initially developed. Sartori argued for his famous "ladder of generality" in which there is a trade-off between the number of cases covered under a concept ("extension") and the dimensionality of that concept ("intention"). Clearly, this is a challenge in the context of denotative categorization. But from an inferential standpoint, the inclusion of new observations can have two results: improved ability to infer the number and nature of clusters and the greater degrees of freedom for expanding the number of dimensions considered, thereby *increasing* intention. In clearly stating the differences between denotive and inferential categorization exercises, we can pinpoint where the statistical literature and, in particular, MBC can make an important contribution.

3 Cluster Analysis and the Social Sciences

Cluster analysis is a well-developed branch of applied statistics that attempts to identify groups in data such that objects within groups are as similar as possible although the differences between groups are maximized. Although classification of similar objects was a prominent task of many fields (e.g., classification of animals and plants, of stars, and of chemical compounds) in the 19th century, the development of statistical techniques to uncover "known underlying dimensions" received major interest in psychology since the 1950s and 1960s. In this section, we briefly review traditional cluster analysis and then go into a detailed discussion of MBC. We contrast MBC with both traditional methods and other data reduction techniques as well as latent variable models like principal components and factor analysis.[5] We are concerned with the class of problems in which the "true" underlying clustering is unknown, to be estimated from the data, contrasted with supervised learning and discriminant analysis in which we have either an algorithmic definition or a training set of known classifications.

3.1 *Hierarchical and Relocation Clustering Methods*

Throughout we will use the term "group" to refer to the true existing groupings of objects and "cluster" to denote the collections of observations identified via some algorithm or statistical model, that is, a cluster

[5] See Zhong and Ghosh (2003) for a more extensive and technical discussion of the relationships between these different classes of models. See Grimmer and King (2011) for an ambitious attempt to integrate all known clustering algorithms.

is an estimated grouping. Cluster analysis has at least one of two objectives: identifying some sort of cluster structure in a set of observations and assigning observations to clusters in some optimal manner. Kaufman and Rousseeuw (2005) provide an accessible introduction to traditional cluster analysis, and several attempts have been made to catalogue the available measures of similarity and dissimilarity (e.g., see Cox and Cox 2001, 11–2).

Within traditional cluster analysis, relocation methods such as k-means require that the analyst posit the number of clusters in the data in advance and then proceed to iteratively move observations among clusters until an optimal allocation can be identified. In hierarchical cluster analysis, the number of groups is unknown. Hierarchical analysis uses intuitively plausible procedures based on various distance metrics to either merge or partition observations into clusters. Hierarchical cluster analysis is commonly "agglomerative," beginning with each object on its own and proceeding to combine them into clusters that maximize within-cluster similarity and between-cluster difference, as determined by a distance metric. Several different metrics can be employed and the literature provides little theoretical guidance about their appropriateness, though Milligan (1980, 1981) surveys Monte Carlo experiments, concluding that Ward's linkage[6] is a useful distance metric.

Hierarchical cluster analysis is primarily an exploratory rather than confirmatory or inferential activity. In fact, Kaufman and Rousseeuw (2005, 37) suggest that "it is permissible to try several algorithms on the same data because cluster analysis is mostly used as a descriptive or exploratory tool. . . . we just want to see what the data are trying to tell us." There are many attributes on which to measure similarity and difference across objects and, given some set of attributes, numerous algorithms for identifying clusters. Furthermore, both hierarchical clustering and k-means generate "hard" solutions that define partitions of the data. There is no foundation in statistical theory on which to prefer a particular clustering solution over another and no possibility of evaluating the uncertainty around a particular observation's assignment to a given cluster. The choice of both the number of clusters to focus on and the substantive interpretations assigned to them is solely the responsibility of the analyst. Referring to the traditional clustering methods, Venables and Ripley (2002, 316) argue that "there are many different clustering methods often giving different answers, and so the danger of over-interpretation is high."

Less commonly recognized by users of these traditional methods is that these tools impose restrictions on the geometry of the clusters they can identify. For example, k-means iteratively moves observations from one cluster to another to minimize the total squared distance from k "centroids" or "prototypes." Ward's linkage, described above, also uses a squared distance loss function. By relying on pairwise Euclidian distances, these methods are effectively constrained to find spherical clusters, something we illustrate below and demonstrated formally by Celeux and Govaert (1992). Ward's linkage, in particular, also tends to generate clusters of approximately equal size (Milligan 1980, 1981). This is a serious shortcoming since many processes of interest do not generate spherical clusters. For example, Dasgupta and Raftery (1998) use a variant of MBC to detect mine fields in image data. Land mines typically line up along a road or other geographical feature rather than being distributed in circles. Social scientists, in asserting that social objects of interest should "hang together," frequently identify clusters that are not circles or spheres (see Fig. 2 below). Nor do we typically have good reason to believe that the clusters should be equal in size.

3.2 *Model-based Clustering*

More recently, mixture models have been applied to the clustering problem (Banfield and Raftery 1993; Bensmail et al. 1997; Dasgupta and Raftery 1998; Fraley and Raftery 1998; Biernacki, Celeux, and Govaert 2000; Fraley and Raftery 2002; Bensmail and Meulman 2003; Fraley and Raftery 2005; Raftery and Dean 2006). This generation of clustering methods assumes that the observed data are generated by some finite mixture of probability distributions. Mixture MBC has four notable advantages over traditional clustering methods. First, MBC derives the partition of the data from an estimated statistical model, thereby enabling "soft" clustering and statements of uncertainty about the resulting classification. Second, the choice of clustering method now becomes a problem of model selection. We have a variety of tools

[6] Ward's (1963) method calculates the distance between clusters as the squared Euclidian distance between the mean vectors of the two divided by the total sum of squares.

derived from statistical theory that can aid us in this task. Third, if we assume that each component of the mixture is a cluster, the model-based approach identifies the number of clusters in the data. Other methods require either a priori assumptions (e.g., k-means) or post hoc subjective decisions (hierarchical clustering). Fourth, MBC can accommodate several cluster shapes not readily implemented in most traditional methods.

Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be the $n \times k$ matrix of n objects measured on k dimensions. The density of $\mathbf{x}$ can then be expressed as a finite mixture of the form

$$f(\mathbf{x}) = \sum_{g=1}^{G} q_g f_g(\mathbf{x}),$$

where G is the number of mixture components, q_g is the proportion of objects in component g, and $f_g(\cdot)$ is the density function for observations in component g. Assuming that all groups are defined by multivariate normal densities, we can substitute $\phi(\mathbf{x}|\theta_g)$ for $f_g(\mathbf{x})$, where $\phi(\cdot|\theta)$ is the multivariate normal density function with parameters $\theta_g = (\mu_g, \Sigma_g)$. The density has the form

$$\phi(\mathbf{x}_i|\mu_g, \Sigma_g) = (2\pi)^{-k/2}|\Sigma_g|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mu_g)'\Sigma_g^{-1}(\mathbf{x}_i - \mu_g)\right]. \tag{1}$$

The model classifies an observation as being in group g if $\tau_g(\mathbf{x}) > \tau_h(\mathbf{x}), \forall\, h \neq g, h \in 1, \ldots, G$, where

$$\tau_g(\mathbf{x}) = \frac{q_g\phi(\mathbf{x}|\theta_g)}{\sum_{h=1}^{G} q_h\phi(\mathbf{x}|\theta_h)},$$

then τ_g can be interpreted as the (posterior) probability that an object belongs to group g. We can now express the full mixture likelihood:[7]

$$\mathcal{L}_{\mathrm{m}}(\theta_1, \ldots, \theta_G; \tau_1, \ldots, \tau_G|\mathbf{x}) = \prod_{i=1}^{n}\sum_{g=1}^{G} \tau_g\phi(\mathbf{x}_i|\theta_g). \tag{3}$$

It is clear from equation (3) that the number of parameters estimated grows rapidly with the number of clusters, G, and the number of dimensions, k. Banfield and Raftery (1993) partially mitigate this problem by placing restrictions on the covariance matrices, Σ_g. Covariance matrices are parameterized using eigenvalue decompositions of the form

$$\tilde{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g^T, \tag{4}$$

where λ_g is the largest eigenvalue of $\tilde{\Sigma}_g$, $\mathbf{D}_g$ is the orthogonal matrix of eigenvectors, and $\mathbf{A}_g$ is a diagonal matrix of scaled eigenvalues. The parameters $\tilde{\theta}_g = (\tilde{\Sigma}_g, \mu_g)$ determine the geometry of the clusters. Specifically, clusters are ellipsoids centered at the mean vector. The decomposition of Σ_g determines other geometric features of the clusters: λ_g determines the cluster's volume, $\mathbf{D}_g$ controls the orientation of the cluster, and $\mathbf{A}_g$ governs the shape of the ellipsoid.

We can modify the complexity of the models estimated by restricting the various elements of the matrix product on the right-hand side of equation (4) to be constant across components. The most restricted version, $\tilde{\Sigma}_g = \lambda I$, implies that the exponential term in equation (1) is $-\frac{1}{2\lambda}\|\mathbf{x}_i - \mu_g\|^2$, that is, the sum of squares/Euclidian distance used in Ward's linkage and k-means. In other words, traditional clustering

[7] An alternative to the mixture likelihood is the classification likelihood. Specifically,

$$\mathcal{L}_{\mathrm{CL}} = \prod_{i=1}^{n}\prod_{g=1}^{G} f_g(\mathbf{x}|\theta_g)^{(\gamma_g)_i}, \tag{2}$$

where γ_g are component labels estimated to maximize equation (2), that is, the maximization is over all possible assignments of X to the components G (Ganesalingam 1989). A key difference is that under the classification likelihood each observation is assigned outright to a specific component *at each iteration* of the numerical maximization of the likelihood (MLE) computation, whereas this does not occur in mixture models. When the relative sizes of the mixing proportions are unknown, as is likely the case for inferential categorization exercises, mixture models are preferred (Celeux and Govaert 1993). See Celeux and Govaert (1993), Fraley and Raftery (1998), and Ganesalingam (1989) for further discussion.

Table 1 Cluster geometries generated by differing parameterizations of the covariance matrices, Σ_g, and implemented in `mclust`

Model	*Distribution*	*Volume*	*Shape*	*Orientation*
λI	Spherical	Equal	Equal	NA
$\lambda_g I$	Spherical	Variable	Equal	NA
$\lambda \mathbf{A}$	Diagonal	Equal	Equal	Along the axes
$\lambda_g \mathbf{A}$	Diagonal	Variable	Equal	Along the axes
$\lambda \mathbf{A}_g$	Diagonal	Equal	Variable	Along the axes
$\lambda_g \mathbf{A}_g$	Diagonal	Variable	Variable	Along the axes
$\lambda \mathbf{DAD}^T$	Ellipsoidal	Equal	Equal	Equal
$\lambda \mathbf{D}_g \mathbf{A} \mathbf{D}_g^T$	Ellipsoidal	Equal	Equal	Variable
$\lambda_g \mathbf{D}_g \mathbf{A} \mathbf{D}_g^T$	Ellipsoidal	Variable	Equal	Variable
$\lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g^T$	Ellipsoidal	Variable	Variable	Variable

methods can be recast as special cases of a mixture of Normals (Celeux and Govaert 1992). Table 1, reproduced from Fraley and Raftery (2007, 7), describes some of the various cluster geometries generated as restrictions on the covariance matrices are relaxed; all these models are currently implemented in the `mclust` library for $\mathcal{R}$.

In fitting the model, the actual cluster to which observation i belongs is treated as missing data. The "complete data", $\mathbf{y}_i$, can be expressed as $\mathbf{y}_i = (\mathbf{x}_i, \mathbf{z}_i)$, where $\mathbf{x}_i$ are the observed data on which we seek to fit the clustering model and $\mathbf{z}_i$ is a G-vector the gth element of which takes on 1 iff i belongs to cluster g and 0 otherwise. Assuming that $\mathbf{z}_i \sim \text{multinom}(\tau_1, \ldots, \tau_G)$, the resulting complete data likelihood is given by

$$\mathcal{L}_{\mathrm{c}} = \prod_{i=1}^{n} \prod_{g=1}^{G} [\tau_g \phi_g[(\mathbf{x}_i|\theta_g)]]^{z_{ig}}. \tag{5}$$

Equation (5) is maximized via expectation maximization (EM; Dempster, Laird, and Rubin 1977). For the M-step, equation (5) is maximized with respect to $(\tau_1, \ldots, \tau_G; \theta_1, \ldots, \theta_G)$, holding $\mathbf{z}$ at $\tilde{\mathbf{z}}$. Given estimates $(\tilde{\tau}_g, \tilde{\theta}_g)$, $\tilde{\mathbf{z}}_g$ is given from the E-step:

$$\frac{\tilde{\tau}_g \phi_g(\mathbf{x}_i|\tilde{\theta}_g)}{\sum_{h=1}^{G} \tilde{\tau}_h \phi_h(\mathbf{x}_i|\tilde{\theta}_h)}. \tag{6}$$

For multivariate Normal mixtures used here, Fraley and Raftery (2002) give closed-form solutions for $\tilde{\tau}_g$ and $\tilde{\mu}_g$: $\tilde{\tau}_g = n_g/n$, where $n_g = \sum_{i=1}^{n} \tilde{z}_{ig}$ and $\tilde{\mu}_g = \left(\sum_{i=1}^{n} \tilde{z}_{ig}\mathbf{x}_i\right)/n_g$. The quantity $(1 - \tilde{\tau}_g)$ is referred to as the "uncertainty" in assigning i to cluster g, that is, $(1 - \tilde{\tau}_g)$ is the misclassification probability implied by the model.

3.2.1 Some challenges: Normality, EM convergence, and singular covariances

Social science data are frequently categorical or integer counts that are clearly not Gaussian. Although, to our knowledge, clustering mixture models with different distributional components have yet to be developed, there are a several options for applied work. First, if the data in question are *all* categorical, as is common with survey data and other sorts of contingency table setups, latent class analysis (Lazarsfeld and Henry 1968; Linzer and Lewis 2011; McCutcheon 1987) defines a set of models nearly identical to those described for MBC but substituting a mixture of multinomials for the Gaussian densities.[8] Second, if we confront data in which some of the k dimensions of $\mathbf{x}$ are not reasonably considered to be Normal, one possible strategy is to combine variables using tools such as principle components analysis (PCA) and then perform MBC on the principle components. As discussed below, the combination of MBC and model selection tools makes MBC well suited to incorporating PCA into clustering problems.

On the technical side, EM is known to be sensitive to starting values and can also have convergence problems. Convergence is most commonly a problem when an estimated covariance matrix approaches

[8] See Blaydes and Linzer (2008) for a recent political science application.

computational singularity as the likelihood diverges to infinity. This problem is known to occur in models with a large number of mixture components and/or in which the covariance matrices are allowed to vary across components (Fraley and Raftery 2005). In our experience, singular covariances are also common when n is small. Bensmail et al. (1997) and Bensmail and Meulman (2003) specify fully Bayesian estimation procedures to avoid this difficulty. Fraley and Raftery (2005) propose an EM estimator for the maximum a posteriori mode (MAP) of component parameters θ_g. Specifically, Fraley and Raftery assume a Bayesian predictive density proportional to

$$\mathcal{L}_{\mathrm{m}}\mathcal{P}(\tau_g, \mu_g, \Sigma_g|\eta),$$

where $\mathcal{P}$ is the prior and η are hyperparameters. Fraley and Raftery (2005) propose conjugate priors,[9] specify diffuse hyperparameter values, and then derive closed-form expressions for parameter estimation via EM analogous to those above. Including the prior drastically improves computational performance and stability. Fraley and Raftery (2005) show that in the absence of convergence problems, estimating the model with the prior yields results very close to those from the EM-MLE procedure.

3.2.2 Model choice and MBC

The challenge of MBC is to select both the number of clusters and the parameterizations of the covariance matrix. Since each combination of these choices represents a (nonnested) statistical model, MBC recasts the clustering problem as one of model selection. We thus have a variety of tools from which to choose. Dasgupta and Raftery (1998) and Fraley and Raftery (1998) recommend using the Bayesian Information Criterion (BIC) approximation to the Bayes's factor for model selection. The BIC is given as

$$\mathrm{BIC} \equiv 2\log\mathcal{L}(\mathbf{x}, \hat{\theta}_G) - m_G \log n,$$

where $\hat{\theta}$ is either the MLE or the MAP (depending on whether a prior has been specified) and m is the number of free parameters in the model. In comparing models, we choose the parameterizations that maximize the BIC. Conventionally, two models with a BIC difference less than 2 are difficult to distinguish, whereas a difference of 10 or greater constitutes strong evidence for favoring one model over another (Kass and Raftery 1995). The BIC penalizes relatively more complex models and so privileges simpler constructs (fewer clusters and/or more constrained covariance structures).

Although some have objected to using the BIC for model selection generally and as an approximation to the Bayes's factor specifically (Gelman and Rubin 1995; Weakleim 1999), the BIC model selection procedure has been shown to perform well in the MBC context. Using asymptotic reasoning, Keribin (2000) proves that, under certain conditions, the BIC generates a consistent estimator of the number of mixture components. The BIC criterion performs well in simulation experiments where the exact mixture distribution is known (Fraley and Raftery 1998; Biernacki, Celeux, and Govaert 2000). Other selection criteria exist, most notably the integrated completed likelihood (ICL) of Biernacki, Celeux, and Govaert (2000). We discuss the ICL in greater detail below.

3.2.3 Variable selection and MBC

Any collection of objects can be measured and classified on a large number of attributes (or dimensions). This is especially true when looking at complex entities, such as nation states, over time. Intuitively, one can imagine that as the number of dimensions approaches the number of objects to be classified, there must be increasingly tight clustering to discern any pattern in the data. Technically, as the number of dimensions increases so does the number of parameters to estimate for θ, imposing constraints on the effective number of dimensions we can consider. In the VoC application below, we have a maximum of 21 cases for any point in time.

In this context, it is important to mention that traditional data reduction techniques and cluster analysis do not easily work together. Chang (1983) proves that clustering information is not monotonically related to the eigenvalues of the principal components. There is no reason to believe that principal components

[9] A Normal prior for the mean vector and inverse Gamma (when $k = 1$) or inverse Wishart (for $k > 1$) for covariance.

with the largest eigenvalues are those providing the most information enabling us to discriminate across classes of objects. This problem is exacerbated when some of the random variables along various dimensions are non-Gaussian. Thus, reducing the dimensionality of the data by selecting principal components with the largest eigenvalues and then performing a cluster analysis, whether mixture model, hierarchical, or relocation clustering, generally is not justified.

The MBC-model selection approach provides one way around this problem. Raftery and Dean (2006) extend the notion of BIC-based model selection to include variable selection.[10] They develop an algorithm in which the data, Y, are partitioned into three sets: variables already selected for clustering (Y_1), variables being considered for inclusion or exclusion from Y_1, denoted Y_2, and all remaining variables (Y_3). The algorithm is initialized by choosing the variable on which there is the most evidence of clustering. At each subsequent step, two models are considered.[11] In the first, Y_2 gives no additional information on clustering *conditional* on Y_1. In the second, the Y_2 does improve clustering. At each step, the models are compared and a variable is included or excluded based on its effect on the BIC, maximized over the number of clusters and model parameterizations.[12] Moreover, the variable selection procedure provides a way in which to use dimension reduction techniques like principal components. The MBC algorithm, applied to principal components, chooses the extracted component with the greatest amount of clustering information rather than the one that maximizes variance "explained", as the eigenvalue criterion does.

3.2.4 Mixture components versus substantive clusters

The lack of a widely accepted formal definition for "cluster" poses a challenge. In the mixture model context, it is commonly assumed that the estimated mixture components are in fact the estimated clusters (Fraley and Raftery 1998). Biernacki, Celeux, and Govaert (2000) challenge this assumption, arguing that the goal of estimating the optimal mixture of Normals is not the same as estimating the number of clusters, thought of as clumps of observations that are both "cohesive and well-separated from the other data" (Baudry et al. 2010, 334). What an analyst may consider to be a cluster may not look like the ellipsoids generated by Normals; indeed, the clusters themselves may be mixtures.

The left panel of Fig. 1 illustrates the situation Biernacki et al. have in mind. There is reason to believe that there are two clusters, with the one in the lower right looking "unusual," perhaps with overlapping components. The center panel displays the MBC-BIC clustering solution with five components. This panel illustrates how the MBC-BIC method picks mixtures that are a good fit to the data but, as Biernacki, Celeux, and Govaert (2000) argue, may generate "too many" components when the goal is to identify clusters. Biernacki, Celeux, and Govaert (2000) propose using the ICL criterion instead. The ICL is given by

$$\mathrm{ICL} \equiv p(\mathbf{x}, \mathbf{z}|G) = \int_{\Theta_G} \mathcal{L}_{\mathrm{c}} \mathcal{P}(\theta|G) \mathrm{d}\theta, \tag{7}$$

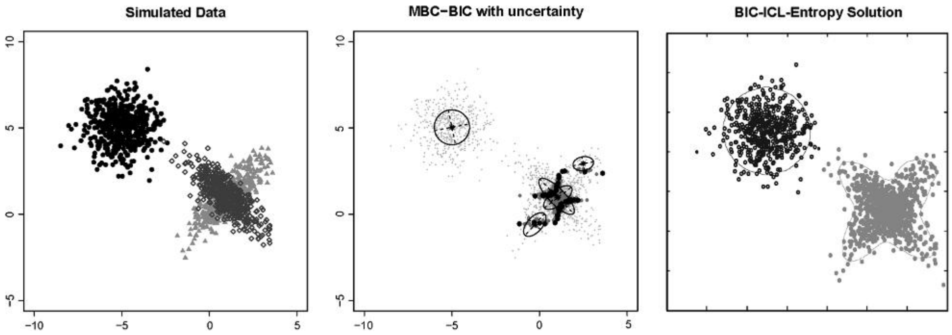


Fig. 1 Simulated data, MBC-BIC, MBC-BIC-entropy methods compared using fake data.

[10] See Dean and Raftery (2010) for an extension of the variable selection algorithm to multinomial mixtures and latent class analysis.

[11] Formally, we are interested in models characterizing $p(Y|\mathbf{z})$, where $\mathbf{z}$ defines cluster membership. Model 1 factors this into $p(Y_3|Y_2, Y_1)p(Y_2|Y_1)p(Y_1|\mathbf{z})$, whereas model 2 posits $p(Y_3|Y_2, Y_1)p(Y_2, Y_1|\mathbf{z})$.

[12] The maximum number of clusters must be set prior to analysis. For our application below, we set this maximum to seven unless otherwise noted. See Raftery and Dean (2006, 176–7) for a detailed elaboration of the algorithm.

where $\mathcal{P}(\theta|G)$ is the conditional prior. The approximation to the ICL is given as

$$\text{ICL} \approx \log \mathcal{L}_c(\mathbf{x}, \hat{\mathbf{z}}|\hat{\theta}, G) + \frac{m_G \log n}{2}.$$

The ICL approximation in equation (3.2.4) is roughly equal to the BIC penalized by the posterior mean entropy (Biernacki, Celeux, and Govaert 2000; Baudry et al. 2010), where entropy is given by

$$\text{Ent}(G) = -\sum_{i=1}^{n}\sum_{g=1}^{G} \hat{\tau}_{ig} \log \hat{\tau}_{ig}, \tag{8}$$

where $\hat{\tau}_{ig}$ is τ evaluated at $\mathbf{x}_i$ and $\hat{\theta}$, given G. The choice of BIC or ICL presents a dilemma well summarized by Celeux (2007, 10):

> ICL favors [choice of G] giving rise to [the] partitioning [of] the data with the greatest evidence. . . But ICL, which is not aiming to discover the true number of mixture components, can underestimate the number of components for simulated data arising from mixture with poorly separated components. . . BIC performs remarkably well to assess the true number of components from simulated data. . . But, for real world data sets, BIC has a marked tendency to overestimate the numbers of components. The reason is that real data sets do not arise from the mixture densities at hand, and the penalty term of BIC is not strong enough to balance the tendency of the loglikelihood to increase with [G] in order to improve the fit of the mixture model.

Baudry et al. (2010) provide a way out by combining the best of both BIC and ICL. The algorithm they propose uses the MBC-BIC approach (along with variable selection if needed) to generate the mixture solution. Denote this solution as $\hat{G}$. They then generate a series of candidate clustering solutions, $\hat{G}, \hat{G} - 1, \ldots, 1$, by successively merging mixture components into one another such that at each stage the components merged generate the largest entropy reduction.[13] Baudry et al. (2010) recommend choosing the clustering solution that has the same number of clusters as the ICL solution or, alternatively, using scree or "elbow" plots of entropy against the number of clusters. Baudry et al. (2010) provide MATLAB code to perform this analysis on MBC-BIC output from both the MATLAB `mixmod` toolbox and $\mathcal{R}$ output from `mclust`.

A few points are worth making. First, the iterative merging of components does not generate a hierarchical clustering; objects assigned to the same cluster at one stage could be assigned to different ones in subsequent stages. Second, each of the candidate clustering solutions fit the data equally well from a likelihood perspective. The BIC is constant across all clustering solutions; all that changes is the cluster labels. Third, the BIC-MBC-entropy procedure provides another way to relax the assumption of Normality. Although the mixture components are all Gaussian, the combinations they generate can be extremely flexible. As an example, the right panel of Fig. 1 displays the results from the MBC-BIC-entropy procedure, combining the components identified in the center panel until we reach two clusters, the number identified using the ICL procedure.

3.2.5 Summary

Older generations of cluster analysis tools are best considered as exploratory. The MBC and model selection approach improves previous efforts by (1) allowing for more flexible clustering geometries based on well-understood parametric distributions; (2) providing a principled way for selecting the optimal clustering solutions by comparing nonnested models via the BIC; and (3) generalizing the variable selection problem to one of model selection, thereby providing guidance on which variables to use, be they original variables or principal components. Recent work provides a suite of tools for using the mixture models to identify non-Gaussian clusters that may stand out in the data.

[13] Specifically, suppose we propose to merge components g and g'. The $\hat{\tau}_g$ stay the same for all clusters save g and g'; the new conditional probability for these clusters is simply $\hat{\tau}_{ig\cup g'} = \hat{\tau}_{ig} + \hat{\tau}_{g'}$. The entropy is $-\sum_{i=1}^{n}\sum_{j\neq g,g'} \hat{\tau}_{ij} \log \hat{\tau}_{ij} + \hat{\tau}_{ig\cup g'} \log \hat{\tau}_{ig\cup g'}$.

In general, then, the steps in a model-based cluster analysis are as follows:

1. Define the set of units and dimensions, **x**. If some random variables in **x** are non-Gaussian, consider transformations or combining with other variables via PCA.
2. Specify the maximum number of components, $G_{\max}$, over which to search.
3. Consider performing a variable selection procedure on the various dimensions of **x**. This is especially relevant if the variables are principal components.
4. Fit the model via EM for each parameterization of the covariance matrix for each number of components up to $G_{\max}$; if some of the models show degeneracy, then be sure to include a prior.
5. Compute the BIC for each model and select the BIC-maximizing model.
6. Iteratively combine components into clusters, examining the change in entropy at each stage. Choice of final clustering solution can be determined by using the ICL-generated number of clusters and plots of the model entropy against the number of clusters.

4 Application

4.1 *"Varieties of Capitalism"*

In attempting to make sense of the remarkable variation in political-economic institutions and outcomes across industrial democracies, there emerged a long and distinguished tradition in political science and economic sociology classifying countries into one of a small number of categories. Table 2 displays the classification of 21 advanced democracies according to the "three worlds of welfare states" of Esping-Andersen (1990), the institutional diversity of contemporary capitalism of Kitschelt et al. (1999), and the "varieties of capitalism" of Hall and Soskice (2001). We will concentrate on the VoC.[14]

The VoC literature claims that the 21 countries comprising the major industrial democracies of Europe, North America, and the Pacific Basin can be assigned the labels of either "liberal market economy" (LME) or "coordinated market economy" (CME). Although the historical and theoretical reasoning for these typologies is outside the scope of this paper, we remark that none of these typologies is driven by the application of an a priori definition. Rather, in the case of the VoC, the country clusters are held to be direct implications of a set of theoretical arguments. In identifying clusters, early work relied on visual heuristic methods based on additive indices for specific time points. Subsequent work (Obinger and Wagschal 2001; Amable 2003; Hicks and Kenworthy 2003; Saint-Arnaud and Bernard 2003; Hall and Gingerich 2009; Tepe, Gottschall, and Kittel 2010) has tried to put the VoC on sounder empirical ground. But these studies using different variables over different time periods and different methods have generated different clustering solutions and interpretations. Thus, the VoC is controversial: are there only two varieties of capitalism? Where should we put Italy? Are these categories immutable, at least over the period from 1980 to the present?

Notwithstanding these unstable results and ignoring Hall and Soskice's own refusal to use the VoC as a categorization scheme, the CME/LME dichotomy has begun to structure both quantitative and qualitative research. On the quantitative side, indicator variables representing whether a particular country is an LME have appeared as regressors (Rueda and Pontusson 2000; Ringe 2006; Taylor 2006; Oliver 2008; Huo and Feng 2010), sometimes in an effort to account for variation that others have cited as determining the initial classification (Hamann and Kelly 2008).[15] On the qualitative side, the VoC logic has been used to justify case selection as well as the dimensions for comparative case study (Thatcher 2004; Campbell and Pedersen 2007; Culpepper 2007).

In short, social scientists attempting to classify rich democracies have employed methods best characterized as exploratory. We observe unstable results and findings that hinge on the analysts' interpretations.

[14] But see Scruggs and Allan (2006, 2008) for a serious challenge to Esping-Andersen's three worlds.

[15] All these analyses are taking place in the context of time series-cross section data and treat cluster membership as time invariant.

Table 2 Twenty-one OECD economies and their categorizations

Country	*Country code*	*Three worlds*	*Types of capitalism*	*Varieties of capitalism*
Australia	AUS	Liberal	LME	LME
Canada	CAN	Liberal	LME	LME
Great Britain	GBR	Liberal	LME	LME
Japan	JPN	Liberal	GCME	NC/C
Switzerland	CHE	Liberal	SCME	CME
United States	USA	Liberal	LME	LME
Austria	AUT	Conservative	SCME	CME
Belgium	BEL	Conservative	SCME	CME
Germany	DEU	Conservative	SCME	CME
France	FRA	Conservative	SCME	NC/C
Italy	ITA	Conservative	SCME	NC/C
Denmark	DNK	Soc. Dem.	NCME	CME
Finland	FIN	Soc. Dem.	NCME	CME
The Netherlands	NLD	Soc. Dem.	SCME	CME
Norway	NOR	Soc. Dem.	NCME	CME
Sweden	SWE	Soc. Dem.	NCME	CME
Greece	GRC	NC/C	NC/C	NC/C
Ireland	IRL	NC/C	LME	LME
New Zealand	NZL	NC/C	LME	LME
Portugal	PRT	NC/C	NC/C	NC/C
Spain	ESP	NC/C	NC/C	NC/C

Note. The country codes are based on ISO 3166. The country classifications are LME, liberal market economy; CME, coordinated market economy; GCME, group coordinated market economy; NCME, national coordinated market economy; SCME, sectoral coordinated market economy; NC/C, not categorized or controversial.

These works implicitly avoid dimensionality problems by relying on additive indicators or data reduction techniques like principle components. Nevertheless, scholars have reified the VoC typology by including it in both regression and case-comparison studies and an explanatory variable.

4.1.1 VoC replications

In this section, we illustrate MBC by replicating two well-known studies purporting to uncover evidence for the VoC. All analysis was performed in $\mathcal{R}$ 2.8.2 (R Core Development Team 2007) using the `mclust` and `clustvarsel` libraries (Fraley and Raftery 2002; Raftery and Dean 2006; Fraley and Raftery 2007). For these replications, we set $G_{\max}$, the maximum number of components over which to search, to eight. When we encounter situations in which the MBC/model selection procedure identifies solutions with either only one cluster or more than six as the best-fitting model, we interpret these as demonstrating the absence of any clustering structure in the data.

Estevez-Abe, Iversen, and Soskice. We begin with an early VoC work that relies on visual heuristic methods to identify clusters. Estevez-Abe, Iversen, and Soskice (2001) argue that VoC theory leads them to expect rich democracies to cluster in two dimensions: the extent to which countries protect the unemployed and the degree of job protection for those already employed (see Fig. 2). The authors construct additive indices of three employment protection (EP) variables and three unemployment protection (UP) variables, respectively (see Table A1 in the Appendix for the data). They then plot the two indices and identify clusters heuristically. This process highlights the interplay between theory, typology, and theory testing. Estevez-Abe et al. did not posit a priori a definition of LME, take measurements of the appropriate criteria, and then assign labels or scores. Rather analysts expected there to be discernable structure in the data based on theory.

We revisit Fig. 2 in two ways. First, we perform two principal components analyses, one on the three EP variables and one on the three UP variables. Initially, we use the first principal component for each set of variables in order to reproduce the analysis of Estevez-Abe et al. as closely as possible. We then perform MBC with and without priors. Figure 3 displays results for the estimation without priors. In panel

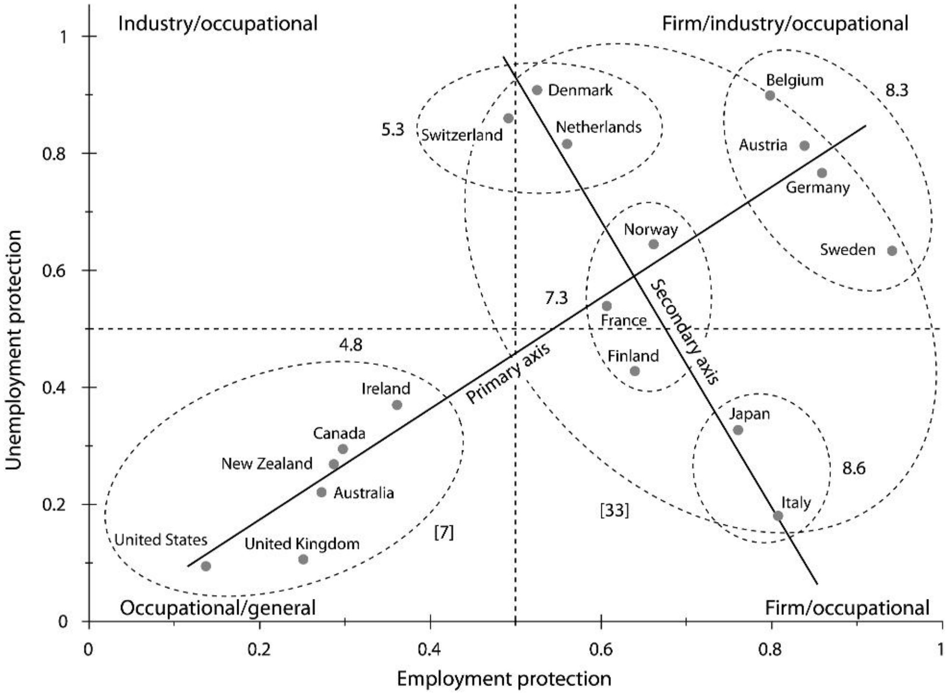


Fig. 2 Heuristic visual clustering. This figure is a replication of figure 4.2 from Estevez-Abe, Iversen, and Soskice (2001, 172).

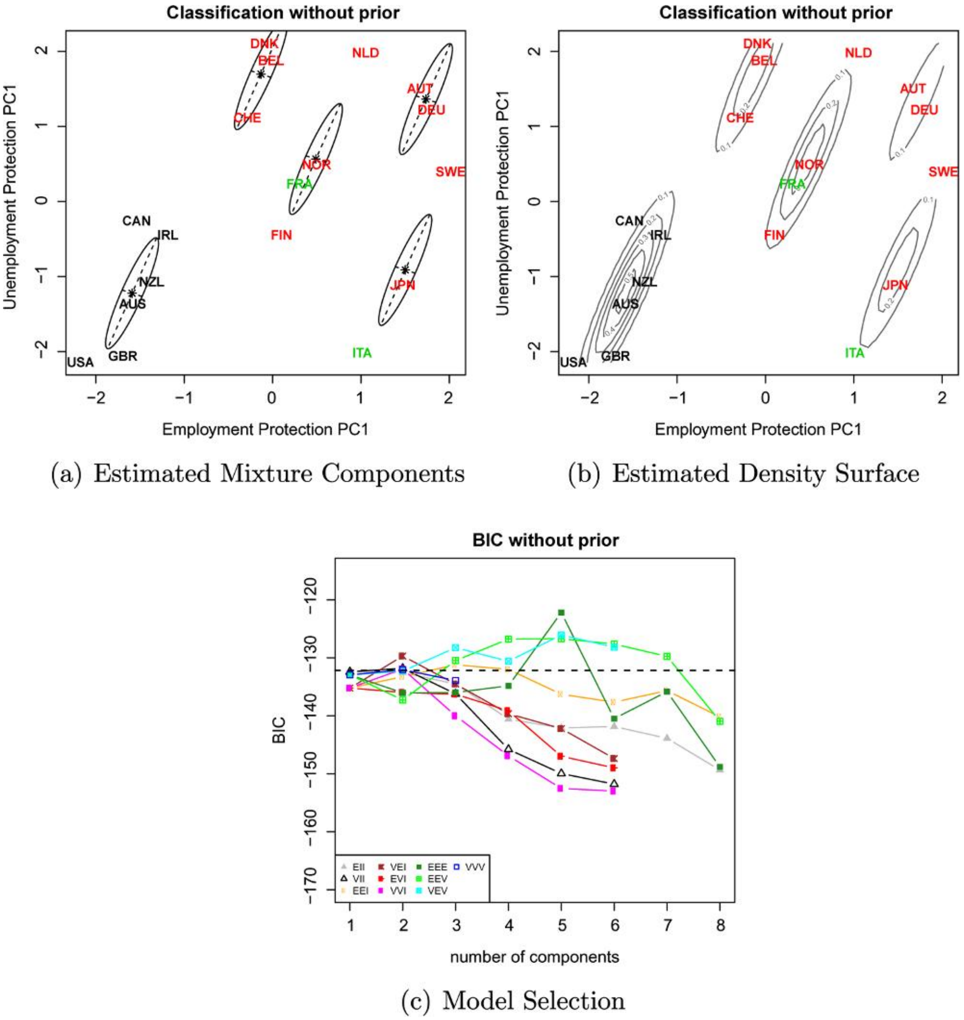
(c), we see that the BIC-maximizing model has five components and is ellipsoidal with equal variance across components. But there is volatility in the BIC and singular covariance matrices among the more complicated models. Panel (a) displays the model's ellipses based on the estimated component mean and variance parameters. Panel (b) displays estimated density contours. The LME cluster is clearly adumbrated in all the figures, with little uncertainty around case assignment. The classification corresponds roughly to the mini-clusters circled by Estevez-Abe et al., but Belgium, Sweden, and the Netherlands are assigned to different components.

Figure 4 displays results for MBC with a prior. Panel (a) shows that the BIC-maximizing model has two components oriented along the axes with equal shape. The classification corresponds to the VoC expectations, if we count controversial cases as CMEs. Figure 4(b) shows the tight estimated density contours. Looking at panel (c), we see that the BICs are smoother across models. The horizontal broken line represents a BIC value 10 less than the BIC value of the best-fitting model. The BIC for the two-component model is close to that for the one-, three-, and four-component models, making it difficult to mount a case that the two-component model is overwhelmingly preferred. That said, the BIC-maximizing three- and four-component models continue to place all the LMEs together in one cluster.[16]

Next we attempt to corroborate these results with weaker assumptions about which variables provide relevant information. We perform variable selection over the six principal components described above, identifying the first principal components of the EP variables and the first two UP principal components. As before MBC with no prior suffers from singularity for some models and selects a model with five components. The resulting classification has an adjusted Rand index (ARI; Hubert and Arabie 1985) of 0.26 when compared against the VoC classification implying that the two categorization schemes have little to do with one another.[17] If we include a prior, then all countries are assigned to one cluster. If,

[16] The ICL criterion also generates a two-cluster solution. Using the BIC-entropy procedure on the three- and four-component BIC solutions leads to the same cluster assignment as in Fig. 4.

[17] The ARI measures the degree of similarity between partitions of sets, ranging from 0 to 1, with 1 representing identical partitions See Steinley (2004) for a discussion of the merits and interpretation of the ARI. He argues that ARI < 0.65 represents "poor" relationship between the two partitions.



(a) Estimated Mixture Components

(b) Estimated Density Surface

(c) Model Selection

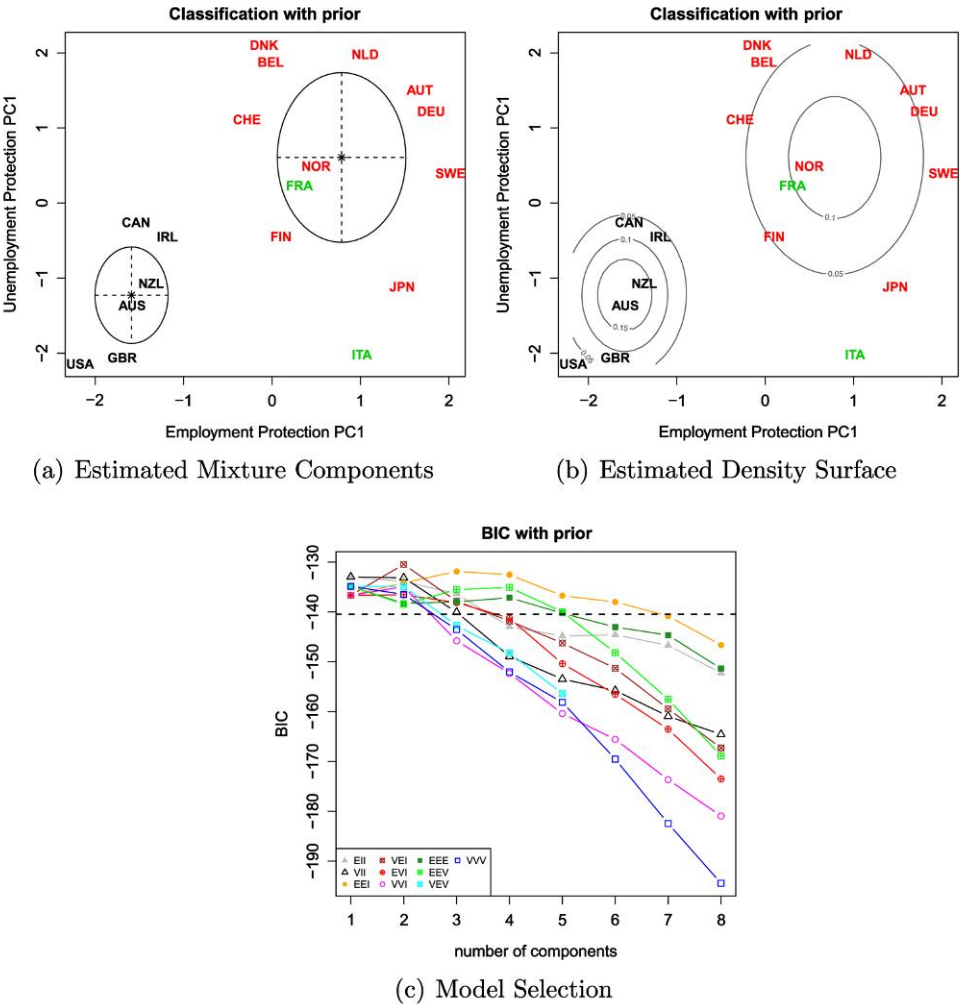
Fig. 3 MBC replication of Estevez-Abe et al. without priors. Country labels in black font are those theorists call LMEs, grey are CMEs and controversially classified in the original VoC typology. The ellipses in the first subfigure are based on the estimated mean and variance parameters for the mixture components. The second subfigure displays the density contours of the mixture model. The third subfigure plots the BIC for different models across values of G.

alternatively, we calculate a principal components decomposition on all six variables together (rather than first breaking them up by EP and UP), variable selection chooses the first and fourth through sixth principal components. Clustering results on these variables are also dissimilar to the VoC.[18]

Hall and Gingerich. In a recent paper, Hall and Gingerich (2009) argue that the VoC are defined by two dimensions: coordination in labor relations and coordination in corporate governance. They identify six variables, three measuring corporate governance and three measuring labor relations. The data and sources are displayed in Table A1. Within each group of three, the authors extract a "principal" factor and argue that countries can be arrayed along these dimensions.[19] We perform a principal components analysis on the corporate governance and labor relations variables, respectively, and then fit two sets of MBC

[18] Again, using the BIC-entropy procedure does not change conclusions here.

[19] Although they are not explicitly attempting to identify clusters, they interpret their results as largely congruent with the VoC theory.



(a) Estimated Mixture Components

(b) Estimated Density Surface

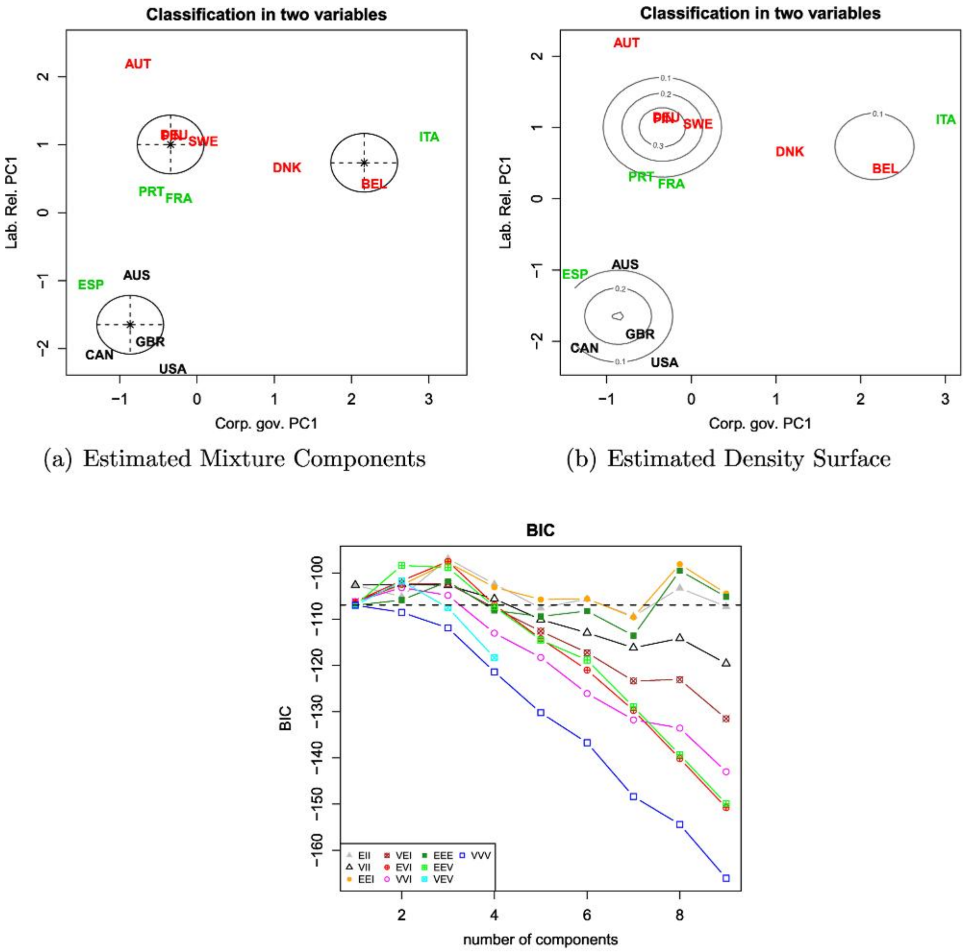
(c) Model Selection

Fig. 4 MBC replication of Estevez-Abe et al. with priors. Country labels in black font are those theorists call LMEs, grey are CMEs and controversially classified in the original VoC typology. The ellipses in the first subfigure are based on the estimated mean and variance parameters for the mixture components. The second subfigure displays the density contours of the mixture model. The third subfigure plots the BIC for different models across values of G.

procedures.[20] In the first set, we follow Hall and Gingerich and examine clustering in the two-dimensional space defined by the first principal components for labor relations and corporate governance, respectively. In the second set, we use variable selection on the labor relations and corporate governance principal components separately to identify the appropriate principal components. The algorithm selected the first and third principal components of each dimension. We then perform MBC over those four variables. Once again we observed degeneracy in the covariance matrices for some models, so all models are fit assuming the appropriate diffuse conjugate prior.

In Fig. 5, we display the classification results from MBC on only the first labor and corporate governance principal components. Panel (a) plots the parameterization of the best-fitting model—spherical, equal volume, and with three components. The estimated probability densities are plotted in panel (b). The classification of Denmark in the Italy-Belgium cluster is somewhat ambiguous here; we barely achieve 95% confidence in its cluster assignment. The BIC plot in panel (c) indicates that model selection is not straightforward here. Models with two and eight components fit the data nearly as well as the selected

[20] Note that several of their variables are indices or bounded below at 0, implying that they cannot be Normal variables. PCA again helps us circumvent this problem.



(a) Estimated Mixture Components

(b) Estimated Density Surface

(c) Model Selection

Fig. 5 MBC replication of Hall and Gingerich for two dimensions. Country labels in black font are those theorists call LMEs, grey are CMEs and controversially classified in the original VoC typology. The ellipses in the first subfigure are based on the estimated mean and variance parameters for the mixture components. The second subfigure displays the density contours of the mixture model. The third subfigure plots the BIC for different models across values of G.

three-component model. When we use variable selection on the labor relations and corporate governance principal components, we again obtain a three-component solution. However, Denmark is now unambiguously clustered with Sweden, Germany, France, etc. The uncertainty around Australia's placement has increased though it is still assigned to the same component as the United States with more than 95% certainty. Neither set of results corresponds well with the canonical VoC classification.[21] Both have ARI values <0.4 when compared to the VoC. Relying on the MBC-entropy procedure, we end up selecting the three-cluster solution with cluster assignment as reported in Fig. 5. This cluster solution bears faint resemblance to the theoretical VoC.

Discussion. We examine a theoretically expected typology, the VoC, using MBC tools to revisit two foundational studies purporting to show evidence of the VoC clustering. We see weak results for the VoC in these two exercises. In the replication of Estevez-Abe et al., we find some evidence for the VoC, but these results depend greatly on which principal components are included in the clustering model; those with the most clustering information generate clustering solutions having little in common with

[21] Nor even with a VoC modified with a "Latin" cluster (Saint-Arnaud and Bernard 2003).

the VoC. In the replication of Hall and Gingerich, we uncover clustering solutions that differ markedly from the VoC. We never observe a two-cluster solution, and Spain is unambiguously clustered with the United States and other LMEs. This latter finding is particularly problematic since no VoC theorist has lumped Spain with "LMEs" and empirical applications of the VoC usually involve indicator variables for LME membership. In both replications, the best-fitting models did not overwhelmingly outperform others on a BIC basis. Several different clustering solutions had similar BIC values, reflecting the difficulty in inferring clustering in several dimensions when we only have 14–18 objects in our data set. Using the MBC-entropy procedure of Baudry et al. for combining mixture components does not alter our conclusions.

Collectively, these authors have theoretical expectations that the VoC clusters should be visible in the variables we examine. Instead, we find unstable results that depend greatly on the combinations of variables we use, even among this highly circumscribed data set.[22] Adding to the mixed evidence of other studies, our MBC analysis indicates that the VoC are not consistently discernable in the data that theorists in the literature consider relevant. Thus, reification of the VoC categories through dummy variables in regression models or as criteria for case selection appears unwarranted.

5 Conclusions

Typologies can emerge from several different research activities: application of a definition, theory building, and testing theoretical implications. Methods for examining typologies as theoretical implication have not been well developed. We argue that MBC provides a statistically principled way to address these issues. Starting with the assumption that the data are generated by a mixture of multivariate Gaussian densities, MBC grounds cluster analysis in probability theory. With the help of approximate Bayes's factors and suitable search algorithms, we are able to provide better estimates of the number of clusters, their shape, composition, *and the level of uncertainty surrounding them* than earlier methods.

When we are trying to *infer* the existence or robustness of a classification scheme, we can amend the received wisdom on "extension" and "intention." Prior discussion held that as typologies are extended to more cases, the dimensionality of the definition must be reduced. Clearly this only holds in the denotive context. When we are attempting to make inference about the existence of clusters, increasing the number of cases can help us in two ways: it can increase the precision of our estimated classification and it can permit us to consider more dimensions on which the objects might vary. When making inference, the curse of dimensionality applies.

MBC also provides a way to guard against reification of typologies, especially those purported to be an implication of theory. Cluster membership is only meaningful as an explanatory construct when we have preexisting empirical evidence of cluster membership for the relevant variables. Asserting cluster membership as a testable proposition in a space with greater dimensionality than the number of cases we have is not meaningful. MBC helps us sharpen our thinking on both tasks.

We applied MBC to a substantive problem. The varieties of capitalism literature is based on a typology that has been asserted as an implication of a set of theoretical arguments. We replicate two important studies in this literature and found that empirical evidence in favor of the VoC clustering is mixed at best. To the extent we can believe there is evidence in favor of the VoC, we must also have strong confidence in the choice of variables over which clustering is supposed to exist. Unfortunately, the literature has not provided anything approaching a consensus about the exact variables defining the VoC. This insight is important for researchers engaged in both typical regression analysis and case-comparison research. We believe that including dummy variables in regression models signifying VoC cluster membership is far too coarse, generally not justified and not a good idea.

Although this negative finding is an important corrective to the VoC literature, we believe that MBC tools will be fruitfully applied to increase our confidence that theoretically inspired typologies are in fact discernible in the data.

[22] A broader literature search turned up over three dozen variables that authors have linked to the VoC.

Data Appendix

Table A1 Data and sources—replications

Variable	*Description*	*Source*
Iversen (2005) and Estevez-Abe, Iversen, and Soskice (2001)		
EP legislation	Index of the “restrictiveness” of individual hiring and firing rules	Iversen (2005, 47)
Collective dismissal protection	Index of restrictions on collective dismissal	Iversen (2005, 47)
Company-based protection	Measure of company-level EP	Iversen (2005, 47)
Net unemployment replacement rates	Net unemployment replacement rates for 40-years-old representative worker	Iversen (2005, 50)
Generosity of benefits	% Gross Domestic Product (GDP) paid in unemployment benefits/(unemployed/ total population)	Iversen (2005, 50)
Definition of suitable jobs	Index of restrictions on definition of suitable job	Iversen (2005, 50)
Hall and Gingerich (2009, 11)		
Shareholder power	Composite measure of legal regulations between ordinary shareholders vis-a-vis managers and dominant shareholders	La Porta et al. (1998)
Dispersion of control	Number of firms that widely held relative to the number with controlling shareholders	La Porta, Lopez-de Silanes, and Shleifer (1999, Table 2)
Size of stock market	Market valuation of equities on the stock exchange as percentage of its GDP	Nestor and Thompson (2001)
Level of wage coordination	Level at which unions coordinate wage claims and employers coordinate wage offer	Layard, Jackman, and Nickell (1991, 52)
Degree of wage coordination	Degree of (strategic) wage bargaining coordinated by unions and employers	OECD (1997, 71)
Labor turnover	Number of employees who had held their jobs for less than one year as a percentage of all employees	OECD (1997, 138)

References

Ahlquist, John S., and Christian Breunig. 2009. *Country clustering in comparative political economy*. Max Planck Institut für Gesellschaftsforschung Discussion Papers (09/5), Cologne, Germany.

Amable, Bruno. 2003. *The diversity of modern capitalism*. New York: Oxford University Press.

Banfield, J. D., and Adrian E. Raftery. 1993. Model-based Gaussian and non-Gaussian clustering. *Biometrics* 48:803–21.

Baudry, Jean-Patrick, Adrian E. Raftery, Gilles Celeux, Kenneth Lo, and Raphael Gottardo. 2010. Combining mixture components for clustering. *Journal of Computational and Graphical Statistics* 19:332–53.

Bensmail, H., G. Celeux, Adrian E. Raftery, and C. P. Robert. 1997. Inference in model-based cluster analysis. *Statistics and Computing* 7:1–10.

Bensmail, H., and J. J. Meulman. 2003. Model-based clustering with noise: Bayesian inference and estimation. *Journal of Classification* 20:49–76.

Biernacki, C., Gilles Celeux, and G. Govaert. 2000. Assessing a mixture model for clustering with integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22:719–25.

Blaydes, Lisa, and Drew A. Linzer. 2008. The political economy of women's support for fundamentalist Islam. *World Politics* 60:576–609.

Bueno de Mesquita, Bruce, Alastair Smith, Randolph M. Siverson, and James D. Morrow. 2003. *The logic of political survival*. Cambridge, MA: MIT Press.

Campbell, John L., and Ove K. Pedersen. 2007. The varieties of capitalism and hybrid success: Denmark in the global economy. *Comparative Political Studies* 40:307–32.

Celeux, Gilles. 2007. Mixture models for classification. In *Advances in data analysis: Proceedings of the 30th Annual Conference of the Gesellschaft fr Klassifikation e.V., Freie Universitt Berlin, March 810, 2006*, ed. Reinhold Decker and Hans J. Lenz, 3–14. New York: Springer.

Celeux, G., and G. Govaert. 1992. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis* 2:73−82.

Celeux, Gilles, and Gerard Govaert. 1993. Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of Statistical Computation and Simulation* 47:127–46.

Chang, Wei-Chien. 1983. On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics* 32:267–75.

Collier, David, Jody Laporte, and Jason Seawright. 2008. Typologies: Forming concepts and creating categorical variables. In *The Oxford handbook of political methodology*, ed. Jeanette M. Box-Steffensmeier, Henry E. Brady, and David Collier, 152–73. Oxford: Oxford University Press.

Collier, David, and James E. Mahoney. 1993. Conceptual stretching revisited: Adapting categories in comparative analysis. *American Political Science Review* 87:845–55.

Cox, Trevor F., and Michael A. A. Cox. 2001. *Multidimensional scaling*. New York: Chapman & Hall.

Culpepper, Pepper D. 2007. Small states and skill specificity: Austria, Switzerland, and interemployer cleavages in coordinated capitalism. *Comparative Political Studies* 40:611–37.

Dasgupta, Abhijit, and Adrian E. Raftery. 1998. Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* 93:294–302.

Dean, Nema, and Adrian E. Raftery. 2010. Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics* 62:11–35.

Dempster, A. P., N. M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39:1–38.

Elman, Colin. 2005. Explanatory typologies in qualitative studies of international politics. *International Organization* 59: 293–326.

Esping-Andersen, Gosta. 1990. *The three worlds of welfare capitalism*. Princeton, NJ: Princeton University Press.

Estevez-Abe, Margarita, Torben Iversen, and David Soskice. 2001. Social protection and the formation of skills: A reinterpretation of the welfare state. In *Varieties of capitalism*, ed. Peter A. Hall, and David Soskice, 145–83. New York: Oxford University Press.

Fraley, Chris, and Adrian E. Raftery. 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal* 41:578–88.

———. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97:611–31.

———. 2005. Bayesian regularization for normal mixture estimation and model-based clustering. Technical report No. 486. Department of Statistics, University of Washington.

———. 2007. MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical report No. 504. Department of Statistics, University of Washington.

Ganesalingam, S. 1989. Classification and mixture approaches to clustering via maximum likelihood. *Journal of the Royal Statistical Society Series C Applied Statistics* 38:455–66.

Geddes, Barbara. 2003. *Paradigms and sand castles*. Ann Arbor, MI: University of Michigan Press.

Gelman, Andrew, and Donald B. Rubin. 1995. Avoiding model selection in Bayesian social research. *Sociological Methodology* 25:165–73.

Grimmer, Justin. 2010. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis* 18(1):1–35.

Grimmer, Justin and Gary King. 2011. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences* 108:2643–50.

Hall, Peter A., and Daniel W. Gingerich. 2009. Varieties of capitalism and institutional complementarities: An empirical analysis. *British Journal of Political Science* 39:449−82.

Hall, Peter A., and David Soskice, eds. 2001. *Varieties of capitalism.* New York: Oxford University Press.

Hamann, Kerstin, and John Kelly. 2008. Varieties of capitalism and industrial relations. In *The Sage handbook of industrial relations*, eds. Paul Blyton, Nicolas Bacon, Jack Fiorito, and Edmund Heery, 129–48. London: Sage.

Hicks, Alexander, and Lane Kenworthy. 2003. Varieties of welfare capitalism. *Socio-Economic Review* 1:27–61.

Hillard, D., S. Purpura, and J. Wilkerson. 2008. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology and Politics* 4(4):31–46.

Hubert, Lawrence, and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2(1):193–218.

Huo, Jingjing, and Hui Feng. 2010. The political economy of technological innovation and employment. *Comparative Political Studies* 43:329–52.

Iversen, Torben. 2005. *Capitalism, democracy, and welfare*. New York: Cambridge University Press.

Kass, R. E., and Adrian E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90:773–95.

Kaufman, Leonard, and Peter J. Rousseeuw. 2005. *Finding groups in data: An introduction to cluster analysis*. New York: Wiley-Interscience.

Keribin, C. 2000. Consistent estimation of the order of mixture models. *Sankhya* 62:49–66.

Kitschelt, Herbert, Peter Lange, Gary Marks, and John D. Stephens. 1999. Convergence and divergence in advanced capitalist democracies. In *Continuity and change in contemporary capitalism*, eds. Herbert Kitschelt, Peter Lange, Gary Marks, and John D. Stephens, 427–60. New York: Cambridge University Press.

Klebanov, Beata Beigman, Daniel Diermeier, and Eyal Beigman. 2008. Lexical cohesion analysis of political speech. *Political Analysis* 16:447–63.

La Porta, R., F. Lopez-de Silanes, and A. Shleifer. 1999. Corporate ownership around the world. *Journal of Finance* 54(2):471–517.

La Porta, R., F. Lopez-de Silanes, A. Shleifer, and R. W. Vishny. 1998. Law and finance. *Journal of Political Economy* 106:1113–55.

Layard, P., R. Jackman, and S. Nickell. 1991. *Unemployment*. New York: Oxford University Press.

Lazarsfeld, P. F., and N. W. Henry. 1968. *Latent structure analysis*. Boston, MA: Houghton Mifflin.

LeDuc, L., R. G. Niemi, and P. Norris. 1996. *Comparing democracies: Elections and voting in global perspective*. Thousand Oaks, CA: Sage Publications.

Lijphart, Arend. 1999. *Patterns of democracy: Government forms & performance in thirty-six countries*. New Haven, CT: Yale University Press.

Linzer, Drew A., and Jeffry B. Lewis. 2011. poLCA: Polytonomous latent class analysis. *Journal of Statistical Software* 42(10):1–29.

Marshall, Monty G., Keith Jaggers, and Ted Robert Gurr. 2004. Polity IV. Technical report. Center for Systemic Peace, University of Maryland.

McCombs, Maxwell. 2004. *Setting the agenda*. Cambridge: Polity.

McCutcheon, A. L. 1987. *Latent class analysis*. Thousand Oaks, CA: Sage.

Milligan, Glenn W. 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika* 45:325–42.

Milligan, Glenn W. 1981. A review of Monte Carlo tests of cluster analysis. *Multivariate Behavioral Research* 16:379–407.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16:372–403.

Nestor, Stilpon and John K. Thompson. 2001. Corporate governance patterns in OECD countries: Is convergence under way? In *Corporate governance in Asia: A comparative perspective*, 19–42. Paris: OECD.

Obinger, Herbert, and Uwe Wagschal. 2001. Families of nations and public policy. *West European Politics* 24(1):99–114.

OECD. 1997. *Employment outlook*. Paris: OECD.

Oliver, Rebecca. 2008. Diverging developments in wage inequality? Which institutions matter? *Comparative Political Studies* 41(12):1551–82.

Pemstein, Daniel, Stephen A. Meserve, and James Melton. 2010. Democratic compromise: A latent variable analysis of ten measures of regime type. *Political Analysis* 18:426–49.

Quinn, Kevin, Burt L. Monroe, Michael Colaresi, Michael Crespin, and Dragomir R. Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 54(1):209–28.

Raftery, Adrian E., and Nema Dean. 2006. Variable selection for model-based clustering. *Journal of the American Statistical Association* 101(473):168–78.

R Core Development Team. 2007. R 2.5.1—A language and environment.

Ringe, Nils. 2006. Policy preference formation in legislative politics: Structures, actors, and focal points. *American Journal of Political Science* 49:731–45.

Rueda, David, and Jonas Pontusson. 2000. Wage inequality and varieties of capitalism. *World Politics* 52:350–83.

Saint-Arnaud, Sebastien, and Paul Bernard. 2003. Convergence or resilience? A hierarchical cluster analysis of the welfare regimes in advanced countries. *Current Sociology* 51:499–527.

Sartori, Giovanni. 1970. Concept misinformation in comparative research. *American Political Science Review* 64:1033–53.

Scruggs, Lyle, and James Allan. 2006. Welfare state decommodification in eighteen OECD countries: A replication and revision. *Journal of European Social Policy* 16(1):55–72.

———. 2008. Social stratification and welfare regimes for the 21st century: Revisiting the three worlds of welfare capitalism. *World Politics* 60(4):642–64.

Steinley, Douglas. 2004. Properties of the Huber-Arabie adjusted rand index. *Psychological Methods* 9:386–96.

Sulkin, Tracy. 2005. *Issue politics in congress*. New York: Cambridge University Press.

Taylor, Mark Zachary. 2006. Empirical evidence against varieties of capitalism's theory of technological innovation. *International Organization* 58:601–31.

Tepe, Markus, Karin Gottschall, and Bernhard Kittel. 2010. A structural fit between states and markets? Public administration regimes and market economy models in the OECD. *Socio-Economic Review* 8:653–84.

Thatcher, Mark. 2004. Varieties of capitalism in an internationalized world: Domestic institutional change in European telecommunications. *Comparative Political Studies* 37:751–80.

Treier, Shawn, and Simon Jackman. 2008. Democracy as latent variable. *American Journal of Political Science* 52(1):201–17.

Vanhanen, Tatu. 2003. Democratization and power resources. 1850–2000.

Venables, W. N., and B. D. Ripley. 2002. *Modern applied statistics with S*. 4th ed. New York: Springer.

Ward, J. H. 1963. Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association* 58:234–44.

Weakleim, David L. 1999. A critique of the Bayesian information criterion for model selection. *Sociological Methods and Research* 27:359–97.

Weber, Max. 1949. Objectivity in social science and social policy. In *The methodology of the social sciences*, eds. Edward A. Shils and Henry A. Finch, 49–112. New York: The Free Press.

Zhong, Shi, and Joydeep Ghosh. 2003. A unified framework for model-based clustering. *Journal of Machine Learning Research* 4:1001–37.