ENCYCLOPEDIA OF COGNITIVE SCIENCE
2003
©Macmillan Reference Ltd

# Mixture Models of Individual Heterogeneity

mixture models # heterogeneity # identification
Stahl, Dale O.
Dale O. Stahl
University of Texas, Austin, TX USA

**[Definition: a model of a population consisting of heterogeneous types of individuals, each type exhibiting a distinct probabilistic behavior, with the probability of each type equal to that type's proportion in the population.**

## Introduction.

That all humans do not behave alike is obvious to even the most casual observer. How to represent this heterogeneity in a model that allows us to predict behavior is far less obvious. The simplist model of heterogeneity assumes that behavior is normally distributed about a mean for the population. An important feature of this model is that there is a single mode of behavior: i.e. if one graphed a histrogram of the behavior, one would find a single hump. We focus here on the countless cases where behavior is "multimodal": a histogram of the behavior would have multiple humps. For example, when voting on the level of federal spending for education, Democrats might be a unimodal population, and Repulicans might be a unimodal population, but the mean desired spending level of these populations would be quite different, so Congress as a whole would be a bimodal population.

Formally, suppose there are $K+1$ unimodal subpopulations, which we will index $k = 0, 1, ..., K$. An individual memberof subpopulation $k$ will be refered to as a type-$k$ individual. Suppose some member of the whole population faces a situation that calls for a behavioral response. Let $s$ denote this situation and all the relevant data about the situation, and let $X(s)$ denote the set of possible behaviors in this situation. Let $P_k(x|s)$ denote the probability that a type-$k$ individual will exhibit behavior $x$ from the set $X(s)$. Note that x could be a whole dynamic sequence of behaviors. Finally, let $\alpha_k$ denote the proportion of the whole population comprising subpopulation $k$ (or equivalently, the probability that a randomly drawn individual from the whole population is type $k$). Then, the probability that a random individual from the whole population will exhibit behavior x in situation s is given by

$$P(x \mid s) \equiv \sum_{k=0}^{K} \alpha_k P_k(x \mid s) \qquad (1)$$

Equation (1) is a canonical "mixture model".

When the probabilistic behavior of each subpopulation, $P_k(x|s)$, is pre-determined, then only the population proportions, $\alpha_k$, need be estimated from the behavioral data. Identifcation only requires that $P_k(x|s) \neq P_j(x|s)$, for $k \neq j$.

In other cases, the probabilistic behavior might be specified parametrically as $P_k(x|s,\beta_k)$, where $\beta_k$ is a vector of parameters for the subpopulation $k$. For example, $\beta_k$ might represent the mean and variance of a normal distribution. Since $\beta_j = \beta_k$ would yield identical probabilistic behavior, this parametric case generally requires identifying restrictions on the parameter space (such as $\beta_j < \beta_k$).

In the next section, we illustrate the application of mixture models with the Level-n model of bounded rationality by Stahl and Wilson (1994,5).

## Level-n Model of Bounded Rationality

Consider a two-player finite normal-form game in which the payoff to player i is $U_{ijk}$ when player i chooses action j and the other player chooses action k. Let $A \equiv \{1, ..., J\}$ denote the set of actions available to both players. The Stahl-Wilson (1994) model begins with the assumption that a proportion, $\alpha_0$ of the population has no understanding of the game and by virtue of the principle of insufficient reason a type-$0$ individual is equally likely to choose any action in $A$. Hence, $P_0(j|s)$ is the uniform distribution over $A$, and $s$ in this context represents the data for the game (i.e. A and U). In the language of Stahl-Wilson, these players are called "level-0" types.

A Bayesian rational player must have a belief about what the other player will do. The simplist non-informative model of other players is that they are level-0 types. A player that believes all others are level-0 types is called a "level-1" type, and chooses an error-prone best-response to this belief. Specifically, define

$$y_{1ij} \equiv \sum_{k=1}^{J} U_{ijk} P_0(k \mid s), \qquad (2)$$

which is the expected payoff to player i when choosing action j against a level-0 player. Then, the probabilistic choice function for a level-1 type is specified logistically as

$$P_{1i}(j \mid s,\beta_1) \equiv exp(\beta_1 y_{1ij}) / \sum_{k=1}^{J} exp(\beta_1 y_{1ik}), \qquad (3)$$

where $\beta_1$ is the precision of a level-1 type; the higher the precision, the higher the likelihood the choice will be the best-response to the belief, and the lower the precision, the more equally probable will be all the actions. Moreover, a logistic choice function has the property that the order of the choice probabilities corresponds to the order of the expected payoffs $y_{1ij}$. The proportion of level-1 types in the population is denoted $\alpha_1$, and for simplicity of presentation we implicitly assume the same proportion for both player 1 and player 2 (i.e. a single population model).

The level-n theory proposes a hierarchy of types in which a level-n type believes that all other players are level-k types with k < n. For simplicity of explanation and as an example, let us assume that a level-2 type believes that all other players are level-1 types. Then, the expected payoff to player i when choosing action j against a level-1 player is

$$y_{2ij} \equiv \sum_{k=1}^{J} U_{ijk} P_{-i1}(k \mid s, \beta_1), \tag{4}$$

where "-*i*" means "the other player, not i". The logistic probabilistic choice function is

$$P_{2i}(j \mid s, \beta) \equiv exp(\beta_2 y_{2ij}) / \sum_{k=1}^{J} exp(\beta_2 y_{2ik}), \tag{5}$$

where $\beta \equiv (\beta_1, \beta_2)$.

This simple three-type, level-n model yields a mixture model of the form:

$$P_i(j \mid s, \alpha, \beta) \equiv \alpha_0 P_0(j \mid s) + \alpha_1 P_{1i}(j \mid s, \beta_1) + \alpha_2 P_{2i}(j \mid s, \beta), \tag{6}$$

where $\alpha \equiv (\alpha_0, \alpha_1)$, and $\alpha_2 = 1-\alpha_0 - \alpha_1$, leaving only four free parameters. This mixture model can be expanded in a straightforward manner to test for the presense of additional types in the population.

## Identification Issues.

Since all the types in eq(6) collapse to uniformly random choice when $\beta_n = 0$, we need to impose identifying restrictions so $P_0(j \mid s)$, $P_{1i}(j \mid s, \beta_1)$, and $P_{2i}(j \mid s, \beta)$ are distinguishable given the sample size of the observed data. A Monte Carlo simulation can be used to determine appropriate lower bounds for the $\beta_n$.

Even having imposed these parameter restrictions, for many games, both level-1 and level-2 types will behave the same, so such games will be inadequate to identify the $\alpha_n$ parameters. The solution to this identification problem is to use a variety of games, so that each type predicts distinctly different patterns of behavior across all the games. Stahl and Wilson (1994) use ten symmetric $3 \times 3$ games; which permits $3^{10}$ (59,049) distinct patterns of choice for an individual, and an underlying space of probabilistic behavior of dimension $2^{10}$. While this approach creates more than enough possibilites to identify the parameters of the proposed mixture model, the curse of dimensionality renders it impossible in practice to use non-parametric methods to characterize the underlying distribution of behavior, and therefore, specificaton tests *vis-a-vis* the true data generation process are hopeless.

## Hypothesis Testing.

On the other hand, likelihood ratio comparisons can be used to test alternative model specifications. For example, the hypothesis that there are no level-2 types in the population sampled is equivalent to restricting $\alpha_2 = 0$. However, there is a complication because 0 is on the boundary of the parameter space (see Self and Liang, 1987; and Feng and McCulloch, 1996). Although the classical regularity conditions (as typically stated in advanced econometrics textbooks) are not met, the maximum likelihood estimators remain consistent. Self and Liang (1987) suggest that the true asymptotic distribution of the likelihood ratio statistic under the null hypothesis is a mixture of chi-squares with 0 and 1 degree of freedom. Since the right tail of the density of any such mixture lies to the left of a chi-square(1) density, the conventional chi-square(1) test would be too conservative, increasing the p-value of the observed statistic and lowering the probability of rejection; hence, a rejection of the null hypthesis ($\alpha_2 = 0$) using the conventional chi-square tests would hold under the true asymptotic distribution.

Alternative theories of behavior naturally suggest different types that can be easily added to the mixture model. For example, when each game has a unique pure-strategy Nash equilibrium, a "Nash" type can be added. If the Nash type is specified as putting probability one on the unique Nash equilibrium, then it is highly likely that one will be able to reject the hypothesis that there are pure Nash types in the population, because just one non-Nash choice by an individual would imply a zero probability of being such a pure Nash type. In general, theories that make extreme predictions are easily rejected, and beg to be augmented with a theory of "errors".

There are two simple and reasonable theories of errors. The first entails uniform trembles: with probability $(1-\varepsilon)$ the individual chooses the Nash equilibrium, and with probability $\varepsilon$ any action is equally likely to be chosen. This specification introduces an additional parameter to be estimated. The second model of errors is prior-based: as in eq(2), we define the expected payoff of each action given the prior belief that all other players will choose the Nash equilibrium. Then, we define the logistic Nash choice function as in eq(3). Again, one parameter is introduced, but now the choice probabilities are positively correlated with the expected payoffs. Haruvy and Stahl (1999) find that the logistic theory fits laboratory data much better than the uniform tremble theory. When a game has multiple Nash equilibria, they find that the logistic theory using the prior belief that each Nash equilibrium action is equally likely fits the data much better than any equilibrium selection criteria (including payoff dominance, risk dominance, and security).

Given the huge variety of behavior that is possible, one might expect that adding most any type to a mixture model will improve the fit. However, we have often failed to reject the hypothesis that certain types are absent. For instance, Stahl and Wilson (1995) failed to reject the absense of a rational expectations type, Haruvy, Stahl and Wilson (1999) failed to reject the absense of a maximin type, and Haruvy and Stahl (1999) failed to reject the absense of payoff-dominance and risk-dominance types. All of these tests had adequate power.

As pointed out, using a variety of games to identify the model parameters creates the potential for enormous behavioral diversity, and highlights the poverty of typical laboratory sample sizes. Given this diversity, it is often possible to reject the hypothesis that the parameters that maximize the likelihood of one small sample (of 10-30 individuals) are the same as the parameters that maximize the likelihood of another similarly sized sample. However, we caution the readers that these rejections are artifacts of overfitting small samples. In the absense of any pre-determined, observeable criteria for distinguishing among different groups of laboratory subjects, we learn nothing useful from separate parameter estimates for each group. Rather, it is better to pool all the data and obtain one estimate for the general population.

## Posterior Probabilies of Individual Types

Given parameter estimates $(\hat{\alpha},\hat{\beta})$, Bayes theorem allows us to compute the posterior probability that any given individual is type-*n,* denoted $\alpha^p_{ni}$.

$$\alpha^p_{ni} = \hat{\alpha}_n P_{ni}(x_i \mid s,\hat{\beta}_n) / \sum_{k=0}^{K} \hat{\alpha}_k P_{ki}(x_i \mid s,\hat{\beta}_k) \ , \qquad (7)$$

where $x_i$ stands for the choices of individual i over all the games. Stahl and Wilson (1995) show how to modify this formula to account for the uncertainty inherent in the parameter estimates. They also find that 38 of 48 participants in their experiments can be identified with one type (i.e. $\alpha^p_{ni}$ > 90% for some n). We suggest that this would hardly have been the case if the level-n model were not capturing a significant part of the true data generating process.

## Constancy of Types

The mixture model discussed and estimated assumes that an individual is one type for all situations he/she faces. The above results on the posterior probabilites supports that assumption. However, we need to consider alternatives to obtain a more rigorous conclusion. One alternative is the hypothesis that an individual draws his type from the population of types (characertized by the $\alpha_k$'s) independently for each game, which implicitly entails that all individuals are *ex ante* alike. This ex ante homoegeneity hypothesis is strongly rejected by our data.

Consider instead the hypothesis that a type-*k* individual is highly likely to behave in the typical way, but with some small probability can behave like another type. Specifically, suppose that with probability (1-$\varepsilon$) the individual behaves like a type-*k*, but with probability $\varepsilon$ is equally likely to behave like any type. For the Stahl-Wilson (1995) data, we found a statistically significant improvement in the maximized log-likelihood with an estimate of $\hat{\varepsilon}$ = 0.05. In other words, individuals appear to be true to one type of behavior 95% of the time, and otherwise tremble to other types of behavior. Since the level-0 type is part of the model, not surprisingly, allowing type trembles lowers the estimated proportion of the population that is type-*0*. Another attractive feature of this type-tremble mixture model is that it feeds easily into the Rule Learning framework of Stahl (2000, 2001).

## Equivalent Mixture Models

For any distribution of behavior, $P(x \mid s)$, there are obviously uncountably many ways to represent that distribution as a mixture model, eq(1). All of these are equivalent mixture models, and as such none can be rejected in favor of any other. Does this fact mean that mixture models are non-falsifiable and hence unworthy of scientific research? The answer is no for three reasons.

First, the true data generating process may be a mixture of types, in which case, clever ways of isolating subpopulations of types would lead to falsifiable predictions. Second, representing the data generating process as a mixture is a constructive approach that starts with archetypes and aggregates to population behavior. This process does produce falsifiable hypotheses about the constituent types, and also provides a practical means of making prediction for novel situations. Third, the curse of dimensionality prevents us from obtaining a full characertization of $P(x \mid s)$, while the mixture model allows us to construct an approximation from simple constituent types.

Furthermore, when faced with two competing mixture models, it is always possible to create an encompassing model and use nested hypothesis tesing to select the best model. See, for example, Haruvy, Stahl and Wilson (1998).

## Individual versus Population Models

While psychologists are primarily interested in the behavior of individuals (even in social settings), other social scientists such as economists and sociologists are more interested in aggregate population behavior. Indeed, in many applications, the economist or sociologist may have only aggregate data. Although theoretically models of individual behavior can be aggregated to produce models of population behavior, there are often informational and computational constraints to exact aggregation. For example, because each individual has a unique history at any point in a repeated game, exact aggregation must keep track of all these unique histories, which grow in number exponentially with time. Models which entail integration over possible histories can be computationally infeasible. There is also the information conservation principle which states that a given sample of data contains only so much information - if you use that data to find the best fit for models of individual behavior, the information gained will not give you the best fit for models of population behavior.

The mixture models discussed above can be applied in a straightforward manner to population data. The only difference is in the construction of the likelihood function for the data. In an individual model, one first computes the likelihood of the behavior of each individual for all the games by type (a product of probabilities over the games), then sums these type-conditional likelihoods, and finally takes the product of these summed likelihoods over all individuals. In a population model, one first computes the sum of the type-conditional probabilistic choice functions for each game, then computes the multinomial probability of the observed aggregate choices using the summed probabilities, and finally takes the product of these multinomial likelihoods over all the games. The latter model does not impose the restriction that an individual's behavior is of one type for all games. Since the likelihood functions are different, the maximum-likelihood estimates of the individual model will differ from the maximum-likelihood estimates of the population model.

## Conclusion

Human behavior is often multimodal, so we need multimodal models to represent and predict such heterogeneous behavior. The mixture model is ideally suited for this purpose. The major challenge is the specification of constituent subpopulation types. In accordance with the scientific method, we advocate theory-based hypothesis generation and testing. Since the best mixture model at any point in time will be an approximation of the true data generating process, there will always be the possibility of discovering a better approximation.

## Bibliography

Feng, Z. and C. McCulloch (1996), "Using Bootstrap Likelihood Ratios in Finite Mixture Models," *Journal of the Royal Statistical Society, ser. B.* 58, 609-617.

Haruvy, E., and D. Stahl, (1999). Empirical tests of equilibrium selection based on player heterogeneity, www.eco.utexas.edu/faculty/Stahl/experimental/HS99.pdf.

Haruvy, E., Stahl, D., and P. Wilson (1998). "Modeling and testing for heterogeneity in observed strategic behavior," *Review of Economic Studies*, forthcoming.

Haruvy, E., D. Stahl, and P. Wilson (1999), "Evidence for Optimistic and Pessimistic Behavior in Normal-form Games," *Economics Letters*, 63, 255-259.

Self, S. and K-Y. Liang (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, 82, 605-610.

Stahl, D. (2000a), "Rule Learning in Symmetric Normal-form Games: Theory and Evidence", *Games and Economic Behavior*, 32, 105-138.

Stahl, D. (2001), "Population Rule Learning in Symmetric Normal-form Games: Theory and Evidence, *Journal of Economic Behavior and Organization*, 45, 19-35.

Stahl, D. and P. Wilson (1994), "Experimental Evidence of Players' Models of Other Players," *Journal of Economic Behavior and Organization*, 25, 309-327.

Stahl, D. and P. Wilson (1995), "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, 10, 213-254.