A Bayesian Method for Characterizing Population Heterogeneity

by

Dale O. Stahl

Malcolm Forsman Centennial Professor

Department of Economics

University of Texas at Austin

December 7, 2018

ABSTRACT

A stylized fact from laboratory experiments is that there is much heterogeneity in human behavior. We present and demonstrate a computationally practical non-parametric Bayesian method for characterizing this heterogeneity. In addition, we define the concept of *behaviorally distinguishable* parameter vectors, and use the Bayesian posterior to say what proportion of the population lies in meaningful regions. These methods are then demonstrated using laboratory data on lottery choices and the rank-dependent expected utility model. We find that 80% of the subject population is not behaviorally distinguishable from the ordinary Expected Utility model.

J.E.L. Classifications: C11, C12, D81

1. Introduction.

A stylized fact from laboratory experiments is that there is much heterogeneity in the subject population. How to characterize that heterogeneity is an active research area among experimentalists and econometricians. The approaches include individual parameter estimation, random coefficient models, mixture models of different types, and Bayesian methods.[1] It is not the intention of this paper to compare all the methods, but rather to present and demonstrate a computationally practical, non-parametric Bayesian method to characterize the heterogeneity in a population of subjects.[2]

Suppose we want to characterize the population distribution of parameter values of a model of behavior. Let $f(x_i \mid \theta)$ denote the model; i.e. $f(x_i \mid \theta)$ gives the likelihood of observed behavior x_i for individual $i \in \{1, ..., N\}$ and $\theta \in \Theta$ is a finite dimensional vector of parameters. Let $g(\theta)$ denote the distribution of θ in a population of individuals. How do we estimate $g(\theta)$ from observed behavior $\underline{x} \equiv \{x_i, 1, ..., N\}$? To motivate why a new method is useful, we will critique in order (i) subject-specific MLEs, (ii) random coefficient methods, (iii) mixture models, and (iv) standard Bayesian methods.

One approach is to find the θ that maximizes $f(x_i \mid \theta)$ for each i, and to treat each MLE $\hat{\theta}_i$ as a random sample from the population. A scatter plot of $\{\hat{\theta}_i, i = 1, \dots, N\}$ gives a view of the sample distribution of θ_i from the population. However, the uncertainty of the MLEs is not represented in such a plot. Standard kernel density estimation methods are inappropriate because they essentially assume a common variance-covariance (Σ) matrix. Estimating Σ_i matrices for each i entails many more parameter estimates, and still any density estimation using these Σ_i matrices would depend upon additional assumptions about the kernel of each i, such as normality: $N(\hat{\theta}_i, \Sigma_i)$.

[1] E.g. see Hey and Orme (1994), Harrison and Rutström (2008, 2009), Wilcox (2008, 2011), Fox, et al. (2011), Conte, Hey and Moffat (2011), and Stahl (2014).

[2] We believe that this Bayesian approach is complementary to Wilcox (2008) and Fox, et al. (2011).

Random coefficient models assume a parametric form for the population distribution: $g(\theta \mid \beta)$, where β is a low-dimensional parameter vector. Typically, $g(\theta \mid)$ is a family of unimodal distributions in which β stands for the mean and Σ matrix. Obviously, these parametric restrictions could be very wrong. For example, the simple scatter plot of the individual MLEs $\{\hat{\theta}_i, i = 1, \ldots, N\}$ may have clusters that suggest the true distribution is multimodal.

One way to embrace the multimodal possibility is a mixture model of K parametric unimodal distributions $g_k(\theta \mid \beta_k)$ for $k \in \{1, ..., K\}$. In addition to the β_k parameters are the mixture parameters $\{\alpha_k \geq 0\}$ such that $\sum_{k=1}^{K} \alpha_k = 1$, so $g(\theta) \equiv \sum_{k=1}^{K} \alpha_k \, g_k(\theta \mid \beta_k)$. Then, the econometric task is to estimate the coefficients $\{(\alpha_k, \beta_k), k = 1, \ldots, K\}$. Since there are uncountably many ways to specify the component distributions $\{g_k(\theta \mid) \mid k = 1, ..., K\}$, one must also provide identifying conditions such as that the component distributions are independent or have non-intersecting supports. Further, the component distributions should have meaningful interpretations such as describing theoretical or behavioral types. This method still suffers from potential mis-specification via the parametric restrictions on the distributions.

To review the standard Bayesian approach, let G denote the space of distributions $g(\theta)$, and let $\Delta(G)$ denote the space of probability measures on G. The standard Bayesian approach requires us to have a prior belief $\mu_0 \in \Delta(G)$. Note that μ_0 is a probability measure on G, whereas g is a point in G and a probability measure on Θ. Given observed behavior $\underline{x} \equiv \{x_i, i=1,...,N\}$, the posterior belief according to Bayes rule is

$$\mu_1(g|x) = \frac{\prod_{i=1}^{N} [\int f(x_i|\theta_i) g(\theta_i) d\theta_i] \mu_0(g)}{\int \prod_{i=1}^{N} [\int f(x_i|\theta_i) g'(\theta_i) d\theta_i] \mu_0(g') dg'} . \qquad (1)$$

Since both G and $\Delta(G)$ are infinite dimensional spaces, in practice this an impossible calculation to carry out exactly.

One method of approximating $\mu_1(g|x)$ is to represent Θ by a finite grid. If for example there are four elements of θ and we want 50 points on each dimension of the grid, then, our grid would have $50^4 = 6{,}250{,}000$ points altogether. A probability distribution $g(\,)$ over Θ would be a

point in the 6.25 million dimensional simplex.[3] Next we might represent $\Delta(G)$ by a grid with only 10 points in each dimension, so that grid would have $10^{6,250,000}$ points. Obviously, this is way beyond computational feasibility.

A commonly employed alternative is to restrict G and $\Delta(G)$ to parametric forms with a small finite number of parameters.[4] For example, to allow for multiple modes in g, G might be specified as a mixture of K normal distributions each with 4+10 (= 14) free parameters[5]. Then $\Delta(G)$ might be specified as a unimodal normal distribution in R^{15K-1}, with $15K-1 + (225K^2-14K)/2$ parameters.[6] However, obviously these restrictions are quite likely to be wrong, seriously bias the results, and still be computationally challenging.[7]

This paper presents a computationally feasible non-parametric alternative within the Bayesian framework. Bayes rule is used to compute posterior distributions $g_i(\theta \mid x_i)$ for individual i based on observed behavior x_i. Assuming the individuals in the data are an unbiased sample from the population, we compute $g^*(\theta \mid \underline{x})$ which is the probability density, conditional on the observed behavior of all individuals in the dataset, that a randomly drawn individual from the population has parameter θ.[8]

Given $g^*(\theta \mid \underline{x})$, we can answer questions such as what percentage of the population has parameter values in set $A \subset \Theta$. Often there are specific parameter values, say θ', that represent interesting types, and we would like to know what percentage of the population is a particular type. Unfortunately, if $g^*(\theta \mid \underline{x})$ is absolutely continuous, the answer is zero. However, what we

[3] Note that almost all the volume of such a high dimensional simplex is very near the boundaries, so a uniform prior would put essentially zero probability on interior points. Therefore, the amount of data and time for Bayesian updating to converge is likely to be longer than the lifetime of the Milky Way galaxy.

[4] E.g. Bolstad (2012).

[5] The variance-covariance matrix in 4 dimensions has 10 degrees of freedom subject also to being positive semi-definite.

[6] The mixture over K modes entails K-1 mixture parameters, so there are $14K + K - 1 = 15K - 1$ dimensions of G. The variance-covariance matrix in 15K-1 dimensions has $(15K-1)15K/2 = (225K^2 - 15K)/2$ free parameters.

[7] The number of variance-covariance parameters could be drastically reduced by assuming zero covariance and even symmetry, but at the potential cost of serious mis-specification.

[8] In section 2 we will comment on why MCMC methods cannot be used to generate samples from g^*.

really want to know is what percentage is similar to θ' in some sense. For this purpose, we formally define the concept of "behavioral distinguishability", which enables us to answer what percent of the population is behaviorally distinguishable from θ' and conversely what percent is behaviorally indistinguishable from θ'.

To demonstrate this new method, we apply it to the Rank-Dependent Expected Utility model and one of the best datasets from laboratory experiments on lottery choices: Hey and Orme (1994; hereafter HO)[9]. Our findings are quite different from results reported by Conti, Hey and Moffatt (2011; hereafter CHM). We demonstrate that this difference is not due to the different econometric methods but due to the different questions being asked.

The paper is organized as follows. Section 2 presents the Bayesian approach. Section 3 presents the RDEU econometric model, and the HO data set. Section 4 presents the results of our Bayesian method. Section 5 develops the formal concept of "behavioral distinguishability", answers pertinent questions about behaviorally distinguishable types, and unveils the reason for the apparent difference with CHM. Section 6 concludes.

2. Our Bayesian Approach.

To develop our Bayesian approach let x_i denote the observed behavior for subject i, and let $f(x_i \mid \theta)$ denote the probability of x_i given parameter vector $\theta \in \Theta$. Given a prior g_0 on θ, by Bayes rule, the posterior on θ is

$$g(\theta \mid x_i) \equiv f(x_i \mid \theta)g_0(\theta)/\int f(x_i \mid z)g_0(z)dz. \tag{2}$$

However, eq(2) does not use information from the other subjects even though those subjects were randomly drawn from a common subject pool. Let N be the number of subjects in the data set. When considering subject i, it is reasonable to use as a prior, not g_0, but

$$g_i(\theta \mid \underline{x}_{-i}) \equiv \frac{1}{N-1}\sum_{h \neq i} g(\theta \mid x_h) \tag{3}$$

[9] Similar analysis was done on the data of Hey (2001) and similar results were obtained.

In other words, having observed the choices ($\underline{x}_{-i}$) of N-1 subjects, $g_i(\theta \mid \underline{x}_{-i})$ is the probability that the N^{th} random draw from the subject pool will have parameter vector θ. We then compute

$$\hat{g}_i(\theta \mid \underline{x}) \equiv f(x_i \mid \theta) g_i(\theta \mid \underline{x}_{-i}) / \int f(x_i \mid z) g_i(z) dz , \quad (4)$$

where $\underline{x}$ denotes the entire N-subject data set. Finally, we aggregate these posteriors to obtain

$$g^*(\theta \mid \underline{x}) \equiv \frac{1}{N} \sum_{i=1}^{N} \hat{g}_i(\theta \mid \underline{x}) . \quad (5)$$

We can interpret $g^*(\theta \mid \underline{x})$ as the probability density that a random draw from the subject pool will have parameter vector θ. Note that eq(5) puts equal weight on each x_i, so we are using each individual's data effectively only once, in contrast to empirical Bayes methods. Also note that while MCMC methods could be used to simulate random draws from each $g(\theta \mid x_i)$, since eq(5) requires that each $\hat{g}_i(\theta \mid \underline{x})$ be properly normalized, MCMC methods cannot be used to simulate random draws from $g^*(\theta \mid \underline{x})$.

When implementing this approach we construct a finite grid on the parameter space Θ and we replace the integrals by summations over the points in that grid. However, we do not need to integrate over the space of distributions $\Delta(G)$, so we avoid the need for a grid on $\Delta(G)$ which would be computationally infeasible.

Since eq(4) uses a prior that is informed by the data of N-1 other individuals, the influence of g_0 in the first step is overwhelmed by the influence of the data. Thus, the specification of g_0 is much less an issue and can be chosen based on computational ease.

3. The Rank-Dependent Expected Utility Model and the HO Data.

a. The Behavioral Model.

The Rank-Dependent Expected Utility (RDEU) model[10] was introduced by Quiggin (1982, 1993). A convenient feature is that it nests EU and Expected Monetary Value. RDEU allows subjects to modify the rank-ordered cumulative distribution function of lotteries as follows. Let $Y \equiv \{y_0, y_1, y_2, y_3\}$ denote the set of potential outcomes of a lottery, where the outcomes are listed in rank order from worst to best. Given rank-ordered cumulative distribution F for a lottery on Y, it is assumed that the individual transforms F by applying a monotonic function H(F). From this transformation, the individual derives modified probabilities of each outcome:

$$h_0 = H(F_0),\ h_1 = H(F_1) - H(F_0),\ h_2 = H(F_2) - H(F_1), \text{ and } h_3 = 1 - H(F_2). \qquad (6)$$

A widely used parametric specification of the transformation function, suggested by Tversky and Kahneman (1992), is

$$H(F_j) \equiv (F_j)^{\beta}/[(F_j)^{\beta} + (1-F_j)^{\beta}]^{1/\beta}\ , \qquad (7)$$

where $\beta > 0$.[11] Obviously, $\beta = 1$ corresponds to the identify transformation, in which case the RDEU model is equivalent to the EU model.

Given value function $v(y_j)$ for potential outcome y_j, the rank-dependent expected utility is

$$U(F) \equiv \sum_j v(y_j)h_j(F). \qquad (8)$$

To confront the RDEU model with binary choice data (F^A *vs*. F^B), we assume a logistic choice function:

$$\text{Prob}(F^A) = exp\{\gamma U(F^A)\}/[\ exp\{\gamma U(F^A)\} + exp\{\gamma U(F^B)\}\ , \qquad (9)$$

where $\gamma \geq 0$ is the precision parameter. $\text{Prob}(F^A)$ gives the probability that lottery F^A is chosen rather than lottery F^B.

As in CHM, we use the Constant Relative Risk Aversion (CRRA) specification in which

[10] This model is the same as the Cumulative Prospect model (Tversky and Kahneman, 1992) restricted to non-negative monetary outcomes.

[11] Alternative specifications are (i) the symmetric Quiggin (1982), and (ii) Prelec (1998). However, the effect of alternative specifications of H() is orthogonal to the purpose of this paper which is the econometric method.

$$v(y_i) = (y_i)^{1-\rho}, \tag{10}$$

where $\rho \leq 1$ is the CRRA measure of risk aversion. Since outcomes in the HO experiment vary only in money (m_i), w.l.o.g. we define $y_i = (m_i - m_0)/(m_3 - m_0)$, so $v(y_0) = 0$ and $v(y_3) = 1$. Hence, the empirical RDEU model entails three parameters: (γ, ρ, β).

Next, to specify the likelihood function for our data, let $\underline{x}_i \equiv \{x_{i1}, ..., x_{iT}\}$ denote the choices of subject i for T lottery pairs indexed by $t \in \{1, ... T\}$, where $x_{it} = 1$ if lottery A was chosen, and $x_{it} = 0$ otherwise. Then the probability of the T observed choices of subject i is the product of the probability of each choice given by eq(9).[12] For notational convenience, let $\theta_i \equiv (\gamma_i, \rho_i, \beta_i)$. Then, in log-likelihood terms:

$$LL(\underline{x}_i, \theta_i) \equiv \sum_{t=1}^{T} \left[x_{it} \, ln\left(Prob[F^A(\theta_i)]\right) + (1 - x_{it}) \, ln\left([1 - Prob[F^A(\theta_i)]]\right) \right]. \tag{11}$$

Then, we define the total log-likelihood of the data as

$$LL(\mathbf{x}, \boldsymbol{\theta}) \equiv \sum_i LL(\underline{x}_i, \theta_i) , \tag{12}$$

where $\boldsymbol{\theta} \equiv \{\theta_i , i = 1, ..., N\}$, and $\mathbf{x} \equiv \{\underline{x}_i, i = 1, ..., N\}$.

b. The HO Data.

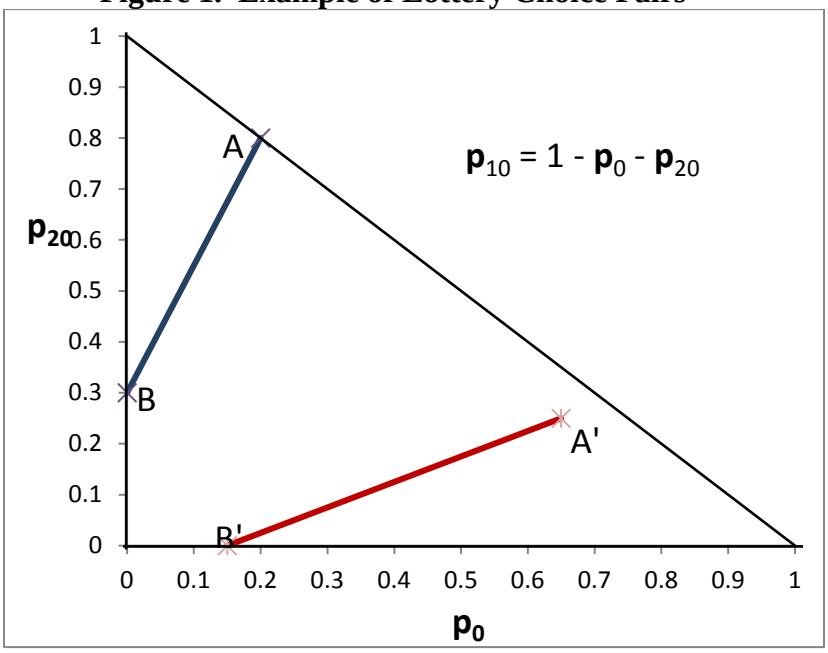
The HO dataset contains 100 unique binary choice tasks.[13] Each task was a choice between two lotteries with three prizes drawn from the set {0£, 10£, 20£, 30£}. A crucial design

[12] As pointed out by Harrison and Swarthout (2014), this specification implicitly assumes the "compound independence axiom". Since we view EU and RDEU as behavioral models, we are comfortable with this implicit assumption.

[13] These 100 tasks were presented to the same subjects again one week later. We do not consider that data here because the test that the same model parameters that best fit the first 100 choices are the same as those that best fit the second 100 choices fails. Possible explanations for this finding are (i) that learning took place between the sessions, (ii) preferences changed due to a change in external (and unobserved) circumstances, and (iii) the subjects did not have stable preferences. Therefore, we focus our attention on the first 100 choice tasks.

factor was the ratio of (i) the difference between the probability of the high outcome for lottery A and the probability of the high outcome for lottery B to (ii) the difference between the probability of the low outcome for lottery A and the probability of the low outcome for lottery B. It is insightful to represent this choice paradigm in a Machina (1982) triangle, as shown in Figure 1.

Figure 1. Example of Lottery Choice Pairs



The ratio for the A-B pair is the slope of the line connecting A and B, which is greater than 1. The ratio for the A'-B' pair is the slope of the line connecting A' and B', which is clearly less than 1. According to EU indifference curves are parallel straight lines with positive slope in this triangle, and the indifference curves of a risk neutral subject would have slope equal to 1. A wide range of ratios was used in order to identify indifference curves and to test the implications of EU (as well as alternative theories). After all choices were completed, one task was randomly selected and the lottery the subject chose was carried out to determine monetary payoffs.[14]

[14] Loomes and Sugden (1998) is a similar study as Hey and Orme (1994), except that their analysis of the data is based on non-parametric tests involving the number of "reversals" and violations of dominance. Hey (2001) reports

One can estimate these parameters for *each* subject in the HO data set. That approach entails (3×80=240) parameters, even without the corresponding variance-covariance matrices. Table 1 gives the population mean and standard deviation of the point estimates.[15] The last column "LL" gives the sum of the individually maximized log-likelihood values. Note that there is substantial heterogeneity across subjects in the parameter estimates for ρ and β.

Table 1. Population Mean and Standard Deviation of Individual Parameter Estimates and the Aggregated LL

p(γ)[a]	**ρ**	**β**	**LL**
0.8826 (0.0921)	0.4465 (0.4405)	1.028 (0.637)	-3007.38

[a] $p(\gamma) \equiv 1/[1 + exp(-0.05\gamma)]$.[16]

These comparisons involve estimates of a large number of parameters. For each individual subject, we obtain point estimates of the parameters, but no confidence interval. One could use a bootstrap procedure to obtain variance-covariance matrices for each individual, but that would be a computationally intense task and entail 6 additional parameters per subject. Further, the estimates for each subject would ignore the fact that the subjects are random draws from of a population of potential subjects and that therefore the behavior of the other subjects contains information that is relevant to each subject. In contrast, the Bayesian approach is better

a similar study with 100 binary tasks repeated over give days. Harrison and Rutström (2009) replicate HO and also run a similar experiment using 30 unique tasks. Bruhin,et al. (2010) also explore heterogeneity, but they elicit certainty equivalents, so the task is arguably different from binary choices as in the other studies.

[15] This is the square root of the variance of the point estimates across subjects; it is not the standard error of individual point estimates.

[16] Note that $p(\gamma)$ is the probability the subject will choose the option with the greater value whenever that value is 5% higher than that of the other option, thereby providing a behavioral interpretation of γ.

suited to extract information from the whole sample population. Consequently, we turn to the Bayesian approach.[17]

4. Implementing our Bayesian Approach.

When implementing our Bayesian method we specify the prior g_0 as follows. For the logit precision parameter, we specify $\gamma = 20ln[p/(1-p)]$ with p uniform on [0.5, 0.999]. In this formulation, p can be interpreted as the probability an option with a 5% greater value will be chosen. Since the mean payoff difference between lottery pairs in the HO data set is about 5%, this is a reasonable scaling factor.[18] ρ is uniform on [-1, 1], and $ln(\beta)$ is uniform on [$-ln(3)$, $ln(3)$].[19] These three distributions are assumed to be independent. For computations, we use a grid of 41x41x41 = 68,921 points.

Since we cannot display a three-dimensional distribution, we present two two-dimensional marginal distributions. Figure 2 shows the marginal on $(p(\gamma), \beta)$. From Figure 2 we see that the distribution is concentrated around $\beta = 0.95$, and that the precision values are large enough to imply that a 5% difference in value is behaviorally significant (i.e. $p(\gamma) > 0.75$).

Figure 2. Marginal of g* on $(p(\gamma), \beta)$.

[17] For another application, see Stahl (2014).

[18] Our results are robust to this specification of the prior on γ.

[19] 95% of the individual MLEs for β lie in this range. Using a wider interval for the prior on β has no noticeable effect on the Bayesian posterior at the cost of more grid points and computational time.

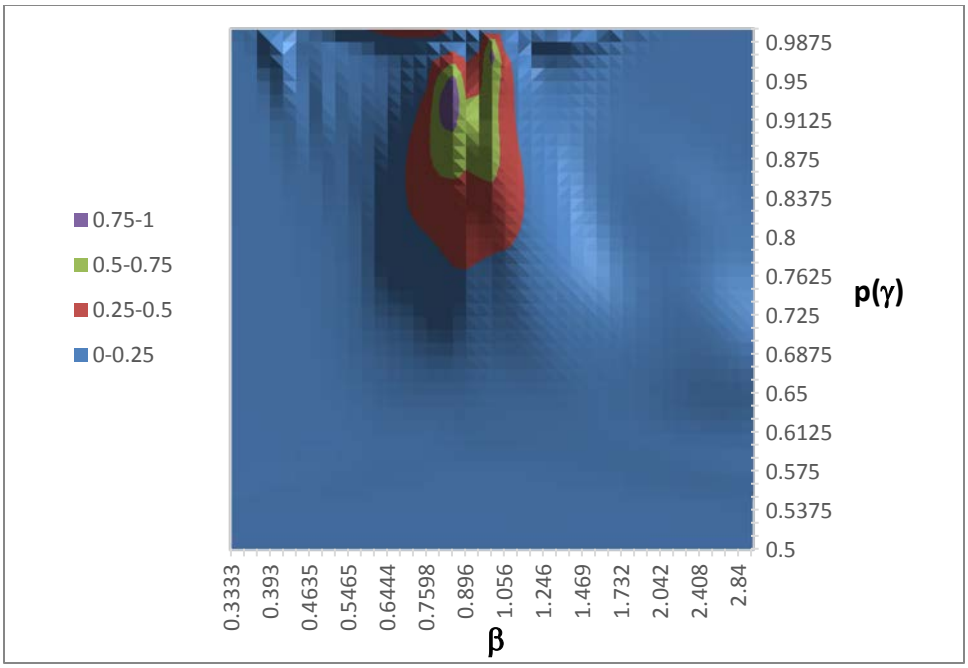
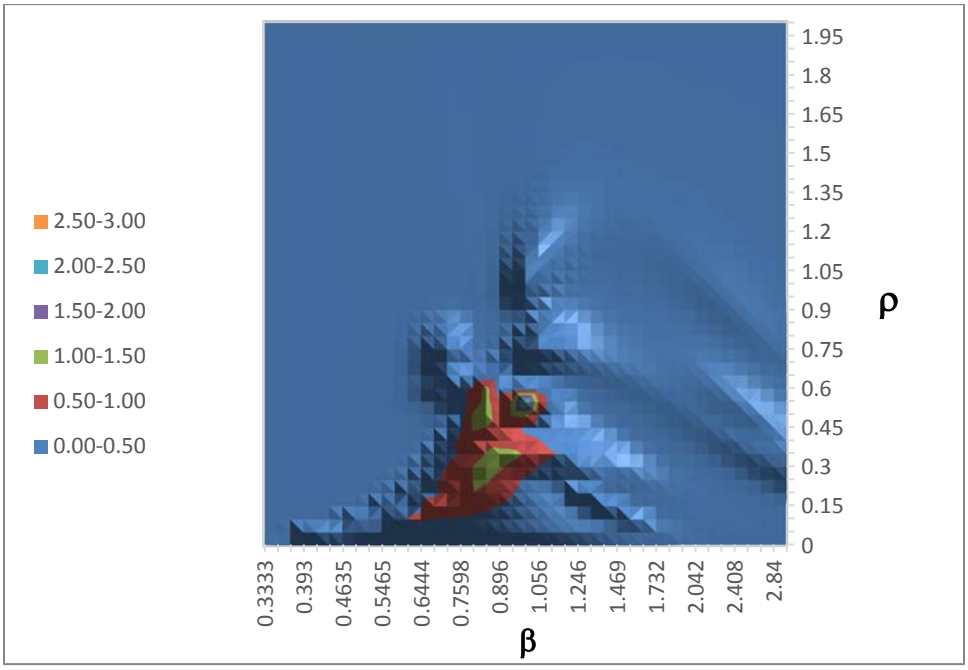


Figure 3 shows the marginal on (ρ, β).

Figure 3. Marginal of g* on (ρ, β).



Given $g^*(\theta \mid \underline{x})$ we can compute several statistics. First, the log-likelihood of the HO data is LL(g*) = -3335.29. In contrast, the log-likelihood of the three-parameter RDEU representative-agent model is -4472.85. Obviously, the heterogeneity implicit in g* fits the data much better than a representative-agent model.[20] Compared to -3007.38 (Table 1), the log-likelihood from the Bayesian method appears to be much worse. However, the direct comparison is inappropriate. LL(g*) is computed as if each subject were drawn independently from g*. In contrast, -3007.38 is the sum of individually computed log-likelihoods using the subject-specific estimated parameters.

The g*-weighted mean of the parameter space is $(\overline{p(\gamma)}, \overline{\rho}, \overline{\beta}) = (0.8556, 0.5815, 1.010)$. . Note that $\overline{\beta} \approx 1$, meaning that on average H(F) is the identity function. Table 2 displays the variance-covariance matrix.

Table 2. Variance-Covariance of g*.

	p(γ)	**ρ**	**β**
p(γ)	0.0085	-0.0054	-0.0150
ρ		0.0851	0.0335
β			0.1638

However, these means and covariances are much less informative when g* is multimodal.

Indeed, we find evidence for multiple modes. A grid point θ is declared a *mode* if and only if it has the highest value of g* in a 7x7x7 cube of nearest neighbors of θ in the grid. The most prominent mode is at p(γ) = 0.999, ρ = 0.85, and β = .681. The next most prominent mode is at p(γ) = 0.975, ρ = 0.45, and β = 1.00. The third most prominent mode is at p(γ) = 0.999, ρ = 0.85, and β = 1.47. Numerous other modes exist but are best described as shallow bumps.

To test for over-fitting, we compute g* based only on the first 50 tasks in the HO data, and use this g* to predict the behavior for the second 50 tasks. We find that the log-likelihood of

[20] One can consider this Bayesian approach as an alternative random parameter model as used by Wilcox (2008). However, in contrast to Wilcox, we assume that each subject draws from this distribution *once* and uses those parameters for all choice tasks, rather than drawing for each choice task. The latter can be viewed as a "diverse" representative agent model, while the former is a heterogeneous agent model.

the latter is -1567.39. In contrast, using individual parameter estimates from just the first 50 tasks, the log-likelihood of the second 50 tasks is -1774.01. This result suggests that the approach of individual parameter estimates is more susceptible to over-fitting and less reliable than the Bayesian approach.

5. Behaviorally Distinguishable Parameter Vectors.

a. Definition.

The most productive use of $g^*(\theta \mid \underline{x})$ is to test hypotheses. For example, we can ask what percent of the subject pool has $\beta = 1$. The answer is 10.4%; however, this number is an artifact of the discrete grid used for computation. Assuming g^* is absolutely continuous, as the grid becomes finer and finer, we would expect the percentage with $\beta = 1$ to approach 0. On the other hand, $\beta = 0.999$ is not meaningfully different. What we want to know is the percentage of the population that is behaviorally indistinguishable in some sense from EU (i.e. $\beta = 1$).

The *behavior* is simply the choice data for a random subject x_i. To assess whether this data was generated by θ or θ', we typically compute the log of the likelihood ratio (LLR): $ln[f(x_i \mid \theta)/f(x_i \mid \theta')]$. A positive LLR means x_i is more likely to have been generated by θ than θ'. However, it is well-known that likelihood-ratio tests are subject to type-I and type-II errors. To compute the expected frequency of these errors, let $X_1 \equiv \{x_i \mid ln[f(x_i \mid \theta)/f(x_i \mid \theta')] < 0\}$. If the data in fact was generated by θ, and $x_i \in X_1$, then a naïve LLR test would yield a type-I error. Similarly, if the data in fact was generated by θ' and $x_i \in X_2$ (the complement of X_1), then a naïve LLR test would yield a type-II error. Hence, the expected frequencies of type-I and type-II errors are respectively:

$$er_1 \equiv \int_{X_1} f(x_i \mid \theta)dx_i \quad \text{and} \quad er_2 \equiv \int_{X_2} f(x_i \mid \theta')dx_i \ . \tag{13}$$

If either of these error rates is too large, we might say that θ and θ' are behaviorally indistinguishable. Classical statistics suggests that a proper test statistic would have these error rates not exceed 5%; to be conservative we will use 10% as our threshold.

Of course, by increasing the number of observations in x_i, we can drive these error rates lower and lower. However, practical considerations often limit the number of observations. In laboratory experiments, boredom, time limitations and budget constraints place severe upper bounds on the number of observations. The HO dataset with 100 tasks is unusually large. Moreover, to test for overfitting we would select a subset, say 50, to use for estimation, and the remaining 50 to assess parameter stability and prediction performance. Therefore, for the illustrative purposes of this paper, we use 50 as a reasonable sample size upon which to judge behavioral distinguishability. With 50 binary choices, there are 2^{50} ($\approx 10^{30}$) possible x_i vectors. Generating all these possible x_i vectors and computing er_1 and er_2 is obviously not feasible. Instead, we generate 1000 x_i vectors from $f(x_i \mid \theta)$ and 1000 from $f(x_i \mid \theta')$.[21] Then, er_1 is approximated by the proportion of x_i generated by $f(x_i \mid \theta)$ that lie in X_1, and er_2 is approximated by the proportion of x_i generated by $f(x_i \mid \theta')$ that lie in X_2. To ensure robustness to the selection of 50 tasks we randomly selected 25 sets of 50 HO tasks and we averaged $max\{er_1, er_2\}$ over these 25 sets of 50 tasks to determine behavioral distinguishability for each (θ, θ') of interest (see next section).

In summary, we define θ and θ' to be *behaviorally distinguishable* if both of the simulated type-I and type-II error rates are less than or equal to 10%, and to be *behaviorally indistinguishable* otherwise.

b. Application to RDEU model and HO Data.

Many questions of interest can be framed in terms of our *behaviorally indistinguishable* relationship on the parameters. To begin, we may want to know what percent of the population is behaviorally indistinguishable from 50:50 random choices (hereafter referred to as Level-0 behavior). Since the latter entails the simple restriction that $\gamma = 0$, we can compute whether $\theta = (\gamma, \rho, \beta)$ is behaviorally distinguishable from $(0, \rho, \beta)$, and then sum $g^*(\gamma, \rho, \beta)$ over all the grid points (γ, β, ρ) that are behaviorally distinguishable from $(0, \rho\ \beta)$. The answer is 99.0% (0.5%), which leaves only 1.0% that are behaviorally indistinguishable from Level-0.

[21] We also made these computations with only 100 simulated x_i vectors, and found virtually the same results. Therefore, we are confident that 1000 simulated x_i vectors are adequate for our purposes.

The question of most interest is what percent are behaviorally indistinguishable from EU. To answer this, we ask how much mass g* puts on the set of parameters (γ, ρ, β) that are behaviorally distinguishable from Level-0 but indistinguishable from (γ, ρ, 1)? The answer is 79.0% (1.7%). The remainder (99.0 - 79.0) 20.0% can be considered RDEU types that are behaviorally distinguishable from Level-0 and EU types.

This conclusion stands in stark contrast to that of CHM who report only 20% EU types and 80% RDEU types. Such a discrepancy requires an explanation. Since CHM used a mixture model, while we used a Bayesian approach, perhaps this specification difference is the underlying cause of the discrepancy.

To investigate this possibility we applied our method for measuring the probability mass that is behaviorally indistinguishable from EU (i.e. $\beta = 1$) to the CHM mixture model on the same data set [Hey (2001); hereafter H01]. That mixture model consists of two types: an RDEU type exactly like our specification, and an EU type (RDEU with β is restricted to be exactly 1). Using the exact same mixture model and parameter estimates as reported in CHM, we computed the implied probability distribution over the RDEU parameters, call it $\varphi_{RDEU}(\rho, \beta)$.[22] We find that φ_{RDEU} puts 0.877 probability mass on parameters (ρ, β) that are behaviorally indistinguishable from EU. Thus, in addition to the mixture coefficient for the EU type of 19.7%, we have 70.4% (= 0.877*80.3%) that are behaviorally indistinguishable from EU, giving a total of 90.1% that are behaviorally indistinguishable EU. Thus, when we ask the same question of the H01 data as we do for the HO data, we find similar answers (90.1% and 80.0% respectively). In other words, both our Bayesian approach and the CHM mixture model approach applied to the H01 data produce similar answers to the same question: what percentage of the population are behaviorally indistinguishable from EU.

To rule out the possibility that this explanation applies only for the H01 data, we confronted the CHM mixture model with the HO data. We implemented the same mixture model as CHM, and found maximum-likelihood estimates for the parameters. As above, we computed the implied probability distribution over the RDEU parameters. We found that φ_{RDEU} puts 0.844 probability mass on parameters (ρ, β) that are behaviorally indistinguishable from

[22] See Appendix for details.

EU. In addition to the mixture coefficient for the EU type of 28.1%, we have 60.7% (= 0.844*71.9%) that are behaviorally indistinguishable from EU, giving a total of 88.8% that are essentially EU, which is larger than the 80.0% we obtained using the Bayesian method but more alike than the mixture coefficient (28.1%).

Thus, it appears that the discrepancy between our conclusions and that of CHM reflects the difference not in the Bayesian vs. mixture model approaches but instead reflects the difference in the questions being asked. We ask what proportion of the population are behaviorally distinguishable from Level-0 but not EU types, while CHM ask what is the MLE of the mixture coefficient for the EU type in a specific parametric mixture model.

A way to understand why these questions are substantively different is to realize that the mixture proportion is just a means of creating a general probability distribution. In principle, there are uncountably many ways to represent a given distribution as a mixture of component distributions. Therefore, a crucial step in estimating a mixture model is the provision of identifying restrictions. Since the RDEU model nests the EU model, we need to specify what region of the parameter space should be identified as the EU region even though those parameters also represent an RDEU model. Surely, just the one-dimensional line in (ρ, β) space with $\beta = 1$ is far too narrow, but that is the implicit identifying restriction when interpreting the mixture parameter as the percentage of EU types. However, when we ask what proportion of the population are EU types, we want to know what proportion are behaviorally indistinguishable from EU types, and not what weight is given to the EU component in a mixture of two parametric distributions that represents the population distribution.

6. Conclusions and Discussion.

This paper has demonstrated the feasibility and usefulness of Bayesian methods when confronting laboratory data, especially when addressing heterogeneous behavior. Specifically we have presented a nonparametric[23] computationally feasible approach. To extend our

[23] That is, the Bayesian posterior is a non-parametric function of the data, albeit the RDEU model is obviously parametric.

approach to models with more parameters, statistical sampling techniques can be employed to tame the curse of dimensionality.[24]

To demonstrate our method, we applied it to the Rank-Dependent Expected Utility model and the Hey and Orme (1994) dataset on lottery choices. Our Bayesian analysis characterized substantial heterogeneity in the subject population. Moreover, it revealed that 80% of the population is behaviorally distinguishable from Level-0 but indistinguishable from EU behavior. The difference between this finding and the opposite finding by others is not due to the econometric methods but due to the question being asked, or equivalently to what we mean by a behavioral type. When asking what proportion of the population are EU types, we argue that we typically want to know what proportion are *behaviorally indistinguishable* from EU behavior.

[24] E.g. see Rubinstein and Kroese (2016).

References

Bolstad, W. (2012). *Understanding Computational Bayesian Statistics*, Wiley, ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/utxa/detail.action?docID=698546.

Bruhin, A., Fehr-Duda, H., and Epper, T. (2010). "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion," *Econometrica,* **78**, 1375-1412.

Conti, A., Hey, J. and Moffatt, P. (2011). "Mixture Models of Choice under Risk", *J. of Econometrics,* **162**,79-88.

Erev, I., Ert, E. and Yechiam, E. (2008). "Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions," *Journal of Behavioral Decision Making,* **21**, 575–597.

Fox, J., Kim, K., Ryan, S. and Bajari, P. (2011). "A Simple Estimator for the Distribution of Random Coefficients," *Quantitative Economics,* **2**, 381-418.

Harrison, G. W. and Rutström, E. (2008). "Risk Aversion in the Laboratory," In J. C. Cox and G. W. Harrison, eds., *Research in Experimental Economics,* **12**, 41-196.

Harrison, G. W. and Rutström, E. (2009). "Expected Utility and Prospect Theory: One Wedding and Decent Funeral," *Experimental Economics,* 12, 133-158.

Hey, J., (2001)."Does Repetition Improve Consistency?" *Experimental Economics*, **4,** 5-54.

Hey, J. and Orme, C. (1994). "Investigating Generalizations of Expected Utility Theory Using Experimental Data," *Econometrica,* 62, 1291-1326.

Kahneman, D. and Tversky, A. (1979). "Prospect Theory: An Analysis of Decision under Risk". *Econometrica,* **47**, 263–291.

Loomes G., and Sugden, R. (1998). "Testing Alternative Stochastic Specifications for Risky Choice," *Economica*, **65**, 581-598.

Machina, M. (2008). "Non-expected Utility Theory," in *The New Palgrave Dictionary of Economics* (2nd edition). Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan.

Prelec, D. (1998), "The Probability Weighting Function," *Econometrica*, 66, 497-527.

Quiggin, J. (1982). "A Theory of Anticipated Utility", *J. of Econ. Behavior and Organization*, **3**, 323-343.

Quiggin, J. (1993). *Generalized Expected Utility Theory: the Rank-Dependent Model*, Kluwer Academic Publishers.

Rubinstein, B. and Kroese, D. (2016). *Simulation and the Monte Carlo Method*, 3rd edition, Wiley & Sons.

Stahl, D. (2014). "Heterogeneity of Ambiguity Preferences," *Review of Economics and Statistics*, **96**, 609-617.

Tversky, A., and Kahneman, D. (1992). "Cumulative Prospect Theory: An Analysis of Decision Under Uncertainty," *Journal of Risk and Uncertainty*, **5**, 297-323.

Wilcox, N. (2008). “Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison”. In J. C. Cox and G. W. Harrison, eds., *Research in Experimental Economics,* **12**, 197-292.

Wilcox, N. (2011). “Stochastically more risk averse:’ A contextual theory of stochastic discrete choice under risk,” *Journal of Econometrics*, **162**, 89–104.

Appendix

A. Behaviorally Distinguishable Types in the H01 Data Using the CHM Mixture Model.

Our implementation of the CHM econometric model on the Hey (2001; hereafter H01) data differs from theirs in only minor ways. First we prefer the logit specification of errors to their probit specification. The former entails a precision parameter γ, while the latter entails a variance parameter V. To compare these parameters, we suggest equating the probability of choosing the lottery with a value 5% more than the value of the other lottery:

$$p(\gamma) \equiv 1/(1 + exp(-0.05\gamma) = \Phi(0.05 \mid 0, V), \quad \text{(A1)}$$

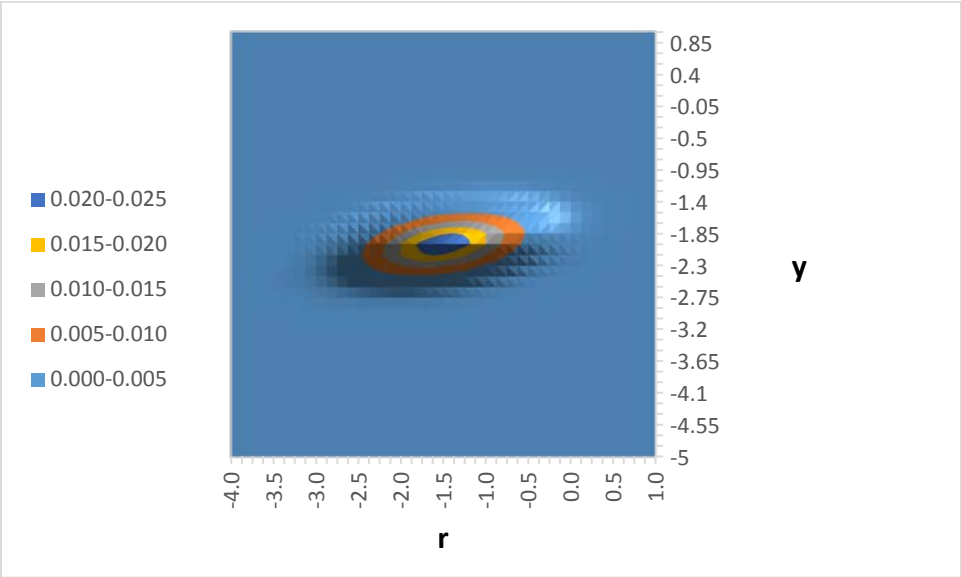
where $\Phi(0.05 \mid 0, V)$ is the cumulative normal distribution at 0.05 given a mean of 0 and variance of V. Thus, for the mixture component of the RDEU type, the CHM estimate of $\sqrt{V}$ = 0.03398 translates to γ = 51.555.

The other five parameters of the RDEU type are $r \equiv ln(1-\rho)$, $y \equiv ln(\beta - 0.2791)$, and the variance-covariance matrix of (r, y). The CHM MLEs are $\hat{r}$ = -0.95425, $\hat{y}$ = -0.55465, and var-cov parameters σ_r = 0.53947, σ_y = 0.24031, and cor(r, y) = 0.33792. The distribution over (r, y) is assumed to be a two-dimensional Gaussian with these means and var-cov matrix.

Second, rather than use statistical sampling from this distribution as CHM did, we create a 41×41 grid in which r ranges over $[\hat{r}-2.5, \hat{r}+2.5]$ and y ranges over $[\hat{y}-3, \hat{y}+3]$, so $(\hat{r}, \hat{y})$ is the midpoint of the grid. We then compute the probability density of the Gaussian distribution (given the CHM parameters) for each grid poin. Let $\varphi(r, y)$ denote this discretized distribution.[1] The figure below is a 2D contour graph of this distribution.

[1] Since we use a bounded range for these integration while CHM do not, it is important to demonstrate that we are not excluding a non-negligible probability mass. To do so, we can simply compute the discretized integral of the two-dimensional Gaussian probability density function $\varphi(r, y)$ over our grid. The result is 99.9996%.

Figure A1. φ(r, y)



As presented in Section 5, for each grid point (r, y), $\rho = 1 - exp(r)$ and $\beta = exp(y) + 0.2791$, we simulate 1000 data samples x from $f(x \mid \gamma, \rho, \beta)$ for a randomly selected set of 50 H01 tasks, and we simulate 1000 data samples from $f(x \mid \gamma, \rho, 1)$ for the same set of 50 H01 tasks, and then we compute the type I and type II rates. Finally, we compute the weighted sum of φ(r, y) over those grid points for which $max\{er_1, er_2\}$ exceeds 10%. This is the percentage of the subject population within the RDEU mixture component that is behaviorally indistinguishable from EU. We repeat this exercise for 24 more randomly selected sets of 50 HO tasks and report the average: 87.7% (3.0%).

Combining this result with the CHM estimate for the mixture parameter for the EU type of 0.19734, the total percentage of the subject population that is behaviorally indistinguishable from EU is given by 0.19734*100% + (1-0.19734)*87.7% = 90.1%.

B. Estimating the CHM Mixture Model on the HO Data.

The CHM mixture model entails 11 parameters. Our implementation of their econometric model on the HO data differs from theirs in only the two ways stated in part A above. That is,(i) we use the logit specification of errors to their probit specification, and (ii) we

use a grid method for integration rather than statistical sampling. Table I displays the maximum likelihood parameter estimates and the maximized log-likelihood for the first session of the HO data (i.e. 100 choice tasks for each of 80 subjects). A subscript "1" indicates a parameter for the EU component, and a subscript "2" indicates a parameter for the RDEU component. In addition, we also give the transformation from the estimated parameters (r, y) to the model parameters (ρ, β).

Table I. MLEs for CHM Mixture Model

	HO first 100	**CHM 2011**
γ_1	77.512	21.895
r_1	-1.5055; $\rho_1 = 0.77808$	-0.76438; $\rho_1 = 0.53438$
σ_{r1}	0.89839	0.32431
γ_2	36.219	51.555
r_2	-0.82503; $\rho_2 = 0.56178$	-0.95425; $\rho_2 = 0.61450$
y_2	-0.45059; $\beta = 0.91635$	-0.55465; $\beta = 0.85337$
σ_{r2}	0.44950	0.53947
σ_{y2}	0.23643	0.24031
cor_2	0.36867	0.33792
w	0.025204	0.01139
$1-\alpha$[2]	0.28146	0.19734
LL	-3363.13	-6708.40[3]

For comparison, the third column lists the MLEs reported in CHM. The first striking difference is the reversal of the precision and risk aversion parameters between the EU and RDEU types. In the HO data, the EU type is more precise and risk averse than the RDEU type, whereas the reverse occurs for the H01 data. The second notable difference is that the mixture parameter for the EU type is 50% greater for the HO data. Also the variance for risk aversion in

[2] In CHM's specification α is the weight on the RDEU component of the mixture, so $1-\alpha$ is the weight on the EU component.

[3] CHM report -6716.50 which is less than our computation; this difference could be attributed to the different computational methods (logit vs. probit, and grid vs. simulated integration).

the EU type is substantially greater for the HO data, while the variance-covariance parameters for the RDEU type are similar for the two data sets. These differences could be due to subject pool differences. Also note that the maximized LL is somewhat less than that found using our Bayesian approach (-3335.29).

C. Behaviorally Distinguishable Types in the HO Data Using the CHM Mixture Model.

Following the procedure described in part A above but using 25 randomly selected sets of 50 HO tasks, we find that the percentage of the subject population within the RDEU mixture component that is behaviorally indistinguishable from EU at the 10% level is 84.4% (2.4%). Combining this result with the estimate for the mixture parameter for the EU type of 0.28146, the total percentage of the subject population that is behaviorally indistinguishable from EU is given by 0.28146*100% + 0.71854*84.4% = 88.8%. These estimates are similar to those for the H01 data.