A Bayesian Method for Characterizing Population Heterogeneity

by

Dale O. Stahl
Malcolm Forsman Centennial Professor
Department of Economics
University of Texas at Austin

March 1, 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 percentage 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.

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 explore 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 \(\theta\) in a population of individuals. How do we estimate \(g(\theta)\) from observed behavior \(x \equiv \{x_i, 1, ..., N\}\)?

One approach is to find the \(\theta\) 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 \(\{\hat{\theta}_i, i = 1, ..., N\}\) gives a view of the sample distribution of \(\theta_i\) 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 (\(\Sigma\)) matrix. Estimating \(\Sigma_i\) matrices for each \(i\) entails many more parameter estimates, and still any density estimation using these \(\Sigma_i\) matrices would depend upon additional assumptions about the kernel of each \(i\), such as normality: \(N(\hat{\theta}_i, \Sigma_i)\).

Random coefficient models assume a parametric form for the population distribution: \(g(\theta \mid \beta)\), where \(\beta\) is a low-dimensional parameter vector. Typically, \(g(\theta \mid \beta)\) is a family of unimodal distributions in which \(\beta\) stands for the mean and \(\Sigma\) matrix. Obviously, these parametric

---

\(^1\) E.g. see Hey and Orme (1994), Harrison and Rutström (2008, 2009), Wilcox (2008, 2011), Fox, et al. (2011), and Stahl (2014).

\(^2\) We believe that this Bayesian approach is complementary to Wilcox (2008) and Fox, et al. (2011).
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 \( K \) distributions \( g(\theta | \beta_k) \) for \( k \in \{1, \ldots, K\} \). In addition, one has the mixture parameters \( \{c_k \geq 0\} \) such that \( \sum_{k=1}^{K} c_k = 1 \), so \( g(\theta) \equiv \sum_{k=1}^{K} c_k g(\theta | \beta_k) \). Then, the econometric task is to estimate the coefficients \( \{c_k, \beta_k, k = 1, \ldots, K\} \). Obviously, this method still suffers from potential mis-specification via the parametric restrictions on the distributions. In addition, there are many specifications of \( K \) distributions (typically called \( types \)) that yield the same aggregated \( g(\theta) \), so without further restrictions, the model is under-identified.

To review the 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 \( \mu_0 \) is a probability measure on \( G \), whereas \( g \) is a point in \( G \) and a probability measure on \( \Theta \). Given observed behavior \( x \equiv \{x_i, i=1,\ldots,N\} \), the posterior belief according to Bayes rule is

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

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

Suppose we first approximate \( \Theta \) by a finite grid. If for example there are four elements of \( \theta \) 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 \( \Theta \) would be a point in the 6.25 million dimensional simplex.\(^3\) Next we might approximate \( \Delta(G) \) by a grid with only 10 points in

\(^3\) 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.

Electronic copy available at: https://ssrn.com/abstract=3346557
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. For example, to allow for multiple modes in \(g\), \(G\) might be specified as a mixture of \(K\) normal distributions each with \(5+15 (= 20)\) free parameters. Then \(\Delta(G)\) might be specified as a unimodal normal distribution in \(R^{21K-1}\), with \(21K-1 + (441K^2-21K)/2\) parameters. However, obviously these restrictions are quite likely to be wrong, seriously bias the results, and still be computationally challenging.

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 each individual based on observed behavior \(x_i\). Assuming the subjects in the data are an unbiased sample from the population, we compute \(g^*(\theta \mid x)\) 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 \(\theta\).

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

---

4 E.g. Bolstad (2012).

5 The variance-covariance matrix in 5 dimensions has 15 degrees of freedom subject also to being positive semi-definite.

6 The mixture over \(K\) modes entails \(K-1\) mixture parameters, so there are \(20K + K - 1 = 21K - 1\) dimensions of \(G\). The variance-covariance matrix in \(21K-1\) dimensions has \((21K-1)21K/2 = (441K^2 - 21K)/2\) 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 \(g^*\).
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).

The paper is organized as follows. Section 2 presents the Bayesian approach. Section 3 presents the encompassing 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”. Section 6 asks and answers pertinent questions about behaviorally distinguishable types. For example, we find that of the subpopulation that is behaviorally distinguishable from uniformly-random, 84.2% is not behaviorally distinguishable from the ordinary Expected Utility (EU) model. Section 7 concludes.

2. Our Bayesian Alternative.

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 \( \theta \), by Bayes rule, the posterior on \( \theta \) is

\[
g(\theta \mid x_i) \equiv \frac{f(x_i \mid \theta)g_0(\theta)}{\int f(x_i \mid z)g_0(z)dz}. \tag{2}
\]

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

\[
g_i(\theta) \equiv \frac{1}{N-1} \sum_{h \neq i} g(\theta \mid x_h) \tag{3}
\]

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

\[
\hat{g}_i(\theta \mid \mathbf{x}) \equiv \frac{f(x_i \mid \theta)g_i(\theta)}{\int f(x_i \mid z)g_i(z)dz}, \tag{4}
\]

where \( \mathbf{x} \) denotes the entire \( N \)-subject data set. Finally, we aggregate these posteriors to obtain
\[ g^*(\theta | x) \equiv \frac{1}{N} \sum_{i=1}^{N} \hat{g}_i(\theta | x). \]  

We can interpret \( g^*(\theta | x) \) as the probability density that a random draw from the subject pool will have parameter vector \( \theta \). 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 \( g(\theta | x_i) \), since eq(5) requires that each \( \hat{g}_i(\theta | x) \) be properly normalized, MCMC methods cannot be used to simulate random draws from \( g^*(\theta | x) \).

When implementing this approach we construct a finite grid on the parameter space \( \Theta \) 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\(^9\) was introduced by Quiggin (1982, 1993). A convenient feature is that it nests EU and Expected Monetary Value (EMV).

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

\(^9\) This model is the same as the Cumulative Prospect model (Tversky and Kahneman, 1992) restricted to non-negative monetary outcomes.
function $H(F)$. From this transformation, the individual derives modified probabilities of each outcome:

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

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

$$
H(F_j) \equiv (F_j^\beta /[F_j^\beta + (1-F_j)^\beta])^{1/\beta}, \quad (2)
$$

where $\beta > 0$. 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

$$
U(F) = \sum_j v(y_j)h_j(F). \quad (3)
$$

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

$$
\text{Prob}(F^A) = \exp\{\gamma U(F^A)\}/[\exp\{\gamma U(F^A)\} + \exp\{\gamma U(F^B)\}], \quad (4)
$$

where $\gamma \geq 0$ is the precision parameter. Prob($F^A$) gives the probability that lottery $F^A$ is chosen rather than lottery $F^B$. As in EU theory, w.l.o.g. we can assign a value of 0 to the worst outcome and a value of 1 to the best outcome. Accordingly, for the data we specify $v_0 \equiv v(y_0) = 0$ and $v_3 \equiv v(y_3) = 1$. This leaves two free utility parameters: $v_1 \equiv v(y_1)$ and $v_2 \equiv v(y_2)$ subject to monotonicity $v_2 \geq v_1$. Hence, the empirical RDEU model entails four parameters: $(\gamma, v_1, v_2, \beta)$. Further, EMV is a special case of EU when $\beta = 1$, $v_1 = 1/3$ and $v_2 = 2/3$.

Next, to specify the likelihood function for our data, let $x_i \equiv \{x_i1, ..., x_iT\}$ 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 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(4).\textsuperscript{10} For notational convenience, let \( \theta_i \equiv (\gamma_i, \upsilon_i, \beta_i, \alpha_i) \). Then, in log-likelihood terms:

\[
LL(x_i, \theta_i) \equiv \sum_{t=1}^{T} x_{it} \ln \left[ \text{Prob}[F_A (\theta_i)] \right] + (1 - x_{it}) \ln \left[ 1 - \text{Prob}[F_A (\theta_i)] \right].
\]  

(5)

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

\[
LL(x, \theta) \equiv \sum_i LL(x_i, \theta_i),
\]  

(6)

where \( \theta \equiv \{ \theta_i, i = 1, ..., N \} \), and \( x \equiv \{ x_i, i = 1, ..., N \} \).

b. The HO Data.

The HO dataset contains 100 unique binary choice tasks.\textsuperscript{11} Each task was a choice between two lotteries with three prizes drawn from the set \{0£, 10£, 20£, 30£\}. A crucial design 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.

\textsuperscript{10} 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.

\textsuperscript{11} 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.
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.\textsuperscript{12}

One can estimate these parameters for each subject in the HO data set. That approach entails (4×80=320) parameters, even without the corresponding variance-covariance matrices.

\textsuperscript{12} 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. 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.
Table 1 gives the population mean and standard deviation of the point estimates\textsuperscript{13}. 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 $\gamma$ and $\beta$.

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

| $\gamma$ | $v_1$    | $v_2$    | $\beta$   | LL       |
|----------|----------|----------|-----------|----------|
| 55.22    | 0.6690   | 0.8371   | 1.048     | -2828.46 |
| (35.56)  | (0.1782) | (0.1039) | (0.736)   |          |

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 12 additional parameters per subject. Further, the estimates for each subject would ignore the fact that the subjects are random draws from 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 suited to extract information from the whole sample population. Consequently, we turn to the Bayesian approach.\textsuperscript{14}

4. Implementing our Bayesian Method.

When implementing our Bayesian method we specify the prior $g_0$ as follows. For the logit precision parameter, we specify $\gamma = 20 \ln[p/(1-p)]$ 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

\textsuperscript{13} This is the square root of the variance of the point estimates across subjects; it is not the standard error of individual point estimates.

\textsuperscript{14} For another application, see Stahl (2014).
chosen. Since the mean payoff difference between lottery pairs in the HO data set is about 5%, this is a reasonable scaling factor.\(^{15}\) \((v_1, v_2)\) is uniform on the unit triangle such that \(v_2 \geq v_1\). \(\ln(\beta)\) is uniform on \([-\ln(3), \ln(3)]\).\(^{16}\) These three distributions are assumed to be independent. For computations, we use a grid of \(41 \times 41 \times 41 \times 21 = 1,477,341\) points.

Since we cannot display a four-dimensional distribution, we present two two-dimensional marginal distributions. Figure 2 shows the marginal on \((p(\gamma), \beta)\), where \(p(\gamma) = 1/[1 + \exp(-0.05\gamma)]\).\(^{17}\) 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. \(p(\gamma) > 2/3\).

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

---

\(^{15}\) The following graphs and results are robust to this specification of the prior on \(\gamma\).

\(^{16}\) 95% of the individual MLEs for \(\beta\) lie in this range. Using a wider interval for the prior on \(\beta\) has no noticeable effect on the Bayesian posterior at the cost of more grid points.

\(^{17}\) Thus, \(p(\gamma)\) 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.
Figure 3 shows the marginal on \((v_1, v_2)\). From Figure 3 we see that the distribution is concentrated along the line \(v_2 = (v_1 + 1)/2\), which implies that utility is essentially linear above 0. We can also see a spike near the EMV point \((1/3, 2/3)\).

![Figure 3. Marginal of g* on (v1, v2).](image)

Given \(g^*(\theta \mid x)\) we can compute several statistics. First, the log-likelihood of the HO data is \(LL(g^*) = -3215.55\). In contrast, the log-likelihood of the four-parameter RDEU representative-subject model is \(-4423.62\). Obviously, the heterogeneity implicit in \(g^*\) fits the data much better than a representative-agent model.\(^{18}\) Compared to -2828.46 (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, -2828.46 is the sum of individually computed log-likelihoods using the subject-specific estimated parameters.

---

\(^{18}\) 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 g*-weighted mean of the parameter space is \((\overline{\gamma}, \overline{v}_1, \overline{v}_2, \overline{\beta}) = (43.3, 0.667, 0.831, 0.993)\); and \(p(\overline{\gamma}) = 0.853\). Further, \((\overline{v}_1 + 1)/2 - \overline{v}_2 = 0.0027\), as anticipated from Figure 2. Also note that \(\overline{\beta} \approx 1\), meaning that on average \(H(F)\) is the identity function. The variance-covariance matrix is

\[
\begin{array}{cccc}
p(\gamma) & v_1 & v_2 & \beta \\
p(\gamma) & 0.0089 & 0.0029 & 0.0028 & -0.0124 \\
v_1 & 0.0311 & 0.0172 & & -0.0184 \\
v_2 & & 0.0113 & -0.0118 & \\
\beta & & & 0.1418 & \\
\end{array}
\]

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

Indeed, we find evidence for multiple modes. A grid point \(\theta\) is declared a mode if and only if it has the highest value of g* in a 7x7x7x7 hypercube of nearest neighbors of \(\theta\) in the grid. The most prominent mode is at \(p(\gamma) = 0.975, v_1 = 0.55, v_2 = 0.80, \) and \(\beta = 1\). The next most prominent mode is at \(p(\gamma) = 0.999, v_1 = 0.70, v_2 = 0.90, \) and \(\beta = 1.12\). The third most prominent mode is at \(p(\gamma) = 0.999, v_1 = 0.85, v_2 = 0.95, \) and \(\beta = 0.681\). 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 the latter is -1538.05. In contrast, using individual parameter estimates from just the first 50 tasks, the log-likelihood of the second 50 tasks is -1851.34. 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.

The most productive use of \( g^*(\theta | x) \) is to test hypotheses. For example, we can ask what percent of the subject pool has \( \beta = 1 \). The answer is 10.5%; 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, what we really want to know is the percent of the population that is behaviorally indistinguishable 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 \( \theta \) or \( \theta' \), we typically compute the log of the likelihood ratio (LLR): \( \ln\left(\frac{f(x_i | \theta)}{f(x_i | \theta')}\right) \). A positive LLR means \( x_i \) is more likely to have been generated by \( \theta \) than \( \theta' \). 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 \( X_1 \equiv \{x_i | \ln(f(x_i | \theta)/f(x_i | \theta')) < 0\} \). If the data in fact was generated by \( \theta \), 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 \( \theta' \) 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:

\[
er_1 = \int_{X_1} f(x_i | \theta)dx_i \quad \text{and} \quad er_2 = \int_{X_2} f(x_i | \theta')dx_i .
\] (9)

If either of these error rates is too large, we might say that \( \theta \) and \( \theta' \) are behaviorally indistinguishable. Classical statistics suggests that a proper test statistic would have these error rates not exceed 5%.

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.
For the tests we want to conduct, generating all these possible $x_i$ vectors and computing $e_{r1}$ and $e_{r2}$ is obviously not feasible. Instead, we generate 1000 $x_i$ vectors from $f(x_i \mid \theta)$ and 1000 from $f(x_i \mid \theta')$. Then, $e_{r1}$ is approximated by the proportion of $x_i$ generated by $f(x_i \mid \theta)$ that lie in $X_1$, and $e_{r2}$ is approximated by the proportion of $x_i$ generated by $f(x_i \mid \theta')$ that lie in $X_2$.

In summary, we define $\theta$ and $\theta'$ to be *behaviorally distinguishable* if both of the simulated type-I and type-II error rates are less than or equal to 5%, and to be *behaviorally indistinguishable* otherwise.

6. **Questions and Answers.**

The questions we are interested in answering are easily framed in terms of our *behaviorally indistinguishable* relationship on the parameters. To begin, we 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 $\theta = (\gamma, u_1, u_2, \beta)$ is behaviorally indistinguishable from $(0, v_1, v_2, \beta)$, and then sum $g^*(\gamma, v_1, v_2, \beta)$ over all the grid points $(\gamma, \beta, v_1, v_2)$ that are behaviorally indistinguishable from $(0, v_1, v_2, \beta)$. The answer is 4.0%, which leaves 96.0% that is behaviorally distinguishable from Level-0. We are not interested in dissecting Level-0 behavior. Therefore, all our subsequent questions are conditional on the parameters being behaviorally distinguishable from Level-0.

Since Figure 3 provides strong evidence that the utility function is Linear Above Zero (LAZ), our next question is what percent of the population is behaviorally distinguishable from Level-0 but behaviorally indistinguishable from LAZ? The latter criteria can be stated as: $(\gamma, v_1, v_2, \beta)$ is behaviorally indistinguishable from $(\gamma, v_1, (1+v_1)/2, \beta)$. The answer is 92.5%. Hence, of the subpopulation that is behaviorally distinguishable from Level-0, 96.4% (= 92.5/96.0) is

---

19 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.
behaviorally indistinguishable from LAZ. We will not further dissect the non-LAZ subpopulation.

Perhaps 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 \((\gamma, v_1, v_2, \beta)\) that are behaviorally distinguishable from Level-0 but indistinguishable from \((\gamma, v_1, (1+v_1)/2, 1)\)? The answer is 78.2%. Hence, of the subpopulation that is behaviorally distinguishable from Level-0, 84.5% (= 78.2/96.0) is behaviorally indistinguishable from EU.

Our fourth and final question concerns the apparent aversion to 0 payoffs. What percent of the population are pure EMVs (i.e. maximize EMV with no aversion for 0 payoffs)? This additional criteria can be stated as: \((\gamma, v_1, v_2, \beta)\) is behaviorally indistinguishable from \((\gamma, 1/3, 2/3, 1)\). The answer is 12.5%. Hence, of the EU and LAZ subpopulation, 17.3% (= 12.5/78.2) have no aversion to $0 and 82.7% are averse to $0. Aversion to 0 is akin to loss aversion\(^{20}\), and the latter is a common result in the psychology literature (e.g. Kahneman and Tversky, 1979; Erev, Ert and Yechiam, 2008).

Figure 4 conveniently gathers these results in a bar graph. The first section labelled Level-0 represents the 4.0% that are behaviorally indistinguishable from Level-0. The second section labelled “not LAZ” represents the 3.5% (=96 – 92.5) that are behaviorally distinguishable from Level-0 and LAZ. The third section labelled “Not Beta=1” represents the 14.3% (= 92.5 – 78.2) that are behaviorally distinguishable from Level-0 and \(\beta=1\) but not from LAZ. The fourth section represents the 12.5% that are pure EMV maximizers. The fifth and final section represents the 65.7% (= 78.2 – 12.5) that are EMV maximizers but with an aversion to 0.

\(^{20}\) With prizes of \(\{0\£, 10\£, 20\£, 30\£\}\), the last three are clear wins, while 0£ is not a win and could easily be interpreted by the DM as losing relative to aspirations.
7. 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 computationally feasible approach. To extend our approach to models with more parameters, statistical sampling techniques can be employed to tame the curse of dimensionality.\textsuperscript{21}

Our Bayesian analysis has characterized substantial heterogeneity in the subject population. On the other hand it has revealed that 78.2% of the population is behaviorally indistinguishable from EU behavior (84.2% of the subpopulation that is behaviorally distinguishable from Level-0). Another interesting finding is that the vast majority of subjects are behaviorally indistinguishable from having a linear utility function from 10£ to 30£, although a majority exhibit an aversion to a 0 payoff.

\textsuperscript{21} 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.

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. 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,” *Econometrica*, 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.

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.

Wilcoxon, 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.

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