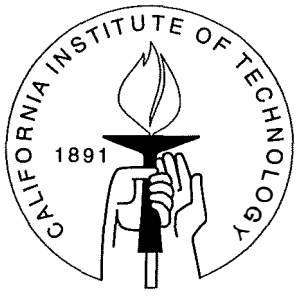
DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91125

UNKNOWN HETEROGENEITY, THE EC-EM ALGORITHM, AND LARGE T APPROXIMATION

Mahmoud A. El-Gamal
University of Wisconsin

David M. Grether



SOCIAL SCIENCE WORKING PAPER 988

October 1996

Unknown heterogeneity, the EC-EM algorithm, and Large T Approximation

Mahmoud A. El-Gamal David M. Grether

Abstract

We study a panel structure with n subjects/entities being observed over T periods. We consider a class of models for each subject's data generating process, and allow for unknown heterogeneity. In other words, we do not know how many types we have, what the types are, and which subjects belong to each type. We propose a large T approximation to the posterior mode on the unknowns through the Estimation/Classification (EC) algorithm of El-Gamal and Grether (1995) which is linear in n, T, and the unknown number of types. If our class of models (likelihood functions) allows for a consistent asymptotically normal estimator under the assumption of homogeneity (number of types $= 1$), then the estimators obtained by our EC algorithm inherit those asymptotic properties as $T \uparrow \infty$ and then as $n \uparrow \infty$ (with a block-diagonal covariance matrix facilitating hypothesis-testing). We then propose a large T approximation to the EM algorithm to obtain posteriors on the subject classifications and diagnostics for the goodness of the large T approximation in the EC stage. If the large T approximation does not seem to be appropriate, then we suggest the use of the more computationally costly EM algorithm, or the - even more costly - full Bayesian updating. We illustrate the procedure with two applications to experimental data on probability assessments within a class of Probit and a class of Tobit models.

Unknown heterogeneity, the EC-EM algorithm, and Large T Approximation

Mahmoud A. El-Gamal David M. Grether *

1 Introduction

Consider a general panel model with unknown heterogeneity. We wish to study observations (y_{it}, x_{it}) for individuals $i = 1, \ldots, n$, and time periods $t = 1, \ldots, T$ within a class of models which assumes that subject i's data is generated via the data generating process:

$$(y_{it}, x_{it})_{t=1}^{T} \sim \prod_{j=1}^{k} \Big(f\Big(\{y_{it}, x_{it}\}_{t=1}^{T}; \theta_j \Big) \Big)^{\delta_{ij}},$$

where $\delta_{ij} \in \{0, 1\}$ and $\sum_{j=1}^{k} \delta_{ij} = 1$, for some likelihood function $f(.; \theta)$ parametrized by $\theta \in \mathbb{R}^d$. [1] The unknown parameters of our model are: $(k, \theta_1, \ldots, \theta_k, \{\delta_{ij}\})$, all of which we wish to estimate. Moreover, once we estimate the number of types k, the actual parameters for each type $(\theta_1, \ldots, \theta_k)$, and the classification of subjects to types $\{\delta_{ij}\}$, we wish to calculate posteriors $\{p_{ij}\}$ for subject i being of type j and use the closeness of the $\{p_{ij}\}$'s to the $\{\delta_{ij}\}$'s as a diagnostic for the goodness of our classification.

We have cast the problem as one of a finite mixture of types, and treated the problem of dealing with heterogeneity in the panel as a classification problem. This contrasts

*Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706-1393, and DHSS, 228-77, Caltech, Pasadena, CA 91125. We acknowledge financial support from NSF grant #SBR-9320497 to the California Institute of Technology. We thank David Stephenson for programming assistance. We also thank participants at the CEME-NSF conference on microeconometrics in Madison, WI, June 1995 and in the econometrics workshops at Arizona, Columbia, Cornell, Northwestern, and Wisconsin, and audiences at Economic Science Association meetings, University of Amsterdam, Bayesian Research Conference at LA, and Behavioral Decision Making at MIT for many useful comments. Any remaining errors are, of course, our own.

[1] Two simplifying assumptions are implicit in this specification. We are implicitly assuming independence across subjects (but not necessarily across time periods). This is made for ease of the exposition. The extensions to possible dependence across individuals will complicate the notation and require unintuitive technical assumptions. The second simplifying assumption we have made is that of θ being finite dimensional. Allowing for an infinite dimensional vector of parameters or nuisance parameters is conceptually straight-forward as long as the proper asymptotics for $k = 1$ hold.

with the approach most commonly used in econometrics following the seminal paper of Heckman and Singer (1984) which uses the continuum mixture model of Lindsay (1981), Lindsay (1983a), Lindsay (1983b). We must note at this point that the setup of Lindsay and of Heckman and Singer is clearly restricted in any given sample to a distribution of types with a finite number of points of support, and the empirical work of Heckman and Singer clearly aims at finding a small number of such types/points of support. Therefore, apart from modeling preferences regarding a continuum or a discrete space of types, the end-result sought by both approaches in any given data set is the same. For a sample of n individuals, the maximum number of points of support in the continuum approach, and the maximum number of types in the discrete approach, is n; and both approaches strive to find a number of types (points of support of the types distribution) $k << n$.

The EC-EM algorithm first presented in El-Gamal and Grether (1995) and discussed below provides us with a computationally simple approximation to the fully Bayesian approach to this problem. The fully Bayesian approach would start by specifying priors $\alpha(k)$ on $\mathbb{N}$, $\lambda_k(\theta_1, \ldots, \theta_k)$ on the k types conditional on a value of k, and $\nu_{k,\theta_1,\ldots,\theta_k}(\{\delta_{ij}\})$ on the classifications of subjects to types given the value of k and the k-types. To simplify the notation, let $\tau = (k, \theta_1, \ldots, \theta_k, \{\delta_{ij}\})$, let the prior on τ be $\eta(.)$, denote the data $z = \{y_{it}, x_{it}\}_{t=1,i=1}^{T,n}$, and let the likelihood function of the data be $f(z|\tau)$. Then, the fully Bayesian approach would calculate the posterior

$$\zeta(d\tau|z) = \frac{\eta(d\tau)f(z|\tau)}{\int_\tau \eta(d\tau)f(z|\tau)}.$$

It is well known that this fully Bayesian approach may be too costly from a computational point of view, especially since the dimensionality of τ grows very quickly (as $d \times k$ for the θ's and $k^n/k!$ for the δ_{ij}'s).

In the statistical mixture of types literature, the number of types k is assumed to be known, and it was proposed (Dempster et al. (1977), Redner and Walker (1984), Little and Rubin (1987)) that one uses the EM algorithm to maximize the expected log posterior (which also can be seen as maximizing the expected log likelihood either asymptotically, or for a flat prior). This provides a computationally cheaper way to obtain an approximation to the mode of the posterior $\zeta(d\tau|z)$. In particular, the classifications $\{\delta_{ij}\}$ are treated as missing data (or nuisance parameters) to be integrated out, and the rest of the parameters are treated as genuine parameters to be estimated by maximizing the expected log posterior (or log likelihood) function through the iterative EM algorithm. When calculating the expected log posterior for a given guess of the parameter values, the expectations of the δ_{ij}'s appear in the form of probabilities that enter linearly, and that are updated using Bayes's rule (see Redner and Walker (1984)). However, the computational burden of this procedure is still quite formidable, for it is known that the EM algorithm is very slow to converge, and even though we only need to maximize over $(\theta_1, \ldots, \theta_k) \in \mathbb{R}^{dk}$ in each M-step, there could be a very large number of such steps. The EM algorithm

will, therefore, give us an approximation of the mode of the expected log posterior by performing a large number of optimizations over $\mathbb{R}^{dk}$. Matters are complicated further by the fact that the optimand function $\mathbb{R}^{dk} \mapsto \mathbb{R}$ has many local optima which requires sophisticated and costly optimization methods in each step of the algorithm, and - if we get trapped in a local optimum in some iteration - can lead to many more iterations before the algorithm converges. As a matter of fact, a single high-dimensional integration for the calculation of the Bayes posterior may indeed be cheaper computationally in this case.

The EC algorithm of El-Gamal and Grether (1995) provides a further approximation to the mode of the posterior function through a single optimization over $\mathbb{R}^{dk}$, and in a manner such that each optimand evaluation requires a number of calculations that grows linearly in k, n and T. The EC algorithm proceeds as follows: for each k, search over $\mathbb{R}^{dk}$, and for each function evaluation at a vector $(\theta_1, \ldots, \theta_k)$, and each individual i, evaluate the k numbers: $l_j = \log f(\{y_{it}, x_{it}\}_{t=1}^{T}; \theta_j)$, and then evaluate the overall likelihood function optimand as $F(\theta_1, \ldots, \theta_k) = \sum_{i=1}^{n} \max_{j \in \{1,\ldots,k\}} l_j$. This procedure is equivalent to maximizing over the δ_{ij}'s and the θ_j's simultaneously.[2] If we interpret the δ_{ij}'s as missing data, the proposed maximization of the likelihood function over the parameters and missing data falls under the criticism in Little and Rubin (1983) of maximizing the likelihood function over the missing data. However, in our panel structure, and as a large T approximation, the EC algorithm replaces the integration in the E-step of the EM algorithm with a maximization over the δ_{ij}'s in the same manner that the EM algorithm replaces the integration over the θ_j's in the fully Bayesian procedure by maximization in each of the M-steps. This appeal to the large T asymptotic results (which we discuss later in the paper), together with the extreme cost-effectiveness over the Bayes method and the EM algorithm, suggests that the EC algorithm can be of substantial usefulness.[3]

For comparability with other recent work on heterogeneity in panel models, we illustrate the use of the EC algorithm in this paper with two applications using limited dependent variables in one case, and trimmed dependent variables in the other. Our Probit and Tobit models with a finite number of types compare with the Probit and Tobit models with fixed effects in two ways. The first is that the EC algorithm uncovers the types in the population instead of presuming a class of types and then testing for homogeneity through the use of fixed effects dummy variables. The second is that the EC

[2] This overall maximiztion is sometimes called "the classification likelihood approach to clustering", see McLachlan and Basford (1988), pp.31–36 and the references therein. The applications of this general idea have been limited by the unavailability of our EC-algorithm which avoids computing the likelihood for all possible classifications of n subjects into k classes, as well as the "fixed T, large n" inconsistency and bias of the resulting estimates discussed below.

[3] Another instance where a large T approximation proved useful in overcoming the classical problems of dealing with slope heterogeneity in the regression framework with dynamic panels is in Pesaran et al. (1996), where they argue that panels with reasonably large T are now easily available, and that ignoring heterogeneity in the slope parameters cannot be justified on the basis of T being too small.

algorithm is stronger in the cases where T is large relative to n, which is exactly where Monte Carlo results show that popular estimation techniques such as Honoré (1992), Honoré (1993) are at their weakest. In this respect, the EC estimators may be seen as a "large T, small n" complimentary tool to the "large n, small T" estimators in the literature. Moreover, as we prove in section 3, our estimator can take as a primitive any extremum estimator with consistency and/or asymptotic normality properties in the case with no heterogeneity, and inherit those properties as $T \uparrow \infty$. In that respect, standard ML estimators such as Tobit and Probit estimators, as well as semi-parametric GMM estimators such as Honoré (1992)'s trimmed least squares and trimmed least absolute deviations estimators, can all be combined with the EC algorithm to estimate the unknown number of types, the actual types, and the subject classifications in any given sample, without the need to overparametrize the model by adding n dummy variables for the fixed effects.

In Section 2, we present the EC estimator and algorithm more rigorously. In Section 3, we prove that EC estimators inherit the consistency and/or asymptotic normality of estimators in the case of no heterogeneity ($k = 1$). In Section 4, we present a hierarchy of large T approximations where each element in the hierarchy is more computationally demanding than those below and more robust to misclassification errors. In Section 5, we calibrate the results of El-Gamal and Grether (1995) on updating rules and investigate their robustness to model parametrization by re-analyzing the data using a Probit class of models. In Section 6, we study a second set of experiments on probability updating using a Tobit class of models. Section 7 concludes the paper.

2 The EC Algorithm

We assume, as a starting point, the availability of some $\sqrt{nT}$-consistent estimator of θ if there is no heterogeneity in the population. Implicit in this assumption is the special case ($n = 1$) for which our estimator would be $\sqrt{T}$-consistent. For much of this paper, we have in mind our $\sqrt{nT}$-consistent estimator being a maximum likelihood estimator, and connections with other statistical procedures can easily be found in this case. However, for the asymptotics of the EC-algorithm estimates studied in the next section, the estimator does not have to be a maximum likelihood estimator. The $\sqrt{nT}$-consistent estimator available to us is assumed to be an extremum estimator, which in the case of no heterogeneity is defined as follows:

$$\hat{\theta}_{nT} = \text{argmax}_{\theta \in \Theta} \sum_{i=1}^{n} \log g\Big(\{y_{it}, x_{it}\}_{t=1}^{T}; \theta_h\Big).$$

This will coincide with a maximum likelihood estimator with independence across individuals if $g(.;\theta) = f(.;\theta)$, but $\log g$ can be specified as any other optimand corresponding

to a semi-parametric procedure as long as the objective function can be decomposed into a sum over the individuals (e.g. least squares, GMM, etc. all satisfy this requirement).

We now consider moving from our extremum estimator in the case of homogeneity ($k = 1$) to allowing for an unknown number of types $(\theta_1, \ldots, \theta_k)$ in the population. Our EC-algorithm simply augments the optimand to allow for an arbitrary number of types k and arbitrary unknown types, and adds a penalty which is increasing in k. Formally, the EC-estimator is defined by:

$$(\hat{k}, \hat{\theta}_1, \ldots, \hat{\theta}_{\hat{k}}) = \text{argmax}_{k', \theta_1, \ldots, \theta_{k'}} \left\{ \sum_{i=1}^{n} \left(\max_{h \in \{1, \ldots, k'\}} \log g\left(\{y_{it}, x_{it}\}_{t=1}^{T}; \theta_h \right) \right) - penalty(n, k') \right\}.$$

Where $penalty(n, k)$ is some non-stochastic increasing function of k for each n, with $m(n) \leq penalty(n, k) - penalty(n, k-1) \leq M(n)$ for some finite positive numbers $m(n), M(n)$. We then estimate π_h, the proportion of subjects of type h, by its estimated sample analog:

$$\hat{\pi}_h = \frac{1}{n} \sum_{i=1}^{n} \hat{\delta}_{ih},$$

where

$$\hat{\delta}_{ih} = \begin{cases} 1 & \text{if } \; h = \arg\max_{l \in \{1, \ldots, \hat{k}\}} \log g\left(\{y_{it}, x_{it}\}_{t=1}^{T}; \theta_h \right); \\ 0 & \text{otherwise.} \end{cases}$$

In principle, for any given k, the maximization over $(\theta_1, \ldots, \theta_k, \{\delta_{ij}\})$ could be done by brute-force, looping over all possible groupings of the n subjects into k groups. In that case, the same optimand written above can be rewritten with the explicit maximization over all possible values of the $\{\delta_{ij}\}$'s written explicitly:

$$(\hat{k}, \hat{\theta}_1, \ldots, \hat{\theta}_{\hat{k}}) = \text{argmax}_{k', \theta_1, \ldots, \theta_{k'}, \{\delta_{ij}\}} \left\{ \sum_{i=1}^{n} \left(\sum_{h=1}^{k'} \delta_{ij} \log g\left(\{y_{it}, x_{it}\}_{t=1}^{T}; \theta_h \right) \right) - penalty(n, k') \right\}.$$

Evaluating the likelihood function for all possible matrices $\{\delta_{ij}\}$ grows with a function of order $k^n/k!$. However, the EC-algorithm we suggest is linear in k, and linear in n. The algorithm performs the EC maximization as follows:

- For each k, call optimization routine for function $F_k(.; \theta_1, \ldots, \theta_k)$:
- For each evaluation of $F_k(.; \theta_1, \ldots, \theta_k)$:
  - Loop over individuals $i = 1, \ldots, n$.
    - For each individual, calculate possible contributions (for $h = 1, \ldots, k$):

$$F_k^{ih}(\{y_{it}, x_{it}\}; \theta_1, \ldots, \theta_k) = \log g(\{y_{it}, x_{it}\}; \theta_h)$$

    * Return contribution $F_k^i(.;\theta_1,\ldots,\theta_k) = \max_h F_k^{ih}(.;\theta_1,\ldots,\theta_k)$.
  - return $F_k(.;\theta_1,\ldots,\theta_k) = \sum_{i=1}^n F_k^i(.;\theta_1,\ldots,\theta_k)$.
- For each k calculate $L_k(.) = \max_{\theta_1,\ldots,\theta_k} F_k(.;\theta_1,\ldots,\theta_k)$.
- Choose $\hat{k}$ to maximize $L_k(.) - penalty(n,k)$.

In the next section, we show that for the consistency of our estimates, any penalty function which is increasing and bounded away from 0 and ∞ will do. In practice, however, we prefer to think of the penalty function as arising from a prior on k, the parameters, and the classifications. In that case, the optimand (maximum log likelihood for k + log prior(n,k)) can be viewed as an asymptotic approximation of the posterior mode over k and the parameters, in the same manner that maximum likelihood estimation in the case of no heterogeneity serves is an asymptotic approximation to the posterior mode (c.f. Jeffreys (1939)). This immediately allows us to draw parallels with information criteria along the lines of Akaike (1974) and Schwarz (1978) in the regression framework, as well as coding theoretic prescriptions along the lines of minimum message length (MML) (Wallace and Boulton (1968)) and minimum description length (MDL) (Rissanen (1978)).[4]

Our approach to unknown heterogeneity is preferable to standard econometric techniques of dealing with heterogeneity for two reasons. The first reason is parsimony. When the algorithm converges to a classification of subjects and estimates of the collection of parameter values for each type, we may get classes of subjects that are not ex ante obvious. For instance, in estimating a wage equation with unknown heterogeneity, we may find a class for subjects who are of a certain sex, race, and educational level. Including dummy variables for the interaction of all three terms in a fixed effects model is unlikely unless we also included dummy variables for each of them (and perhaps also all pair-wise interactions). Fixed and random effects models require prior knowledge of the possible sources of heterogeneity, and the only way to uncover complex sources is initially to include many ex post unnecessary effects. If one thinks of avoiding this loss in parsimony by conducting some sort of specification search eliminating lower order terms that turn out to be insignificant, we come to the second reason our procedure is preferable. A specification search where fixed effects that turn out to be statistically insignificant are removed

[4]The latter two approaches are quite similar, though their authors recommend somewhat different methods of implementation. In both cases, the objective is to describe a data set by the shortest possible message. In general, given a sample from a known discrete distribution $P\{.\}$, the optimum word length for a datum x_i is $-\log_j(P\{X = x_i\})$ (where the base j is equal to the number of symbols in the alphabet used to give the data) so that the expected word length is set equal to the entropy (information) of the distribution. In practice, the distribution is not known, so the message contains two parts , one of which describes the model (a hypothesized distribution) and the second gives the likelihood of the data under that model. There is then a trade-off as more complex models yield shorter summaries of the data, but take longer messages to describe (see Wallace and Freeman (1987) and Rissanen (1987)).

poses serious statistical problems of "pre-testing", which make the interpretation of later statistical tests of significance etc. questionable (even asymptotically).

The EC-algorithm bears close resemblance to a number of procedures in the statistical and engineering literatures. In particular, the k-means literature (c.f. Pollard (1982) and the references therein) in statistics, and the "unsupervised learning" algorithms in the neural networks literature (most notably, the so called winner-takes-all and self-organizing map networks, c.f. Kohonen (1990) and the references therein), deal with a similar problem of classifying n objects into k classes using some objective function. The problems in these literatures are typically coded in such a way that the objective function to minimize is the average Euclidian distance between the elements of each class and the center of that class. In the case of having a $\sqrt{nT}$-consistent and asymptotically normal estimator of the parameters in the case of $k = 1$, one may (for each k) estimate a parameter value $\hat{\theta}_i$ for each individual and run one of the above mentioned algorithms to cluster point estimates into k groups. The centers of the resulting k groups may then be used as initial conditions for the EC optimization. Due to the $\sqrt{T}$-consistency of the individual estimates, the clusters will asymptotically (as $T \uparrow \infty$) be consistent estimates of the k classes, and the EC algorithm will not move from its initial point. Moreover, for finite but large T, this initialization of the search can help circumvent the problem of too many local optima for our EC objective function (generically, each classification of the subjects will provide at least one local optimum for the criterion function). Of course, we would not consider the k-clustering from n point estimates for the n-subjects an end product, since hypothesis tests on the resulting point estimates of cluster centers will be difficult to conduct. However, by using those algorithms as providers of intelligent initial conditions for the search, we can increase our confidence in finding a global optimum, and reduce the amount of calculation required to achieve it (although, of course, we should still use randomization methods to reduce the probability of getting trapped in a local optimum).

In the introduction, we discussed the relationship between our EC algorithm and the EM algorithm used in the mixture of types statistical literature. The numerical advantages of the EC algorithm are unquestionably large, but the justification for making those cuts in computational costs rests heavily on our belief in the goodness of the large T approximation. We recall that the EM algorithm is itself a large T approximation of the expected log posterior mode, where the likelihood of each subject's data under the parameters of each class enters weighted by our posterior on the classification of that subject to the corresponding rule. Asymptotically, those posteriors $p_{ij} = E[\delta_{ij}] \rightarrow \delta_{ij}$, and the EC and EM algorithms will produce the same results. We therefore suggest a compromise application of the EM algorithm once our EC estimates are obtained as a diagnostic for the goodness of our large T approximation. We suggest fixing the EC estimates $(\hat{k}, \hat{\theta}_1, \ldots, \hat{\theta}_k)$, and iterating on the E-step of the EM algorithm (replacing the estimates that would be obtained in each costly M-step by the EC estimates as an approximation). We can then use the resulting posteriors on the classifications as a

diagnostic for the goodness of our large T approximation. The closer those posteriors are to 0's and 1's, the more confident we can be that our asymptotic approximation is sufficient. This issue will be discussed in more detail in Section 4 below.

3 Asymptotics of the EC-Estimator

Before proving that our EC estimates inherit the desirable asymptotic properties of our extremum estimator under no heterogeneity, we need to discuss what our asymptotic approximation is designed to accomplish. First of all, we state all asymptotic results as the number of observations per individual T goes to infinity, and then as the number of individuals n goes to infinity. The order in which we take those limits is crucial for the validity of our asymptotic theorem, but what it means for finite (but large) sample approximation is less obvious. A large T approximation (that is $T \uparrow \infty$ first) does not mean that T must be larger than n, or grow with n in some pre-determined fashion. For instance, there may be some level $t^\star$ that allows us to classify individuals non-stochastically (for example, we may be able to identify species by a finite number of characteristics), in that case, any $t > t^\star$ will be enough to justify our large T approximation. How large T needs to be to justify our approximation clearly depends on the application, and hence it is reasonable to return to the data ex post and use a diagnostic (such as our large T approximation to the EM algorithm) to determine if the large T approximation is justified in that particular sample.

The second possible problem we may encounter with our large sample approximations centers around the interpretation of $n \uparrow \infty$. One may ask: what if the number of types k itself grows with n? The result we prove below for the consistent estimation of k is perhaps better understood within the framework of Fisher consistency (with a large but finite maximal sample size N which serves as an upper bound on k). Another way to interpret that result of the consistency of our estimate of k is to assume that all the types are represented in our finite sample, and consistency simply means that as the number of observations per individual goes to infinity, we will not group together individuals who belong to different types, or create spurious types for individuals who belong to one of the estimated types. As the number of individuals goes to infinity, the proportions of each type in our sample then become more exact estimates of the correct π_j's, hence the consistency of those estimates. Keeping in mind the nature of large sample approximations that we have in mind, and the interpretations of "$T \uparrow \infty$ and then $n \uparrow \infty$" discussed here, we now turn to the technical statement of the EC estimate asymptotics.

We shall prove the consistency and asymptotic normality of our EC-estimates under the following assumptions:

(A.1) For $k = 1$, $\hat{\theta}_1$ is $\sqrt{nT}$-consistent and asymptotically normal.

(A.2) For each individual i belonging to group h, $(1/T)g(\{y_{it}, x_{it}\}_{t=1}^{T}; \theta) \rightarrow G_h(\theta)$ in probability uniformly in θ in a neighborhood of θ_h, the unique global maximizer of the non-stochastic limit function $G_h(.)$.

Assumption (A.1) says that we have as our primitive a consistent and asymptotically normal estimator in the case with no heterogeneity, and we wish to show that those asymptotic properties will be inherited by our EC-estimates under the assumption of unknown heterogeneity (or, potentially, the estimates may show that there indeed is no heterogeneity ($\hat{k} = 1$)). Sufficient conditions for (A.1) may be obtained along the lines of Amemiya (1985), Assumptions (A)-(C), p. 111. Assumption (A.2) is a standard assumption under which one can prove consistency and asymptotic normality of extremum estimators (see e.g. Amemiya (1985), Assumptions (A)-(C), p. 110).

Theorem:

Under (A.1) and (A.2), taking limits as $T \uparrow \infty$ and then as $n \uparrow \infty$:

1. $\hat{k}$ is a consistent estimator of k,
2. $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ are $\sqrt{nT}$-consistent and asymptotically normal, with a block-diagonal variance-covariance matrix,
3. The $\hat{\delta}_{ij}$'s are consistent estimators of the δ_{ij}'s.
4. The $\hat{\pi}_j$'s are consistent estimates of the π_j's.

Proof:

Let us consider any candidate estimate $\hat{k}$ for the number of groups. Given any two individuals i, i' such that $\delta_{ih} = \delta_{i'h} = 1$, as $T \uparrow \infty$, let $\hat{\theta}_T^i$ and $\hat{\theta}_T^{i'}$ be the estimates of θ based only on the data of i and i' respectively. Then, by assumption (A.1) and the triangle inequality, $\|\hat{\theta}_T^i - \hat{\theta}_T^{i'}\| \rightarrow 0$ in probability as $T \uparrow \infty$. Moreover, the contributions of the data generated by i and i' to the overall optimand both converge to the same non-stochastic function $TG_h(\theta)$. Therefore, whatever candidate vector of parameters $(\hat{\theta}_1, \ldots, \hat{\theta}_{\hat{k}})$ we consider, as $T \uparrow \infty$, the same element of that vector which maximizes individual i's contribution to the objective function (which is in this case $\max_{\theta \in \{\hat{\theta}_1, \ldots, \hat{\theta}_{\hat{k}}\}} G_h(\theta)$) must in the limit also maximize individual i''s contribution). Therefore, if $\hat{k} \leq k$, individuals who belong to the same group will asymptotically (as $T \uparrow \infty$) be classified together.[5]

Now, $\hat{k} < k$ means that some individuals belonging to different classes must have been grouped together. Let I be a group of individuals who use rule h, and who are combined with another group that uses rule h'. Let the estimated parameter for the

[5] If $\hat{k} > k$, then two groups may have identical parameters, and subjects of the same type may be split across those identical groups. As shown below, the penalty function would prune away such extra types bringing $\hat{k} = k$.

combined group be $\hat{\theta}_{\hat{h}}$. The increase in the value of our objective function from adding one more rule is asymptotically, as $T \uparrow \infty$, greater than $\sum_{i \in I}[\log(g(\{y_{it}, x_{it}\}_{t=1}^{T}; \theta_h) - \log(g(\{y_{it}, x_{it}\}_{t=1}^{T}; \theta_{\hat{h}}))]$. For any given n, as T gets larger, by assumption (A.2), this term eventually exceeds $T[G_h(\theta_h) - G_h(\theta_{\hat{h}})]$ and hence diverges, eventually exceeding $M(n)$ (the bound on the incremental penalty for adding an extra rule). Hence, for each value of n, as $T \uparrow \infty$, the probability of $\hat{k} < k$ goes to zero. On the other hand, if $\hat{k} > k$, then as $T \uparrow \infty$, the gain from keeping the extra $\hat{k} - k$ rules will converge in probability to 0, eventually falling below $m(n)$, and we shall discard those extra rules. Therefore, for any given n, as $T \uparrow \infty$, $\operatorname{plim}_{T \uparrow \infty} \hat{k} = k$ (this would be almost-sure convergence if we had strong consistency for $k = 1$). Since this holds for all n, it follows that $\hat{k}$ is a consistent estimate of k, and consequently that the δ_{ij}'s are also consistent estimators.

Now that we have established that as $T \uparrow \infty$, the number of classes k is consistently estimated by $\hat{k}$, and the individuals are consistently classified with the other individuals who use the same rule, the rest of the results follow trivially. The consistency of $\hat{\pi}_1, \ldots, \hat{\pi}_{\hat{k}}$ as $n \uparrow \infty$ follows directly from the strong law of large numbers and the consistency of the δ_{ij}'s. The consistent and asymptotically normal estimation of the vector $(\theta_1, \ldots, \theta_k)$ also follows immediately from the assumed consistency and asymptotic normality for the case with $k = 1$ (since each θ_h is being estimated only from the data generated by individuals from group h). Moreover, since subjects asymptotically contribute a score (non-zero gradient to the optimand function) only to the θ_h for their group, the variance-covariance matrix for the estimates $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ is asymptotically block-diagonal. ■

4 Computational Issues and a Hierarchy of Large T Approximations

As seen in section 3, the most important consequence of the large T approximation (used to prove the asymptotic properties of our EC estimators) is the fact that each individual is asymptotically classified with the appropriate group with probability 1. A measure of the goodness of the large T approximation which lends legitimacy to using the EC estimates, therefore, can be constructed from the posteriors on subject classification. If all the posteriors on subject classifications were very close to 1 for the rule to which they were classified (and 0's for all other rules), then we would conclude that the large T approximation is satisfactory (since the probability of more data changing our classification is very small).

The calculation of posteriors on subject classifications may be obtained by interpreting the unknown classifications as missing data and using the EM-algorithm of Dempster et al. (1977) to maximize expected log-likelihood as applied by Redner and Walker (1984), Little and Rubin (1983), and Little and Rubin (1987) to mixture of types models. In general, the EM algorithm provides an approximation to the mode of the Bayesian

posterior over parameters and missing data (classifications in our case) by maximizing the expected log-likelihood where the expectation is calculated by integrating out the missing data. Let Y be the observed data, Z be the missing data, then the EM-algorithm proceeds as follows:

initialization: Start with a guess θ^1.

E-Step: For iteration s, calculate the expected log likelihood:

$$Q(\theta, \theta^s) = \int_Z \log\Big[f(\theta|Y,Z)\Big]\; p(Z|\theta^s, Y)\; dZ$$

M-Step: Set $\theta^{s+1} = \text{argmax}_{\theta\in\Theta} Q(\theta, \theta^s)$.

Convergence: Repeat until $\|\theta^{s+1} - \theta^s\|$ (or $|Q(\theta, \theta^{s+1}) - Q(\theta, \theta^s)|$) is small.

In the case where the missing data are the classification $\{\delta_{ij}\}$'s, the integration in the E-step becomes very simple: Let π_j be the proportion of subjects of type j. Then, start with guesses π_j^0, $\theta^1 = (\theta_1^1, \ldots, \theta_k^1)$, and for the τ^{th} iteration,

$$p(\{\delta_{ij} = 1\}|\theta^s, Y) = p_{ij}^s = \frac{\pi_j^{s-1} f(\theta^s|Y_i, \{\delta_{ij} = 1\})}{\sum_{l=1}^k \pi_l^{s-1} f(\theta^s|Y_i, \{\delta_{il} = 1\})},$$

set

$$\pi_j^s = \sum_{l=1}^k p_{ij}^s,$$

and calculate

$$Q(\theta, \theta^s) = \sum_{i=1}^n \sum_{j=1}^k p_{ij}^s \log\Big[f(Y_i|\theta_j)\Big].$$

As we mentioned above, using the EM-algorithm to calculate the posteriors on the classifications as well as to provide a small sample correction of possible EC-estimate biases (due to misclassifications) by re-maximizing over the coefficients in each M-step may be costly. We therefore suggest the use of a large T approximation to the EM-algorithm by assuming that in each M-step, the parameter estimates will remain unchanged. The resulting approximate EM algorithm, therefore, produces posteriors on each subject's classification that are consistent with the EC-algorithm parameter estimates and priors equal to the average posterior for each group. In other words, letting l_{ij} be the likelihood of group j's parameters under subject i's data, we impose the restrictions that the posterior on subject i belonging to group j should be:

$$p_{ij} = \frac{\pi_j l_{ij}}{\sum_{j'=1}^k \pi_{j'} l_{ij'}},$$

where

$$\pi_j = \frac{1}{n}\sum_{i=1}^{n} p_{ij}.$$

The large T approximation to the EM-algorithm is used to obtain the fixed point $\{\pi_j\}$ and $\{p_{ij}\}$ satisfying those two equations. We start with the proportion of subjects allocated to each rule as a guess for the π_j's, and then iterate on the two equations till convergence. This provides us with the collection of posteriors $\{p_{ij}\}$. We now need a collective measure for how close these posteriors are to 1's and 0's. We propose the measure

$$\text{ANE}(k) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} p_{ij}\ \log_k(p_{ij}),$$

which is the average over all subjects of the entropy in the posterior for that subject normalized by using $\log_k$ so that the value 1 denotes a uniform prior. The closer ANE is to 0, the better is our large T approximation, and the closer it is to 1, the worse is our approximation.

Figure 1 summarizes our discussion of the hierarchy of large T approximations, where at the bottom of the hierarchy, estimates of the parameters for each individual provide consistent estimates of the parameters for each individual, and therefore provide a good starting point for EC-estimation searches. Once the EC-estimates are obtained, we can use the large T approximation to the EM algorithm to obtain approximate posteriors on the subject classifications and compute ANE's to check the reliability of the large T approximation (and resulting EC-estimates) in any given study. If the ANE's suggest that the EC-estimates are not reliable, then we would need to calculate the EM-estimates, and if the M-steps of that algorithm suggest that the expected log-likelihood function is too flat around the estimates to make them reliable, then the fully Bayesian procedure described in Section 1 would be appropriate. We note that each step in the hierarchy provides better estimates/posterior distributions, but is more computationally intensive. Let $\theta_j \in \Theta \subset \Re^d$, and let the number of types be k, then the computational requirements of the procedures summarized in Figure 1 are as follows:

Bayes's rule: Integrate over $\mathbb{R}^{dk} \times \{0,1\}^{O(k^n/k!)}$.

EM-algorithm: For each iteration, perform $\mathbb{R}^{dk}$ optimization (which may factor into k separate $\mathbb{R}^d$ optimizations), and one Bayesian updating of nk posteriors. Number of iterations unknown ex ante, multiple local optima for each optimization.

Large T approximation to the EM-algorithm: In addition to the requirements for the EC-algorithm, calculate nk likelihoods and iterate on Bayes's formula to find fixed point π_j's and p_{ij}'s (very fast).

EC-algorithm: One maximization over an $\Re^{dk}$ dimensional function with many local maxima.

Separate estimates: n optimizations, each over $\mathbb{R}^d$

In El-Gamal and Grether (1995), as well as in the two applications in this paper, the EC-estimates and large T approximation to the EM algorithm produce ANE's which suggest that there is no need to employ the full EM-algorithm, which is known to be very slow to converge, even to a local optimum.

5 Data and Results I: Calibration of Earlier Results using a Probit Class of Models

To illustrate the ideas described in this paper, we first re-analyze the data in Grether (1980) and El-Gamal and Grether (1995). The reason for re-analyzing this data for the third time is two-fold. First, the analysis in Grether (1980) was carried out with a simple Logit model, whereas our analysis in El-Gamal and Grether (1995) was carried out using a class of discrete models and error structure devised specifically for the experiments. Whereas customizing a class of models for each application is feasible (and often desirable), we recognize that most econometric studies use existing estimation techniques and transform their problem to fit those techniques. We shall analyze the data in this paper using a class of Probit models with unknown heterogeneity to illustrate the usefulness and feasibility of the EC-EM algorithm in uncovering arbitrary types of heterogeneity within any class of likelihood functions. The second reason for re-analyzing this data set is to compare the qualitative features of the grouping of subjects discovered by the algorithm using the class of Probits and that of El-Gamal and Grether (1995) using the discrete class of models, thus investigating the robustness of our results to the parametrization of the behavioral models.

The data of Grether (1980) was collected to study three specific hypotheses about how individuals update probability assessments given evidence. The experimental design used in that paper (for details see Grether (1980) or El-Gamal and Grether (1995)) constructed a probability updating environment by inducing prior probabilities on two urns, making repeated sampling with replacement from one of the urns, and then asking the subjects to state which urn they believed was the more likely source for that empirical sample. A simple Probit model of the following type captures many of the candidate decision rules:

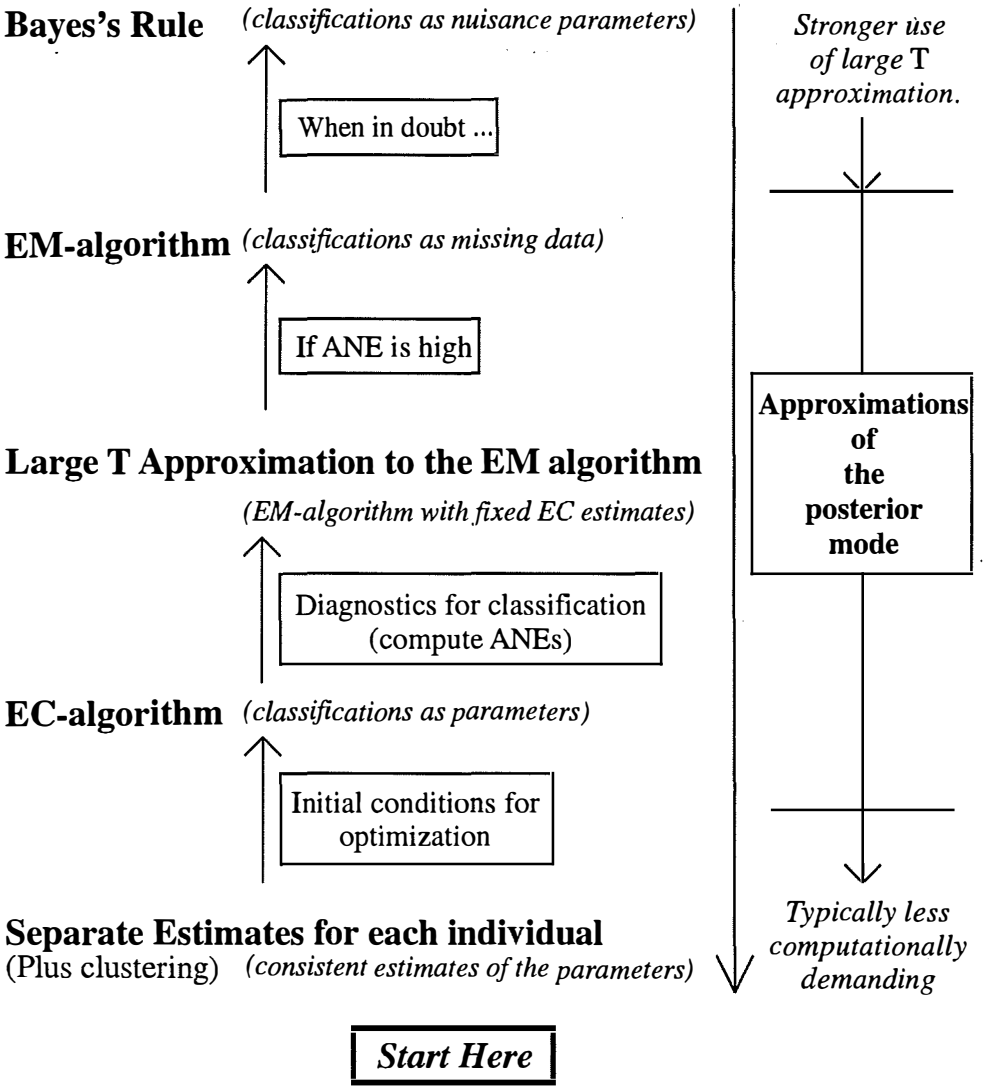
$$y^{\star}_{it} = \beta_0 + \beta_1 \log(LR_{it}) + \beta_2 \log(PO_{it}) + \epsilon_{it},$$

$$y_{it} = \begin{cases} 1 & \text{if } y^{\star}_{it} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

$$\epsilon_{it} \text{ i.i.d. } N(0,1),$$

where LR_{it} is the log likelihood ratio between the two urns, and PO_{it} is the log prior odds induced for subject i in task t.

Road Map For Large T Approximations



Start Here

If we interpret the unobservable $y^\star$ as a proxy for the log posterior odds in a subject's mind, then the three hypotheses of interest (with added noise) are as follows:

- Bayes's rule: $\beta_1 = \beta_2$.
- The representativeness heuristic (Kahneman and Tversky (1972)): $\beta_1 > \beta_2$ (over-weighting the evidence).
- Conservatism (Edwards (1982)): $\beta_1 < \beta_2$ (over-weighting the prior).

The results in Grether (1980) (using a Logit specification without allowing for subject heterogeneity) were a rejection of $\beta_1 = \beta_2$ in favor of $\beta_1 > \beta_2$, i.e. finding strong evidence for the representativeness heuristic hypothesis. In El-Gamal and Grether (1995), using our EC-EM algorithm with a discrete class of models, we found evidence for all three hypotheses. The dominant hypothesis (in terms of the largest number of subjects being classified to that model of behavior) was Bayes's rule, followed by the representativeness heuristic. There was also evidence for conservatism as the third most prominent rule, but the gain in likelihood from adding a third rule was not sufficient to warrant it.

The discrete class of models of El-Gamal and Grether (1995) had its advantages and disadvantages compared to the class of Probit models used in this paper. The Probit class of models allows for a more general error structure: In the discrete class of models, the probability of making an error was modeled as the probability of trembles independent of prior information and evidence. In the Probit class of models, the prior information and data together with the β's determine the strength with which an agent of type β and facing that prior information and evidence, should choose urn A in the absence of errors. In order to make an error, the draw ϵ_{it} should be sufficiently large to overcome the prior information and data. Therefore, fixing the variance of ϵ_{it}, by scaling up the β's the probability of "making an error" in judgement gets smaller; also the stronger is the evidence from the combination of prior and data, the lower the probability of making an error in judgement.

A second advantage of the Probit class of models is the existence of a continuum of possible degrees of representativeness and conservatism. Also, since our EC procedure produces a block-diagonal covariance matrix for the coefficients of the collection of Probits we estimate, testing within rules and across rules is very straight-forward. Finally, the Probit class of models allows for two rules which are identical except for a lower probability of making an error (e.g. $\beta_1^1 = 2\beta_1^2$, and $\beta_2^1 = 2\beta_2^2$). On the other hand, the Probit class of models suffers some shortcomings when compared to the discrete class of El-Gamal and Grether (1995). First, by lacking the discreteness of the latter, it is impossible to guarantee achieving a global maximum for the likelihood function. In fact, we generically have as many local maxima as there are ways to cluster n individuals into k groups. Second, each optimization of the likelihood function with k-Probits becomes

very costly due to the existence of many local optima and the need to employ "global optimization" routines. This makes the use of the EM-algorithm prohibitively expensive even as a diagnostic and small T bias correction mechanism as suggested in El-Gamal and Grether (1995). In this paper, we shall use a large T approximation to the EM-algorithm as a diagnostic but not as a small sample bias correction.

In Tables I.1-I.7 of Appendix I, we report the results of applying our EC algorithm to the class of Probit models. The first column of each table shows the number of rules k being estimated in each row of that table. The second column reports the number of individuals classified to each of the k rules. The next three columns report the coefficient estimates for β_0, β_1 and β_2 for that rule, with standard errors in parentheses below each coefficient. The next column reports a t value for the test of the Bayes rule hypothesis: $\beta_1 - \beta_2 = 0$ (a large positive t rejects in favor of representativeness, a large negative t rejects in favor of conservatism). The next two columns report the actual value of maximized log likelihood followed by the value of our information criterion (the value is written in bold face for the k chosen by the algorithm) with imposed priors $\beta_0 \sim N(0, 1/16)$, $\beta_1 \sim N(1, 1/16)$, $\beta_2 \sim N(1, 1/16)$, $\mathrm{P}(k = l) = 1/2^l$, and all classifications having equal prior probability.[6] By inspecting the 7 tables, we find the same qualitative results we found in El-Gamal and Grether (1995) dominating most tables.[7] In particular:

- In most tables, the information criterion dictates two rules which the t-tests identify as Bayes's rule and the representativeness heuristic.
- Conservatism is the third most prominent rule discovered by the EC algorithm, but the information criterion does not allow us to include it.

The last column in Tables 1-7 reports the Average Normalized Entropy numbers (ANE's) described in section 4. For reference, the highest ANE in Appendix I's Tables 1-7 is 0.4122, which is smaller than the ANE for $k = 2$, and $p = \{0.915, 0.085\}$, and the ANE of 0.225 for all schools pooled is smaller than the ANE for $p = \{0.96, 0.04\}$. It is clear from Tables 1-7 that the ANE is small, suggesting that our large T approximation is quite good, and allowing us to save the great expenses required to run the EM-algorithm

[6]The number of classifications is equal to $k^n/k! + h(k, n)$ where $h(1, n) = h(2, n) = 0$, $h(3, n) = 1/2$, $h(4, n) = -2^{n-2} + 1/3$. More generally, $h(k, n)$ can be directly calculated from the Stirling numbers of the scond kind, for details see Cohen (1978), pp. 118-134. For the values of k and n in this paper, the $k^n/k!$ term dominates $h(k, n)$ by many orders of magnitude, and the results in all the tables are unchanged up to many decimal places if we ignore the latter term.

[7]The estimates reported in this paper were not obtained following our suggestion of clustering individual point estimates first and then running the EC algorithm. The estimates reported here were produced through a combination of randomization over initial conditions and occasional randomization in the middle of a deterministic non-gradient based optimization algorithm. For that purpose, we used a version of the *amoeba()* routine in Press et al. (1988) adjusted to allow for randomization to reduce the probability of being trapped in a local maximum. The best results (in terms of log likelihood) from a large number of *greedy* searches with initial and intermittent randomization are reported in these tables.

or full Bayes procedure. In Table 8, we report likelihood ratio tests for the constrained hypotheses of homogeneity across schools and across payment schemes and strongly reject the homogeneity hypotheses.

6 Data and Results II: Multiple Updating Experiments and a Tobit class of Models

The second data set we analyze in this paper was collected in three schools: UCLA, California State University - Northridge, and Pasadena City College. The design of this second set of experiments differs from the first set that we analyzed above in two main aspects: (i) subjects are faced with multiple batches of draws from one of two urns, and the BDM method (Becker et al. (1964)) is used to elicit posterior beliefs on the urns after each batch of data, thus allowing the possibility of uncovering recency effects in the updating rules used by subjects; and (ii) the populations in the two urns are chosen so that the outcomes of the draws can never exactly mimic one of the parent populations, thus reducing the availability of the representativeness heuristic as a simple tool for the subjects. In this set of experiments, the two urns each had 10 balls, and priors were induced as probability $x/10$ of the first urn being used to draw batches of data. Observations were then drawn in batches of 4 draws with replacement. A variety of priors were induced for different tasks, and a variety of compositions for the first urn were chosen with the second urn always having the opposite composition (i.e. if urn A has composition 3 and 7 of the two types of balls, then urn B has compsition 7 and 3, respectively). For more details on this experimental data, see Grether (1992) where they were first described and analyzed under the assumption of subject homogeneity.

Posterior beliefs were elicited from subjects on the unit interval by asking them to choose a point in the [0,1] interval on a grid of 0.01's. The model we wish to use to analyze the data for each individual's responses again includes Bayes's law as the notable special case:

$$\log(PostO_{it}) = \beta_0 + \beta_3 \log(PO_{it}) + \beta_2 \log(LR_{it}) + \beta_3 \log(LR_{it-1}) + \epsilon_{it},$$

with $\epsilon_{it} \sim$ i.i.d. $N(0, \sigma^2)$. This model forces us to truncate observations of posteriors away from 0 and 1 (observations of 0's and 1's exist in the sample) in order for the left hand side variable to always be defined. Since the nearest observations to the extreme points were at 0.01 and 0.99, we opted to truncate the posteriors at 0.005 and 0.995, and estimate this model using Tobits. In this case, Bayes's rule would set $\beta_0 = 0$, and $\beta_1 = \beta_2 = \beta_3$; a larger β_1 would indicate conservatism, and a smaller β_1 would indicate representativeness; and $\beta_2 > \beta_3$ would indicate a recency effect. We again allow for multiple types in the population, and a type in this case would be defined by a quintuple $(\beta_0, \beta_1, \beta_2, \beta_3, \sigma)$, with σ measuring the degree of idiosyncrasy in the determination of subjects' posterior beliefs.

We must note at this point that, unlike the Probits case where the β parameters were only identified up to the scaling factor σ, the Tobit class of models analyzed in this section allows us to identify all parameters. In particular, we can compare the values of $(\beta_1, \beta_2, \beta_3)$ to their theoretical values of $(1, 1, 1)$ under the Bayesian model. Parameters greater than unity would indicate over-reaction, and parameters less than unity would indicate insensitivity to prior or observational evidence. Many of the estimated types of subjects that we characterize as "conservative" are not just weighting prior evidence more than observational evidence, they (e.g. group 1 in UCLA, group 2 in PCC, and group 3 in the pooled sample) give excessive weight to prior evidence in an absolute sense (in comparison to Bayes's rule). In contrast, group 6 in the pooled sample exhibit extreme recency effects (β_3 not significantly different from zero, and β_2 much larger), coupled with extreme sensitivity to the latest observational evidence ($\beta_2 > 1$).

The tables in Appendix B report our estimates of the Tobit models described above. The data was analyzed using a GAUSS program, utilizing its MAXLIK4.0 likelihood maximization subroutine, and initializing the likelihood maximization search with 100 random initial conditions. We then selected the best local optimum to which the algorithm converged from the random initial conditions as our best estimate of the global maximum of the Tobit likelihood function. The most striking features in Tables II.1–II.4 is the strong evidence of recency effects in the data, with only one of the estimated decision rules in Cal. State Northridge deviating from the recency effects pattern.

7 Concluding Remarks

It has long been argued in the Psychology literature that subjects use different problem representations and decision rules for different experiments (cf. Wagenaar et al. (1988)). Such "framing effects" have also been acknowledged in many experimental results in Economics (cf. Tversky and Kahneman (1986)). In the two applications of the EC-EM algorithm in this paper, we see that when the representativeness heuristic was easily available in the first set of experiments, it was the second most dominant decision rule picked by our EC algorithm, whereas in the second set of experiments designed to make that heuristic more difficult to use, conservatism was the dominant decision rule. This raises a number of interesting questions about the mapping from experimental design / problem solving situation to the context effects on decision rules. For instance, would subjects who have been exposed to problems where they can use the representativeness heuristic easily continue to use it when its use is made more difficult through a change in the urn compositions? Would they revert to using Bayes's rule if it is equally easy or easier from a computational standpoint? Indeed, despite the appearance of studying the dynamics of learning in both applications in this paper, our study is - in another sense - quite static, since we assume that decision rules are "hard-wired" for the subjects, and our investigation limits itself to uncovering those decision rules used by the subjects in

our sample. The richer and more dynamic question regarding the dynamics of learning in probabilistic settings would investigate the mapping from experimental design (or, more generally, problem solving context) to the "choice" of decision rules based on computational ease, etc. This seems to be a fruitful direction for future research, and we plan to follow that line of reasoning in future experiments.

Appendix I: Probit Results

Table I.1: UCLA								
k	N	Constant	Log(LR)	Log(PO)	t($\beta_1 - \beta_2 = 0$)	Log Lik.	Info. Cr.	ANE
1	97	0.0650	1.2453	1.3206	-1.01	-751.67	**-759.30**	0.0000
		(0.0383)	(0.0530)	(0.0613)				
2	50	0.0391	1.2304	0.7578	4.41	-696.83	-778.63	0.496
		(0.0503)	(0.0661)	(0.0795)				
	47	0.1021	1.0688	2.0081	-7.86			
		(0.0654)	(0.0977)	(0.1263)				
3	19	0.1879	0.5401	0.4175	0.79	-657.13	-784.71	0.3098
		(0.0729)	(0.0932)	(0.1153)				
	25	-0.0387	2.3541	1.2730	3.68			
		(0.1084)	(0.3759)	(0.1895)				
	53	0.2943	1.1760	1.9090	-6.13			
		(0.0632)	(0.1022)	(0.1169)				

Table I.2: PCC								
k	N	Constant	Log(LR)	Log(PO)	t($\beta_1 - \beta_2 = 0$)	Log Lik.	Info. Cr.	ANE
1	67	-0.0725	1.1057	0.5106	5.92	-442.37	-449.99	0.0000
		(0.0486)	(0.0533)	(0.0793)				
2	45	-0.0225	1.9566	0.6690	7.90	-384.35	**-445.37**	0.4122
		(0.0745)	(0.1682)	(0.1290)				
	22	-0.1169	0.4459	0.5298	-0.53			
		(0.0756)	(0.0806)	(0.1260)				
3	43	-0.0045	1.9873	0.6178	7.96	-375.63	-470.17	0.3622
		(0.0773)	(0.1824)	(0.1366)				
	5	0.0615	-0.2706	1.0145	-3.67			
		(0.1606)	(0.1734)	(0.2773)				
	19	-0.0594	0.6377	0.2850	1.99			
		(0.0847)	(0.0950)	(0.1389)				

Table I.3: Occidental College								
k	N	Constant	Log(LR)	Log(PO)	t($\beta_1 - \beta_2 = 0$)	Log Lik.	Info. Cr.	ANE
1	56	0.0395	1.1080	0.8583	2.69	-451.78	-459.39	0.0000
		(0.0503)	(0.0617)	(0.0860)				
2	44	0.1418	1.8076	1.2799	4.48	-376.28	**-429.69**	0.1347
		(0.0746)	(0.1310)	(0.1433)				
	12	-0.0557	0.1343	0.2929	-0.89			
		(0.0920)	(0.1099)	(0.1546)				
3	9	-0.0025	-0.1297	0.4855	-3.06	-352.15	-434.83	0.1827
		(0.1041)	(0.1221)	(0.1749)				
	22	0.2558	1.6681	1.7944	-0.77			
		(0.1169)	(0.2064)	(0.2568)				
	25	0.0742	1.9983	0.6802	6.77			
		(0.0966)	(0.1632)	(0.1622)				

Table I.4: CSULA								
k	N	Constant	Log(LR)	Log(PO)	t($\beta_1 - \beta_2 = 0$)	Log Lik.	Info. Cr.	ANE
1	37	-0.0494	0.7470	0.5544	1.88	-385.58	-393.20	0.0000
		(0.0524)	(0.0534)	(0.0871)				
2	13	-0.0188	0.2666	0.0892	1.10	-337.20	-377.43	0.2544
		(0.0811)	(0.0820)	(0.1380)				
	24	-0.1619	1.5641	1.2292	2.55			
		(0.0782)	(0.1119)	(0.1433)				
3	10	0.0146	1.4340	2.0341	-2.58	-314.81	**-376.61**	0.2599
		(0.1448)	(0.2289)	(0.3338)				
	16	-0.0558	0.3302	0.2042	0.85			
		(0.0738)	(0.0760)	(0.1250)				
	11	-0.1861	1.9672	0.4063	4.09			
		(0.1852)	(0.3388)	(0.2158)				
4	8	-.0164	.9996	1.3219	-1.93	-302.79	-399.10	0.2316
		(0.0965)	(0.1294)	(0.1847)				
	14	-0.5114	3.1806	3.2037	-0.06			
		(0.2626)	(0.7642)	(0.8912)				
	7	-.0683	0.2291	-.1442	2.00			
		(0.0920)	(0.0949)	(0.1558)				
	8	-0.2330	2.0811	0.0528	4.61			
		(0.2400)	(0.3885)	(0.2579)				

Table I.5: All Schools - Payed acc. to outcome								
k	N	Constant	Log(LR)	Log(PO)	t($\beta_1 - \beta_2 = 0$)	Log Lik.	Info. Cr.	ANE
1	125	-0.0863	1.2768	0.9846	4.26	-954.10	**-961.71**	0.0000
		(0.0342)	(0.0436)	(0.0535)				
2	58	0.2925	1.1969	1.8332	-5.28	-864.86	-966.07	0.2973
		(0.0661)	(0.1093)	(0.1335)				
	67	-0.1525	1.1373	0.3616	8.37			
		(0.0438)	(0.0532)	(0.0696)				
3	71	-0.0972	1.4015	0.5528	7.72	-812.76	-971.24	0.1290
		(0.0498)	(0.0899)	(0.0776)				
	45	0.2744	1.1195	2.0787	-7.05			
		(0.0735)	(0.1196)	(0.1435)				
	9	0.0088	0.0710	-0.0387	0.53			
		(0.1036)	(0.1221)	(0.1654)				

Table I.6: All Schools - Payed a flat fee								
k	N	Constant	Log(LR)	Log(PO)	t($\beta_1 - \beta_2 = 0$)	Log Lik.	Info. Cr.	ANE
1	132	0.0467	1.0059	0.8614	2.41	-1128.46	-1136.07	0.0000
		(0.0303)	(0.0319)	(0.0510)				
2	88	0.1180	1.8013	1.2615	6.11	-964.19	**-1070.25**	0.2251
		(0.0491)	(0.1014)	(0.0967)				
	44	0.0320	0.3719	0.6100	-2.55			
		(0.0464)	(0.0490)	(0.0795)				
3	59	0.1796	1.6844	1.6436	0.43	-927.01	-1093.20	0.3671
		(0.0635)	(0.1091)	(0.1265)				
	40	0.0367	0.3577	0.4048	-0.47			
		(0.0488)	(0.0520)	(0.0845)				
	33	-0.0648	2.6453	1.2019	5.19			
		(0.1016)	(0.3478)	(0.1708)				

Table I.7: All Schools - Pooled								
k	N	Constant	Log(LR)	Log(PO)	t($\beta_1 - \beta_2 = 0$)	Log Lik.	Info. Cr.	ANE
1	257	0.0008	1.0723	0.9076	3.78	-2091.76	-2098.98	0.0000
		(0.0222)	(0.0250)	(0.0367)				
2	204	0.0530	1.3221	1.0482	4.84	-1886.03	-2078.72	0.2393
		(0.0294)	(0.0525)	(0.0533)				
	53	0.0353	0.2613	0.5215	-2.99			
		(0.0422)	(0.0461)	(0.0713)				
3	62	0.0147	0.4792	0.0850	4.97	-1739.86	**-2043.32**	0.2468
		(0.0388)	(0.0430)	(0.0658)				
	104	-0.0549	1.8934	0.7810	10.42			
		(0.0470)	(0.1081)	(0.0808)				
	91	0.2297	1.1815	1.7968	-7.46			
		(0.0463)	(0.0770)	(0.0932)				
4	79	0.1708	1.2336	1.5287	-3.43	-1701.11	-2084.87	0.2360
		(0.0478)	(0.0831)	(0.0924)				
	25	0.5001	0.8913	3.2054	-10.15			
		(0.1190)	(0.1561)	(0.2885)				
	92	0.0247	1.9516	0.7497	9.65			
		(0.0506)	(0.1286)	(0.0874)				
	61	0.0283	0.3685	0.3194	0.60			
		(0.0394)	(0.0440)	(0.0668)				

Table I.8: Tests of Homogeneity

No. of Rules	Across Schools			pay vs. flat fee		
	χ^2	d.f.	p-value	χ^2	d.f.	p-value
1	120.52	9	0.000	18.40	3	0.000
2	182.74	18	0.000	113.96	6	0.000
3	80.28	27	0.000	0.18	9	1.000

Appendix II: Tobit Results

Table II.1: UCLA								
k	N	Constant	Log(PO)	Log(LR)	$\log(\text{LR})_{-1}$	σ	Info. Cr.	ANE
1	27	0.4084	0.7923	0.5784	0.2818	1.6326	-913.06	0.0000
		(0.0775)	(0.1384)	(0.0439)	(0.0502)	(0.0553)		
2	10	0.6791	0.9930	0.8310	0.3897	2.3958	-834.60	0.0100
		(0.1890)	(0.3363)	(0.1077)	(0.1247)	(0.1433)		
	17	0.2596	0.6755	0.4574	0.2427	0.9867		
		(0.0589)	(0.1052)	(0.0333)	(0.0379)	(0.0402)		
3	6	0.4318	1.7151	0.6074	0.4640	1.0477	**-827.07**	0.0381
		(0.1057)	(0.1887)	(0.0599)	(0.0701)	(0.0757)		
	7	0.7349	0.3642	0.8938	0.3730	2.6819		
		(0.2532)	(0.4500)	(0.1447)	(0.1667)	(0.1942)		
	14	0.2342	0.5931	0.4389	0.1858	0.9658		
		(0.0635)	(0.1134)	(0.0360)	(0.0409)	(0.0432)		
4	14	0.2341	0.5931	0.4389	0.1858	0.9658	-839.280.0072	
		(0.0635)	(0.1134)	(0.0360)	(0.0409)	(0.0432)		
	4	0.4336	1.4622	0.5426	0.5042	0.7884		
		(0.0972)	(0.1736)	(0.0551)	(0.0646)	(0.0680)		
	4	1.6984	1.4745	0.6746	0.1159	1.9856		
		(0.2468)	(0.4419)	(0.1400)	(0.1599)	(0.1797)		
	5	-0.1994	0.2058	1.0187	0.6154	2.4797		
		(0.2797)	(0.4948)	(0.1598)	(0.1894)	(0.2185)		

Table II.2: PCC								
k	N	Constant	Log(PO)	Log(LR)	$\log(\text{LR})_{-1}$	σ	Info. Cr.	ANE
1	44	0.3440	0.7930	0.4441	0.0969	2.1731	-1770.23	0.0000
		(0.1411)	(0.2233)	(0.0396)	(0.0437)	(0.0579)		
2	28	0.3760	0.8056	0.2032	0.0551	1.4203	-1648.31	0.0802
		(0.1151)	(0.1826)	(0.0320)	(0.0359)	(0.0444)		
	16	0.3204	0.7360	1.0013	0.1655	3.1017		
		(0.3409)	(0.5373)	(0.1002)	(0.1063)	(0.1531)		
3	13	0.4412	0.4916	0.6126	0.0729	1.3510	**-1618.62**	0.0729
		(0.1598)	(0.2529)	(0.0450)	(0.0497)	(0.0625)		
	17	0.2478	0.8665	0.0957	0.0152	1.2537		
		(0.1309)	(0.2073)	(0.0362)	(0.0413)	(0.0500)		
	14	0.3549	1.0732	0.8343	0.2542	3.5283		
		(0.4150)	(0.6554)	(0.1202)	(0.1297)	(0.1919)		
4	20	0.2794	0.8210	0.1906	0.0423	1.2010	-1890.64	0.0466
		(0.1147)	(0.1818)	(0.0320)	(0.0356)	(0.0439)		
	2	2.8394	-2.1162	4.4012	2.1045	11.009		
		(5.0474)	(7.5266)	(1.7537)	(1.5953)	(3.6993)		
	10	0.2410	0.5268	0.9617	0.1331	1.5124		
		(0.2074)	(0.3268)	(0.0610)	(0.0647)	(0.0856)		
	12	0.5207	1.0760	0.2830	0.0291	2.6643		
		(0.3301)	(0.5222)	(0.0929)	(0.0994)	(0.1368)		

Table II.3: Cal-State Northridge								
k	N	Constant	Log(PO)	Log(LR)	$\log(\text{LR})_{-1}$	σ	Info. Cr.	ANE
1	20	0.1125	-0.0376	0.5566	0.2476	2.2254	**-761.73**	0.0000
		(0.1267)	(0.4287)	(0.0592)	(0.0540)	(0.0923)		
2	16	0.2234	0.3844	0.4658	0.2090	1.6108	-792.42	0.0085
		(0.1013)	(0.2709)	(0.0452)	(0.0431)	(0.0701)		
	4	-0.7753	-2.3149	1.3261	0.5148	4.5465		
		(0.6490)	(1.6138)	(0.3137)	(0.2741)	(0.5560)		

Table II.4: Pooled Tobits								
k	N	Constant	Log(PO)	Log(LR)	$\log(\text{LR})_{-1}$	σ	Info. Cr.	ANE
5	20	0.1400	0.8962	0.4206	0.4922	1.0583	-3100.88	0.0799
		(0.0576)	(0.1028)	(0.0284)	(0.0303)	(0.0403)		
	29	0.4241	0.4579	0.7168	-0.0084	1.2282		
		(0.0748)	(0.1317)	(0.0328)	(0.0438)	(0.0478)		
	25	0.8112	1.1733	0.2387	0.1167	2.4109		
		(0.1781)	(0.2927)	(0.0698)	(0.0744)	(0.1107)		
	21	0.2889	0.7946	0.1099	0.0277	1.1259		
		(0.0726)	(0.1223)	(0.0271)	(0.0292)	(0.0404)		
	16	-0.1087	0.3370	1.4669	0.4300	3.6386		
		(0.2463)	(0.4348)	(0.1223)	(0.1145)	(0.2028)		
6	9	0.2681	1.1125	0.3922	0.6525	0.9759	**-3073.18**	0.0660
		(0.0783)	(0.1495)	(0.0377)	(0.0402)	(0.0555)		
	23	0.3061	0.6816	0.5422	0.1070	1.0238		
		(0.0530)	(0.0918)	(0.0249)	(0.0264)	(0.0356)		
	20	0.2957	0.7649	0.1012	0.0205	1.1337		
		(0.0770)	(0.1300)	(0.0280)	(0.0299)	(0.0417)		
	14	0.2403	0.8268	1.0787	0.1988	1.7311		
		(0.1230)	(0.2113)	(0.0575)	(0.0557)	(0.0861)		
	14	0.9036	0.8072	0.1991	0.0287	2.3601		
		(0.1889)	(0.3207)	(0.0701)	(0.0757)	(0.1106)		
	11	-0.1659	0.1387	1.4793	0.5587	4.4393		
		(0.3814)	(0.7808)	(0.1792)	(0.1711)	(0.3167)		
7	17	0.7618	1.1708	0.1515	0.0231	1.9046	-3175.99	0.0513
		(0.1388)	(0.2364)	(0.0512)	(0.0628)	(0.0792)		
	26	0.2914	0.7311	0.5149	0.1038	1.0530		
		(0.0517)	(0.0888)	(0.0240)	(0.0254)	(0.0344)		
	12	0.2729	1.3738	0.4781	0.6021	1.1365		
		(0.0791)	(0.1496)	(0.0387)	(0.0411)	(0.0572)		
	11	-0.2041	0.5801	0.8071	0.2285	3.1809		
		(0.2558)	(0.4393)	(0.1120)	(0.1143)	(0.1840)		
	10	0.4543	0.2801	1.1770	0.1036	1.6332		
		(0.1549)	(0.2537)	(0.0673)	(0.0643)	(0.0957)		
	4	0.2866	0.9193	3.3771	1.9149	8.6346		
		(1.2067)	(2.6383)	(0.8454)	(0.6980)	(1.6559)		
	11	0.2453	0.3699	0.0675	0.0082	0.9214		
		(0.0846)	(0.1455)	(0.0306)	(0.0348)	(0.0454)		

References

Akaike, H. 1974. A new look at the statistical idenitification model. *IEEE Transactions on Automatic Control* 19:716–723.

Amemiya, T. 1985. *Advanced econometrics.* Cambridge, MA: Harvard University Press.

Becker, G., M. DeGroot, and J. Marschak. 1964. Measuring utility by a single response sequential method. *Behavioral Science* 9:226–232.

Cohen, D. 1978. *Basic techniques of combinatorial theory.* New York: John Wiley and Sons.

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

Edwards, W. 1982. Conservatism in human information processing. In Kahneman, D., P. Slovic, and A. Tversky, eds., *Judgment Under Uncertainty: Heuristic and Biases*, pages 359–369. Cambridge University Press, Cambridge.

El-Gamal, M. and D. Grether. 1995. Are people Bayesian? uncovering behavioral strategies. *Journal of the American Statistical Association* 90 (432):1137–1145.

Grether, D. 1980. Bayes rule as a descriptive model: The representativeness heuristic. *Quarterly Journal of Economics* pages 537–557.

Grether, D. 1992. Testing bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior and Organization* 17(2):31–58.

Heckman, J. and B. Singer. 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52:271–320.

Honoré, B. 1992. Timmed lad and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica* 60:533–565.

Honoré, B. 1993. Orthogonality conditions for tobit models with fixed effects and lagged dependent variables. *Journal of Econometrics* 59:35–61.

Jeffreys, H. 1939. *Theory of probability.* London: Oxford University Press.

Kahneman, D. and A. Tversky. 1972. Subjective probability: A judgment of representativeness. *Cognitive Psychology* 3:430–454.

Kohonen, T. 1990. The self-organizing map. *Proceedings of the IEEE* 78, No. 9:1464–1480.

Lindsay, B. 1981. Properties of the maximum likelihood estimator of a mixing distribution. In Tallie, C., ed., *Statistical Distributions in Scientific Work*, pages 95–105. Kluwer Academic Publishers, Hingham Mass.

Lindsay, B. 1983a. The geometry of mixture likelihoods, part i. *Annals of Statistics* 11:86–94.

Lindsay, B. 1983b. The geometry of mixture likelihoods, part i. *Annals of Statistics* 11:783–792.

Little, R. and D. Rubin. 1983. On jointly estimating parameters and missing data. *The American Statistican* 37:218–220.

Little, R. and D. Rubin. 1987. *Statistical analysis with missing data*. New York: Wiley.

McLachlan, G. and K. Basford. 1988. *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker, Inc.

Pesaran, H., R. Smith, and K. Im. 1996. Dynamic linear models for heterogeneous panels. In Matyas, L. and P. Sevestre, eds., *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, pages 145–195. Kluwer Academic Publishers, Boston.

Pollard, D. 1982. Quantization and the method of k-means. *IEEE Transactions on Information Theory* IT-28, No. 2:199–205.

Press, W., B. Flannery, S. Teukolsky, and W. Vetterling. 1988. *Numerical recipes in C: The art of scientific computing*. Cambridge: Cambridge University Press.

Redner, R. and H. Walker. 1984. Mixture densities, maximum likelihood and the EM algorithm. *Siam Review* 26, no.2:195–239.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica* 14:465–471.

Rissanen, J. 1987. Stochastic complexity. *Journal of the Royal Statistical Society B* 49(3):223–239 and 252–265.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6,2:461–464.

Tversky, A. and D. Kahneman. 1986. Rational choice and the framing of decisions. *Journal of Business* 59(4):S251–S278.

Wagenaar, W., G. Keren, and S. Lichtenstein. 1988. Islanders and hostages: Deep and surface structure of decision problems. *Acta Psychologica* 67:175–189.

Wallace, C. and D. Boulton. 1968. An information measure for classification. *computer Journal* 11:185–195.

Wallace, C. and P. Freeman. 1987. Estimation and inference by compact coding. *Journal of the Royal Statistical Society B* 49(3):240–265.