Optimal Categorization*

Erik Mohlin†

April 15, 2011.

Abstract

This paper provides a model of categorizations that are optimal for the purpose of making predictions. In the beginning of each period a subject observes a two-dimensional object in one dimension and wants to predict the object's value in the other dimension. The subject uses a categorization that partitions the space of objects into categories. She has a data base of objects that were observed in both dimensions in the past. The subject determines what category the new object belongs to on the basis of observation of its first dimension. She predicts that its value in the second dimension will be equal to the average value among the past observations in the corresponding category. At the end of each period the second dimension is observed. The optimal categorization minimizes the expected prediction error. The main result is that the optimal number of categories is determined by a trade-off between (a) decreasing the size of categories in order to enhance category homogeneity, and (b) increasing the size of categories in order to enhance category sample size.

Keywords: Categorization; Priors; Prediction; Similarity-Based Reasoning.
JEL codes: C72.

*This paper has benefited from comments by Ola Andersson, Stefano Demichelis, Tore Ellingsen, Drew Fudenberg, Philippe Jehiel, Topi Miettinen, Robert Östling, Rani Spiegler, Tomasz Strzalecki, and Jörgen Weibull, as well as participants at presentations at the Third Nordic Workshop in Behavioral and Experimental Economics in Copenhagen, November, 2008, SUDSWEc in Uppsala, May 2009 and the Stockholm School of Economics. Financial support from the Jan Wallander and Tom Hedelius Foundation is gratefully acknowledged.

†E-mail: e.mohlin@ucl.ac.uk. Mail: Department of Economics, University College London, Gower Street, London WC1E 6BT, United Kingdom.

1 Introduction

Numerous psychological studies have demonstrated the importance of categorical reasoning for human cognition in general.[1] Categorical thinking also matters in many economic contexts: Consumers categorize goods and services when deciding what to purchase, and this leads to segmentation of markets (Smith (1965)). Firms may respond with marketing strategies that take advantage of the consumers' categorizations (Punj and Moon (2002)).In financial markets, investors engage in "style investing", the practice of allocating funds among classes of assets rather than to individual assets (Bernstein (1995)). Rating agencies categorize firms in order to reflect the probability that a firm will default on its debt, an activity whose importance was highlighted by the recent financial crisis (Coval et al. (2009)).

In the psychological literature it is widely acknowledged that an important function of categories is to facilitate predictions (Anderson (1990)). Prediction on the basis of categorical reasoning is relevant in situations where one has to predict the value of a variable on the basis of one's previous experience with similar situations, but where the past experience does not necessarily include any situation that is identical to the present situation. One may then divide the experienced situations into categories, such that situations in the same category are similar to each other. When a new situation is encountered one determines what category this situation belongs to, and the past experiences in this category are used to make a prediction about the current situation. These predictions can be computed in advance, thereby facilitating a fast response.

In this paper I ask which categorizations are optimal in the sense that they minimize prediction error – a notion that is made precise below. In particular, I study the optimal number of categories without imposing any exogenous costs and benefits of the number of categories. Instead both costs and benefits are derived endogenously from the objective of making accurate predictions. The advantage of fine grained categorizations is that objects in a category are similar to each other. The advantage of coarse categorizations is that a prediction about a category is based on a large number of observations. Comparative statics reveal how the optimal categorization depends on the number of observations, as well as on the frequency of objects with different properties. So far there are only a few explicit models of categorization in economics. The question of optimality has rarely been discussed, and those who have done so, e.g. Fryer and Jackson (2008), assume an exogenous number of categories. The relevant literature is discussed in sections 4.3 and 5.

The focus on optimal categorizations is based on evolutionary considerations. Many categorizations are acquired early in life, through socialization and education, or because they are innate. From an evolutionary perspective we would expect humans to employ categorizations that generate predictions that induce behavior that maximize fitness. It

[1] For overviews of the voluminous literature see e.g. Laurence and Margolis (1999), or Murphy (2002).

seems reasonable to assume that fitness is generally increasing in how accurate the predictions are. For instance, a subject encountering a poisonous plant will presumably be better off if she predicts that the plant is indeed poisonous, rather than nutritious. For this reason we would expect that humans have developed, and passed on, categorizations that are at least approximately optimal, in the sense that they tend to minimize prediction error in the relevant environments. Such a categorization will be called *ex ante optimal*. In a similar way one would expect that, over time, a profession has developed categorizations that are ex ante optimal for the kind of data that members of that profession tend to encounter. Other categorizations are developed only after experience has been accumulated – e.g. for some area of investigation where one did not have useful concepts before. In this case we would expect evolution to have endowed us with heuristics or algorithms that allow us to form categorizations that organize our experience in way that tends to minimize prediction error. Categorizations that attain this goal will be called *ex post optimal*.

The model is centered on a subject who lives for a certain number of periods. First she goes through a learning phase and then a prediction phase. In each period of the *learning phase* she observes an object, represented by a vector (x, y). All objects are independently drawn from the same distribution, and are stored in a data base. A categorization is a set of categories which together partition the set of objects. Each object's category membership is determined by its x-value. In the beginning of each period of the *prediction phase* the subject encounters a new object and observes the x-value but not the y-value. The y-value has to be predicted with the help of the object's x-value and the data base of past experiences. The new object is put in one of the categories on the basis of its x-value. The empirical mean y-value, of the previously experienced objects in that category, serves as prediction for the y-value of the new object. At the end of the period, after the prediction has been made, the y-value is revealed and the information is added to the data base. In the case of categorizations that are acquired prior to accumulating a data base, the model assumes that the subject is endowed with a categorization at the beginning of the learning phase, and this categorization is kept fixed for the subject's whole life time. In the case of categorizations that are formed after a data base has been accumulated, a categorization is assumed to be formed in the prediction phase, and used for one prediction only.[2]

It has been discussed whether categorization presupposes a notion of similarity or not (see Goldstone (1994) and Gärdenfors (2000)). The model presented in this paper is

[2]The model reflects to findings regarding predictions based on categorization: First, predictions about a particular category are generally formed only on the basis of objects that were put into that category in the past, not on the basis of objects that were put into other categories (Malt et al. (1995) and Murphy and Ross (1994)). Second, a prediction about a particular object is generally based only on what category the object belongs to, and does not take into account within-category correlations between properties. This means that roughly the same prediction is made for all objects in the same category (Krueger and Clement (1994)).

neutral in this respect. The x-dimension may, but need not, be endowed with a metric. The y-dimension is taken to be the real line and prediction error is measured as the squared difference between the prediction and the actual y-value of the object. Using the probability density function over the set of objects one can define the (unconditional) *expected prediction error* of a categorization. In this case expectation is taken over the set of data bases that the subject may encounter in the future. One can also define the *expected prediction error conditional on a given data base*. In this case expectation is taken only over the next observation. The unconditional expected prediction error is minimized by an *ex ante optimal categorization*, i.e. a categorization that is optimal prior to a data base has been accumulated. The expected prediction error conditional on a given data base is minimized by the *ex post optimal categorization*, i.e. a categorization that is optimal for predicting next observation, given the current data base.

Note that the set-up does not presume the existence of any natural kinds, in the sense of Quine (1969). There does not have to exist an objectively true categorization "out there". The optimal categorization is a framework we impose on our environment in order to predict it. The fact that we use categorizations is taken as given in this paper, though section 4.4 uses the model to discuss why this might be the case.

As an example of a categorization that is acquired very early on, think of color concepts. The subset of the spectrum of electromagnetic radiation that is visible to the human eye allows for infinitely fine grained distinctions. More precisely all colors can be described as points in the space of the three dimensions hue, saturation and lightness. However, in every day reasoning and discourse we seem to employ only a coarse color classification, using words such as red and green. Presumably the color categorizations that were developed and passed on to new generations were successful in the kind of environments that we faced.[3] As an example of categorizations that are formed after a data base has been accumulated, one may think of the many categorizations and classifications that science has produced.

The two kinds of categorization are probably often combined. Think of a physician who encounters a new patient in each period. The x-value could represent information about a patient's personal characteristics such as weight, blood-pressure, or aspects of the patient's medical history. The y-value could represent some dimension of the patient's future health. First the physician goes to medical school and learns a set of categories while observing various patients' characteristics together with their subsequent health state. Later she works in a hospital: In the beginning of each period she receives information about a patient's personal characteristics, and has to make a prediction about some aspect of the patient's health. In order to make such a prediction she assigns the new patient

[3] We could have sliced up the space of colors differently. Indeed there are cultures where even the basic color categories are different from the ones used by speakers of the English language. Still, behind the variation there seems to be certain principles of classification that are stable between cultures (see Kay and Maffi (1999) and references therein). The model presented in this paper allows for the existence of multiple optimal categorizations.

to a category and predicts that the outcome for this patient will be like the empirical average outcome among previous patients in that category. At the end of each period she can observes the outcome for the current patient. Eventually she might have accumulated sufficiently many observations to motivate the development of a refined categorization on her own.

The main result of this paper is that optimal (both the ex ante and ex post optimal) number of categories is determined by a trade-off between the value of within-category similarity of objects and the value of having many stored observations in each category. Increasing the number of categories has two effects. (a) The average size of each category decreases and thus the differences between objects that belong to the same category will be relatively small. (b) The average number of experienced objects in each category decreases. Thus generalizations about a category are based on a smaller sample, making inferences from observed objects to future cases less reliable. Note that this trade-off does *not* depend on any exogenous cost of categories. The trade-off sheds light on the phenomenon of basic-level categories, which has received much attention from psychologists; the most salient level of categorization is neither the most fine-grained, nor the most general level of categorization (Rosch et al. (1976)). The dominant view within psychology is that the number of categories (the coarseness of the categorization) is determined by another trade-off. Like in this paper, the benefit of small categories is supposed to be within-category homogeneity of objects. But, unlike this paper, the benefit of having a fewer larger categories is supposed to be that one needs to observe fewer properties of an object in order to categorize it as belonging to a large category (Medin (1983)). The model also predicts that experts will have a more fine grained conceptual structure than laymen (Tanaka and Taylor (1991)). Furthermore, comparative statics with respect to the distribution of objects with different properties show that (i) the larger the variability in the y-dimension, the larger is the optimal number of categories, and (ii) the more frequent objects in one subset of the x-dimension are, the larger is the optimal number of categories in that subset. In particular, assuming that the relationship between x- and y-values is given by a linear regression model, the optimal number of categories is decreasing in the variance of the error term and increasing in the slope of the regression line. The model can be extended in various ways. The set Y may be multidimensional, and different subject may then weigh the different dimensions differently. A subject's cost of prediction errors may vary with x. The model also has implications for the case of a fixed number of categories.

It should be emphasized that the inference, from properties of objects in the data base, to the unobserved property of the present object, is *not* Bayesian. In particular, the subject does not have a prior about an object's properties before it is categorized. On the contrary, the model of this paper is intended to shed some light on how priors are generated. When an object is categorized, the data base is used to form a point prediction about the new object, in a non-Bayesian, frequentist way. This point prediction should be interpreted as a prior (point) belief. Binmore (2007) and Gilboa et al. (2008) have argued for the need to complement Bayesian decision theory with a theory of belief formation that accounts for how priors are formed.

Categories are closely related to concepts. Categories can be said to be defined by concepts in the sense that an object belongs to a category if an only if it falls under the corresponding concept. Conversely, categorization is one of the most important functions of concepts. One might suggest that we use categories because language is categorical and say that a categorization is optimal if it is induced by a language that is optimal in some sense. Language is undoubtedly important in shaping our concepts and categories, but concepts seem to have come prior to language in evolution – there are animals that use concepts even though they do not use language – and children can use certain concepts before they have a language.[4] Therefore I suggest that we try to explain the use of categories without reference to language. In addition to this, a language usually allows many different categorizations of the same subject matter. Therefore the optimal categorization is under-determined by the demands of communication.

The rest of the paper is organized as follows. Section 2 describes the model and defines prediction error and optimality. The results are developed in Section 3 presents the results, regarding ex ante and ex post optimality, and discusses extensions. Section 4 discusses the results and applications. Related literature is reviewed in section 5, and section 6 concludes. All proofs are in the appendix, section 7.

2 Model

2.1 Subject and Objects

A subject lives for T periods; first a learning phase of $L < T$ periods, and then a prediction phase of $T - L$ periods. In each period $t \in \{1, ..., T\}$ she encounters an object, which is represented by a point $v_t = (x_t, y_t)$ in a two-dimensional space $V = X \times Y$, where $Y = \mathbb{R}$. The set X may be a closed interval $[a, b] \subseteq \mathbb{R}$, or some arbitrary finite set. Hence, the set X need not be endowed with a metric, thus allowing for categorizations not based on similarity, but in this case X is assumed to be finite in order to assure the existence of a solution.

All objects are drawn independently according to a continuous probability density function $f : V \to [0, 1]$, satisfying $f(v) > 0$ for all $v \in V$. In order to abstract from trivialities I will assume that if $X = [a, b] \subseteq \mathbb{R}$ then $\mathbb{E}[y|x] \neq \mathbb{E}[y|x']$ for some $x, x' \in X$, and if X is finite then $\mathbb{E}[y|x] \neq \mathbb{E}[y|x']$ for all $x, x' \in X$, $x \neq x'$.

[4] Regarding animals there is evidence that pigeons have concepts, at least in a way that enables them to categorize objects (Herrnstein et al. (1976)). There are also studies indicating that rhesus monkeys (Hauser et al. (1997)) have simple numerical concepts. Regarding children Franklin et al. (2005) provides evidence that toddlers have a pre-linguistic understanding of color concepts.

Experienced objects are stored in a data base, so at the beginning of any period $t > 1$ the subject has a data base $v^{t-1} = (v_1, ..., v_{t-1}) \in V^{t-1}$. In each period $t \in \{1, ..., L\}$ of the learning phase the subject observes each object in both dimensions. In the beginning of each period $t \in \{L+1, ..., T\}$ of the prediction phase she observes the x-value, x_t, of an object v_t, and not its y-value, y_t. She makes a prediction about y_t on the basis of x_t, and the data base v^{t-1}. At the end of the period uncertainty is resolved; the subject observes y_t, and updates the data base. Thus learning does not only occur in the learning phase but continues through the whole life time.

2.2 Categories

A category C_i is a subset of V. A categorization is a finite set of categories $C = \{C_1, ..., C_k\}$ that constitutes a partitioning of V. Let X_i be the projection of C_i onto X. Since the category membership of an object only depends on the object's x-value, the collection of sets $\{X_1, ..., X_k\}$ form a partitioning of X, and we can write $C_i = X_i \times Y$. Each set X_i is assumed to be the union of finitely many intervals.[5] The relative size of categories is constrained by some (small) number $\rho \in (0, 1)$ such that $\Pr(x \in X_i) / \Pr(x \in X_j) > \rho$ for all i and j. For the case of a finite number of categories this implies that all categories have positive probability. When the number of categories goes to infinity (requiring that X is infinite) the assumption implies that no category becomes relatively infinitely larger than another category. Furthermore, the number of categories per objects is bounded by some arbitrarily large but finite κ, i.e. $k/T < \kappa$. This assumption is only made in order to assure existence of a solution when T is small. When T is sufficiently large, existence can be proved without this assumption. The set of categorizations satisfying these assumptions, the feasible categorizations, is denoted Ψ. When studying optimality of categorizations that are acquired prior to a data base all categorizations in Ψ will be considered. When studying optimality of categorizations of a given data base v^{t-1}, I will restrict attention to categorizations in which all categories are non-empty under v^{t-1}. Let $\Psi(v^{t-1})$ denote this set of feasible categorizations.

It might seem problematic to assume that categories in the same categorization are mutually exclusive, since we have many categories that are not mutually exclusive. This is the case for hierarchically organized concepts such as the two categories of stone and granite. However, we generally do not use such overlapping categories for the same prediction tasks. If I am interested in whether an object will burn when thrown on the fire I might categorize the object as made of stone rather than wood, and infer that it will not burn. In this context it is useless to know whether the object is of granite or not. But if I

[5] If categories are only composed of one interval the categories are required to be convex. Gärdenfors (2000) argues that we should expect the extension of natural concepts to be convex on the grounds that convex concepts are easier to learn than non-convex concepts. Still, I will work with the more general assumption that the categories are the union of finitely many intervals.

want to build a house it may be useful to employ a narrower categorization of materials, since granite is more solid than e.g. limestone.

2.3 Prediction

For each category $C_i \in C$, and for date t, the subject has a prediction $\hat{y}_{it}$ about the y-value of objects in that category. As discussed above, it will be assumed that the prediction equals the mean of all previously experienced objects in that category. Let

$$D_{it} = \{s \in \mathbb{N} : s < t \wedge v_s \in C_i\}.$$

This is the set of dates, prior to date t, at which objects in category C_i were observed. Let $m_{it} = |D_{it}|$, so that $\sum_{i=1}^{k} m_{it} = t - 1$, for all t. Thus at date $t > L$ the prediction for category i is

$$\hat{y}_{it} = \begin{cases} \frac{1}{m_{it}} \sum_{s \in D_{it}} y_s & \text{if } m_{it} > 0 \\ \hat{y}_t & \text{if } m_{it} = 0 \end{cases}, \tag{1}$$

where

$$\hat{y}_t = \frac{1}{t-1} \sum_{s=1}^{t-1} y_s. \tag{2}$$

This definition says that if the data base does not contain any objects in the category that object v_t belongs to, then the prediction for this object is made on the basis of all objects currently in the data base. This seems like a natural assumption, but there are alternatives: For instance one could assume that there is some fixed prediction error associated with empty categories, and this would not affect the results of the paper. Alternatively one could modify the model and assume that the subject is endowed with at least one object in each category, in period 1. Again, all the results will go through under this alternative assumption.

2.4 Prediction Error and Optimality

For any object v_t that the subject may encounter at date t, there is a unique category C_i such that $v_t \in C_i$. For any data base $v^{t-1} \in V^{t-1}$ that the subject may have at date t the prediction y_{it} is then determined according to (1) and (2). Given a categorization C and a data base v^{t-1}, the *prediction error* associated with a new object $v_t \in C_i$ is defined as the squared Euclidean distance between the predicted value $\hat{y}_{it}$ and the true value y_t, i.e.

$$PE\left(C, v_t, v^{t-1}\right) = \left(y_t - \hat{y}_{it}\right)^2. \tag{3}$$

At the beginning of period t the data base v^{t-1} has been accumulated, but object v_t has not been observed yet. Given a categorization C, one might ask what the expected prediction error associated with v_t is. The answer is given by taking expectation over all objects $v_t \in V$. That is, *conditional on a data base* v^{t-1} the *expected prediction error* of categorization C at date t is

$$EPE\left(C, v^{t-1}\right) = \mathbb{E}\left[PE\left(C, v_t, v^{t-1}\right) | v^{t-1}\right]. \tag{4}$$

Furthermore, by taking expectation also over data bases $v^{t-1} \in V^{t-1}$, one obtains the *unconditional expected prediction error* of categorization C at date t;

$$EPE\left(C, t\right) = \mathbb{E}\left[PE\left(C, v_t, v^{t-1}\right)\right]. \tag{5}$$

Summing over the $T - L$ prediction tasks that the subject has to perform, one can define the *total expected prediction error* of a categorization C as

$$EPE\left(C, T, L\right) = \frac{1}{T-L} \sum_{t=L+1}^{T} EPE\left(C, t\right) \tag{6}$$

With these equations one may define the two notions of optimal categorizations that will be the focus of this paper. For the case of categories that are acquired before a data base has been accumulated, the relevant notion optimality is the following:

Definition 1 *A categorization $C \in \Psi$ is optimal prior to data, or **ex ante optimal**, if it minimizes $EPE\left(C, T, L\right)$.*

The set of such ex ante optimal categorizations is

$$\Psi^* = \arg\min_{C \in \Psi} EPE\left(C, T, L\right).$$

Let $k^*_{\min}$ (and $k^*_{\max}$) be the smallest (and largest) number of categories among the ex ante optimal categorizations. That is $k^*_{\min} = \arg\min_{C \in \Psi^*} |C|$ and $k^*_{\max} = \arg\max_{C \in \Psi^*} |C|$.

For the case of categories that are developed after a data base has been acquired we restrict attention to categorizations $\Psi\left(v^{t-1}\right)$ whose categories are non-empty categories under v^{t-1}. The relevant notion of optimality for categorizations that are developed conditional on a data base is:

Definition 2 *A categorization $C \in \Psi\left(v^{t-1}\right)$ is optimal conditional on a data base v^{t-1}, or **ex post optimal**, if it minimizes $EPE\left(C, v^{t-1}\right)$.*

3 Results

3.1 Preliminary Results

In order to derive an expression for $EPE\left(C, v^{t-1}\right)$, the expected prediction error conditional on a data base v^{t-1}, note that, for $X = [a, b]$,

$$\Pr\left((x, y) \in C_i\right) = \Pr\left(x \in X_i\right) = \int_{x \in X_i} \int_{y \in Y} f\left(x, y\right) dx dy,$$

and define

$$f\left(y | x \in X_i\right) = \frac{1}{\Pr\left(x \in X_i\right)} \int_{x \in X_i} f\left(x, y\right) dx.$$

In case X is finite the integral over X_i in these two expressions is replaced by a summation. Also define $Var\left(y_i\right) = Var\left(y | x \in X_i\right)$. Using this one can show.

Lemma 1 *The expected prediction error for a categorization C, conditional on a data base v^{t-1}, is*

$$EPE\left(C, v^{t-1}\right) = \sum_{i=1}^{k} \Pr\left(x \in X_i\right) \left(Var\left(y_i\right) + \left(\hat{y}_{it} - \mu_i\right)^2\right).$$

This expression reveals the basic trade-off that determines the ex post optimal number of categories. The term $Var\left(y_i\right)$ measures how similar (with respect to the y-dimension) different objects in category C_i are. The term $\left(\hat{y}_{it} - \mu_i\right)^2$ measures how close the prediction is to the actual average of category C_i. The ex post optimal categorization strikes a balance between the goal of having a low within category variance and the goal of estimating the category mean correctly. As we will see below, the same basic trade-off determines the ex ante optimal number of categories.

In order to derive $EPE\left(C, t\right)$, fix the date t and take expectation of $EPE\left(C, v^{t-1}\right)$ with respect to the data bases of size $t - 1$:

Lemma 2 *The (unconditional) expected prediction error for a categorization C, at time t, is*

$$\begin{aligned} EPE\left(C, t\right) = & \sum_{i=1}^{k} \Pr\left(x \in X_i\right) Var\left(y_i\right) \left(1 + \sum_{r=1}^{t-1} \Pr\left(m_{it} = r\right) \frac{1}{r}\right) \\ & + \sum_{i=1}^{k} \Pr\left(x \in X_i\right) \Pr\left(m_{it} = 0\right) \mathbb{E}\left[\left(\hat{y}_t - \mu_i\right)^2 | m_{it} = 0\right], \end{aligned}$$

where m_{it} has a binomial distribution

$$\Pr(m_{it} = r) = \binom{t-1}{r} (\Pr(x \in X_i))^r (1 - \Pr(x \in X_i))^{t-1-r}.$$

It will be fruitful to decompose the within-category variance (with respect to the y-dimension), $Var(y_i)$, into the contribution of the *within-category average conditional variance*

$$\mathbb{E}[Var(y|x)|x \in X_i] = \int_{x \in X_i} \frac{f(x)}{\Pr(x \in X_i)} Var(y|x)\, dx,$$

and, what I will call, the *within-category variance of the conditional expected value*

$$Var(\mathbb{E}[y|x]|x \in X_i) = \int_{x \in X_i} \frac{f(x)}{\Pr(x \in X_i)} \left(\mathbb{E}[y|x] - \int_{x \in X_i} \frac{f(x)}{\Pr(x \in X_i)} \mathbb{E}[y|x]\, dx \right)^2 dx.$$

(Again, in case X is finite the integral over X_i in these two expressions above is replaced by a summation.) The within-category variance is the sum of the within-category average conditional variance, and the within-category variance of the conditional expected value;[6]

$$Var(y_i) = \mathbb{E}[Var(y|x)|x \in X_i] + Var(\mathbb{E}[y|x]|x \in X_i). \tag{7}$$

Before providing results regarding optimal categorizations, I establish that such categorizations exist:

Proposition 1 *For any t, there exist a solution to the problem of minimizing $EPE(C,t)$, with respect to $C \in \Psi$. For any L and T, there exist a solution to the problem of minimizing $EPE(C,T,L)$, with respect to $C \in \Psi$. For any v^{t-1} there exist solution to the problem of minimizing $EPE(C, v^{t-1})$ with respect to $C \in \Psi(v^{t-1})$.*

It can be noted that there is no guarantee that any of these solutions are unique, thus allowing for a (mild) form of conceptual relativism.

3.2 Ex Ante Optimal Categorizations

The following proposition describes how the optimal categorization changes with length of the learning and prediction phases.

Proposition 2 ***(a)*** *If $T \to \infty$ then $k^*_{\max}/T \to 0$ and $k^*_{\min} \to |X|$.* ***(b)*** *There are finite L' and T', with $L' < T'$, such that if $L' < L < T < T'$, then $k^*_{\max} < L$.*

[6] It is a standard result that $\mathbb{E}[Var(y|x)] = Var(\mathbb{E}[y|x])$. Conditioning on $x \in X_i$ is straightforward.

Part (a) establishes that as T goes to infinity it is optimal to let the number of categories increase too, but a slower rate, so that the number of categories per object goes to zero. In the case of a finite X this result is a direct consequence of the finiteness of X but in the case of an infinite X this is a non-trivial result. Note that part (a) implies if the learning phase, or the prediction phase, is sufficiently long, then all optimal categorizations have more than one category. Part (b) says that if the learning phase is sufficiently long in relation to the prediction phase then all optimal categorizations have a smaller number of categories than the number of observations made during the learning phase. In total, proposition 2 provides an explanation for why we typically employ categorizations that are neither maximally fine grained – with one object in each category – nor maximally general – with all object in the same category. This is discussed further section 4.1.

Now consider two subjects 1 and 2, with different total number of observations. Denote their numbers of observations by T_1 and T_2 respectively. (In section 4.1 subjects 1 and 2 are interpreted as being a layman and an expert, respectively.) It is a corollary of the previous proposition that if the differences between the two subjects are large enough then it is optimal for individual 1 to have fewer categories than individual 2. Write $k^*_{\min}(T)$ and $k^*_{\max}(T)$ to make the dependence of $k^*_{\min}$ and $k^*_{\max}$ on T explicit.

Corollary 1 *For any T_1, if $k^*_{\max}(T_1) < |X|$, then there is a T' such that if $T_2 > T'$, then $k^*_{\max}(T_1) < k^*_{\min}(T_2)$.*[7]

The next three propositions concern the relationship between the density $f(x, y)$ and the optimal categorization. The first result considers the marginal density over X, i.e. $f(x)$. Write $k^*_{\min}(f)$ and $k^*_{\max}(f)$ to make the dependence of $k^*_{\min}$ and $k^*_{\max}$ on f explicit.

Proposition 3 *Restrict attention to the categorization of a proper subset $E \subsetneq X$. Consider two densities f_0 and f_1, such that for all $x \in E$ it holds that $f_0(y|x) = f_1(y|x)$ and $\alpha f_0(x) = f_1(x)$, for some $\alpha > 1$. The lowest optimal number of categories in E is at least as large with f_1 as with f_0, i.e. $k^*_{\min}(f_0) \leq k^*_{\min}(f_1)$.*

The more common objects from one subset of X are, the more fine-grained should the optimal categorization for that subset be. This is a generalization of the result in Fryer and Jackson (2008), to the effect that less frequent objects will be categorized more coarsely. Their result assumes a fixed number of categories, whereas mine does not. They relate the result to the possibility that ethnic minorities will be categorized more coarsely than majorities. This will tend to lead to more stereotypical predictions about the minority than about the majority.

The next result concerns the effect of the conditional variance, $Var(y|x)$, on the optimal categorization.

[7] Note that the restriction $k^*_{\max}(T_1) < |X|$ is always satisfied for infinite X.

Proposition 4 *Consider two densities f_0 and f_1, such that $f_0(x) = f_1(x)$, $\mathbb{E}_{f_0}[y|x] = \mathbb{E}_{f_1}[y|x]$ and $Var_{f_1}(y|x) > Var_{f_0}(y|x)$ for all $x \in X$. The lowest optimal number of categories ($k^*_{\min}$) is at least as large with f_1 as with f_0, i.e. $k^*_{\min}(f_0) \leq k^*_{\min}(f_1)$.*

We saw above that $Var(y_i)$ is the sum of $\mathbb{E}[Var(y|x)|x \in X_i]$ and $Var(\mathbb{E}[y|x]|x \in X_i)$. Proposition 4 concerns comparative statics with respect to the former term. Comparative statics with respect to the latter term requires more detailed assumptions about the distribution f.[8] For this reason I now restrict attention to the following special case: Suppose $X = [0,1]$ (and $Y = \mathbb{R}$ as before) and suppose that the relation between X and Y is described by the classical linear regression model;

$$y = \alpha + \beta x + z, \tag{8}$$

where $z \sim N(0, \sigma^2)$. Furthermore assume that x is uniformly distributed on X. Assume also that the subject only makes one prediction during her life, i.e. $T - L = 1$ (extension to $T - L > 1$ is straightforward but does not add insight). Finally, for simplicity, also assume that subjects are endowed with one observation in each category already during the learning phase, as mentioned in section 2.3. (The results become more tractable with this assumption but the general insight is unaltered.) Under these assumptions we have the following result:

Proposition 5 *For any T and L the number of categories in the optimal categorization is unique and all categories are convex with the same length along the x-axis. The optimal number of categories is increasing in β and decreasing in σ^2.*

Recall that for the linear regression model it holds that

$$\beta = \frac{Cov(x,y)}{Var(y)},$$

so increasing covariance of x and y increases the optimal number of categories. Increasing the conditional variance of y decreases the optimal number of categories. This result is very intuitive: If the covariance is large then the categories have to be narrow in order to keep the heterogeneity of objects in each category within limits. If the variance of y is large then (in line with proposition 4) the categories have to be broad in order to contain enough objects to allow reasonably accurate estimates of the means of each category.

[8] Without any assumption on f one will only be able to make observations like the following: Consider a categorization C' and a refinement C'', of C'. Suppose that $Var(\mathbb{E}[y|x]|x \in X_i') \geq \phi Var(\mathbb{E}[y|x]|x \in X_j^{''})$, $\phi > 1$, for any two categories $C_i' \in C'$ and $C_j'' \in C''$ such that $C_j'' \subseteq C_i'$. This will be the case if the intervals over which $\mathbb{E}[y|x]$ is constant, are sufficiently small. If $Var(\mathbb{E}[y|x]|x \in X_j^{''})$ is increased sufficiently much, for all $C_j'' \in C''$, then categorization C'' will yield a lower expected prediction error than categorization C'.

Proposition 5 can be extended to the case of a finite X even when there is no metric on X, provided that we can infer an order on X from $\mathbb{E}[y|x]$, as described by (8).

3.3 Ex Post Optimal Categorizations

Some categorizations are formed only after a data base has been accumulated. In the introduction it was argued that humans might have evolved an ability to form categorizations that tend to minimize expected prediction error conditional on the accumulated data base. That is, humans would have access to algorithms or heuristics that takes a given data base as input and delivers an approximately (ex post) optimal categorization as output, without using any information that is not in the given data base. As analysts we might be willing to abstract from these heuristics and assume that subjects act *as if* they optimized on the basis of knowledge of f. However, a more realistic approach would specify some heuristic that could potentially be used to find approximately ex post optimal categorizations. In this section I present results along both these lines.

The notion of ex post optimal categorizations is defined with reference to a given data base v^{t-1}. This means that the optimal categorization may look very different depending on the particular data base. The results presented in this section are therefore formulated in terms of how changes in the model's parameters influences the *probability* that a categorization yields a lower expected prediction error $EPE(C, v^{t-1})$ than another categorization. With this phrasing, it turns out that one can prove results that are fairly direct counterparts to the results provided for ex ante optimal categorizations (though in some cases more restrictive assumptions are applied).

What heuristics might a categorizing subject use to form categorizations given a data base? Computing the expected prediction error $EPE(C, v^{t-1})$ requires knowledge of the distribution f. The subject should not be assumed to know f, because if the subject did know f, there would be no need to base predictions on categorization, rather than using knowledge of the density f directly. However, the subject could use the data base to compute some estimator of $EPE(C, v^{t-1})$, and then pick a categorization that minimizes this value – possibly within an a priori restricted set of categorizations. The following estimator could be used:

Definition 3 *Let $\Psi(v^{t-1})$ denote the set of categorizations in which all categories have at least two elements ($m_{it} \geq 2$) given the data base v^{t-1}. The sample prediction error for a categorization $C \in \Psi(v^{t-1})$, conditional on a data base v^{t-1}, is*

$$\widehat{EPE(C, v^{t-1})} = \sum_{i=1}^{k} \frac{m_{it}}{t-1} \left(1 + \frac{1}{m_{it}}\right) s_{it}^2,$$

where

$$s_{it}^2 = \frac{1}{m_{it} - 1} \sum_{s \in D_{it}} (y_s - \hat{y}_{it})^2 .$$

The motivation for the definition of $\widehat{EPE(C, v^{t-1})}$ comes from the following observation, which follows directly from the facts that $\mathbb{E}[(\hat{y}_{it} - \mu_i)^2] = Var(y_i)/m_{it}$, and $\mathbb{E}[s_{it}^2] = Var(y_i)$.

Lemma 3 *For a given categorization C, and a data base v^{t-1}. If expectation is taken over all databases with have the same number of objects in each category as v^{t-1} has,*[9] *then*

$$\mathbb{E}\left[EPE\left(C, v^{t-1}\right)\right] = \sum_{i=1}^{k} \Pr(x \in X_i) \left(1 + \frac{1}{m_{it}}\right) Var(y_i),$$

and

$$\mathbb{E}\left[\widehat{EPE(C, v^{t-1})}\right] = \sum_{i=1}^{k} \frac{m_{it}}{t-1} \left(1 + \frac{1}{m_{it}}\right) Var(y_i).$$

The lemma implies that if the actual fraction of objects in each category $(m_{it}/(t-1))$ is equal to the probability of receiving an object in the corresponding category $(\Pr(x \in X_i))$ then $EPE(C, v^{t-1})$ and $\widehat{EPE(C, v^{t-1})}$ have the same expected value.

We are now in a position to state results both regarding the actual expected prediction error and the estimated expected prediction. More specifically, each of the following results compares a categorization C' with a refinement C''. Since C'' is a refinement of C' we have $k' < k''$. Changing parameters of the model influences the probability that C'' will yield a lower *expected prediction error* than C'. Changing the parameters also influences the probability that C'' will yield a lower *estimated expected prediction error* than C'. It turns out that these effects are very similar, indicating that $\widehat{EPE(C, v^{t-1})}$ is a useful estimator.

The first result regards the effect of varying the size of the data base, it corresponds to proposition 2 about ex ante optimality.

Proposition 6 ***(a)*** *For any categorization C' with $k' < |X|$ categories, there is a refinement C'' with $k'' > k'$ categories such that for any $\delta \in (0, 1)$, there is a t', such that if $t > t'$ then*

$$\Pr\left(EPE\left(C', v^{t-1}\right) > EPE\left(C'', v^{t-1}\right)\right) > \delta. \tag{9}$$

(b) *Restrict attention to databases such that $m_{it} \geq 2$ for all categories in C' and C''. The statement in (a) holds if (9) is replaced with*

$$\Pr\left(\widehat{EPE(C', v^{t-1})} > \widehat{EPE(C'', v^{t-1})}\right) > \delta. \tag{10}$$

[9] That is, take expectation over the data bases in the set $\left\{\tilde{v}^{t-1} \in V^{t-1} | \forall i : C_i \in C \Rightarrow \tilde{m}_{it} = m_{it}\right\}$.

Part (a) says that by increasing the size of the data base, we can increase the probability that C'' yields a higher expected prediction error than the refinement C'', meaning that C'' is closer to being optimal than C' is. By increasing the size of the data base sufficiently much we can push this probability arbitrarily close to one. Note that the proposition can be applied iteratively so that it holds also if C'' is replaced by any refinement of C''. This implies that, for any given number $k < |X|$, by increasing the size of the data base we are increasing the probability that the optimal categorizations will have at least k categories. Part (b) goes on to state that this same relationship holds for the estimated expected prediction error, provided that we restrict attention to data bases for which both C' and C'' are well-defined.

To see why it is necessary to formulate the proposition in probabilistic terms, consider the following example which shows that adding an observation to a data base may sometimes lead the optimal number of categories to decrease. Similar examples may be constructed for minimization of estimated expected prediction error.

Example 4 *Assume* $X = [0,1]$,$Y = \mathbb{R}$, $Var\,(y|x) = \sigma^2$, $f\,(x) = 1$, *and*

$$\mathbb{E}\,[y|x] = \begin{cases} 0.5 & \text{if } x < 0.5 \\ 0.1 & \text{if } x \geq 0.5 \end{cases}.$$

Consider the data base $v = \{(0.1, 0.6)\,, (0.2, 0.6)\,, (0.7, 0)\}$. *Compare a categorization* C' *consisting of only one category, with a categorization* C'' *that divides* X *into two categories* $C_1 = [0, 0.5) \times \mathbb{R}$ *and* $C_2 = [0.5, 1] \times \mathbb{R}$. *It is straightforward to compute* $EPE\,(C', v^{t-1}) = \sigma^2 + 0.2^2 + 0.1^2 = \sigma^2 + 0.05$ *and* $EPE\,(C'', v^{t-1}) = \sigma^2 + 0.1^2 = \sigma^2 + 0.01$. *Thus* C'' *is the ex ante optimal categorization. Now suppose one object* $(0.8, -0.6)$ *is added to the data base. We now have* $EPE\,(C', v^{t-1}) = \sigma^2 + 0.2^2 + 0.15^2 = \sigma^2 + 0.0625$ *and* $EPE\,(C'', v^{t-1}) = \sigma^2 + 0.5 \cdot 0.1^2 + 0.5 \cdot 0.4^2 = \sigma^2 + 0.085$. *Hence* C' *is the new ex ante optimal categorization. The intuition behind this result is that the added object is such an outlier that it needs to be "neutralized" in a larger sample, which is achieved by merging the categories.*

The next proposition tells us what happens if we restrict attention to categorizing a subset E of X, and vary the density on E.

Proposition 7 *Restrict attention to the categorization of a proper subset* $E \subsetneq X$. *Consider two densities* f_0 *and* f_1, *such that for all* $x \in E$; $f_0\,(y|x) = f_1\,(y|x)$ *and* $\alpha f_0\,(x) = f_1\,(x)$, *for some* $\alpha > 0$. ***(a)*** *For any categorization* C' *with* $k' < |E|$ *categories, there is a categorization* C'' *with* $k'' > k'$ *categories such that for any* $\delta \in (0,1)$, *there is a* t', *such that for any* $t > t'$ *there is an* $\alpha'\,(t)$ *such that if* $\alpha > \alpha'\,(t)$ *then*

$$\Pr\left(EPE\left(C', v^{t-1}\right) > EPE\left(C'', v^{t-1}\right)\right) > \delta, \tag{11}$$

and if $\alpha = \alpha'(t)$ then the above does not hold. Moreover, $\alpha'(t)$ is decreasing in t. ***(b)*** *Restrict attention to databases such that $m_{it} \geq 2$ for all categories in C' and C''. The statement in (a) holds if (11) is replaced with*

$$\Pr\left(\widehat{EPE(C', v^{t-1})} > \widehat{EPE(C'', v^{t-1})}\right) > \delta. \tag{12}$$

In other words, if t is large enough then we can increase the probability that C'' yields a lower expected prediction error (and estimated expected prediction error) than C' by increasing the density over E, as parameterized by α.

The effect of changing the conditional variance is described by the following proposition. Compared to proposition 4 it makes the stronger assumption that the variance of y conditional on x is normally distributed. The assumption is made for reasons of tractability, and it could be conjectured that a similar result would hold in its absence.

Proposition 8 *Consider a density f such that $y|x \sim N(\mathbb{E}[y|x], \sigma^2)$ for all $x \in X$.* ***(a)*** *For any categorization C' with $k' < |X|$ categories, there is a refinement C'' with $k'' > k'$ categories such that, in the set of databases with $m_{it} \geq 1$ for all categories $C_i \in C''$, the probability*

$$\Pr\left(EPE\left(C'', v^{t-1}\right) < EPE\left(C', v^{t-1}\right)\right), \tag{13}$$

is weakly increasing in $Var(y|x)$.[10] ***(b)*** *Restrict attention to databases such that $m_{it} \geq 2$ for all categories in C' and C''. The statement in (a) holds if (13) is replaced with*

$$\Pr\left(\widehat{EPE(C'', v^{t-1})} < \widehat{EPE(C', v^{t-1})}\right). \tag{14}$$

By increasing the variance of y conditional on x one can increase the probability that C'' yields a higher expected prediction error (and estimated expected prediction error) than C'. Thus by increasing the variance one may increase the probability that the optimal categorizations have less than k categories, for any given $k > 1$.

Finally, if we make the same assumptions as for proposition 5 then we obtain the following result. Compared to proposition 5 this result is weaker since it assumes convex categories, rather than deriving them as a result.

Proposition 9 ***(a)*** *Restrict attention to categorizations with convex categories.*[11] ***(a)*** *For any categorization C' there is a refinement C'' such that, in the set of databases with $m_{it} \geq 1$ for all categories $C_i \in C''$, the probability*

$$\Pr\left(EPE\left(C'', v^{t-1}\right) < EPE\left(C', v^{t-1}\right)\right), \tag{15}$$

[10] If one did not assume $y|x \sim N(\mathbb{E}[y|x], \sigma^2)$. the result would still hold for large enough t, as a consequence of the central limit theorem.

[11] The probability that convex categories outperform non-convex categories, can be made arbitrarily large by increasing t.

is weakly increasing in σ^2, and weakly decreasing in β. ***(b)*** *Restrict attention to databases such that $m_{it} \geq 2$ for all categories in C' and C''. The statement in (a) holds if (15) is replaced with*

$$\Pr\left(\widehat{EPE\left(C'', v^{t-1}\right)} < \widehat{EPE\left(C', v^{t-1}\right)}\right). \tag{16}$$

By increasing the variance the slope β one can increase the probability that C'' yields a lower expected prediction error (and estimated expected prediction error) than C'. Thus by increasing β one may increase the probability that the optimal categorizations have more than k categories, for any given $k < |X|$. The effect of changing the conditional variance is, of course, the same as in proposition 8.

Propositions 6-9 indicate that $\widehat{EPE\left(C, v^{t-1}\right)}$ might be a reasonable guide to choices between different ways of categorizing a given data base. However, it might it might be too cognitively demanding and time consuming to compute $\widehat{EPE\left(C, v^{t-1}\right)}$ for all categorizations in $\Psi\left(v^{t-1}\right)$. For this reason the set of categories that one chooses from may be restricted in some way. There are many restrictions that one could reasonably use.[12] For instance one could use the following procedure that restricts the set of categorizations to those that can be obtained by successively refining some initial categorization:

- Start out with an initial categorization C'. This could either be some prior categorization that needs to be updated or it could be the trivial categorization with only one category.
- Pick a category $C'_i \in C'$, and perform the following test: Create a new categorization C'' by splitting C'_i into two categories C''_{i1} and C''_{i2}.
  - If either C''_{i1} or C''_{i2} contain less than two objects (given v^{t-1}), pick another category C'_j $(j \neq i)$ in C', and do the same test with that category.
  - If both C''_{i1} and C''_{i2} contain at least two objects (given v^{t-1}), compute the estimator of the expected prediction error for both C' and C''.
    * If $\widehat{EPE\left(C'', v^{t-1}\right)} > \widehat{EPE\left(C', v^{t-1}\right)}$ then pick another category C'_j $(j \neq i)$ in C', and do the same test with that category.
    * If $\widehat{EPE\left(C'', v^{t-1}\right)} < \widehat{EPE\left(C', v^{t-1}\right)}$ then let C'' be the new benchmark categorization.
- Continue this process until splitting categories further either would result in categories with less than two objects or would result in an increase in estimated expected prediction error.

[12] Alternatively, if the subject already has some prior categorization (that needs to be updated) then the set $\Psi\left(v^{t-1}\right)$ could be restricted to only include that prior categorization together with categorizations that are in some way similar to that categorization.

The process ends in a finite number of steps. Since it restricts the choice of categorizations there is no guarantee that it will deliver a categorization that is optimal within a larger set of categories. However propositions 6-9 indicate that, among the categorizations that it considers, it will tend to pick categorizations in a way that is similar to what minimization of expected prediction error dictates.

Categorizations that are formed in response to a data base may, of course, be modified as more data is accumulated. Eventually, the subject might suspect that the old categorization could be improved to such an extent that it is worth the trouble of going through the above procedure again. In such a setting of occasional re-categorizations, it is likely that the succession of categorizations will display some path dependence, in the sense that the order of observations matter for the end result. The reason is that the suggested procedure takes the current categorization as starting point and only considers certain refinements thereof. The data that has been accumulated before the first categorization is formed will therefore have a decisive influence on how the categorization is subsequently modified in response to added data points.

3.4 Extensions

3.4.1 Interest-dependent predictions

The cost of a prediction error has been assumed to be independent of x, only depending on the distance between the predicted and the actual y-value. More realistically it could be that predictions associated with some set $E \subseteq X$ are considered more important than predictions in some other set $F \subseteq X$. It is easy to extend the model to handle this possibility. One may simply add a function $w : X \to [0, 1]$ such that $w(x)$ measures the importance of predictions associated with x. It is straightforward to verify that changes in $w(x)$ will have much the same effect as changes in $f(x)$. Increasing the importance of predictions in a set E will tend to increase the optimal number of categories in that set.

3.4.2 Multi-dimensional X and Y

The model can be extended to allow for prediction of many different attributes of an object, represented by a vector $y_t \in Y = \mathbb{R}^n$. The easiest way of doing this is to let the prediction error, whose expectation should be minimized, be a weighted "city-block" metric. Let $y(j)$ denote the j^{th} component of y and let $z(j)$ be the weight put on the j^{th} dimension. The prediction error may be defined as $PE(C, v_t, v^{t-1}) = \sum_{j=1}^{n} z(j)(y_t(j) - \hat{y}_{it}(j))^2$. A similar modeling choice is made by Fryer and Jackson (2008), who also refer to empirical evidence on the psychological relevance of the city-block metric. With this specification all of the results presented above hold for a multi-dimensional Y. The weights may differ between subjects, allowing for another form of interest-dependent predictions.

It was assumed that either X is finite or X is a closed interval on the real line. When X

is a finite set it is inconsequential whether objects are multidimensional or not. It is more complicated to allow for an infinite set X that is not one-dimensional. The difficulty lies in proving that an optimal categorization exists. The problem can be handled by restricting attention to finding an optimal categorization within a finite subset of categorizations.

3.4.3 Fixed number of categories

It has been shown that a subject basing predictions on categories might be better off using a coarse rather than a fine categorization. Still, the number of categories may be restricted for other reasons which are external to the model (e.g. some cognitive cost that is increasing in the number of categories). The model can be adapted to deliver predictions about how the optimal categorization divides X into an exogenously given number of categories. For simplicity I confine the discussion to ex ante optimality.

First, note that $\sum_{r=1}^{t-1} \Pr(m_{it} = r)/r \rightarrow 0$ as $t \rightarrow \infty$, implying that $EPE(C,t)$ approaches $\sum_{i=1}^{k} \Pr(x \in X_i) Var(y_i)$. The term $\sum_{r=1}^{t-1} \Pr(m_{it} = r)/r$ is decreasing in $\Pr(x \in X_i)$. It follows that, as t increases it becomes less important that categories with a high variance $Var(y_i)$ are also categories with a high $\Pr(x \in X_i)$. In the limit as $t \rightarrow \infty$, the optimal categorization is completely determined by $Var(y_i)$. In this case the optimal categorization simply maximizes the sum of within category variances. Second, suppose that the subject categorizes a proper subset $E \subseteq X$ and its complement separately. Restrict attention to the categorization of E and consider the effect of increasing $f(x)$ for all $x \in E$. The results depend on whether the total number of categories, or the number of categories in E, is fixed. If the number of categories in E is allowed to vary, but the total number of categories is fixed, then it follows from proposition 3 that increasing $f(x)$ for all $x \in E$ will result in a larger optimal number of categories in E. If the number of categories in E is fixed, then increasing $f(x)$ for all $x \in E$ will have an effect similar to that of increasing t, since it will decrease the term $\sum_{r=1}^{t-1} \Pr(m_{it} = r)/r$. Third, increasing $Var(y|x)$ for all x will make expected variance $\mathbb{E}[Var(y|x)|x \in X_i]$, more decisive than the variance of the expected mean $Var(\mathbb{E}[y|x]|x \in X_i)$, in determining the optimal categorization. Finally, in the context of the linear regression model studied in propositions 5 and 9 changing σ^2 or β will not affect the optimal categorization, since it affects $Var(y_i)$.equally for all categories.

4 Discussion

4.1 Psychological applications

In studies of concepts and categorization with hierarchically organized concepts (e.g. animal – bird – robin) it is found that there is a privileged level in the hierarchy, called the basic level. Generally this level is named spontaneously in categorization tasks, learned first by children, and is in other ways salient (Rosch et al. (1976)). The basic level is neither the most general level nor the most detailed level (e.g. bird rather than the superordinate category animal or the subordinate category robin). If categories are useful because they facilitate prediction then it might seem that it should be optimal to have an infinitely fine grained conceptual structure, since the narrower a category is, the more precise are the predictions that can be inferred from category membership. For example, if something is categorized as a bird then one can infer that it lays eggs and has wings, but if it is categorized (more finely) as a penguin, then one can also infer that it cannot fly but can dive. Since we do not use infinitely fine-grained category structures there must be some other factor that decreases the benefit of narrow categories. I have argued that such a factor is constituted by the need to have a sufficiently large sample in each category to generalize from. The dominant view in psychology has instead been that the cost of fine grained categorizations has to do with the difficulty of categorizing objects into fine grained categories: In order to make a more fine-grained categorization one has to observe more properties of an object (Medin (1983), and Jones (1983)). It is difficult to come up with a clean test between these two explanations. The reason is that lower level categories both contain less objects and are associated with more stringent conditions for application. Experimentally one could try to find a superordinate category and a subordinate category which are equally easy to apply. The conventional psychological explanation would then predict that the basic level will not be the superordinate of these two categories. In contrast, explanation suggested in this paper would predict the superordinate category to be basic if the subordinate category contains too few exemplars, or is associated with too much variance. The explanations are probably best viewed as complementary. Both may describe forces that shape our categorizations.

Experts tend to have a more fine grained conceptual structure than laymen (Tanaka and Taylor (1991), Johnson and Mervis (1998)). This can be explained in the present model, with the help of proposition 1. Consider a layman with a learning phase of length L_1 and a prediction phase of $T_1 - L_2$ periods. Suppose the optimal number of categories for this person is k_1. An expert is distinguished by that she goes through more extensive training, L_2, or a longer prediction phase $T_2 - L_1$, than the layman The model predicts that if these differences between an expert and a layman, then it is optimal for the expert to have larger number of categories than the layman; $k_2 > k_1$. This may also explain why some populations use a more fine-grained category structure than other populations: For instance, people in traditional subsistence cultures tend to have more specific biological categories than e.g. American college students (Berlin et al. (1973), Solomon et al. (1999)). Needless to say there are other possible explanations for this phenomenon.

4.2 Heterogeneous priors

As mentioned in the introduction the model can be seen as describing a way of generating priors. This interpretation allows us to distinguish various sources of heterogeneous priors. Clearly, different experience, in the form of different databases will lead to different predictions, for a given common categorization. More importantly even subjects with the same experience may arrive at different priors, if they use different categorizations. In the case of ex ante categorizations this might be due to the fact that they have learned categorizations from cultures or trades which have developed in response to different distributions f. In the case of ex post categorizations is may be due to the subjects having had different data bases at time they formed their categorizations, even if subsequent observations have made their data bases identical. Furthermore different interests, as discussed in section 3.4, and as represented by different weights w (on X) and z (on Y), will lead to different optimal categorizations.

4.3 Categorization in Game Theory

There are some interesting game theoretic studies of the impact of coarse reasoning on decision making and equilibrium: Jehiel (2005) develops a notion of analogy based expectations equilibrium for extensive form games. Players bundle together the nodes of the opponents into analogy classes in order to predict the opponents' behavior. A player expects the same behavior in all nodes in an analogy class. In equilibrium these expectations are correct on average. The equilibrium is parameterized by the analogy classes, which are exogenous. Jehiel and Samet (2007) define a notion of valuation equilibrium. Players bundle their own strategies into different similarity classes, when predicting their own payoffs. The same payoff is expected for each strategy in the same similarity class. Again, in equilibrium the expectations are correct on average. The similarity classes are exogenous, even though Jehiel and Samet discuss the possibility of endogenizing them. There is also a literature, starting with Dow (1991) that examines the optimal way to partition a state space in the face of limited memory. In these models the number of cells in the partition is determined exogenously by the bound on memory, and the subject has a prior on the state space.

The results obtained above could potentially be used as a way of endogenizing the number of cells, and possibly also the structure of the cells, in such models. At least, the results could motivate restrictions on the set of feasible partitions. The full model developed above is not suitable for these purposes. Rather it can be taken as inspiration for simpler modeling devices that can be applied more readily to game theoretic contexts. Propositions 6 and 7 imply that optimal categories should not contain too few observations. This could motivate a restriction that each cell of a feasible partition needs to be reached with at least some minimum probability in equilibrium.[13] Moreover, in accordance with proposition 8) this threshold probability should be increasing in the expected variance $\mathbb{E}\left[Var\left(y|x\right)|x \in X_i\right]$, and in line with line with proposition 9 it should be decreasing in the variance of the expected mean $Var\left(\mathbb{E}\left[y|x\right]|x \in X_i\right)$. Analyzing the

[13] Such a restriction is mentioned by Jehiel (2005), note 22.

implications of such restrictions is beyond the scope of this paper.[14]

4.4 Why use Categories?

This paper builds on the assumption that we use categories to make predictions. The assumption is based on a substantial body of psychological research establishing that categorical thinking permeates human cognition, including prediction. Nevertheless, one might ask why we use categorizations rather than some other method for making predictions. In particular one might suggest that one could employ some form of similarity based reasoning instead, for instance as formalized by kernel-based regressions. On this approach the prediction of y conditional on x will be a weighted average of nearby observations, where the weights put on an observation (x', y') is a decreasing function of the distance between x and x'. One obvious limitation of this approach is that it will not work unless the subject has access to some metric on X. In contrast, as shown above, prediction based on categorization is possible even when there is no such metric available. The objects are then grouped solely on the basis of their y-values. In essence such a categorization creates a similarity relation – objects sharing the same category are similar and objects not sharing the same category are dissimilar. The question of whether categorizations are based on similarity relations or not is subject to debate (insert ref), and will not be discussed further here. However, even if one is willing to assume that subjects have a on X there are some further potential shortcomings of kernel based predictions, compared with predictions based on categorizations.

Presumably categorizations are used in order to facilitate fast predictions; when facing a new object the subject simply puts the object in a category and uses the corresponding prediction. Of course, the subject might decide to devote more time to the prediction problem, but in that case the categorization-based prediction might be modified by other modes of prediction-making. Hence category-based and kernel-based predictions should be compared for the case when predictions are produced in a relatively fast and automatic way. In this case predictions have to be computed in advance. Note that this line of reasoning does not depend of on whether the categorization was acquired before the subject had accumulated a data base, or formed on the basis of an accumulated data base. Once a categorization is in place predictions are computed in advance.

A subject basing predictions of categories will use something like the following procedure: At the beginning of period t the subject has stored k pairs $(\hat{y}_{it}, m_{it})$ of predictions and samples sizes, one for each category. She then observes x_t, identifies C_i, such that $x_t \in C_i$ and predicts $\hat{y}_{it}$. At the end of period t she observes y_t and uses it to computes an updated prediction $\hat{y}_{it+1} = (m_{it}\hat{y}_{it} + y_t) / (m_{it} + 1)$ for category C_i. and replaces $(\hat{y}_{it}, m_{it})$ with $(\hat{y}_{it+1}, m_{it+1})$.

[14] It might be fruitful to note the analogy between the thresholds discussed here and the lexicographic costs of states that have been imposed in the literature on games played by finite automata (Abreu and Rubinstein (1988)).

In contrast, a subject basing predictions on kernel based estimation will use a procedure akin to the following: At the beginning of period t the subject has stored $t-1$ different objects (x, y) and a number of $|X|$ predictions $\hat{y}_t|x$. She then observes x_t, and uses the corresponding prediction $\hat{y}_t|x_t$. At the end of period t she adds the observation (x_t, y_t) to her memory. She computes an updated prediction $\hat{y}_{t+1}|x$ for each x within some distance η of the observed x_t. (Observation x_t has positive weight only in the computations of predictions $\hat{y}_t|x'$ for x' that are within distance η from x_t.)

In conclusion, for the case of categories and similarity relations that are acquired prior to data, the kernel-based procedure has at least three drawbacks. (1) It requires the subject to store a larger number of predictions; $|X|$ rather than k predictions. (2) It requires the subject to update a larger number of predictions after each new observation; in the kernel based procedure the subject has to update all predictions associated with values x that are within distance η of the new observation x_t; in the category based procedure only k predictions are updated. (3) It stores more information about observations; $t-1$ individual observations rather than k pairs $(\hat{y}_{it}, m_{it})$.

4.5 Philosophy of Science Applications

4.5.1 Interplay of Observation and Categorization

Consider a scientist who wants both to develop new theories, in the form of new concepts, and to make new observations, in order to make better predictions. The scientist can influence what observations she makes by choosing what experiments to perform. For a given categorization the subject chooses the numbers of observations in different categories $\{m_{it}\}_{i=1}^{k}$, under the restriction that $\sum_{i=1}^{k} m_{it} = t-1$, and $m_{it} \geq 1$ for all i. She would like to minimize expected prediction error

$$
\begin{aligned}
\mathbb{E}\left[EPE\left(C, v^{t-1}\right)|\{m_{it}\}_{i=1}^{k}\right] &= \sum_{i=1}^{k} \Pr\left(x \in X_i\right)\left(Var\left(y_i\right) + \mathbb{E}\left[\left(\hat{y}_i - \mu_i\right)^2 |m_{it}\right]\right) \\
&= \sum_{i=1}^{k} \Pr\left(x \in X_i\right)\left(1 + \frac{1}{m_{it}}\right) Var\left(y_i\right).
\end{aligned}
$$

The first equality uses lemma 1 and the second equality uses the same logic as the proof of lemma 2. Immediately one sees that, for a given categorization, it is optimal to choose to make observations in categories that have a large probability mass and/or a large variance. It is also evident that the marginal benefit of new observations is decreasing.

Forming new categories and concepts is intellectually demanding. More specifically it seems reasonable to assume that this cost is not continuous in the magnitude of the conceptual change; even small changes are likely to involve some strictly positive minimal costs. Thus there is reason not to perform re-categorization continuously. Rather, it is optimal to perform such activities occasionally. Directly after a re-categorization the benefit of new observations is high and efforts are rationally directed at performing experiments within the framework of the current categorization. After a while the marginal benefit of new observations has declined sufficiently much to make it more valuable to invest efforts in category formation rather than observation. In this way periods of observation with fixed categories and periods of re-categorization will alternate. Note the similarity with how Kuhn (1970) describes the interaction of normal science and scientific revolutions.

4.5.2 Predicates Suitable for Induction

The model may serve to clarify intuitions regarding Goodman's "new riddle of induction" (Goodman (1955)). Let $X = \{green, blue\} \times \{t < t^*, t \geq t^*\}$, were t^* is some future date. Consider two categorizations $C' = \{green, blue\}$, and $C'' = \{grue, bleen\}$, where an object v is grue if it is green and observed before t^* of blue and observed after t^*, and an object is bleen if blue before t^* or green after t^*. Two subjects accumulate observations of emeralds. Subject 1 uses categorization C' and hence registers each emerald as green. Subject 2 uses categorization C'' so each emerald is recorded as being grue. The subjects would like to predict the color of emeralds observed after date t^*. Both subjects predict that emeralds observed after t^* will be like the emeralds observed before t^*. This means that subject 1 predicts that they will be green (or bleen in the terminology of subject 2) and subject 2 predicts that they will be grue (or blue in the terminology of subject 1). Assuming that emeralds will in fact continue to be green after t^* (they will change from grue to bleen in the terminology of subject 2) we can compute prediction errors. Regardless of whether we evaluate prediction error with $Y = \{green, blue\}$ or $Y = \{grue, bleen\}$ subject 1 will make correct predictions while subject 2 will make incorrect predictions. The problem that Goodman points out is that in advance there is no way of distinguishing the predicates blue/green from grue/bleen. The solution hinted at here is to use second-order inductions, as discussed by Quine (1969): Inductions involving predicates like grue/bleen tend to yield large prediction errors. Hence we make a second-order inductive inference to the effect that we should avoid such concepts when engaging in first-order inductive inference.

5 Related Literature

Fryer and Jackson (2008) consider a notion of optimal categorization. Their model has similarities with the present model; objects are represented as vectors in some space of features, and the prediction about a new object in a category is based on the average of past objects in that category. But there are also some important differences: First, the number of categories is exogenously given. Second, although the purpose of categorization is to generate predictions Fryer & Jackson do not define optimality in terms of minimization of prediction error. Instead they define the optimal categorization as the one that minimizes the sum of within-category differences between objects that have already been encountered. Third, the probability of encountering different objects is not modeled.[15] As a consequence the trade-off that is central to the present paper cannot be formulated within their framework.

Al-Najjar and Pai (2010) develop a model of coarse decision making, which is applied to categorization. It should be noted that the first version of their paper is dated December 2008, whereas the first version of my paper was presented before that, on November 14, 2008, at the Third Nordic Workshop in Behavioral and Experimental Economics in Copenhagen. Like in my paper a subject categorizes two-dimensional objects with respect to one dimension in order to predict the other dimension. With the help of Vapnik Chervonenkis theory they provide results regarding the set of categorizations that guarantee a worst case predictive performance below some threshold. Worst case is taken both over the set of distributions and over the set of categorizations. Their results are effectively about what I have called ex ante categorization.

Peski (2010) studies categorization in a quite different setting. There are infinitely many properties and objects. A state of the world specifies which of the properties that each object has. Properties are modeled as discrete, and similarity is defined in terms of sharing properties. Peski argues for the usefulness of categorization by compares predictions based on Bayesian updating with predictions based on a categorization algorithm. The analysis depends crucially on a symmetry assumption: The Bayesian prior over the states of the world is symmetric, in the sense that it is invariant with respect to relabeling of objects and properties. Under this assumption predictions based on the categorization algorithm will asymptotically approach the predictions based on Bayesian updating. Thus a Bayesian subject with a symmetric prior will expect to asymptotically perform approximately the same regardless of whether she uses Bayesian updating or follows a categorization algorithm. If the state of nature is in fact drawn from a symmetric distribution, then a subject following the categorization algorithm will asymptotically make predictions that are no worse than the predictions made by a subject who knows the distribution. In my model, there is no subjective prior, and the true distribution need not be symmetric. It should also be noted that Peski's results are asymptotic, relying on a sufficient data condition, which states that the number of observations asymptotically becomes infinitely much larger than the number of distinct features in the data base of past observations (although there are infinitely many properties and objects). In my model categorization is instead a consequence of scarcity of data.

Other, more distantly related models of categorization are due to Mullainathan (2002) and Anderson (1991). Gilboa and Schmeidler (2003) model case-based predictions; given a data base of past cases the subject's task is to rank the likelihood of different outcomes in a new case. Gilboa et al. (2006) provide an axiomatization of a similarity based prediction

[15] When relating their optimality criterion to utility maximization, they simply assume that the distribution of future objects will be the same as the empirical distribution of already encountered objects.

rule for the case of predicting a real-valued variable y. The rule states that the value of y in the case at hand will be equal to a similarity weighted average of that variable in the past cases. The axiomatization assures that there exists such a similarity function if and only if the probability rankings made by the subject, given various data bases, satisfies certain axioms. The axiomatization only tells us that a similarity function exists, not what it looks like. One may view a categorization as a special similarity function that treat all cases in the same category as exactly similar to each other and treat a case in a category as completely dissimilar to any case outside that category. I restrict attention to the set of such similarity functions, and instead of providing an axiomatization I seek to characterize the optimal such function.

Pothos and Chater (2002) put forward an account of categorization that builds on the idea that the simplest categorization, as measured by code length, will be preferred. In their model there is a trade-off between reducing the number of categories (thereby simplifying the representation of similarity of categories) and reducing the number of objects within each category (thereby simplifying the representation of the within-category similarity). On this account the optimal categorization is one that maximizes simplicity, rather than predictive success (see Chater (1999) on the link between simplicity and prediction).

In the field of machine learning there are several models related to categorization. The approach most relevant to the kind of unsupervised categorization studied in this paper is cluster analysis (for a review see e.g. Jain et al. (1999)). In cluster analysis one seeks to partition a set of objects in a way that maximizes some measure of within cluster similarity and between cluster dissimilarity. Still, there are important differences compared to the present paper. In cluster analysis one would not define optimality with respect to the underlying distribution generating the data base. This stands in contrast with the definitions of both ex ante and ex post optimality. It does not conflict with minimization of the estimator of expected prediction error. However, the cluster algorithms that I know of do not cluster a set of two-dimensional objects by defining clusters with respect to one dimension (X) and evaluating the clusters with respect to the other dimension (X), in contrast with categorization based on the estimated expected prediction error.

6 Conclusion

I have provided a framework for the study of optimal categorization for the purpose of making predictions. The optimal number of categories is endogenous to the model. A small category results in smaller variance of objects in that category. A large category leads to a large number of experienced objects in the category, thus improving the precision of the predictions of the category mean. Thus the optimal categorization strikes a balance between fitting and over-fitting. This can explain the fact that the privileged level of categorization – the so-called basic level – is neither the coarsest nor the finest one.

Comparative statics yield several predictions about how the optimal categorization varies with the number of observations and the distribution of objects. It would be interesting to test experimentally some of the predictions of the model that have not been tested before, such as the predictions that the optimal number of categories are increasing in the variance of the density.

7 Appendix

All proof are given for the case of an infinite set $X = [a, b]$, and extending the results to the case of a finite X is straightforward, unless stated otherwise.

7.1 Preliminaries

Proof of Lemma 1. We have

$$\begin{aligned} EPE\left(C, v^{t-1}\right) &= \sum_{i=1}^{k} \int_{(x,y)\in C_i} f(x,y)\left(y-\hat{y}_{it}\right)^2 d(x,y) \\ &= \sum_{i=1}^{k} \int_{y\in Y} \left(\int_{x\in X_i} f(x,y)\, dx\right)\left(y-\hat{y}_{it}\right)^2 dy \\ &= \sum_{i=1}^{k} \int_{y\in Y} \Pr\left(x\in X_i\right) f\left(y|x\in X_i\right)\left(y-\hat{y}_{it}\right)^2 dy, \end{aligned}$$

where the last equality uses the definition of $f\left(y|x\in X_i\right)$. Note that

$$\left(y-\hat{y}_{it}\right)^2 = \left(y-\mu_i\right)^2 + \left(\hat{y}_{it}-\mu_i\right)^2 - 2\left(y-\mu_i\right)\left(\hat{y}_{it}-\mu_i\right).$$

Using this we have

$$\begin{aligned} EPE\left(C, v^{t-1}\right) = &\sum_{i=1}^{k} \Pr\left(x\in X_i\right)\left(\int_{y\in Y} f\left(y|x\in X_i\right)\left(y-\mu_i\right)^2 dy + \left(\hat{y}_{it}-\mu_i\right)^2\right) \\ &- \sum_{i=1}^{k} \Pr\left(x\in X_i\right) 2\left(\int_{y\in Y} f\left(y|x\in X_i\right) y dy - \mu_i\right)\left(\hat{y}_{it}-\mu_i\right) \end{aligned}$$

The desired result follows from the facts that the second factor on the right hand side is equal to zero, and

$$\int_{y\in Y} f\left(y|x\in X_i\right)\left(y-\mu_i\right)^2 dy = Var\left(y|x\in X_i\right) = Var\left(y_i\right).$$

■

Proof of Lemma 2. We have

$$\begin{aligned}
EPE\left(C,t\right) &= \mathbb{E}\left[EPE\left(C,v^{t-1}\right)\right] \\
&= \sum_{i=1}^{k}\Pr\left(x\in X_i\right)Var\left(y_i\right) \\
&+ \sum_{i=1}^{k}\Pr\left(x\in X_i\right)\sum_{r=1}^{t-1}\Pr\left(m_{it}=r\right)\mathbb{E}\left[\left(\hat{y}_{it}-\mu_i\right)^2|m_{it}=r\right] \\
&+ \sum_{i=1}^{k}\Pr\left(x\in X_i\right)\Pr\left(m_{it}=0\right)\mathbb{E}\left[\left(\hat{y}_{t}-\mu_i\right)^2|m_{it}=0\right].
\end{aligned}$$

The number of objects in a category, m_{it}, has a binomial distribution as follows

$$\Pr\left(m_{it}=r\right)=\binom{t-1}{r}\left(\Pr\left(x\in X_i\right)\right)^r\left(1-\Pr\left(x\in X_i\right)\right)^{t-1-r}.$$

If $r>0$ then $E\left[\hat{y}_{it}|m_{it}=r\right]=\mu_i$, so

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{y}_{it}-\mu_i\right)^2|m_{it}=r\right] &= Var\left(\hat{y}_{it}|m_{it}=r\right) \\
&= \sum_{j=1}^{r}\frac{1}{r^2}Var\left(y_i|m_{it}=r\right) \\
&= \frac{1}{r}Var\left(y_i\right).
\end{aligned}$$

Plugging this into the expression above yields the desired result. ■

Proof of Proposition 1. (i) First consider minimization of $EPE\left(C,t\right)$. Since $t\leq T$ we require $k<\kappa T$. For any T let $\Psi\left(\kappa,\iota\right)\subseteq\Psi$ be the set of categorizations such that $k<\kappa T$ and such that the number of unconnected subsets of each category is uniformly bounded above by ι. Any categorization $C\in\Psi\left(\kappa,\iota\right)$ with k categories can be described by a set of $T\kappa\iota-1$ points on $[a,b]$ together with a mapping from the induced $(T\kappa\iota)$ subintervals to the set $\{1,2,...,k\}$. Take any mapping ν from subintervals to $\{1,2,...,k\}$. Choosing a categorization among the categorizations that are consistent with the mapping ν is equivalent to choosing a point z in the compact set

$$Z=\left\{z\in[a,b]^{T\kappa\iota-1}:z_j\leq z_{j+1}\forall j\in\{1,...,T\kappa\iota-2\}\right\}.$$

Furthermore, since f is continuous in x, $EPE\left(C,t\right)$ is continuous in z. Hence by Weierstrass' maximum theorem there exists a solution $z^*(\nu)$ to the problem of minimizing $EPE(C,t)$ with respect to categorizations that are consistent with the mapping ν. This was for a given mapping ν from subintervals to $\{1,2,...,k\}$. Since there are only a finite number of mappings from $T\kappa\iota$ subintervals to the set $\{1,2,...,k\}$, the desired result follows for minimization of $EPE(C,t)$.

(ii) Extension to $EPE(C,T,L)$ is straightforward.

(iii) Finally consider minimization of $EPE(C,v^{t-1})$. Recall that by definition all categories in $\Psi(v^{t-1})$ are non-empty under v^{t-1}. Since v^{t-1} is finite, the set $\Psi(v^{t-1})$ is therefore finite. Existence of a solution is therefore trivial. ■

7.2 Ex ante optimality

The following two lemmata will be used in the proof of proposition 2.

Lemma 4 *Let E and F be disjoint intervals. We have*

$$\Pr(x \in E \cup F)\, Var(y|x \in E \cup F) - \sum_{I \in \{E,F\}} \Pr(x \in I)\, Var(y|x \in I) \geq 0,$$

with equality if and only if $\mathbb{E}[y|x \in E] = \mathbb{E}[y|x \in F]$.

Proof of Lemma 4. Note

$$\Pr(x \in I)\, Var(y|x \in I) = \int_{y \in Y} \Pr(x \in I)\, f(y|x \in I)\,(y - \mathbb{E}[y|x \in I])^2\, dy,$$

and

$$\Pr(x \in I)\, f(y|x \in I) = \int_{x \in I} f(x,y)\, dx,$$

for $I \in \{E, F, E \cup F\}$. Using this one can show

$$\begin{aligned} &\Pr(x \in E \cup F)\, Var(y|x \in E \cup F) - \sum_{I \in \{E,F\}} \Pr(x \in I)\, Var(y|x \in I) \\ &= \sum_{I \in \{E,F\}} \Pr(x \in I) \int_{y \in Y} f(y|x \in I)\left((y - \mathbb{E}[y|x \in E \cup F])^2 - (y - \mathbb{E}[y|x \in I])^2\right) dy. \end{aligned}$$

That the left hand side is weakly positive follows from the fact that the function

$$q(z) = \int_{y \in Y} f(y|x \in I)\,(y - z)^2\, dy.$$

is minimized at $z = \mathbb{E}[y|x \in I]$. The weak inequality holds with equality if and only if $\mathbb{E}[y|x \in E \cup F] = \mathbb{E}[y|x \in E] = \mathbb{E}[y|x \in F]$. ■

Lemma 5 *If $k/t \geq \gamma$ then*

$$\sum_{r=1}^{t-1} \Pr\left(m_{it} = r\right) \frac{1}{r} > \frac{\gamma\rho}{(1 - 1/t)}.$$

Proof of Lemma 5. Since r is binomially distributed we have

$$\mathbb{E}\left[m_{it}\right] = \sum_{r=0}^{t-1} \Pr\left(m_{it} = r\right) r = (t-1) \Pr\left(x \in X_i\right).$$

This implies

$$\sum_{r=1}^{t-1} \Pr\left(m_{it} = r\right) r = (t-1) \Pr\left(x \in X_i\right) - \Pr\left(m_{it} = 0\right) \cdot 0 = (t-1) \Pr\left(x \in X_i\right)$$

Since $g\left(x\right) = \frac{1}{x}$ is concave, Jensen's inequality implies

$$\sum_{r=1}^{t-1} \Pr\left(m_{it} = r\right) \frac{1}{r} \geq \frac{1}{\sum_{r=1}^{t-1} \Pr\left(m_{it} = r\right) r} = \frac{1}{(t-1) \Pr\left(x \in X_i\right)}. \tag{17}$$

Let $p_{\max} = \max_i \Pr\left(x \in X_i\right)$ and $p_{\min} = \min_i \Pr\left(x \in X_i\right)$. Note that $p_{\max} < 1 - (k-1)\, p_{\min}$ and since $p_{\min} > \rho p_{\max}$, we have $p_{\max} < 1 - (k-1)\, \rho p_{\max}$ or equivalently

$$p_{\max} < \frac{1}{((k-1)\,\rho + 1)} = \frac{1}{k\rho - \rho + 1}.$$

Since $\rho \in (0,1)$ this implies $p_{\max} < 1/k\rho$. Using this in (17) we get

$$\frac{1}{(t-1) \Pr\left(x \in X_i\right)} \geq \frac{1}{(t-1)\, p_{\max}} > \frac{k\rho}{(t-1)}.$$

Use $k/t \geq \gamma$ to obtain

$$\frac{k\rho}{(t-1)} \geq \frac{t\gamma\rho}{(t-1)} = \frac{\gamma\rho}{(1-1/t)}.$$

■

Proof of Proposition 2. The proof of part (b) is very similar to the proof of part (a), and therefore omitted. The proof of part (a) is as follows:

(i) Assume $k = \sqrt{t}$. If $t \to \infty$ then $k = \sqrt{t} \to \infty$ and $k/t = 1/\sqrt{t} \to 0$. Moreover, by the assumption $\min_i \Pr\left(x \in X_i\right) > \rho \max_i \Pr\left(x \in X_i\right)$ $t \to \infty$ implies $\Pr\left(x \in X_i\right) \to 0$

for all i. Write

$$\sum_{i=1}^{k} \Pr\left(x \in X_i\right) Var\left(y_i\right) = \sum_{i=1}^{k} \left(\int_{x \in X_i} f\left(x\right) dx \right) \int_{y \in Y} f\left(y | x \in X_i\right) \left(y - \mu_i\right)^2 dy.$$

For any t, let all sets X_i be intervals of length $\left(b - a\right)/k$. (Note that it is not sufficient to let $\Pr\left(x \in X_i\right) \to 0$, we need the categories to be convex). If $k \to \infty$ then the right hand side approaches

$$\int_{x \in X} f\left(x\right) \left(\int_{y \in Y} f\left(y|x\right) \left(y - \mathbb{E}\left(y|x\right)\right)^2 dy \right) dx = \int_{x \in X} f\left(x\right) Var\left(y|x\right) dx. \tag{18}$$

Moreover, note that if $k/t \to 0$ then, for then for all i,

$$\sum_{r=1}^{t-1} \Pr\left(m_{it} = r\right) \frac{1}{r} \to 0,$$

Hence if $t \to \infty$

$$EPE\left(C, t\right) \to \int_{x \in X} f\left(x\right) Var\left(y|x\right) dx, \tag{19}$$

and thus

$$EPE\left(C, T, L\right) \to \frac{1}{T - L} \sum_{t=L+1}^{T} \int_{x \in X} f\left(x\right) Var\left(y|x\right) dx = \int_{x \in X} f\left(x\right) Var\left(y|x\right) dx.$$

It follows that for any $\varepsilon > 0$ there are finite numbers k', L' and T' such that if $L > L'$ or $T > T'$, then there is a categorization with $k > k'$, in such that

$$\left| EPE\left(C, T, L\right) - \int_{x \in X} f\left(x\right) Var\left(y|x\right) dx \right| < \varepsilon.$$

(ii) Assume that there is some κ such that $k \leq \kappa$. If $t \to \infty$ then

$$EPE\left(C, t\right) \to \sum_{i=1}^{k} \Pr\left(x \in X_i\right) Var\left(y_i\right).$$

The continuity of f and the assumption that $\mathbb{E}\left[y|x\right] \neq \mu$ for some x, together with lemma 4, implies that

$$\sum_{i=1}^{k} \Pr\left(x \in X_i\right) Var\left(y_i\right) > \int_{x \in X} f\left(x\right) Var\left(y|x\right) dx.$$

Hence not allowing $k \to \infty$ as $t \to \infty$ is suboptimal.

(iii) Now restrict attention to the set of categorizations with $k/t \geq \gamma > 0$. By lemma 5 we have

$$EPE\,(C,t) > \sum_{i=1}^{k} \Pr\,(x \in X_i)\,Var\,(y_i)\left(1 + \frac{\gamma\rho}{(1-1/t)}\right). \tag{20}$$

Let $t \to \infty$. By $t \leq k/\gamma$ this implies $k \to \infty$. As before this also implies $\Pr\,(x \in X_i) \to 0$ for all i. In this limit the right hand side of the above equation is minimized by using convex categories. The reason is that since f is continuous, $\max_{x,x \in X_i} |Var\,(y|x) - Var\,(y|x')|$ approaches zero if X_i is convex, whereas this need not be the case of categories are not convex. If categories are convex and $t \to \infty$, then the right hand side of (20) approaches

$$\int_{x\in X} f\,(x)\,Var\,(y|x)\,dx\,(1+\gamma\rho)\,.$$

Thus, for any ε there is some t' such that if $t > t'$ then

$$\left|\sum_{i=1}^{k} \Pr\,(x \in X_i)\,Var\,(y_i)\left(1 + \frac{\gamma\rho}{(1-1/t)}\right) - \int_{x\in X} f\,(x)\,Var\,(y|x)\,dx\,(1+\gamma\rho)\right| < \varepsilon$$

This implies that there is some t' such that if $t > t'$ then

$$EPE\,(C,t) > \int_{x\in X} f\,(x)\,Var\,(y|x)\,dx\,(1+\gamma\rho)\,.$$

Comparing this with (19) we see that it is suboptimal to restrict attention to categorizations with $t \leq k/\gamma$. ■

Proof of Corollary 1. The proof is omitted since it follows fairly straightforwardly from proposition 2. ■

Proof of Proposition 3. Write $EPE_{E,f}\,(C,t)$ to make the dependence upon f explicit. Suppose C' is an optimal categorization of E at date t given f_0, i.e. $C' \in \arg\min_{C\in\Psi} EPE_{E,f_0}\,(C,t)$, and suppose that there is no other optimal categorization with a lower number of categories. This categorization C' strikes an optimal balance between the goal of having a few large categories in order to minimize the factors $\sum_{r=1}^{t-1} \Pr\,(m_{it} = r)\,\frac{1}{r}$ and $\Pr\,(m_{it} = 0)$ (one of each for each category), and the goal of having many small categories in order to minimize the factors $Var\,(y_i)$ (one for each category). Decreasing the number of categories will lead to an increase in at least one of the factors $Var\,(y_i)$ and a decrease in at least one of the factors $\sum_{r=1}^{t-1} \Pr\,(m_{it} = r)\,\frac{1}{r}$ and $\Pr\,(m_{it} = 0)$. The former effect will dominate the latter so that the total effect will be an increase in prediction error – otherwise C' would not be an optimal categorization with a minimal number of categories.[16]

Now suppose one uses the same categorization C' when the distribution is f_1 (rather than f_0). All the factors $\sum_{r=1}^{t-1} \Pr(m_{it}=r)\frac{1}{r}$ and $\Pr(m_{it}=0)$ are smaller under f_1 than under f_0. But we have

$$\frac{f(x)}{\Pr(x \in X_i)} = \frac{f(x)}{\int_{x \in X_i} f(x)\,dx},$$

and

$$\frac{f_1(x)}{\int_{x \in X_i} f_1(x)\,dx} = \frac{\alpha f_0(x)}{\int_{x \in X_i} \alpha f_0(x)\,dx} = \frac{f_0(x)}{\int_{x \in X_i} f_0(x)\,dx}.$$

so from the expressions for $\mathbb{E}[Var(y|x)\,|x \in X_i]$ and $Var(\mathbb{E}[y|x]\,|x \in X_i)$ together with equation (7), one sees that all the factors $Var(y_i)$ and $\mathbb{E}\left[(\hat{y}_t-\mu_i)^2\,|m_{it}=0\right]$ are the same under f_0 and f_1. Also all the factors $\mathbb{E}\left[(\hat{y}_t-\mu_i)^2\,|m_{it}=0\right]$ are unaffected. Hence, keeping C' fixed, the only difference between f_0 and f_1 is that the factors $\sum_{r=1}^{t-1} \Pr(m_{it}=r)\frac{1}{r}$ and $\Pr(m_{it}=0)$, are smaller under f_1 than under f_0. Since it was suboptimal to decrease the number of categories relative to C' under f_0 it must be (even more) suboptimal to decrease the number of categories relative to C' under f_1. ■

Proof of Proposition 4. Write $EPE_{f_0}(C,t)$ and $EPE_{f_1}(C,t)$ to make the dependence on the distribution explicit. Note that $f_0(x) = f_1(x) = f(x)$ for all x, so that $\mathbb{E}[y|x]$, and $\Pr(m_{it}=r)$ are the same for f_0 and f_1 for all x, i, t, and r.

(i) Use (7) and note that $Var(\mathbb{E}[y|x]\,|x \in X_i)$ is the same under f_0 and f_1. It follows that

$$\begin{aligned}
Var_{f_1}(y_i) - Var_{f_0}(y_i) &= \left(\mathbb{E}_{f_1}[Var_{f_1}(y|x)\,|x \in X_i] - \mathbb{E}_{f_0}[Var_{f_0}(y|x)\,|x \in X_i]\right) \\
&= \mathbb{E}[Var_{f_1}(y|x) - Var_{f_0}(y|x)\,|x \in X_i] \\
&= \frac{1}{\Pr(x \in X_i)} \int_{x \in X_i} f(x)\left(Var_{f_1}(y|x) - Var_{f_0}(y|x)\right)dx.
\end{aligned}$$

(ii) For $j \in \{0,1\}$ we have

$$\begin{aligned}
\mathbb{E}_{f_j}\left[(\hat{y}_t-\mu_i)^2\,|m_{it}=0\right] &= \mathbb{E}_{f_j}\left[(\hat{y}_t-\mu)^2\,|m_{it}=0\right] \\
&\quad + (\mu-\mu_i)^2 + 2(\mu-\mu_i)^2\left(\mathbb{E}[\hat{y}_t|m_{it}=0]-\mu\right),
\end{aligned}$$

[16] The effect on the factors $\mathbb{E}\left[(\hat{y}_t-\mu_i)^2\,|m_{it}=0\right]$ of increasing the number of categories is ambiguous, but if these terms are decreased by increasing the number of categories it still must be the case that the total effect on expected prediction error, of increasing the number of categories, is positive.

and

$$\begin{aligned}
\mathbb{E}_{f_j}\left[(\hat{y}_t-\mu)^2\,|m_{it}=0\right] &= Var_{f_j}\left(\hat{y}_t|m_{it}=0\right)\\
&= \frac{1}{t-1}Var_{f_j}\left(y|x\notin X_i\right)\\
&= \frac{1}{t-1}\left(\mathbb{E}\left[Var_{f_j}\left(y|x\right)|x\notin X_i\right]+Var\left(\mathbb{E}\left[y|x\right]|x\notin X_i\right)\right).
\end{aligned}$$

so

$$\begin{aligned}
&\mathbb{E}_{f_1}\left[(\hat{y}_t-\mu_i)^2\,|m_{it}=0\right]-\mathbb{E}_{f_0}\left[(\hat{y}_t-\mu_i)^2\,|m_{it}=0\right]\\
&=\mathbb{E}_{f_1}\left[(\hat{y}_t-\mu)^2\,|m_{it}=0\right]-\mathbb{E}_{f_0}\left[(\hat{y}_t-\mu)^2\,|m_{it}=0\right]\\
&=\frac{1}{t-1}\left(\mathbb{E}\left[Var_{f_1}\left(y|x\right)|x\notin X_i\right]-\mathbb{E}\left[Var_{f_0}\left(y|x\right)|x\notin X_i\right]\right)\\
&=\frac{1}{t-1}\mathbb{E}\left[Var_{f_1}\left(y|x\right)-Var_{f_0}\left(y|x\right)|x\notin X_i\right]\\
&=\frac{1}{\Pr\left(x\in X_i\right)}\frac{1}{t-1}\left(\int_{x\notin X_i}f\left(x\right)\left(Var_{f_1}\left(y|x\right)-Var_{f_0}\left(y|x\right)\right)dx\right).
\end{aligned}$$

(iii) From (i) and (ii) it follows that

$$EPE_{f_1}\left(C,t\right)-EPE_{f_0}\left(C,t\right)=M_1+M_2, \tag{21}$$

where

$$M_1=\sum_{i=1}^{k}\left(\int_{x\in X_i}f\left(x\right)\left(Var_{f_1}\left(y|x\right)-Var_{f_0}\left(y|x\right)\right)dx\right)\left(1+\sum_{r=1}^{t-1}\Pr\left(m_{it}=r\right)\frac{1}{r}\right),$$

and

$$M_2=\frac{1}{t-1}\sum_{i=1}^{k}\Pr\left(m_{it}=0\right)\left(\int_{x\notin X_i}f\left(x\right)\left(Var_{f_1}\left(y|x\right)-Var_{f_0}\left(y|x\right)\right)dx\right).$$

For any categorization C' with $k>1$, note that both M_1 and M_2 can be decreased by changing to some categorization with fewer categories. Suppose that C' is an optimal categorization for t, and f_1, i.e. $C'\in\arg\min_{C\in\Psi}EPE_{f_1}\left(C,t\right)$, and suppose that there is no other optimal categorization with a lower number of categories. From the right hand side of (21) we see that any categorization that minimizes $EPE_{f_0}\left(C,t\right)$ will have at least as many categories as C'. ■

Proof of Proposition 5. We assume that the subject observes at least one object in each category during the learning phase. Thus the maximal number of categories in a category is now $t-1-(k-1)=t-1$ rather that $t-1$. This requires us to revise the expression for the expected prediction error as follows:

$$EPE(C,t)=\sum_{i=1}^{k}\Pr(x\in X_i)\,Var(y_i)\left(1+\sum_{r=0}^{t-k}\Pr(m_{it}=r)\frac{1}{r+1}\right),$$

where m_{it} has a binomial distribution

$$\Pr(m_{it}=r)=\binom{t-k}{r}(\Pr(x\in X_i))^r(1-\Pr(x\in X_i))^{t-k-r}.$$

Note that the assumption that x is uniformly distributed on X, implies that $f(x)=1$ for all $x\in X$, and hence that $f(x,y)=f(y|x)$. Then derive the variance of y in interval $A_i=[a_i,b_i)$. We have

$$\Pr(x\in A_i)=b_i-a_i,$$

and

$$f(y|x\in A_i)=\frac{\int_{x\in A}f(y,x)\,dx}{\Pr(x\in A_i)}=\frac{1}{b_i-a_i}\int_{x\in A_i}f(y|x)\,dx,$$

and

$$\mathbb{E}(y|x\in A_i)=\left(\alpha+\beta\frac{(a_i+b_i)}{2}\right).$$

Using this we get, after a fair amount of manipulation,

$$Var(y_i)=\frac{\beta^2(b_i-a_i)^2}{12}+\sigma^2. \tag{22}$$

(a) Now we show that the optimal categories are intervals on the x-axis. Take a categorization C where not all categories are convex. Without loss of generality one can assume that there is a category C_α such that $X_\alpha=\cup_{s=1}^{S}[a_s,b_s)$, with $b_s<a_{s+1}$. Let $EPE_\alpha(C,t)$ denote the expected prediction error for objects in this category;

$$EPE_\alpha(C,t)=Var(y_\alpha)\left(1+\sum_{r=0}^{t-k}\Pr(m_{\alpha t}=r)\frac{1}{r+1}\right).$$

Consider a categorization C' that is a modification of C such that $X'_\beta=[a_1,b)$ where $b=a_1+\sum_{s=1}^{S}(b_s-a_s)$. The other categories are only moved to the right so that if, under categorization C the point $p>a_1$ was a boundary point between two categories then, under categorization C' this boundary is located at the point $p+\sum_{s=1}^{S}(b_s-a_s)$.

Let $EPE_\beta(C', t)$ denote the expected prediction error for objects in category $C_\beta \in C'$;

$$EPE_\beta(C,t) = Var(y_\beta)\left(1 + \sum_{r=0}^{t-k} \Pr(m_{\beta t} = r)\frac{1}{r+1}\right).$$

Since $\Pr(m_{\alpha t} = r) = \Pr(m_{\beta t} = r)$ and $Var(y_\alpha) > Var(y_\beta)$ we have $EPE_\beta(C', t) < EPE_\alpha(C, t)$. From equation 22 we see that the expected prediction error for objects in the other categories are unaffected so $EPE(C', t) < EPE(C, t)$. Hence the categorization with a convex category is better than the one with a non-convex category.

We have shown that an optimal categorization with k categories has $X_i = [a_i, b_i)$ for $i \in \{1, ..., k-1\}$ and $X_k = [a_k, b_k] = [a_k, 1]$. Letting $d_i = b_i - a_i$, we seek a categorization that minimizes

$$EPE(C,t) = \sum_{i=1}^{k}(b_i - a_i)\left(\frac{\beta^2 (b_i - a_i)^2}{12} + \sigma^2\right)\left(1 + \sum_{r=0}^{t-k}\Pr(m_{it} = r)\frac{1}{r+1}\right),$$

where

$$\Pr(m_{it} = r) = \frac{(t-k)!}{r!\,(t-k-r)!}((b_i - a_i))^r (1 - (b_i - a_i))^{t-k-r}.$$

Since $EPE(C,t)$ is quadratic in $b_i - a_i$ it is optimal to have $b_i - a_i = 1/k$ for all i. Since we have assumed $T - L = 1$ this finishes the proof of (a).

(b) With $b_i - a_i = 1/k$ for all i, the probability $\Pr(m_{it} = r)$ is the same for all i so write $\Pr(m_{it} = r) = \Pr(m_t = r)$. We have

$$EPE(C,t) = \left(\frac{\beta^2}{12}\left(\frac{1}{k}\right)^2 + \sigma^2\right)\left(1 + \sum_{r=1}^{t-k}\Pr(m_t = r)\frac{1}{r+1}\right).$$

Let C' and C'' be categorizations with k' and k'' categories respectively. It is easy to verify that

$$EPE(C'', t) - EPE(C', t) = \beta^2 M_1 + \sigma^2 M_2.$$

where

$$M_1 = \frac{1}{12}\left(\left(\frac{1}{k''}\right)^2\left(1 + \sum_{r=0}^{t-k''}\Pr(m''_{it} = r)\frac{1}{r+1}\right) - \left(\frac{1}{k'}\right)^2\left(1 + \sum_{r=0}^{t-k'}\Pr(m'_{it} = r)\frac{1}{r+1}\right)\right)$$

and

$$M_2 = \left(\sum_{r=0}^{t-k''}\Pr(m''_{it} = r)\frac{1}{r+1} - \sum_{r=0}^{t-k'}\Pr(m'_{it} = r)\frac{1}{r+1}\right)$$

Note that $M_2 > 0$. Thus $EPE\left(C'', t\right) - EPE\left(C', t\right)$ is increasing in σ^2. If $M_1 > 0$ then $EPE\left(C'', t\right) > EPE\left(C', t\right)$ for all β and σ^2. If $M_1 < 0$ then $EPE\left(C'', t\right) - EPE\left(C', t\right)$ is decreasing in β and attains negative values if β is large enough. ■

7.3 Ex post optimality

Lemma 6 ***(a)*** *For any categorization C*

$$P \lim_{t\to\infty} EPE\left(C, v^{t-1}\right) = \sum_{i=1}^{k} \Pr\left(x \in X_i\right) Var\left(y_i\right).$$

(b) *Consider an initial database v^{t_0-1} and any categorization $C \in \Psi\left(v^{t_0-1}\right)$. Add $t - t_0$ objects to the data base.*

$$P \lim_{t\to\infty} \widehat{EPE\left(C, v^{t-1}\right)} = \sum_{i=1}^{k} \Pr\left(x \in X_i\right) Var\left(y_i\right).$$

Proof of Lemma 6. **(a)** Consider a category $C_i \in C$. Suppose that $m_{it} \geq 2$. It can be verified that

$$\left(\hat{y}_{it} - \mu_i\right)^2 = \left(\frac{1}{m_{it}} \sum_{s\in D_{it}} y_s\right)^2 + \mu_i^2 - 2\mu_i \frac{1}{m_{it}} \sum_{s\in D_{it}} y_s.$$

Let $m_{it} \to \infty$. Since, for each category i, $\{y_s\}$ is an i.i.d. sequence with $\mathbb{E}\left[y_s\right] = \mu_i$ we can use Kinchine's law of large numbers and Slutsky's lemma to conclude that

$$P \lim_{m_{it}\to\infty} \left(\hat{y}_{it} - \mu_i\right)^2 = \left(\mu_i^2 + \mu_i^2 - 2\mu_i\mu_i\right) = 0.$$

In other words, for any $\varepsilon > 0$ and $\delta \in (0, 1)$, there is an M such that if $m_{it} > M$ then

$$\Pr\left(\left(\hat{y}_{it} - \mu_i\right)^2 < \varepsilon\right) > \delta^{1/2}.$$

Moreover, for any M there is a t' such that if $t > t'$ then

$$\Pr\left(m_{it} > M\right) > \delta^{1/2}.$$

This implies that, for any $\varepsilon > 0$ and $\delta \in (0, 1)$, there is a t' such that if $t > t'$ then

$$\Pr\left(\left(\hat{y}_{it} - \mu_i\right)^2 < \varepsilon\right) > \delta.$$

Thus, for any for any $\varepsilon > 0$ and $\delta \in (0,1)$, there is a t' such that if $t > t'$ then

$$\Pr\left(\sum_{i=1}^{k} \Pr(x \in X_i)(\hat{y}_{it} - \mu_i)^2 < \varepsilon\right) > \delta.$$

Since $EPE(C, v^{t-1}) \geq 0$, the desired result follows.

(b) Similar to the proof of (a), using the standard result $P\lim_{m_{it}\to\infty} s_{it}^2 = Var(y_i)$. ■

Proof of Proposition 6. **(a)** Let C' be a categorization with k' categories. Since $\mathbb{E}[y|x]$ is not constant across X, there exists a $\gamma > 0$, and a categorization C'' with $k'' > k'$ categories, such that

$$\sum_{i=1}^{k'} \Pr(x \in X_i') Var(y_i) - \sum_{i=1}^{k''} \Pr(x \in X_i'') Var(y_i) = \gamma.$$

By lemma 6;

$$P\lim_{t\to\infty}\left[EPE\left(C', v^{t-1}\right) - EPE\left(C'', v^{t-1}\right)\right] = \gamma.$$

The desired result follows.

(b) Analogous to (a). ■

Proof of Proposition 7. Follows from lemma 6 in a way similar to the proof of proposition 6. For a given t, increasing α leads to an increase in the probability that $m_{it} > M$. ■

Proof of Proposition 8. **(a)** Without loss of generality suppose that C' has one category, named 0, and C'' has two categories, named 1 and 2. Without loss of generality assume that C'' is chosen so that

$$Var(\mathbb{E}[y|x] \,|x \in X_i) < Var(\mathbb{E}[y|x] \,|x \in X_0) \tag{23}$$

for $i \in \{1,2\}$. Fix $m_{1t} \geq 1$ and $m_{2t} \geq 1$ and only consider data bases with these numbers of objects in each category. We have

$$\begin{aligned} EPE\left(C'', v^{t-1}\right) - EPE\left(C', v^{t-1}\right) &= \Pr(x \in X_1)\left(Var(y_1) + (\hat{y}_{1t} - \mu_1)^2\right) \\ &+ \Pr(x \in X_2)\left(Var(y_2) + (\hat{y}_{2t} - \mu_2)^2\right) \\ &- \Pr(x \in X_0)\left(Var(y_0) + (\hat{y}_t - \mu_0)^2\right). \end{aligned}$$

The assumption that $Var(y|x)$ is constant across X implies

$$\begin{aligned}\mathbb{E}\left[Var(y|x)|x\in X_0\right] &= \Pr(x\in X_1)\,\mathbb{E}\left[Var(y|x)|x\in X_1\right]\\ &\quad+\Pr(x\in X_2)\,\mathbb{E}\left[Var(y|x)|x\in X_2\right],\end{aligned}$$

Together with the decomposition of $Var(y_i)$ given by (7) this implies

$$\begin{aligned}EPE\left(C'',v^{t-1}\right)-EPE\left(C',v^{t-1}\right) &= \Pr(x\in X_1)\left(Var\left(\mathbb{E}[y|x]\,|x\in X_1\right)+\left(\hat{y}_{1t}-\mu_1\right)^2\right)\\ &\quad+\Pr(x\in X_2)\left(Var\left(\mathbb{E}[y|x]\,|x\in X_2\right)+\left(\hat{y}_{2t}-\mu_2\right)^2\right)\\ &\quad-\Pr(x\in X_0)\left(Var\left(\mathbb{E}[y|x]\,|x\in X_0\right)+\left(\hat{y}_{t}-\mu_0\right)^2\right)\\ &= M_1+M,\end{aligned}$$

where

$$M_1=\sum_{i\in\{1,2\}}\Pr(x\in X_i)\left(Var\left(\mathbb{E}[y|x]\,|x\in X_i\right)-Var\left(\mathbb{E}[y|x]\,|x\in X_0\right)\right).$$

and

$$M=\sum_{i\in\{1,2\}}\Pr(x\in X_i)\left(\left(\hat{y}_{it}-\mu_1\right)^2-\left(\hat{y}_t-\mu_0\right)^2\right)$$

Note that M_1 is independent of $Var(y|x)$. We can rewrite M as

$$M=\Pr(x\in X_1)\left(\frac{\sigma_1^2}{m_{1t}}\right)Z_1+\Pr(x\in X_2)\left(\frac{\sigma_2^2}{m_{2t}}\right)Z_2-\Pr(x\in X)\left(\frac{\sigma_0^2}{m_{0t}}\right)Z_0$$

Where

$$Z_i=\left(\frac{\hat{y}_{it}-\mu_i}{\sigma_i/\sqrt{m_{it}}}\right)^2,$$

for $i\in\{0,1,2\}$. Since $f(y|x)$ is normally distributed $f(y|x\in X_i)$ is normally distributed, with some variance σ_i^2. Then $\hat{y}_{it}$ (being the average of i.i.d. draws) is normally distributed with variance σ_i^2/m_{it}, and

$$\frac{\hat{y}_{it}-\mu_i}{\sigma_i/\sqrt{m_{it}}}\sim N(0,1) \tag{24}$$

It follows that

$$\left(\frac{\hat{y}_{it}-\mu_i}{\sigma_i/\sqrt{m_{it}}}\right)^2\sim\chi^2_{(1)},$$

for $i \in \{0,1,2\}$.[17] Using (7), M can be further decomposed as $M = M_2 + M_3$ where

$$M_2 = \sum_{i\in\{1,2\}} \Pr(x \in X_i) \left(\frac{Var(\mathbb{E}[y|x]\,|x \in X_i)}{m_{it}} Z_i - \frac{Var(\mathbb{E}[y|x]\,|x \in X_i)}{m_{0t}} Z_0 \right),$$

and

$$M_3 = Var(y|x) \sum_{i\in\{1,2\}} \Pr(x \in X_i) \left(\frac{1}{m_{it}} Z_i - \frac{1}{m_{0t}} Z_0 \right).$$

Thus we have found that

$$EPE\left(C'', v^{t-1}\right) - EPE\left(C', v^{t-1}\right) = M_1 + M_2 + M_3,$$

where M_1 and M_2 are independent of $Var(y|x)$. Note that the assumption (13) implies $M_1 > 0$. It also implies that, for any realization (z_1, z_2, z_0) of (Z_1, Z_2, Z_0), it holds that if $M_3 > 0$ then $M_2 > 0$. Pick any realization (z_1, z_2, z_0). If $M_3 > 0$ then $M_2 > 0$ and hence $EPE(C'', v^{t-1}) - EPE(C', v^{t-1}) > 0$ independently of $Var(y|x)$. If $M_3 < 0$ then $EPE(C'', v^{t-1}) - EPE(C', v^{t-1})$ is increasing in $Var(y|x)$. This latter fact shows that the probability of $EPE(C'', v^{t-1}) > EPE(C', v^{t-1})$ is increasing in $Var(y|x)$. This was for given numbers m_{1t} and m_{2t}, but the same reasoning holds for any choice of $m_{1t} \geq 1$ and $m_{2t} \geq 1$.

(b) Without loss of generality suppose that C' has one category, named 0, and C'' has two categories, named 1 and 2. Without loss of generality assume that C'' is chosen so that $Var(\mathbb{E}[y|x]\,|x \in X_i) < Var(\mathbb{E}[y|x]\,|x \in X_0)$ for $i \in \{1,2\}$. Fix $m_{1t} \geq 2$ and $m_{2t} \geq 2$ and only consider data bases with these numbers of objects in each category. We have

$$\begin{aligned}\widehat{EPE(C'', v^{t-1})} - \widehat{EPE(C', v^{t-1})} &= \frac{1}{t-1}\left((m_{1t}+1)\, s_{1t}^2 + (m_{2t}+1)\, s_{2t}^2 - (m_{0t}+1)\, s_{0t}^2 \right) \\ &= \frac{1}{t-1}\left(\frac{m_{1t}+1}{m_{1t}-1}\sigma_1^2 Z_1 + \frac{m_{2t}+1}{m_{2t}-1}\sigma_2^2 Z_2 - \frac{m_{0t}+1}{m_{0t}-1}\sigma_0^2 Z_0 \right),\end{aligned}$$

where

$$Z_i = \frac{m_{it}-1}{\sigma_i^2} s_{it}^2.$$

Since $y|x \sim N(\mathbb{E}[y|x], \sigma^2)$ we have $Z_i \sim \chi^2_{(m_{it}-1)}$. Note that this distribution is independent of σ_i^2. We can write

$$\widehat{EPE(C'', v^{t-1})} - \widehat{EPE(C', v^{t-1})} = \frac{1}{t-1}(M_1 + M),$$

[17] If $m_{it} \to \infty$ is large then (24) holds by virtue of the central limit theorem, even if $y|x$ is not normally distributed.

where

$$M_1 = \frac{m_{1t}+1}{m_{1t}-1} Var\left(\mathbb{E}\left[y|x\right]|x \in X_1\right) Z_1 + \frac{m_{2t}+1}{m_{2t}-1} Var\left(\mathbb{E}\left[y|x\right]|x \in X_2\right) Z_2$$
$$- \frac{m_{0t}+1}{m_{0t}-1} Var\left(\mathbb{E}\left[y|x\right]|x \in X_0\right) Z_0,$$

and

$$M_2 = Var\left(y|x\right)\left(\frac{m_{1t}+1}{m_{1t}-1} Z_1 + \frac{m_{2t}+1}{m_{2t}-1} Z_2 - \frac{m_{0t}+1}{m_{0t}-1} Z_0\right).$$

Note that M_1 is independent of $Var\left(y|x\right)$. The assumption (13) implies that, for any realization (z_1, z_2, z_0) of (Z_1, Z_2, Z_0), it holds that if $M_2 < 0$ then $M_1 < 0$. If $M_2 < 0$ then $M_1 < 0$ so that $\widehat{EPE\left(C'', v^{t-1}\right)} < \widehat{EPE\left(C', v^{t-1}\right)}$ regardless of $Var\left(y|x\right)$. If $M_2 > 0$ and $M_1 > 0$ then $\widehat{EPE\left(C'', v^{t-1}\right)} > \widehat{EPE\left(C', v^{t-1}\right)}$ regardless of $Var\left(y|x\right)$. If $M_2 > 0$ and $M_1 < 0$ then $\widehat{EPE\left(C'', v^{t-1}\right)} - \widehat{EPE\left(C', v^{t-1}\right)}$ is increasing in $Var\left(y|x\right)$. This latter fact shows that the probability of $\widehat{EPE\left(C'', v^{t-1}\right)} > \widehat{EPE\left(C', v^{t-1}\right)}$ is increasing in $Var\left(y|x\right)$. This was for given numbers m_{1t} and m_{2t}, but the same reasoning holds for any choice of $m_{1t} \geq 2$ and $m_{2t} \geq 2$. ■

Proof of Proposition 9. **(a)** The proof is similar to the proof of proposition 8. Without loss of generality suppose that C' has one category, named 0, and C'' has two categories, named 1 and 2. Fix $m_{1t} \geq 1$ and $m_{2t} \geq 1$ and only consider data bases with these numbers of objects in each category. Similar to the proof of proposition 8 we can write

$$EPE\left(C'', v^{t-1}\right) - EPE\left(C', v^{t-1}\right) = M_1 + M_2$$

where, using (22),

$$M_1 = \sum_{i \in \{1,2\}} \Pr\left(x \in X_i\right)\left(Var\left(\mathbb{E}\left[y|x\right]|x \in X_i\right) - Var\left(\mathbb{E}\left[y|x\right]|x \in X_0\right)\right)$$
$$+ \sum_{i \in \{1,2\}} \Pr\left(x \in X_i\right)\left(\frac{Var\left(\mathbb{E}\left[y|x\right]|x \in X_i\right)}{m_{it}} Z_i - \frac{Var\left(\mathbb{E}\left[y|x\right]|x \in X_0\right)}{m_{0t}} Z_0\right)$$
$$= \sum_{i \in \{1,2\}} \Pr\left(x \in X_i\right)\frac{\beta^2}{12}\left(\left(1 + \frac{1}{m_{it}} Z_i\right) d_i^2 - \left(1 - \frac{1}{m_{0t}} Z_0\right) d_0^2\right),$$

and

$$M_2 = \sum_{i\in\{1,2\}} \Pr(x \in X_i) \left(\frac{\mathbb{E}[Var(y|x)|x \in X_i]}{m_{it}} Z_i - \frac{\mathbb{E}[Var(y|x)|x \in X_0]}{m_{0t}} Z_0 \right)$$
$$= \sum_{i\in\{1,2\}} \Pr(x \in X_i) \sigma^2 \left(\frac{1}{m_{it}} Z_i - \frac{1}{m_{0t}} Z_0 \right).$$

Note that, for any realization (z_1, z_2, z_0) of (Z_1, Z_2, Z_0); if $M_2 < 0$ then $M_1 < 0$. Pick any realization (z_1, z_2, z_0). If $M_2 < 0$ then $M_1 < 0$ so $EPE(C'', v^{t-1}) - EPE(C', v^{t-1}) < 0$ regardless of β and σ^2. If $M_2 > 0$ and $M_1 > 0$ then $EPE(C'', v^{t-1}) - EPE(C', v^{t-1}) > 0$ regardless of β and σ^2. If $M_2 > 0$ and $M_1 < 0$ then $EPE(C'', v^{t-1}) - EPE(C', v^{t-1})$ is decreasing in β and increasing in σ^2.

(b) From the proof of proposition 8, using the expressions for $Var(\mathbb{E}[y|x]|x \in X_i)$ and $\mathbb{E}[Var(y|x)|x \in X_0]$ we have

$$M_1 = \frac{\beta^2}{12} \left(d_1^2 \frac{m_{1t}+1}{m_{1t}-1} Z_1 + d_2^2 \frac{m_{2t}+1}{m_{2t}-1} Z_2 - d_0^2 \frac{m_{0t}+1}{m_{0t}-1} Z_0 \right),$$

and

$$M_2 = \sigma^2 \left(\frac{m_{1t}+1}{m_{1t}-1} Z_1 + \frac{m_{2t}+1}{m_{2t}-1} Z_2 - \frac{m_{0t}+1}{m_{0t}-1} Z_0 \right).$$

We know the effect of σ^2 from proposition 8. To see the effect of β, note that if $M_2 < 0$ then $M_1 < 0$ so that $\widehat{EPE(C'', v^{t-1})} < \widehat{EPE(C', v^{t-1})}$ regardless of β. If $M_2 > 0$ and $M_1 > 0$ then $\widehat{EPE(C'', v^{t-1})} > \widehat{EPE(C', v^{t-1})}$ regardless of β. If $M_2 > 0$ and $M_1 < 0$ then $\widehat{EPE(C'', v^{t-1})} - \widehat{EPE(C', v^{t-1})}$ is decreasing in β. ■

References

Abreu, D. and Rubinstein, A. (1988), 'The structure of nash equilibrium in repeated games with finite automata', *Econometrica* **56**, 1259–1282.

Al-Najjar, N. I. and Pai, M. (2010), Coarse decision making. Manuscript.

Anderson, J. R. (1990), *The Adaptive Character of Thought*, Erlbaum, Hillsdale, NJ.

Anderson, J. R. (1991), 'The adaptive nature of human categorization', *Psychological Review* **98**(3), 409–429.

Berlin, B., Breedlove, D. and Raven, P. (1973), 'General principles of classification and nomenclature in folk biology', *American Anthropologist* **74**, 214–242.

Bernstein, R. (1995), *Style Investing*, Wiley, New York.

Binmore, K. (2007), Making decisions in large worlds. Working paper, University College, London.

Chater, N. (1999), 'The search for simplicity: A fundamental cognitive principle?', *Quarterly Journal of Experimental Psychology* **52**, 273–302.

Coval, J. D., Jurek, J. and Stafford, E. (2009), 'The economics of structured finance', *Journal of Economic Perspectives* **23**(1), 3–25.

Dow, J. (1991), 'Search decisions with limited memory', *Review of Economic Studies* **58**, 1–14.

Franklin, A., Clifford, A., Williamson, E. and Davies, I. (2005), 'Color term knowledge does not affect categorical perception in toddlers', *Journal of Experimental Child Psychology* **90**, 114–141.

Fryer, R. and Jackson, M. O. (2008), 'A categorical model of cognition and biased decision making', *The B.E. Journal of Theoretical Economics (Contributions)* **8**(1), 1–42.

Gärdenfors, P. (2000), *Conceptual Spaces: The Geometry of Thought*, MIT Press, Cambridge, MA.

Gilboa, I., Lieberman, O. and Schmeidler, D. (2006), 'Empirical similarity', *Review of Economics and Statistics* **88**, 433–444.

Gilboa, I., Postlewaite, A. and Schmeidler, D. (2008), 'Probabilities in economic modeling', *Journal of Economic Perspectives* **22**, 173–188.

Gilboa, I. and Schmeidler, D. (2003), 'Inductive inference: An axiomatic approach', *Econometrica* **71**, 1–26.

Goldstone, R. L. (1994), 'The role of similarity in categorization: Providing a groundwork', *Cognition* **52**, 125–157.

Goodman, N. (1955), *Fact, Fiction, and Forecast*, Harvard University Press, Cambridge, MA.

Hauser, M., MacNeilage, P. and Ware, M. (1997), 'Numerical representations in primates', *Proceeding of the National Academy of the Sciences* **93**, 1514–1517.

Herrnstein, R. J., Loveland, D. H. and Cable, C. (1976), 'Natural concepts in pigeons', *Journal of Experimental Psychology: Animal Behavior Processes* **2**, 285–302.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999), 'Data clustering: A review', *ACM Computing Surveys* **31**, 264–323.

Jehiel, P. (2005), 'Analogy-based expectation equilibrium', *Journal of Economic Theory* **123**, 81–104.

Jehiel, P. and Samet, D. (2007), 'Valuation equilibrium', *Theoretical Economics* **2**, 163–185.

Johnson, K. E. and Mervis, C. B. (1998), 'Impact of intuitive theories on feature recruitment throughout the continuum of expertise', *Memory and Cognition* **26**(2), 382–401.

Jones, G. Y. (1983), 'Identifying basic categories', *Psychological Bulletin* **94**, 423–428.

Kay, P. and Maffi, L. (1999), 'Color appearance and the emergence and evolution of basic color lexicons', *American Anthropologist* **101**(1), 743–760.

Krueger, J. and Clement, R. (1994), 'Memory-based judgments about multiple categories', *Journal of Personality and Social Psychology* **67**, 35–47.

Kuhn, T. S. (1970), *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago.

Laurence, S. and Margolis, E. (1999), Concepts and cognitive science, *in* E. Margolis and S. Laurence, eds, 'Concepts: Core Readings', MIT Press, Cambridge, MA, pp. 3–81.

Malt, B. C., Ross, B. H. and Murphy, G. L. (1995), 'Predicting features for members of natural categories when categorization is uncertain', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**, 646–661.

Medin, D. L. (1983), Structural principles of categorization, *in* B. Shepp and T. Tighe, eds, 'Interaction: Perception, Development and Cognition', Erlbaum, Hillsdale, NJ, pp. 203–230.

Mullainathan, S. (2002), Thinking through categories. Mimeo, MIT.

Murphy, G. L. (2002), *The Big Book of Concepts*, MIT Press, Cambridge, MA.

Murphy, G. L. and Ross, B. H. (1994), 'Predictions from uncertain categorizations', *Cognitive Psychology* **27**, 148–193.

Peski, M. (2010), 'Prior symmetry, similarity-based reasoning, and endogenous categorization', *Journal of Economic Theory* **146**, 111–140.

Pothos, E. M. and Chater, N. (2002), 'A simplicity principle in unsupervised human categorization.', *Cognitive Science* **26**, 303–343.

Punj, G. and Moon, J. (2002), 'Positioning options for achieving brand association: A psychological categorization framework', *Journal of Business Research* **55**, 257–283.

Quine, W. V. O. (1969), Natural kinds, *in* 'Ontological Relativity and Other Essays', Columbia Univ. Press.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D. and Boyles-Brian, P. (1976), 'Basic objects in natural categories', *Cognitive Psychology* **8**, 382–439.

Smith, W. (1965), 'Product differentiation and market segmentation as alternative marketing strategies', *Journal of Marketing* **3-8.**, 3–8.

Solomon, K., Medin, D. and Lynch, E. (1999), 'Concepts do more than categorize.', *Trends in Cognitive Science* **3**, 99–105.

Tanaka, J. W. and Taylor, M. (1991), 'Object categories and expertise: Is the basic level in the eye of the beholder', *Cognitive Psychology* **23**, 457–482.