Optimal Categorization*

forthcoming in the Journal of Economic Theory

Erik Mohlin†
University of Oxford

April 3, 2014.

Abstract

This paper studies categorizations that are optimal for the purpose of making predictions. A subject encounters an object (x, y). She observes the first component, x, and has to predict the second component, y. The space of objects is partitioned into categories. The subject determines what category the new object belongs to on the basis of x, and predicts that its y-value will be equal to the average y-value among the past observations in that category. The optimal categorization minimizes the expected prediction error. The main results are driven by a bias-variance trade-off: The optimal size of a category around x, is increasing in the variance of y conditional on x, decreasing in the variance of the conditional mean, decreasing in the size of the data base, and decreasing in the marginal density over x.

Keywords: Categorization; Priors; Coarse Reasoning; Similarity-Based Reasoning; Case-Based Reasoning; Regression Trees.
JEL codes: D83; C72.

*This paper has benefited from comments by Ola Andersson, Stefano Demichelis, Tore Ellingsen, Drew Fudenberg, Philippe Jehiel, Topi Miettinen, Robert Östling, Ron Peretz, Roland Poellinger, Rani Spiegler, Tomasz Strzalecki, and Jörgen Weibull, as well as participants at presentations at the Third Nordic Workshop in Behavioral and Experimental Economics, SUDSWEc 2009, the Stockholm School of Economics, the Stony Brook Workshop on Bounded Rationality, Decisions Games and Logic 12, and the London School of Economics. Comments from an anonymous associate editor and an anonymous referee improved the paper. Financial support from the Jan Wallander and Tom Hedelius Foundation, and the European Research Council, Grant no. 230251, is gratefully acknowledged.

†Nuffield College and Department of Economics, University of Oxford. Address: Nuffield College, New Road, Oxford OX1 4PX, United Kingdom. E-mail: erik.mohlin@nuffield.ox.ac.uk.

1 Introduction

Numerous studies in psychology and cognitive science have demonstrated the fundamental role played by categorical reasoning in human cognition. In particular, categorical reasoning facilitates prediction.[1] Prediction on the basis of categories 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 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 rapid response.

Assuming that we use categorizations to make predictions, I ask which categorizations that are optimal, in the sense that they minimize prediction error.[2] The optimal number of categories is derived without imposing any exogenous costs or benefits of the number of categories. Instead costs and benefits arise endogenously from a bias-variance tradeoff that is inherent to 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 many observations.

The focus on optimal categorizations stems from evolutionary considerations.[3] 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 behaviour that maximize fitness. It 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. Hence, we would expect humans to have developed, and passed on, categorizations that are at least approximately optimal, in the sense that they tend to minimize prediction error

[1] For overviews of the voluminous literature, see e.g. Laurence and Margolis (1999), or Murphy (2002). Regarding categorization and prediction, see Anderson (1991). Categorical thinking matters in economic contexts: Consumers categorize products (Smith 1965), investors engage in "style investing" (Bernstein 1995), rating agencies categorize firms w.r.t. default risk (Coval et al. 2009).

[2] In section 5.1 I argue that categorization-based prediction is less cognitively demanding than other forms of similarity-based reasoning, such as kernel-based estimation.

[3] One might suggest 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, but concepts seem to have come prior to language in evolution; there are animals who use concepts even though they do not use language (see e.g. Herrnstein et al. 1976), and children can use certain concepts before they have a language (see e.g. Franklin et al. 2005). This suggests that we need to explain the use of categories without reference to language.

in the relevant environments. Such categorizations will be called *ex ante optimal*.[4] Other categorizations are developed only after a data base of experiences has been accumulated. We would expect evolution to have endowed us with heuristics or algorithms that allow us to form categorizations that organize our experience in a way that tends to minimize prediction error, conditional on the data base. Categorizations that attain this goal will be called *ex post optimal*.[5]

As an example of a categorization that is acquired very early on, think of colour concepts. The subset of the spectrum of electromagnetic radiation that is visible to the human eye allows for infinitely fine-grained distinctions. However, in every day reasoning and discourse we employ a coarse colour classification, using words such as red and green. Presumably the colour categorizations that were developed and passed on to new generations were successful in the kind of environments that we faced.[6] As an example of categorizations that are formed after a data base has been accumulated, one may think of the many classifications that science has produced. The two modes of categorization are often combined. Think of a physician who first 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: She receives information about a patient and predicts some aspect of the patient's health, based on a categorizations and her past experience. Then she observes the outcome for the patient. Eventually she might have accumulated sufficient experience to motivate the development of a refined categorization on her own.

An object is modelled as a vector (x, y), in finite-dimensional Euclidean space. (Extension to more general metric spaces is possible.) A subject has a data base consisting of objects that were observed in all dimensions in the past. She has to predict the y-dimension of a new object, based on that object's value in the x-dimensions. All objects are drawn independently from the same probability distribution, whose density is assumed to be continuous. The distribution may represent a mixture of distributions that are relevant for the subject. A categorization partitions the set of x-values. A new object is categorized solely on the basis of its x-value, and the empirical mean y-value of the previously experienced objects in that category (and that category only), serves as prediction for the y-value of the new object.

Prediction error is measured as squared difference between the predicted and actual y-value of an object. Using the probability density function over the set of objects one can define the (unconditional) *ex ante expected prediction error* of a categorization. In this case expectation is taken over the set of data bases that the subject may encounter.

[4] One might ask why the exact distribution of objects is not transmitted between generations. I will simply take it as an empirical fact that many categorizations are transmitted between generations. This indicates that there are some factors that make it infeasible or inefficient to transmit detailed information about the distribution.

[5] See Chater (1996) on the relationship between simplicity and likelihood in perceptual organization.

[6] For inter-cultural comparisons, see Kay and Maffi (1999) and references therein.

One can also define the *ex post expected prediction error* conditional on a given data base. In this case expectation is taken only over the new observation. The unconditional expected prediction error is minimized by an *ex ante optimal categorization*. The expected prediction error conditional on a data base is minimized by the *ex post optimal categorization*.

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. The subject does not have a prior. On the contrary, the model of this paper is intended to shed some light on how priors are generated (cf. Binmore 2007 and Gilboa et al. 2008). Relatedly, 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.[7] The x-dimensions may correspond to basic sensory input, as in the colour categorization example, or to some abstract dimension, as in the example of scientific categorization. For a discussion of similarity along such abstract dimensions, and how it builds on similarity along more concrete dimensions, see Gärdenfors (2000).

The main results of this paper are driven by a bias-variance (or fitting- vs. overfitting) trade-off**.** Increasing the number of categories has two effects. (a) The average size of each category decreases and thus the within-category differences between objects will be smaller. (b) The average number of experienced objects in each category decreases, thereby making inferences from observed objects to future cases less reliable. It follows that if the number of observations is increased, then the optimal number of categories is also increased, but at a slower rate. Comparative statics with respect to the distribution of objects reveals that (i) the larger the variance of y conditional on x, the smaller is the optimal number of categories, (ii) the larger the variance of the mean of y conditional on x, the larger is the optimal number of categories, and (iii) the more frequent objects in one subset of the x-dimension are, the larger is the optimal number of categories in that subset.

The results may explain the phenomenon of basic-level categories; the most salient level of categorization is neither the most fine-grained, nor the most general level of categorization (Rosch et al. 1976). For instance, bird is more salient than either the superordinate category animal or the subordinate category robin. The model also explains why experts will have a more fine-grained conceptual structure than laymen (Tanaka and Taylor 1991), and why minorities are categorized more coarsely than the majority (thereby generalizing a result of Fryer and Jackson 2008).

Although the main purpose of this paper is to model and investigate a psychological phenomenon, the results are relevant for some areas of statistical learning theory, in particular regression trees. Furthermore the results on optimal categorizations lie in between those previously obtained for, on the one hand, asymptotically optimal adaptive

[7] In this respect the paper builds on ideas that have been around since Kant (1781/87).

kernels for non-parametric regression analysis (Fan and Gijbels 1992), and on the other hand, asymptotically optimal adaptive histograms for non-parametric density estimation (Kogure 1987).

There are only a few explicit models of categorization in economics and the question of optimality has rarely been discussed. Fryer and Jackson (2008) consider a notion of optimal categorization. In their model the number of categories is exogenously given, and they do not define optimality in terms of minimization of prediction error. Moreover, the probability of encountering different objects is not modelled. Consequently the trade-off that is central to the present paper does not obtain within their framework. In independent work, Al-Najjar and Pai (2012) develop a model of coarse decision making, which is applied to categorization. Some of their results are similar to mine, but their focus and methodology is different: They use Vapnik-Chervonenkis theory to find categorizations whose worst case prediction error is below some threshold. Their set-up is essentially confined to what I call ex ante categorization.

Peski (2010) takes on the task of investigating when categorization may be an optimal tool for generating predictions. He compares predictions based on Bayesian updating with predictions based on a categorization algorithm. If the Bayesian prior over the states of the world is symmetric, then a Bayesian subject will expect to asymptotically perform approximately the same with Bayesian updating as with a categorization algorithm. If the state of nature is indeed drawn from a symmetric distribution, then a subject following the categorization algorithm will asymptotically make predictions that are no worse than those made by a subject who knows the distribution.

Gilboa et al. (2006) provide an axiomatization of a similarity-based prediction rule. The axiomatization tells us when a similarity function exists, but not what it looks like. A categorization is a psychologically relevant similarity function that treats all cases in one category as exactly similar to each other and treats a case in a category as completely dissimilar to any case outside that category. I seek to characterize the optimal such function.

There are a number of recent studies of categorization in game theoretic-contexts, e.g. Jehiel (2005), Jehiel and Samet (2007), and Azrieli (2009).[8] The results obtained in this paper may potentially be used to endogenise the categorizations in such models.

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; first for ex ante, and then for ex post, categorization. Section 4 explores the relation to statistical learning theory. Section 5 discusses the results and some applications. Section 6 concludes. All proofs are in the appendix.

[8] Mengel (2012) studies the evolutionary dynamics of categorization, assuming a fixed cost per category.

2 Model

2.1 Subject and Objects

An *object* is an attribute vector

$$v = (x, y) \in \mathcal{V} = \mathcal{X} \times \mathcal{Y},$$

where $\mathcal{Y} = \mathbb{R}$, and $\mathcal{X}$ is a compact subset of $\mathbb{R}^n$.[9] Objects are drawn i.i.d. according to an absolutely continuous cumulative distribution function $F : \mathcal{V} \to [0, 1]$. The probability density function, $f : \mathcal{V} \to \mathbb{R}_+$, is assumed to be continuous and bounded. Define the marginal density $f_X(x) = \int_{y \in \mathcal{Y}} f(x, y)\, dy$, and assume $f_X(x) > 0$, for all $x \in \mathcal{X}$. Similarly, define $f_Y(y) = \int_{x \in \mathcal{X}} f(x, y)\, dx$.

A *subject* first goes through a learning phase and accumulates a data base $v^t = (v_1, ..., v_t) \in \mathcal{V}^t$ consisting of t objects that have been observed in all dimensions. Then the subject encounters a new object $v_{t+1} = (x_{t+1}, y_{t+1}) \in \mathcal{V}$ and observes x_{t+1}, but not y_{t+1}. She makes a prediction about y_{t+1} on the basis of x_{t+1} and the data base v^{t+1}.[10]

From the point of view of the subject, an unobserved object is a random variable $V = (X, Y)$. Conditional on having observed x, the y-value is a random variable $Y|(X = x)$, with density $f_Y(y|x) = f(x, y)/f_X(x)$. It is convenient to normalise the conditional random variable $Y|(X = x)$ by expressing it as the sum of the conditional mean and a random term with mean zero. Assume that $|\mathbb{E}[Y|X = x]| < \infty$ for all $x \in \mathcal{X}$. Let $m(x) = \mathbb{E}[Y|X = x]$, and $\sigma^2(x) = Var(Y|X = x)$. Define $\varepsilon(x) = Y|(X = x) - \mathbb{E}[Y|X = x]$. It follows, for all $x \in \mathcal{X}$, that $\mathbb{E}[\varepsilon(x)] = 0$, $Var(\varepsilon(x)) = \sigma^2(x)$, and

$$Y|(X = x) = m(x) + \varepsilon(x).$$

We may also treat the conditional expectation as a random variable. Let $m(X) = \mathbb{E}[Y|X]$, $\sigma^2(X) = Var(Y|X)$, and $\varepsilon(X) = Y - \mathbb{E}[Y|X]$, so that

$$Y = m(X) + \varepsilon(X).$$

In order to abstract from trivialities, it is assumed that $|\mathcal{X}| = \infty$, and that $m(x) \neq m(x')$ for some $x, x' \in \mathcal{X}$. Some results will be specialized to the case of $\mathcal{X} = [a, b] \subseteq \mathbb{R}$. This allows us to derive stronger results, and facilitates a comparison with results from non-parametric estimation.

[9] The assumption that $\mathcal{X}$ is compact and $\mathcal{X} \subseteq \mathbb{R}^n$ is made for technical convenience. The results can be extended to a more general metric space $\mathcal{X}$.

[10] In a previous version of this paper it was assumed that, following the learning phase, the subject goes through a prediction phase consisting of several periods, during which she both makes predictions and accumulated more data. The insights from this, more complicated, specification are the same as for the present specification.

2.2 Categories

A category $\mathcal{C}_i$ is a Borel measurable subset of $\mathcal{V}$ such that $\mathcal{C}_i = \mathcal{X}_i \times \mathcal{Y}$. Thus $\mathcal{X}_i$ is the projection of $\mathcal{C}_i$ onto $\mathcal{X}$, implying that the category membership of an object only depends on the object's x-value. A *categorization* is a finite set of categories $\mathcal{C} = \{\mathcal{C}_1, ..., \mathcal{C}_k\}$ that constitutes a partitioning of $\mathcal{V}$. Let $k = |\mathcal{C}|$. Note that $\{\mathcal{X}_1, ..., \mathcal{X}_k\}$ partitions $\mathcal{X}$.

The probability that an object belongs to category i is $p_i = \int_{x \in \mathcal{X}_i} f_X(x)\, dx > 0$, the conditional marginal density of y in $\mathcal{X}_i$ is $f_Y(y|x \in \mathcal{X}_i) = \int_{x \in \mathcal{X}_i} f(x,y)\, dx/p_i$, and the conditional marginal density of x in $\mathcal{X}_i$ is $f_X(x|x \in \mathcal{X}_i) = f_X(x)/p_i$. The relative size of categories is constrained by some (small) number $\rho \in (0,1)$ such that, for all i and j, if $p_i \geq p_j$ then $p_j \geq \rho p_i$. For a finite number of categories this simply means that all categories have positive probability. When the number of categories goes to infinity, the constraint implies that no category becomes relatively infinitely larger than another category. The set of categorizations satisfying the assumptions above, called feasible categorizations, is denoted Ψ.[11] For a given data base v^t the set of feasible categorizations with non-empty categories is $\Psi(v^t)$. Denote the within-category variance $Var(Y_i) = Var(Y|X \in \mathcal{X}_i)$, and the within-category mean $\mu_i = \mathbb{E}[m(X)|X \in \mathcal{X}_i]$.

In the context of ex ante optimal categorizations, a categorization is formed before the learning phase, i.e. before any observation has been made. In the context of ex post optimal categorizations, a categorization is formed after a data base has been accumulated.

2.3 Prediction Error and Optimality

I assume that, given a data base v^t, each category $\mathcal{C}_i$ is associated with a unique (point) prediction $\hat{y}_{it}$. Thus, the same prediction is made for all objects in the same category. I take this to be a basic feature of categorical reasoning. It is also in line with the empirical evidence of Krueger and Clement (1994). An alternative is discussed in section 5.1.

Let $\hat{y}_t(x)$ be the prediction for the category to which the object $v = (x, y)$ belongs. The *prediction error* associated with v is defined as the squared difference between the predicted value $\hat{y}_t(x)$ and the true value y, i.e.

$$PE\left(\mathcal{C}, v, v^t\right) = \left(y - \hat{y}_t(x)\right)^2. \tag{1}$$

Before the object v has been observed it is a random variable $V = (X, Y)$. One might ask for the expected prediction error associated with V. Conditional on a data base v^t

[11] Hierarchically organized concepts, such as the two categories of stone and granite, are *not* mutually exclusive. 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 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.

the *ex post expected prediction error* of categorization $\mathcal{C}$ is

$$EPE\left(\mathcal{C}, v^t\right) = \mathbb{E}\left[\left(Y - \hat{y}_t\left(X\right)\right)^2\right]. \tag{2}$$

Ex ante the data base that will have been accumulated at time t is a random variable V^t, and hence the resulting prediction is also a random variable $\hat{Y}_t$. The unconditional, or *ex ante, expected prediction error* of categorization $\mathcal{C}$ at date t is

$$EPE\left(\mathcal{C}, t\right) = \mathbb{E}\left[\left(Y - \hat{Y}_t\left(X\right)\right)^2\right]. \tag{3}$$

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

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

For categorizations that are developed conditional on a data base, define:

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

Note that minimization is defined for categories that are nonempty for v^t. It will prove illuminating to define point-wise notions of expected prediction error. Conditional on v^t, the *point-wise ex post expected prediction error,* is

$$EPE\left(\mathcal{C}, v^t\right)\left(x\right) = \mathbb{E}\left[\left(Y - \hat{y}_t\left(X\right)\right)^2 | X = x\right]. \tag{4}$$

By taking expectation also over the data base V^t, one obtains the *point-wise ex ante, expected prediction error*

$$EPE\left(\mathcal{C}, t\right)\left(x\right) = \mathbb{E}\left[\left(Y - \hat{Y}_t\left(X\right)\right)^2 | X = x\right]. \tag{5}$$

In the non-parametric statistics literature a number of criteria based on squared error have been used. In the context of (univariate) non-parametric regression Härdle (1990) defines conditional integrated square error $ISE = \int \left(m\left(x\right) - \hat{y}\left(x\right)\right)^2 f_X\left(x\right) w\left(x\right) dx$, and conditional mean integrated square error, $MISE = \int \mathbb{E}\left[\left(m\left(x\right) - \hat{y}\left(x\right)\right)^2\right] f_X\left(x\right) w\left(x\right) dx$, where $w\left(x\right)$ is some weighting function. If $\sigma^2\left(x\right) = 0$ and $w\left(x\right) = 1$ for all x then ISE coincides with $EPE\left(\mathcal{C}, v^t\right)$ and $MISE$ coincides with $EPE\left(\mathcal{C}, t\right)$.

2.4 Prediction

In the previous section, ex ante and ex post optimal categorizations were defined for given predictions $\hat{y}_{it}$. The prediction rule remains to be specified. In line with the general focus on optimality one may look for a prediction rule that minimizes expected prediction error. Specifically, in the case of ex post categorization, one may look for predictions $\{\hat{y}_{it}\}_{i=1}^{k}$ that minimize $EPE\left(\mathcal{C}, v^t\right)$, for a given categorization $\mathcal{C}$. Thus, for each category i the optimal prediction $\hat{y}_{it}$ solves

$$\min_{\hat{y}'_{it}} \mathbb{E}\left[\left(Y - \hat{y}'_{it}\right)^2 | X \in \mathcal{X}_i\right]. \tag{6}$$

It is basic result of statistical decision theory that mean square error is minimized when the prediction is equal to the conditional mean (see e.g. Hastie et al. 2009). In the current set-up this means that the solution to (6) is

$$\hat{y}_{it} = \mathbb{E}\left[Y | X \in \mathcal{X}_i\right] = \mu_i.$$

Similarly, ex ante prediction error is minimized by setting $\hat{Y}_{it} = \mu_i$. The true mean μ_i is not observed, but the sample mean $\bar{y}_{it}$ is an unbiased estimator of the true mean. For this reason the within-category mean $\bar{y}_{it}$ will be used as prediction, as long as the category is non-empty. The choice of predictor for empty categories will turn out to be unimportant for the results.[12] For simplicity I assume that the prediction for an empty category is the mean across all observations, $\bar{y}_t$. Formally let

$$D_{it} = \{s \in \mathbb{N} : s \leq t \wedge v_s \in \mathcal{C}_i\}.$$

This is the set of dates, at which objects in category $\mathcal{C}_i$ were observed. Let $n_{it} = |D_{it}|$, so that $\sum_{i=1}^{k} n_{it} = t$, for all t. At date t the prediction for category i is

$$\hat{y}_{it} = \begin{cases} \bar{y}_{it} = \frac{1}{n_{it}} \sum_{s \in D_{it}} y_s & \text{if } n_{it} > 0 \\ \bar{y}_t = \frac{1}{t} \sum_{s=1}^{t} y_s & \text{if } n_{it} = 0 \end{cases}. \tag{7}$$

Note that the prediction $\hat{y}_{it}$, for a non-empty category i, only uses the part of the data base that belongs to category i. This captures the basic fact that predictions about a particular category tend to be formed only on the basis of objects that were put into that category in the past, and not on the basis of objects that were put into other categories (see Malt et al. 1995 and Murphy and Ross 1994).

[12] For instance one could simply assume that the subject is endowed with at least one object in each category. Alternatively one could assume that there is some fixed prediction error associated with empty categories. All the results of the paper will go through under these alternative assumptions.

3 Results

3.1 Preliminary Results

The first two results provide expressions for ex post and ex ante prediction error.

Lemma 1 *The* ex post expected prediction error *for a categorization* $\mathcal{C} \in \Psi(v^t)$*, conditional on a data base* v^t*, is,* pointwise *for* $x \in \mathcal{X}_i$*,*

$$EPE\left(\mathcal{C}, v^t\right)(x) = \sigma^2(x) + \left(m(x) - \bar{y}_{it}\right)^2,$$

and, globally

$$EPE\left(\mathcal{C}, v^t\right) = \sum_{i=1}^{k} p_i \left(Var\left(Y_i\right) + \left(\bar{y}_{it} - \mu_i\right)^2\right).$$

The expression for $EPE(\mathcal{C}, v^t)$ reveals the fundamental bias-variance trade-off. The squared bias-term $\left(\bar{y}_{it} - \mu_i\right)^2$ measures how close the prediction is to the actual average in category $\mathcal{C}_i$. The variance-term $Var(Y_i)$ measures how similar (with respect to the y-dimension) different objects in category $\mathcal{C}_i$ are.

Note that ex ante the number of observations in category i at date t is a random variable N_{it}. For ex ante prediction error we find:

Lemma 2 *The unconditional* ex ante expected prediction error *for a categorization* $\mathcal{C} \in \Psi$*, at time* t*, is,* pointwise *for* $x \in \mathcal{X}_i$*,*

$$EPE(\mathcal{C}, t)(x) = \sigma^2(x) + \left(m(x) - \mu_i\right)^2 + Var\left(\bar{Y}_{it}\right),$$

and, globally

$$EPE(\mathcal{C}, t) = \sum_{i=1}^{k} p_i \left(Var\left(Y_i\right) + Var\left(\bar{Y}_{it}\right)\right),$$

where

$$Var\left(\bar{Y}_{it}\right) = Var\left(Y_i\right) \sum_{r=1}^{t} \Pr\left(N_{it} = r\right) \frac{1}{r} + \Pr\left(N_{it} = 0\right) \mathbb{E}\left[\left(\bar{Y}_t - \mu_i\right)^2 | N_{it} = 0\right],$$

and $N_{it} \sim Bin(t, p_i)$*.*

It will be useful to decompose the within-category variance, $Var(Y_i)$, into the contribution of the *within-category average conditional variance*

$$\sigma_i^2 \equiv \int_{x \in \mathcal{X}_i} f_X\left(x | x \in \mathcal{X}_i\right) \sigma^2(x)\, dx,$$

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

$$Var\left(m\left(X\right)|X\in\mathcal{X}_i\right)=\int_{x\in\mathcal{X}_i}f_X\left(x|x\in\mathcal{X}_i\right)\left(m\left(x\right)-\mu_i\right)^2dx.$$

It is straightforward to verify that within-category variance is the sum of the within-category average conditional variance, and the within-category variance of the conditional expected value;

$$Var\left(Y_i\right)=\sigma_i^2+Var\left(m\left(X\right)|X\in\mathcal{X}_i\right). \tag{8}$$

A final preliminary result concerns the existence of optimal categorizations.

Proposition 1 *Suppose $\mathcal{X}=[a,b]\subseteq\mathbb{R}$. Let $\Psi\left(\iota\right)$ be the set of categories such that each category is the union of at most ι disconnected subsets. There is a $\bar{t}$ such that if $t>\bar{t}$ then there exists a solution to the problem of minimizing $EPE\left(\mathcal{C},t\right)$, w.r.t. $\mathcal{C}\in\Psi\cap\Psi\left(\iota\right)$, and there exists a solution to the problem of minimizing $EPE\left(\mathcal{C},v^t\right)$ w.r.t. $\mathcal{C}\in\Psi\left(v^t\right)\cap\Psi\left(\iota\right)$.*

For a general set $\mathcal{X}\subseteq\mathbb{R}^n$ we cannot be sure that there exists a solution in Ψ or $\Psi\left(v^t\right)$. This need not trouble us very much since the actual choice between categorizations may be implemented by an algorithm that effectively restricts attention to a finite subset of categorizations, or (in the case of regression trees) restricts attention to categorizations that are hyper-rectangles in $\mathbb{R}^n$, which are generated by successive splits of intervals on $\mathbb{R}$.

It can be noted that in general there will be no guarantee that a solution is unique, thus allowing for a (mild) form of conceptual relativism.

It is easy to see that $EPE\left(\mathcal{C},v^t\right)$, and $EPE\left(\mathcal{C},t\right)$ are continuous in the distribution f. Mohlin (2014*b*) defines a metric on partitions which implies that prediction errors $EPE\left(\mathcal{C},v^t\right)$, and $EPE\left(\mathcal{C},t\right)$ are also continuous in the decision variable $\mathcal{C}$.

3.2 Ex Ante Optimal Categorizations

Intuitively, as t increases one may obtain a reasonable approximation of $m\left(x\right)$ by increasing the number of categories k, though at a slower rate than t. The following lemma formalizes this intuition:

Lemma 3 *For any $\varepsilon>0$ there are $\bar{t}>0$ and $\delta>0$, such that if $t>\bar{t}$ and*

$$EPE\left(\mathcal{C},t\right)-\inf_{\mathcal{C}'\in\Psi}EPE\left(\mathcal{C}',t\right)<\delta,$$

then $\mathcal{C}$ satisfies $1/k<\varepsilon$, and $k/t<\varepsilon$.

This result does not presume the existence of a minimum, and hence it is phrased in terms of the infimum. When existence is guaranteed, the lemma says that minimization

of $EPE(\mathcal{C},t)$ implies that if the number of observations goes to infinity ($t \to \infty$) then it is optimal to let the number of categories go to infinity too ($k \to \infty$), but at a slower rate ($k/t \to 0$). The result provides an explanation for why we typically employ categorizations that are neither maximally fine-grained nor maximally coarse. This is discussed further in section 5.2, along with other psychological applications of the model.

If f is three times differentiable, Taylor approximations may be used to derive the asymptotically optimal width of categories in the case $\mathcal{X} = [a,b] \subseteq \mathbb{R}$.[13] The following theorem can be viewed as the main result of the paper. For the case of $\mathcal{X} = [a,b] \subseteq \mathbb{R}$ it summarizes all the comparative statics results that will later be formulated for a general set $\mathcal{X}$, and for the setting of ex post categorizations. For simplicity, assume a fixed design model with data such that $n_{it} = tp_i$. Mohlin (2014*a*) generalises the result to a multidimensional setting, and a random design model.

Theorem 1 *Suppose $\mathcal{X} = [a,b] \subseteq \mathbb{R}$ and f three times differentiable. Restrict attention to convex categories and let $h(x)$ denote the width of the category to which x belongs.*[14] *Asymptotically, as $t \to \infty$, if $m'(x) > 0$ and $\mathbb{E}[s|s \in \mathcal{X}_i] \neq x$, then the pointwise ex ante expected prediction error, $EPE(\mathcal{C},t)(x)$, is minimized by*

$$h^*(x) = \left(\frac{\sigma^2(x)}{2tf_X(x)(\delta m'(x))^2} \right)^{\frac{1}{3}}, \tag{9}$$

where $\delta \in (0,1]$ is such that $|\mathbb{E}[s|s \in \mathcal{X}_i] - x| = \delta h(x)$. The associated pointwise ex ante expected prediction error is

$$EPE(\mathcal{C},t)(x) = \sigma^2(x) + O\left(t^{-\frac{2}{3}}\right). \tag{10}$$

The number of categories should increase at a rate proportional to the cubic root of the number of observations. The expression for $h^*(x)$ reveals a a number of comparative statics results that hold asymptotically: The size of the optimal categories is decreasing in the density $f_X(x)$, the number of observations t, and the curvature $m'(x)^2$ of the conditional mean. It is increasing in the variance $\sigma^2(x)$.

In what follows these comparative statics results are extended to the general case where $\mathcal{X}$ need not be an interval on the real line. We start by describing how the ex ante optimal categorization is affected by the size of the data base. This is merely a corollary to lemma 3 and mainly stated for the purpose of comparison with the ex post case below.

Proposition 2 *For any categorization $\mathcal{C}'$ there is a refinement $\mathcal{C}''$ and a time $\bar{t}$ such that if $t > \bar{t}$ then $EPE(\mathcal{C}',t) > EPE(\mathcal{C}'',t)$.*

[13] If f is Lipschitz continuous we may derive an upper bound on the expected prediction error, and an expression for the associated width h^*. The comparative statics resulting from this exercise exactly parallel those for the Taylor approximation. A formal statement and proof is available upon request.

[14] See Gärdenfors (2000) for arguments as to why to expect natural concepts to be convex.

The next two propositions concern the relationship between the density $f(x,y)$ and the optimal categorization. We saw above (equation 8) that $Var(Y_i)$ is the sum of σ_i^2 and $Var(m(X)|X \in \mathcal{X}_i)$. In order to study comparative statics with respect to these two terms some more detailed assumptions are made which allow for parameterization of the compared distributions.

Proposition 3 *Consider categorization under a density f, such that for all $x \in \mathcal{X}$, and some density $\tilde{f}$; $f_X(x) = \tilde{f}_X(x)$, $\sigma_f^2(x) = \sigma^2$, and $m_f(x) = \mu + \beta\left(m_{\tilde{f}}(x) - \mu\right)$.*

For any categorization $\mathcal{C}''$ with $k'' \geq 2$ categories, there is a coarsening $\mathcal{C}'$ such that $EPE(\mathcal{C}'',t) - EPE(\mathcal{C}',t)$ is increasing in σ^2, and there is a $\bar{\sigma}^2$ such that if $\sigma^2 > \bar{\sigma}^2$ then $EPE(\mathcal{C}'',t) > EPE(\mathcal{C}',t)$.

For any categorization $\mathcal{C}'$, there is a refinement $\mathcal{C}''$ and $\bar{t}$ such that if $t > \bar{t}$ then $EPE(\mathcal{C}'',t) - EPE(\mathcal{C}',t)$ is decreasing in β, and there is a $\bar{\beta}$ such that if $\beta > \bar{\beta}$ then $EPE(\mathcal{C}'',t) < EPE(\mathcal{C}',t)$.

That is, the larger variance, the better the coarse categorization fares. In contrast, the larger variance of the conditional mean, the worse the coarse categorization fares (assuming that t is large enough).

In order to study comparative statics w.r.t. the marginal density $f_X(x)$ we need to restrict attention to categorization of a proper subset $\mathcal{E} \subseteq \mathcal{X}$ (because total probability over $\mathcal{X}$ must sum to one).

Proposition 4 *Restrict attention to the categorization of a proper subset $\mathcal{E} \subseteq \mathcal{X}$ with $\Pr(x \in \mathcal{E}) > 0$. Consider categorization under a density f, such that for all $x \in \mathcal{E}$, and some density $\tilde{f}$; $f_Y(y|x) = \tilde{f}_Y(y|x)$ for all y, and $f_X(x) = \alpha\tilde{f}_X(x)$, for $\alpha > 0$. For any categorization $\mathcal{C}'$ there is refinement $\mathcal{C}''$ and a $\bar{t}$ such that if $t > \bar{t}$ and $\alpha \geq 1$ then $EPE(\mathcal{C}'',t) - EPE(\mathcal{C}',t)$ is decreasing in α, and there is an $\bar{\alpha} > 1$ such that if $\alpha > \bar{\alpha}$ and $t > \bar{t}$, then $EPE(\mathcal{C}'',t) < EPE(\mathcal{C}',t)$.*

Assuming that t is large enough, the more frequent objects from one subset of $\mathcal{X}$ are, the more fine-grained should the optimal categorization for that subset be.

3.3 Ex Post Optimal Categorizations

In the introduction it was argued that humans might have evolved some algorithms or heuristics that take a given data base as input and deliver an approximately (ex post) optimal categorization as output. As analysts we might be willing to abstract from the heuristics and assume that subjects act *as if* they minimized $EPE(\mathcal{C}, v^t)$ 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, when the subject does *not* know f. In this section I present results along both these lines.

The ex post optimal categorization may look very different depending on what data base that is accumulated. The results presented in this section are therefore formulated in terms of how changes in the model's parameters influence the *probability* that the ex post optimal categorizations will have certain properties. With this phrasing, it turns out that one can prove results that are fairly direct counterparts to the results proved for ex ante optimal categorizations.

What heuristics might a categorizing subject use to form categorizations given a data base? A the subject could use the data base to compute some estimator of $EPE\left(\mathcal{C},v^t\right)$, and then pick a categorization that minimizes this value – possibly within an a priori restricted set of categorizations. I would like to suggest the following estimator:

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

$$\widehat{EPE\left(\mathcal{C},v^t\right)}=\sum_{i=1}^{k}\frac{n_{it}}{t}\left(1+\frac{1}{n_{it}}\right)s_{it}^2,\qquad s_{it}^2=\frac{1}{n_{it}-1}\sum_{s\in D_{it}}\left(y_s-\bar{y}_{it}\right)^2.$$

The motivation for this definition comes from the following observation, which follows directly from the facts that $\mathbb{E}[(\hat{Y}_{it}-\mu_i)^2]=Var\left(Y_i\right)/n_{it}$, and $\mathbb{E}\left[s_{it}^2\right]=Var\left(Y_i\right)$.

Lemma 4 *For a given categorization $\mathcal{C}$ and data base v^t with an allocation of observations to categories $\{n_{1t},n_{2t},...,n_{kt}\}$, such that $n_{it}\geq 2$ for all i, let $\tilde{\mathcal{V}}^t\left(\mathcal{C},v^t\right)$ be the set of data bases $\tilde{v}^t$ such that $\tilde{n}_{it}=n_{it}$ for each category i in $\mathcal{C}$. If expectation is taken over $\tilde{\mathcal{V}}^t\left(\mathcal{C},v^t\right)$, then*

$$\mathbb{E}\left[EPE\left(\mathcal{C},v^t\right)\right]=\sum_{i=1}^{k}p_i\left(1+\frac{1}{n_{it}}\right)Var\left(Y_i\right),$$

and

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

The lemma implies that if the actual fraction of objects in each category, n_{it}/t, is equal to the probability of receiving an object in the corresponding category, p_i, then $EPE\left(\mathcal{C},v^t\right)$ and $\widehat{EPE\left(\mathcal{C},v^t\right)}$ 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 states how the categorizations that minimize $EPE\left(\mathcal{C},v^t\right)$ and the categorizations that minimize $\widehat{EPE\left(\mathcal{C},v^t\right)}$ are likely to be affected by changes in different parameters. The first result regards the effect of varying the size of the data base. It corresponds to proposition 2 in the case of ex ante optimality.

Proposition 5 ***(a)*** *For any categorization $\mathcal{C}'$ there is a refinement $\mathcal{C}''$ such that for any $\delta \in (0,1)$, there is a $\bar{t}$, such that if $t > \bar{t}$ then, in the set of databases with $n_{it} \geq 1$ for all categories $\mathcal{C}_i \in \mathcal{C}''$,*

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

(b) *Restrict attention to databases such that $n_{it} \geq 2$ for all categories $\mathcal{C}_i \in \mathcal{C}''$. The statement in (a) holds if (11) is replaced with*

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

For any categorization $\mathcal{C}'$, part (a) states that by increasing the size of the data base, we can ensure that there is a refinement of $\mathcal{C}'$, called $\mathcal{C}''$, which is arbitrarily likely to outperform $\mathcal{C}'$. By increasing the size of the data base sufficiently much we can push this probability arbitrarily close to one. Part (b) goes on to state that the same relationship holds for the estimated expected prediction error.

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.

Example 2 *Assume $\mathcal{X} = [0,1]$,$\mathcal{Y} = \mathbb{R}$, $\sigma^2(x) = \sigma^2$, $f_X(x) = 1$, and*

$$m(x) = \begin{cases} 0.5 & \text{if } x < 0.5 \\ 0.1 & \text{if } x \geq 0.5 \end{cases}.$$

Consider the data base $v = \{(0.1, 0.6), (0.2, 0.6), (0.7, 0)\}$. Compare a categorization $\mathcal{C}'$ consisting of only one category, with a categorization $\mathcal{C}''$ that divides $\mathcal{X}$ into two categories $\mathcal{C}_1 = [0, 0.5) \times \mathbb{R}$ and $\mathcal{C}_2 = [0.5, 1] \times \mathbb{R}$. It is straightforward to compute $EPE(\mathcal{C}', v^t) = \sigma^2 + 0.05$ and $EPE(\mathcal{C}'', v^t) = \sigma^2 + 0.01$. Thus $\mathcal{C}''$ is the ex post optimal categorization. Now suppose one object $(0.8, -0.6)$ is added to the data base, so that $EPE(\mathcal{C}', v^t)\,\sigma^2 + 0.0625$ and $EPE(\mathcal{C}'', v^t) = \sigma^2 + 0.085$. Hence $\mathcal{C}'$ is the new ex post 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 effect of changing the conditional variance, and the variance of the conditional mean, is described by the following proposition. Compared to proposition 3 it makes the stronger assumption that y is normally distributed conditional on x.

Proposition 6 *Consider categorization under a density f, such that for all $x \in \mathcal{X}$, and some density $\tilde{f}$; $f_X(x) = \tilde{f}_X(x)$, $\sigma_f^2(x) = \sigma^2$, $\varepsilon(x) \sim N(m(x), \sigma^2)$, and $m_f(x) = \mu + \beta\left(m_{\tilde{f}}(x) - \mu\right)$.*[15]

[15] If one did not assume normality, this would still hold for large enough t, by the central limit theorem.

(a) *For any categorization $\mathcal{C}'$ there is a refinement $\mathcal{C}''$ such that, in the set of databases with $n_{it} \geq 1$ for all categories $\mathcal{C}_i \in \mathcal{C}''$,*

$$\Pr\left(EPE\left(\mathcal{C}'', v^t\right) > EPE\left(\mathcal{C}', v^t\right)\right), \tag{13}$$

is increasing in σ^2 and decreasing β.

(b) *Restrict attention to databases such that $n_{it} \geq 2$ for all categories $\mathcal{C}_i \in \mathcal{C}''$. The statement in (a) holds if (13) is replaced with*

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

By increasing the variance σ^2, or by decreasing the variance of the conditional mean $Var\left(m\left(X\right)|X \in \mathcal{X}\right)$, one can increase the probability that $\mathcal{C}''$ yields a higher expected prediction error, and higher estimated expected prediction error, than $\mathcal{C}'$.

The next proposition tells us what happens if we restrict attention to categorizations of a subset $\mathcal{E}$ of $\mathcal{X}$, and vary the density on $\mathcal{E}$.

Proposition 7 *Restrict attention to the categorization of a proper subset $\mathcal{E} \subseteq \mathcal{X}$ with $\Pr\left(x \in \mathcal{E}\right) > 0$. Consider categorization under a density f, such that for all $x \in \mathcal{E}$, and some density $\tilde{f}$; $f_Y\left(y|x\right) = \tilde{f}_Y\left(y|x\right)$ for all y, and $f_X\left(x\right) = \alpha\tilde{f}_X\left(x\right)$, for $\alpha > 0$.*

(a) *For any categorization $\mathcal{C}'$ there is a refinement $\mathcal{C}''$ such that for any $\delta \in (0,1)$, there is a $\bar{t}$, such that for any $t > \bar{t}$ there is a an $\bar{\alpha}\left(t\right) > 1$ such that if $\alpha > \bar{\alpha}\left(t\right)$ then, in the set of databases with $n_{it} \geq 1$ for all categories $\mathcal{C}_i \in \mathcal{C}''$,*

$$\Pr\left(EPE\left(\mathcal{C}', v^t\right) > EPE\left(\mathcal{C}'', v^t\right)\right) > \delta, \tag{15}$$

while if $\alpha \in [1, \bar{\alpha}\left(t\right))$ then (15) does not hold. ***(b)*** *Restrict attention to databases such that $n_{it} \geq 2$ for all categories $\mathcal{C}_i \in \mathcal{C}''$. The statement in (a) holds if (15) is replaced with*

$$\Pr\left(\widehat{EPE\left(\mathcal{C}'\right)}, v^t) > \widehat{EPE\left(\mathcal{C}''\right)}, v^t)\right) > \delta. \tag{16}$$

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

Propositions 5-7 indicate that $\widehat{EPE\left(\mathcal{C}, v^t\right)}$ is a reasonable guide to choices between different ways of categorizing a given data base. However, it might be too cognitively demanding and too time consuming to compute $\widehat{EPE\left(\mathcal{C}, v^t\right)}$ for all categorizations in $\hat{\Psi}\left(v^t\right)$. For this reason the subject's choice may be restricted in some way, for example by follow an algorithm such as those developed for regression trees; see section 4.1**.**

3.4 Extensions

3.4.1 Interest-dependent Predictions

The cost of a prediction error has been assumed independent of x. More realistically it could be that predictions associated with some values of x are more important than others. To model this one may simply add a function $w : \mathcal{X} \to [0,1]$ 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 $f_X(x)$. Increasing the importance of predictions in a set $\mathcal{E}$ will increase the optimal number of categories in $\mathcal{E}$.

3.4.2 Multi-dimensional $\mathcal{Y}$

The model can be extended to allow for prediction of many different attributes of an object, represented by a vector $y \in \mathcal{Y} = \mathbb{R}^n$. The easiest way of doing this is to let the prediction error 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 $PE(\mathcal{C}, v_t, v^t) = \sum_{j=1}^{n} z[j] \left(y_t[j] - \hat{y}_t[j](x_t)\right)^2$. With this specification all of the results presented above hold for a multi-dimensional $\mathcal{Y}$. The weights may differ between subjects, allowing for another form of interest-dependent predictions.

4 Relation to Other Fields

The prediction problem studied in this paper is closely related to prediction and estimation problems studied the literature on statistical learning, machine learning, and non-parametric estimation. It is customary to distinguish supervised and non-supervised learning (Hastie et al. 2009). The problem studied in this paper falls within *supervised learning*: The learner has a data base and her task is to estimate the conditional distribution $f_Y(y|x)$ or some property thereof, such as $m(x)$. Within supervised learning the most closely related area is non-parametric regression using regression trees.[16]

4.1 Regression trees

Regression trees were introduced by Morgan and Sonquist (1963) (see also Breiman et al. 1984). By a successive procedure of binary splits, the set $\mathcal{X} \subseteq \mathbb{R}^n$ is split into a number of hyper-rectangles. These regions function like categories. The prediction for each

[16] In *unsupervised learning* the learners task is to estimate a joint distribution $f(x_1, x_2)$, or some aspect thereof. The kind of unsupervised learning closest to optimal categorization is *cluster analysis*: A set of objects is partitioned in a way that maximizes some measure of within-cluster similarity and between cluster-dissimilarity. The most important difference compared to the model of optimal categorization is the same set of dimensions are used to both to define and to evaluate clusters, whereas I define categorizations in terms of one set of dimensions and evaluate them in terms of another dimension.

region/category is equal to the sample mean of data points in that region/category. In each step of the splitting procedure, one dimension $\mathcal{X}_i \subseteq \mathbb{R}$ and one splitting point $s \in \mathcal{X}_i$ is used to divide $\mathcal{X}$ into two halves. Thus the sequence of splits can be represented as a binary tree τ. The criterion for evaluating a tree is usually based on the residual sum of squares $R(\tau)$, with the addition of a complexity cost α per split k. Thus one selects a tree that minimizes $R_\alpha(\tau) = R(\tau) + \alpha k$. The value of α is determined by some cross-validation procedure.

The main differences between regression trees and the model of ex post categorization is that I propose to use an estimate of expected prediction error criterion, $\widehat{EPE}(\mathcal{C}, v^t)$, which takes care of the bias-variance trade-off without any need for the addition of a complexity parameter α. Lemma 4 showed that the proposed estimator is unbiased in a certain sense. To the best of my knowledge there are no similar results available for regression trees. Moreover I am not aware of comparative statics results of the kind presented here. Finally, there is no counterpart to ex ante optimal categorization in the literature on regression trees.

4.2 Kernels and Histograms

Non-parametric regression aims at estimating the conditional mean $m(x)$. Fan and Gijbels (1992) derive the following expression for the locally adaptive asymptotically optimal kernel bandwidth (where optimal is understood in the sense of minimizing asymptotic integrated mean-square error),

$$h_K^*(x) = q_K \left(\frac{\sigma^2(x)}{t f_X(x) \left(m''(x)\right)^2} \right)^{1/5}, \tag{17}$$

where q_K is a constant independent of x. This is remarkably similar to the expression (9) for the optimal category width $h^*(x)$, in theorem 1. One difference is that the curvature of the conditional mean enters through the second derivative $m''(x)$ here, compared to the first derivative $m'(x)$ above. Another difference is that speed at which the bandwidth vanishes is slower than the speed at which the optimal categorization binwidth vanishes.

Within density estimation histograms is the closest related approach. Kogure (1987) derives the following expression for the locally adaptive asymptotically optimal (minimizing asymptotic integrated mean-square error) bin width,

$$h_H^*(x) = q_H \left(\frac{f_X(x)}{t \left(f_X'(x)\right)^2} \right)^{1/3}, \tag{18}$$

where q_H is a constant independent of x. In contrast to (9) and (17) this expression is independent of the variance $\sigma^2(x)$, and the optimal width at x is increasing in density $f_X(x)$, rather than decreasing. Interestingly, the width of the bins optimally vanishes at

the rate $O\left(t^{-1/3}\right)$, exactly as in (9).

In conclusion, the ex ante optimal categorization derived in theorem 1 lies between the results derived for kernels in regression analysis and histograms in density estimation. This reflects the fact that while the prediction *task* of the optimal categorization approach is closer to that of regression analysis, the *tool* (the categorization) is more similar to histograms.

It should be noted that the close parallel between asymptotically optimal kernels and histograms, and optimal categorizations, is only present in the ex ante setting.

4.3 Rational Inattention

In the rational inattention framework (Sims 2003) an agent tries to predict the state of the world an the basis of some signal. The joint distribution of the state and signal is chosen by the agent, subject to an entropy constraint. Translated to my notation the state of the world is y and the signal is the category $\mathcal{C}_i \in \mathcal{C}$ to which $x \in \mathcal{X}$ belongs. By choosing a categorization of $\mathcal{X}$ the subject effectively chooses a joint distribution of signals and states. The joint distribution is also determined by the size of the data base, which therefore plays a role similar to that of the entropy bound in the rational inattention framework. It might be considered an advantage of the optimal categorization approach, that it models a well-documented and important reasoning process in a more precise way.

5 Discussion

5.1 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. Nevertheless, one might ask why we use categorizations rather than some other method for making predictions. In particular one might suggest some smoother form of similarity-based reasoning, for instance as formalized by kernel-based estimation. In that case 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'. Observations which are not located within some distance η from x receive zero weight.

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, which has been computed in advance. Hence category-based and kernel-based predictions should be compared for the case when predictions are produced in a relatively fast and automatic way.

A subject basing predictions of categories will use something like the following procedure: At the beginning of period $t+1$ the subject has stored k pairs $(\bar{y}_{it}, n_{it})$ of predictions

and samples sizes, one for each category. She then observes x_{t+1}, identifies the category $\mathcal{C}_i$, such that $x_{t+1} \in \mathcal{C}_i$ and predicts $\bar{y}_{it}$. At the end of the period she observes y_{t+1} and uses it to compute an updated prediction $\bar{y}_{it+1} = (n_{it}\bar{y}_{it} + y_{t+1}) / (n_{it} + 1)$ for category $\mathcal{C}_i$, and replaces $(\bar{y}_{it}, n_{it})$ with $(\bar{y}_{it+1}, n_{it+1})$.

In contrast, a subject basing predictions on kernel based estimation will use a procedure akin to the following: At the beginning of period $t + 1$ the subject has stored t different objects (x, y) and a number of predictions $\hat{y}_t|x$, one for each x in a sufficiently fine grid over $\mathcal{X}$. She then observes x_{t+1}, and uses the corresponding (closest) prediction $\hat{y}_t|x_{t+1}$. At the end of the period she adds the observation (x_{t+1}, y_{t+1}) to her memory. She computes an updated prediction $\hat{y}_{t+1}|x$ for each x (in the grid) within distance η of the observed x_{t+1}.

In conclusion, the kernel-based procedure has at least three drawbacks. (1) The kernel-based procedure requires the subject to store a larger number of predictions. (2) The kernel-based procedure requires the subject to update a larger number of predictions after each new observation. (3) The kernel-based procedure requires the subject to store more information about observations.

5.2 Psychological Applications

5.2.1 Basic Level Categories

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 animal or robin). The model put forward in this paper suggests that the reason that we do not use the finest categorization as our basic level is 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 categorize something as belonging to a very narrow category one must observe many properties of an object, something that may be inconvenient or impossible (Medin 1983, and Jones 1983).

5.2.2 Experts and Laymen

Experts tend to have a more fine-grained conceptual structure than laymen (Tanaka and Taylor 1991). This can be explained by proposition 2, which predicts that people with a larger data base optimally have a larger number of categories. This is also consistent with the fact that people in traditional subsistence cultures tend to have more specific biological categories than e.g. American college students (Berlin et al. 1973).

5.2.3 Discrimination and Stereotypes

Propositions 4 and 7 are generalizations 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.

6 Conclusion

This paper provides a framework for the study of optimal categorization for the purpose of making predictions. The optimal number, and the optimal shape, of categories, is endogenous to the model. There are a number of extensions and applications that would be interesting to pursue in the future: The model (or a simplification thereof) may be adapted to handle categorizations in game-theoretic context. It would also be interesting to apply the optimal categorization framework to questions studied in the literature on rational inattention. Experimentally it is desirable to test the comparative static predictions of the optimal categorization model. Furthermore, the framework might potentially be applied to questions from the philosophy of science. On a more speculative note, the sample prediction error estimator could perhaps be used in evaluation of regression trees.

7 Appendix

7.1 Preliminaries

Proof of Lemma 1. *Pointwise*: Note $\int_y y f_Y(y|x)\,dx = m(x)$. For $x \in \mathcal{X}_i$ we have

$$\begin{aligned} EPE\left(\mathcal{C}, v^t\right)(x) &= \int_y f_Y(y|x)\left(y - \bar{y}_{it}\right)^2 dy \\ &= \int_y f_Y(y|x)\left(\left(y - m(x)\right)^2 + \left(m(x) - \bar{y}_{it}\right)^2 - 2\left(y - m(x)\right)\left(m(x) - \bar{y}_{it}\right)\right) dy \\ &= \sigma^2(x) + \left(m(x) - \bar{y}_{it}\right)^2 \end{aligned}$$

Global: The result can be found using a method similar to that for the pointwise expression, using $y - \bar{y}_{it} = y - \mu_i - (\bar{y}_{it} - \mu_i)$. ■

Proof of Lemma 2. *Pointwise*: Suppose $x \in \mathcal{X}_i$. Start by noting

$$EPE(\mathcal{C}, t)(x) = \mathbb{E}\left[EPE\left(\mathcal{C}, v^t\right)(x)\right] = \sigma^2(x) + \mathbb{E}\left[\left(m(x) - \bar{Y}_{it}\right)^2\right].$$

Using $m(x) - \bar{Y}_{it} = m(x) - \mu_i - (\bar{Y}_{it} - \mu_i)$ we have

$$\mathbb{E}\left[\left(m(x) - \bar{Y}_{it}\right)^2\right] = Var\left(\bar{Y}_{it}\right) + \left(m(x) - \mu_i\right)^2.$$

Furthermore

$$Var\left(\bar{Y}_{it}\right) = \sum_{r=1}^{t} \Pr(N_{it} = r)\,\mathbb{E}\left[\left(\bar{Y}_{it} - \mu_i\right)^2 | N_{it} = r\right] + \Pr(N_{it} = 0)\,\mathbb{E}\left[\left(\bar{Y}_t - \mu_i\right)^2 | N_{it} = 0\right].$$

Note that if $r \geq 1$ then $E\left[\bar{Y}_{it} | n_{it} = r\right] = \mu_i$, so

$$\mathbb{E}\left[\left(\bar{Y}_{it} - \mu_i\right)^2 | n_{it} = r\right] = Var\left(\bar{Y}_{it} | N_{it} = r\right) = \frac{1}{r} Var(Y_i).$$

It is evident that the number of objects in a category, N_{it}, has a binomial distribution with parameters p_i and t.

Globally: The result can be found using the same methods as for the pointwise expression. Alternatively use $EPE(\mathcal{C}, t) = \int_{x \in \mathcal{X}} EPE(\mathcal{C}, t)(x)\, f_X(x)\,dx$. ■

Proof of Proposition 1. I only prove the result for minimization of $EPE(\mathcal{C}, t)$. The claim regarding $EPE\left(\mathcal{C}, v^t\right)$ is verified in a similar way. For any t let $\Psi(\kappa, \iota) \subseteq \Psi$ 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 $\mathcal{C} \in \Psi(\kappa, \iota)$

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, $EPE(\mathcal{C}, t)$ is continuous in z. Hence by Weierstrass' maximum theorem there exists a solution $z^*(\nu)$ to the problem of minimizing $EPE(\mathcal{C}, t)$ 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 $\{1, 2, ..., k\}$, there is a solution in the set $\Psi(\kappa, \iota)$. This result holds for any t. Lemma 3 below implies that as $t \to \infty$, any categorization that results in an $EPE(\mathcal{C}, t)$ that is arbitrarily close to $\inf_{\mathcal{C}' \Psi} EPE(\mathcal{C}', t)$, satisfies the condition $k < \kappa t$. ■

7.2 Ex Ante Optimality

Lemma 3 is proved with the help of lemmas 5 and 6.

Lemma 5 *Let $\mathcal{E}$ and $\mathcal{F}$ be disjoint intervals. If $\mathbb{E}[Y|X \in \mathcal{E}] \neq \mathbb{E}[Y|X \in F]$ then then there is some $\varepsilon > 0$ such that*

$$\Pr(X \in \mathcal{E} \cup \mathcal{F}) Var(Y|X \in \mathcal{E} \cup \mathcal{F}) - \sum_{\mathcal{I} \in \{\mathcal{E}, \mathcal{F}\}} \Pr(X \in \mathcal{I}) Var(Y|X \in \mathcal{I}) > \varepsilon.$$

Proof of Lemma 5. Follows fairly straightforwardly from the fact that $\mathbb{E}\left[(Y - z)^2 | X \in \mathcal{I}\right]$ is minimised by $z = \mathbb{E}[Y|X \in \mathcal{I}]$ for any $\mathcal{I} \subseteq \mathcal{X}$. ■

Lemma 6 *If $k/t \geq \gamma$ then $\sum_{r=1}^{t} \Pr(N_{it} = r)/r > \gamma\rho/(1 - 1/t)$.*

Proof of Lemma 6. Since r is binomially distributed we have

$$\sum_{r=1}^{t} \Pr(N_{it} = r) r = \sum_{r=0}^{t} \Pr(N_{it} = r) r - \Pr(N_{it} = 0) \cdot 0 = \mathbb{E}[N_{it}] = tp_i.$$

Since $g(x) = 1/x$ is concave, Jensen's inequality implies

$$\sum_{r=1}^{t} \Pr(N_{it} = r) \frac{1}{r} \geq \frac{1}{\sum_{r=1}^{t} \Pr(N_{it} = r) r} = \frac{1}{tp_i}. \tag{19}$$

Let $p_{\max} = \max_i p_i$ and $p_{\min} = \min_i p_i$. Note that $p_{\max} < 1 - (k-1) p_{\min}$. Since $p_{\min} \geq \rho p_{\max}$ 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 these relationships in (19) we get $1/tp_i \geq 1/tp_{\max} > k\rho/t$. Use $k/t \geq \gamma$ to obtain the desired result. ■

Proof of Lemma 3. **(i)** Suppose $t \to \infty$, $k \to \infty$, and $k/t \to 0$. Note that from the assumption $\min_i p_i > \rho \max_i p_i$ it follows that $k \to \infty$ implies $p_i \to 0$ for all i, and $k/t \to 0$ implies $n_{it} \to \infty$ for all i. Write

$$\sum_{i=1}^{k} p_i Var(Y_i) = \sum_{i=1}^{k} \left(\int_{x \in \mathcal{X}_i} f_X(x)\, dx \right) \int_{y \in \mathcal{Y}} f_Y(y | x \in \mathcal{X}_i) (y - \mu_i)^2 \, dy.$$

Restrict attention to categories such that, for all i; $\sup_{x,x' \in X_i} \|x - x'\| \to 0$ as $p_i \to 0$. (For instance, this is accomplished with convex categories.) If $k \to \infty$ then the right hand side approaches

$$\int_{x \in \mathcal{X}} f_X(x) \left(\int_{y \in \mathcal{Y}} f_Y(y|x) (y - m(x))^2 \, dy \right) dx = \int_{x \in \mathcal{X}} f_X(x) \sigma^2(x)\, dx.$$

Moreover, note that if $k/t \to 0$ then, for then for all i, $\sum_{r=1}^{t} \Pr(N_{it} = r)/r \to 0$. Hence if $t \to \infty$ then $EPE(\mathcal{C}, t) \to \int_{x \in \mathcal{X}} f_X(x) \sigma^2(x)\, dx$, and thus for any $\varepsilon > 0$ there are finite numbers k' and $\bar{t}$ such that if $t > \bar{t}$, then there is a categorization with $k > k'$, such that

$$\left| EPE(\mathcal{C}, t) - \int_{x \in \mathcal{X}} f_X(x) \sigma^2(x)\, dx \right| < \varepsilon, \tag{20}$$

(ii) Restrict attention to categorizations such that $k \leq \kappa$, for some finite $\kappa > 0$. If $t \to \infty$ then $EPE(\mathcal{C}, t) \to \sum_{i=1}^{k} p_i Var(Y_i)$. The continuity of f and the assumption that $m(x) \neq \mu$ for some x, together with lemma 5, implies that there is an $\varepsilon > 0$ such that

$$\sum_{i=1}^{k} p_i Var(Y_i) > \int_{x \in \mathcal{X}} f_X(x) \sigma^2(x)\, dx + \varepsilon, \tag{21}$$

Comparing (21) with (20) reveals that 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 6 we have

$$EPE(\mathcal{C}, t) > \sum_{i=1}^{k} p_i Var(Y_i) \left(1 + \frac{\gamma\rho}{(1 - 1/t)} \right). \tag{22}$$

Let $t \to \infty$. By $t \leq k/\gamma$ this implies $k \to \infty$. As before this also implies $p_i \to 0$ for all i. For each category i let $\sup_{x,x' \in X_i} \|x - x'\| \to 0$ as $p_i \to 0$. It follows that since f is continuous, $\sup_{x,x' \in \mathcal{X}_i} |\sigma^2(x) - \sigma^2(x')|$ approaches zero. Categories may be chosen so that, as $t \to \infty$, the right hand side of (22) approaches $\int_{x \in \mathcal{X}} f_X(x) \sigma^2(x) dx (1 + \gamma\rho)$. Thus, for any ε there is some $\bar{t}$ such that if $t > \bar{t}$ then

$$\left| \sum_{i=1}^{k} p_i Var(Y_i) \left(1 + \frac{\gamma\rho}{(1 - 1/t)}\right) - \int_{x \in \mathcal{X}} f_X(x) \sigma^2(x) dx (1 + \gamma\rho) \right| < \varepsilon.$$

This implies that there is some $\bar{t}$ such that if $t > \bar{t}$ then

$$EPE(\mathcal{C}, t) > \int_{x \in \mathcal{X}} f_X(x) \sigma^2(x) dx (1 + \gamma\rho). \tag{23}$$

Comparing (23) with (20) we see that it is suboptimal to restrict attention to categorizations with $t \leq k/\gamma$. ■

7.2.1 Ex Ante Optimality for $\mathcal{X} = [a, b] \subseteq \mathbb{R}$

Proof of Theorem 1. Proposition 1 guarantees existence of a solution asymptotically. By lemma 3 the solutions must satisfy $k \to \infty$ and $k/t \to 0$ as $t \to \infty$. This allows us to use lemma 2, to write, for $x \in \mathcal{X}_i$,

$$EPE(\mathcal{C}, t)(x) = \sigma^2(x) + (m(x) - \mu_i)^2 + Var(Y_i) \frac{1}{p_i t}. \tag{24}$$

Also note

$$Var(Y_i) = \int_{s \in \mathcal{X}_i} \left[\sigma^2(s) + (m(s) - \mu_i)^2\right] f_X(s | s \in \mathcal{X}_i) ds. \tag{25}$$

From the proof of lemma 3 we know that asymptotically $\min_{\mathcal{C}'} EPE(\mathcal{C}', t)(x)$ can be attained with convex categories. Let h_i be the width of $\mathcal{X}_i$. A Taylor approximation yields

$$\begin{aligned} p_i &= \int_{a_i}^{b_i = a_i + h_i} \left[f_X(x) + f_X'(x)(s - x) + \frac{1}{2} f_X''(x)(s - x)^2 + ... \right] ds \\ &= f_X(x) h_i + f_X'(x) \left(\int_{a_i}^{b_i = a_i + h_i} (s - x) ds \right) + \frac{1}{2} f_X''(x) \int_{a_i}^{b_i = a_i + h_i} (s - x)^2 ds + ..., \end{aligned}$$

Further one may calculate

$$\int_{a_i}^{b_i = a_i + h_i} (s - x) ds = \left(\frac{h_i}{2} + a_i - x \right) h_i = \left(\frac{h_i}{2} + O(h_i) \right) h_i = O\left(h_i^2\right).$$

Thus

$$p_i = f_X(x)\, h_i + f'_X(x)\, O\left(h_i^2\right) + O\left(h_i^3\right) = f_X(x)\, h_i,$$

and consequently

$$\frac{1}{p_i t} = \frac{1}{f_X(x)\, h_i t} + O\left(\frac{1}{h_i^2 t}\right). \tag{26}$$

Next consider μ_i. Provided that $\int_{s\in\mathcal{X}_i} s f_X(s|s\in\mathcal{X}_i)\, ds = \mathbb{E}[s|s\in\mathcal{X}_i] \neq x$, a Taylor approximation yields,

$$\begin{aligned}\mu_i &= m(x) + m'(x)\left(\int_{a_i}^{b_i=a_i+h_i} s f_X(s|s\in\mathcal{X}_i)\, ds - x\right) \\ &\quad + \frac{1}{2}m''(x)\int_{a_i}^{b_i=a_i+h_i}(s-x)^2 f_X(s|s\in\mathcal{X}_i)\, ds + ... \\ &= m(x) + m'(x)\left(\mathbb{E}[s|s\in\mathcal{X}_i] - x\right) + m''(x)\, O\left(h_i^2\right) + O\left(h_i^3\right).\end{aligned}$$

Thus,

$$(m(x) - \mu_i)^2 = \left(m'(x)\,\delta h_i\right)^2 = m'(x)^2 (\delta h_i)^2 + O\left(h_i^3\right). \tag{27}$$

Next use this to find

$$\int_{s\in\mathcal{X}_i} (m(s) - \mu_i)^2 f_X(s|s\in\mathcal{X}_i)\, ds = O\left(h_i^2\right)\int_{s\in\mathcal{X}_i} m'(s)^2 f_X(s|s\in\mathcal{X}_i)\, ds. \tag{28}$$

Finally, yet another Taylor approximation yields

$$\int_{s\in\mathcal{X}_i} \sigma^2(s)\, f_X(s|s\in\mathcal{X}_i)\, ds = \sigma^2(x) + \sigma^{2\prime}(x)\, O(h_i). \tag{29}$$

Using (25)-(29) in (24) gives

$$EPE(\mathcal{C},t)(x) = \sigma^2(x) + \frac{\sigma^2(x)}{f_X(x)\, h_i t} + m'(x)^2\,\delta^2 h_i^2 + O\left(h_i^3 + \frac{1}{t}\right). \tag{30}$$

The first order condition for minimization of (30) w.r.t. h_i yields (9). Evaluating (30) at (9) yields (10). ■

7.2.2 Ex Ante Optimality for $\mathcal{X} \subseteq \mathbb{R}^n$

Proof of Proposition 2. Omitted since it follows fairly directly from lemma 3. ■

Proof of Proposition 3. Without loss of generality, assume that $\mathcal{C}'$ has one category, named 0, and $\mathcal{C}''$ has two categories, named 1 and 2. Also without loss of generality assume $Var(m(X)|X\in\mathcal{X}_i) < Var(m(X)|X\in\mathcal{X})$, for $i\in\{1,2\}$. Since $\sigma^2(x) = \sigma^2$ for

all $x \in \mathcal{X}$ the decomposition (8) becomes $Var(Y_i) = \sigma^2 + Var(m(X)|X \in \mathcal{X}_i)$. It is not difficult to show that

$$Var(m(X)|X \in \mathcal{X}_i) = \beta^2 Var_{\tilde{f}}(m(X)|X \in \mathcal{X}_i) \tag{31}$$

and similarly $Var(m(X)|X \notin \mathcal{X}_i) = \beta^2 Var_{\tilde{f}}(m(X)|X \notin \mathcal{X}_i)$. Using (31) and (**??**) one may write (after quite some manipulation),

$$\Delta \equiv EPE(\mathcal{C}'', t) - EPE(\mathcal{C}', t) = \sum_{i=1,2} p_i (M_1(i) + M_2(i) + M_3(i)), \tag{32}$$

where

$$M_1(i) = \sigma^2 \left(\sum_{r=1}^{t} \Pr(N_{it} = r) \frac{1}{r} + \frac{1}{t} \Pr(N_{it} = 0) \right) - \sigma^2 \left(\sum_{r=1}^{t} \Pr(N_{0t} = r) \frac{1}{r} + \frac{1}{t} \Pr(N_{0t} = 0) \right),$$

$$\begin{aligned} M_2(i) = {} & \beta^2 Var_{\tilde{f}}(m(X)|X \in \mathcal{X}_i) \left(1 + \sum_{r=1}^{t} \Pr(N_{it} = r) \frac{1}{r} \right) \\ & - \beta^2 Var_{\tilde{f}}(m(X)|X \in \mathcal{X}_0) \left(1 + \sum_{r=1}^{t} \Pr(N_{0t} = r) \frac{1}{r} \right) \\ & + \beta^2 \frac{1}{t} \Pr(N_{it} = 0) \left(Var_{\tilde{f}}(m(X)|X \notin \mathcal{X}_i) \right) \\ & - \beta^2 \frac{1}{t} \Pr(N_{0t} = 0) \left(Var_{\tilde{f}}(m(X)|X \notin \mathcal{X}_0) \right), \end{aligned}$$

$$M_3(i) = \Pr(N_{it} = 0) \left((\mu_{-i} - \mu_i)^2 \right) - \Pr(N_{0t} = 0) \left((\mu_{-0} - \mu_0)^2 \right).$$

Changing σ only affects $M_1(i)$ and changing β only affects $M_2(i)$. Note that $M_1(i) > 0$, so that Δ is increasing in σ^2. Since $Var_{\tilde{f}}(m(X)|X \in \mathcal{X}_0) < Var_{\tilde{f}}(m(X)|X \in \mathcal{X}_0)$, we have $M_2(i) < 0$ for large enough t (using lemma 3). Thus, there is a $\bar{t}$ such that if $t > \bar{t}$ then Δ is decreasing in β. ■

Proof of Proposition 4. We study the optimal way to categorize $\mathcal{E} \subseteq \mathcal{X}_i$ separately from $\mathcal{X} - \mathcal{E}$. Without loss of generality, assume that $\mathcal{C}'$ has one category in $\mathcal{E}$, named 0, and $\mathcal{C}''$ has two categories in $\mathcal{E}$, named 1 and 2. We may now use (32) from the proof of proposition (3). Since $Var_{\tilde{f}}(m(X)|X \in \mathcal{X}_0) < Var_{\tilde{f}}(m(X)|X \in \mathcal{X}_0)$, there is a $\bar{t}$ such that if $t > \bar{t}$ and $\alpha \geq 1$ then then $M_2(i) < 0$ and $|M_2(i)| > |M_1(i)| + |M_3(i)|$, for $i \in \{1, 2\}$. Thus, there is an $\bar{\alpha} > 1$ such that if $\alpha > \bar{\alpha}$ and $t > \bar{t}$ then $EPE(\mathcal{C}'', t) < EPE(\mathcal{C}', t)$. ■

7.3 Ex Post Optimality

Proposition 5 is proved with the help of lemma 7.

Lemma 7 ***(a)*** *For any categorization $\mathcal{C}$; $P\lim_{t\to\infty} EPE(\mathcal{C}, v^t) = \sum_{i=1}^{k} p_i Var(Y_i)$.* ***(b)*** *Consider an initial database v^{t_0-1} and any categorization $\mathcal{C} \in \Psi(v^{t_0-1})$. If $t - t_0$ objects are added to the data base then $P\lim_{t\to\infty} \widehat{EPE(\mathcal{C}, v^t)} = \sum_{i=1}^{k} p_i Var(Y_i)$.*

Proof of Lemma 7. (a) Consider a category $\mathcal{C}_i \in \mathcal{C}$. Suppose that $n_{it} \geq 1$. It can be verified that

$$\left(\bar{Y}_{it} - \mu_i\right)^2 = \left(\frac{1}{n_{it}} \sum_{s \in D_{it}} Y_s\right)^2 + \mu_i^2 - 2\mu_i \frac{1}{n_{it}} \sum_{s \in D_{it}} Y_s.$$

Let $n_{it} \to \infty$. Since, for each category i, $\{Y_s\}$ is an i.i.d. sequence with $\mathbb{E}[Y_s] = \mu_i$ we can use Kinchine's law of large numbers and Slutsky's lemma to conclude that $P\lim_{n_{it}\to\infty} \left(\bar{Y}_{it} - \mu_i\right)^2 = \mu_i^2 + \mu_i^2 - 2\mu_i\mu_i = 0$. In other words, for any $\varepsilon > 0$ and $\delta \in (0,1)$, there is an $\bar{n}$ such that if $n_{it} > \bar{n}$ then $\Pr\left(\left(\bar{Y}_{it} - \mu_i\right)^2 < \varepsilon\right) > \delta^{1/2}$. Moreover, for any $\bar{n}$ there is a $\bar{t}$ such that if $t > \bar{t}$ then $\Pr(N_{it} > \bar{n}) > \delta^{1/2}$. This implies that, for any $\varepsilon > 0$ and $\delta \in (0,1)$, there is a $\bar{t}$ such that if $t > \bar{t}$ then $\Pr\left(\left(\bar{Y}_{it} - \mu_i\right)^2 < \varepsilon\right) > \delta$. Thus, for any $\varepsilon > 0$ and $\delta \in (0,1)$, there is a $\bar{t}$ such that if $t > \bar{t}$ then

$$\Pr\left(\sum_{i=1}^{k} p_i \left(\bar{Y}_{it} - \mu_i\right)^2 < \varepsilon\right) > \delta.$$

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

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

Proof of Proposition 5. (a) Let $\mathcal{C}'$ be a categorization with k' categories. Since $m(x)$ is not constant across $\mathcal{X}$, there exists a $\gamma > 0$, and a categorization $\mathcal{C}''$ with $k'' > k'$ categories, such that

$$\sum_{i=1}^{k'} \Pr\left(X \in X_i'\right) Var(Y_i) - \sum_{i=1}^{k''} \Pr\left(X \in X_i''\right) Var(Y_i) = \gamma.$$

By lemma 7; $P\lim_{t\to\infty} \left[EPE(\mathcal{C}', v^t) - EPE(\mathcal{C}'', v^t)\right] = \gamma$. The desired result follows.

(b) Analogous to (a). ■

Proof of Proposition 6. Without loss of generality suppose that $\mathcal{C}'$ has one category, named 0, and $\mathcal{C}''$ has two categories, named 1 and 2, such that, for $i \in \{1,2\}$,

$$Var\left(m\left(X\right)|X \in \mathcal{X}_i\right) < Var\left(m\left(X\right)|X \in \mathcal{X}\right). \tag{33}$$

(a) Fix $n_{1t} \geq 1$ and $n_{2t} \geq 1$ and restrict attention to data bases with these numbers of objects in each category. Since $\sigma^2\left(x\right) = \sigma^2$ for all $x \in \mathcal{X}$ the decomposition (8) becomes $Var\left(Y_i\right) = \sigma^2 + Var\left(m\left(X\right)|X \in \mathcal{X}_i\right)$. Using this and (31) from above, one may write (after some manipulation),

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

where

$$M_1 = \beta^2 \sum_{i \in \{1,2\}} p_i \begin{pmatrix} Var_{\tilde{f}}\left(m\left(X\right)|X \in \mathcal{X}_i\right)\left(1 + \frac{1}{n_{it}} Z_i\right) \\ -Var_{\tilde{f}}\left(m\left(X\right)|X \in \mathcal{X}_0\right)\left(1 + \frac{1}{n_{0t}} Z_0\right) \end{pmatrix},$$

and

$$M_2 = \sigma^2 \sum_{i \in \{1,2\}} p_i \left(\frac{1}{n_{it}} Z_i - \frac{1}{n_{0t}} Z_0\right),$$

and, for $i \in \{0,1,2\}$,

$$Z_i = \left(\frac{\bar{Y}_{it} - \mu_i}{\sqrt{Var\left(Y_i\right)/n_{it}}}\right)^2,$$

Since $f_Y\left(y|x\right)$ is normally distributed, $f_Y\left(y|x \in \mathcal{X}_i\right)$ is normally distributed (with variance $Var\left(Y_i\right)$), and $\bar{Y}_{it}$ (being the average of i.i.d. draws) is normally distributed with variance $Var\left(Y_i\right)/n_{it}$. Hence $\sqrt{Z_i} \sim N\left(0,1\right)$ and $Z_i \sim \chi^2_{(1)}$, for $i \in \{0,1,2\}$.

It is clear that Z_1 and Z_2 are independent. However, each of Z_1 and Z_2 are correlated with Z_0. We need to know how the probability $\Pr\left(Z_i > Z_0\right)$ depends on the parameters σ^2, and β. It can be verified that $Z_i > Z_0$ is equivalent to.

$$\frac{\sigma^2 + \beta Var_{\tilde{f}}\left(m\left(X\right)|X \in \mathcal{X}_0\right)}{\sigma^2 + \beta Var_{\tilde{f}}\left(m\left(X\right)|X \in \mathcal{X}_i\right)} > \frac{\left(n_{1t} + n_{2t}\right)}{n_{it}} \frac{\left(\bar{Y}_{0t} - \mu_0\right)^2}{\left(\bar{Y}_{it} - \mu_i\right)^2}.$$

By (33), the left hand side is decreasing in σ^2 and increasing in β. Thus $\Pr\left(Z_i > Z_0\right)$, as well as $\Pr\left(M_1 > 0\right)$ and $\Pr\left(M_2 > 0\right)$, is is increasing in σ^2 and decreasing in β.

For any given realization (z_1, z_2, z_0) of (Z_1, Z_2, Z_0), if $M_2 < 0$ then $M_1 < 0$. Pick a realization (z_1, z_2, z_0). If $M_2 < 0$ and $M_1 < 0$ then $\Delta < 0$ regardless of β and σ^2. If $M_2 > 0$ and $M_1 > 0$ then $\Delta > 0$ regardless of β and σ^2. If $M_2 > 0$ and $M_1 < 0$ then Δ is decreasing in β and increasing in σ^2.

Thus by increasing σ^2 and decreasing β we shift probability from the event that $M_1 < 0$ and $M_2 < 0$ (implying $\Delta < 0$) to the event that $M_1 < 0$ and $M_2 > 0$, and from the event that $M_1 < 0$ and $M_2 > 0$ the the event that $M_1 > 0$ and $M_2 > 0$ (implying $\Delta > 0$). Moreover, for the event that $M_1 < 0$ and $M_2 > 0$, increasing σ^2 and decreasing β increases Δ.

This was for given numbers n_{1t} and n_{2t}. The same reasoning holds for any choice of $n_{1t} \geq 1$ and $n_{2t} \geq 1$.

(b) Fix $n_{1t} \geq 2$ and $n_{2t} \geq 2$ and only consider data bases with these numbers of objects in each category. We can write

$$\hat{\Delta} \equiv \widehat{EPE\left(\mathcal{C}''\right)}, v^t) - \widehat{EPE\left(\mathcal{C}'\right)}, v^t) = \frac{1}{t}\left(M_1 + M_2\right),$$

where

$$M_1 = \beta^2 \left(\sum_{i=1,2} \frac{n_{it}+1}{n_{it}-1} Var_{\tilde{f}}\left(m\left(X\right)|X \in \mathcal{X}_i\right) Z_i - \frac{n_{0t}+1}{n_{0t}-1} Var_{\tilde{f}}\left(m\left(X\right)|X \in \mathcal{X}_0\right) Z_0 \right),$$

and

$$M_2 = \sigma^2 \left(\sum_{i=1,2} \frac{n_{it}+1}{n_{it}-1} Z_i - \frac{n_{0t}+1}{n_{0t}-1} Z_0 \right),$$

and

$$Z_i = \frac{n_{it}-1}{Var\left(Y_i\right)} s_{it}^2.$$

Since $y|x \sim N\left(m\left(x\right), \sigma^2\right)$ we have $Z_i \sim \chi^2_{(n_{it}-1)}$. As before Z_1 and Z_2 are independent, while each of Z_1 and Z_2 are correlated with Z_0. It can be verified that $Z_i > Z_0$ is equivalent to.

$$\frac{\sigma^2 + \beta Var_{\tilde{f}}\left(m\left(X\right)|X \in \mathcal{X}_0\right)}{\sigma^2 + \beta Var_{\tilde{f}}\left(m\left(X\right)|X \in \mathcal{X}_i\right)} > \frac{n_{0t}-1}{n_{it}-1} \frac{s_{0t}^2}{s_{it}^2}.$$

By (33), the left hand side is decreasing in σ^2 and increasing in β. Thus $\Pr\left(Z_i > Z_0\right)$ is increasing in σ^2 and decreasing in β.

For any realization (z_1, z_2, z_0) of (Z_1, Z_2, Z_0), if $M_2 < 0$ then $M_1 < 0$. If $M_2 < 0$ and $M_1 < 0$ then $\hat{\Delta} < 0$ regardless of σ^2 and β. If $M_2 > 0$ and $M_1 > 0$ then $\hat{\Delta} > 0$ regardless of σ^2 and β. If $M_2 > 0$ and $M_1 < 0$ then $\hat{\Delta}$ is increasing in σ^2. This shows that the probability of $\hat{\Delta} > 0$ is increasing in σ^2 and decreasing in β. The same reasoning holds for any choice of $n_{1t} \geq 2$ and $n_{2t} \geq 2$. ■

Proof of Proposition 7. Similar to the proof of proposition 5, therefore omitted. ■

References

Al-Najjar, N. I. and Pai, M. (2012), Coarse decision making. Manuscript.

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

Azrieli, Y. (2009), 'Categorizing others in a large game', *Games and Economic Behavior* **67**(2), 351–362.

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.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and Regression Trees*, Wadsworth, New York.

Chater, N. (1996), 'Reconciling simplicity and likelihood principles in perceptual organization', *Psychological Review* **103**, 566–581.

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

Fan, J. and Gijbels, I. (1992), 'Variable bandwidth and local linear regression smoothers', *The Annals of Statistics* **20**, 2008–2036.

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.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning Theory*, Springer, New York.

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

Härdle, W. (1990), *Applied Nonparmetric Regression*, Cambridge University Press, Cambridge, England.

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.

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

Kant, I. (1781/87), *Critique of Pure Reason*, Macmillan, London. Translation: Kemp Smith, N., 1963.

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

Kogure, A. (1987), 'Asymptotically optimal cells for a histogram', *Annals of Statistics* **15**, 1023–1030.

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

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.

Mengel, F. (2012), 'Learning across games', *Games and Economic Behavior* **74**(2), 601–619.

Mohlin, E. (2014*a*), Asymptotically optimal regression trees. Mimeo.

Mohlin, E. (2014*b*), A metric for measurable partitions. Mimeo.

Morgan, J. M. and Sonquist, J. A. (1963), 'Problems in the analysis of survey data, and a proposal', *Journal of the American Statistical Association* **58**, 415–434.

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.

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.

Sims, C. A. (2003), 'Implications of rational inattention', *Journal of Monetary Economics* **50**(3), 665–690.

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

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.