What causes category-shifting in human semi-supervised learning?

Bryan R. Gibson (bgibson@wisc.edu)
Department of Computer Sciences
1210 W. Dayton St.
Madison, WI 53706 USA

Timothy T. Rogers (ttrogers@wisc.edu)
Department of Psychology
1202 W. Johnson Street
Madison, WI 53706 USA

Chuck W. Kalish (cwkalish@wisc.edu)
Department of Educational Psychology
1025 W. Johnson Street
Madison, WI 53706 USA

Xiaojin Zhu (jerryzhu@cs.wisc.edu)
Department of Computer Sciences
1210 W. Dayton St.
Madison, WI 53706 USA

Abstract

In a categorization task involving both labeled and unlabeled data, it has been shown that humans make use of the underlying distribution of the unlabeled examples. It has also been shown that humans are sensitive to shifts in this distribution, and will change predicted classifications based on these shifts. It is not immediately obvious what causes these shifts – what specific properties of these distributions humans are sensitive to. Assuming a parametric model of human categorization learning, we can ask which parameters or sets of parameters humans fix after exposure to labeled data and which are adjustable to fit subsequent unlabeled data. We formulate models to describe different parameter sets which humans may be sensitive to and a dataset which optimally discriminates among these models. Experimental results indicate that humans are sensitive to all parameters, with the closest model fit being an unconstrained version of semi-supervised learning using expectation maximization.

Keywords: Categorization; Semi-Supervised Learning; Cognitive Modeling

Introduction

The ability of human beings to learn and generalize category structure has been of perennial interest to cognitive science. This ability has often been studied using supervised learning experiences, where the learner is provided only with labeled examples – that is, with correct information about category membership in each learning trial.

Real-world category learning is somewhat different: while we may learn an item's category membership directly on occasion, in most experiences we simply observe objects in the environment and make implicit inferences about their category membership. That is, most of our worldly experience is unlabeled. The joint use of labeled and unlabeled data is sometimes called *semi-supervised learning* (SSL), and is a topic of considerable interest in machine learning where a range of different approaches have been developed for various learning problems (Zhu & Goldberg, 2009). A key insight from this work has been that combined use of labeled and unlabeled examples can produce quite different category structures, and in many cases more accurate structures, than learning from the labeled items alone.

The last few years have provided substantial evidence that human category learning in the lab can be strongly influenced by the distribution of unlabeled examples. In a seminal study, Zhu, Rogers, Qian, and Kalish (2007) had participants classify a set of novel, visually complex objects lying varying along a single dimension. Following a short supervised learning experience with a single item from each category, participants acquired a category boundary approximately midway between the two labeled items. Subsequently they classified a large number of additional items sampled from a bimodal distribution along the single stimulus dimension, without receiving any feedback. This unlabeled distribution was selected so that the trough between the two modes lay some distance from the learned boundary between classes. After exposure to this distribution, participants had shifted their beliefs about the location of the category boundary, aligning it with the trough in the unlabeled distribution – a behavior predicted by a simple parametric SSL model.

Subsequent work has shown that such *category shifts* – changes to beliefs about category structure arising from unlabeled learning experiences – can be quite dramatic. For instance in one study, a majority of participants ended up misclassifying the very item that had been directly taught during the initial supervised learning phase, after exposure to a dramatically shifted unlabeled distribution (C. Kalish, Kim, & Young, 2012). Other work has shown that the temporal ordering of unlabeled items can also change the acquired category structure (Zhu et al., 2010); that young children are more susceptible to influences from unlabeled data (C. W. Kalish, Zhu, & Rogers, 2014); that exposure to unlabeled distributions can lead to acquisition of quite counter-intuitive category structures (Gibson, Zhu, Rogers, Kalish, & Harrison, 2010); and that, despite receiving no feedback, people will revise a (completely accurate) classification rule learned on the fully labeled data after exposure to unlabeled examples (C. W. Kalish, Rogers, Lang, & Zhu, 2011; Lake & McClelland, 2011).

Note that, while SSL has been observed in many differing scenarios, there have been instances where the addition of unlabeled information has not impacted behavior (Vandist, De Schryver, & Rosseel, 2009; McDonnell, Jew, & Gureckis, 2012). Clearly more work is necessary to fully understand how humans make use of combinations of labeled and unlabeled data during category learning.

In this paper we consider the causes behind the category-shifts observed in semi-supervised learning studies of the kind initially described by Zhu et al. (2007) – that is, in studies where initial category structures are learned from fully

supervised experience, then those structures are observed to change after exposure to unlabeled examples. We consider two general hypotheses.

Under the first, the shifts happen because, during the initial supervised phase, participants notice and track one or more parameters of the distribution from which the labeled items are sampled, then seek to maintain a category structure that preserves the noticed parameter. For instance, in Zhu et al.'s (2007) study, the supervised phase involved learning about just two examples (one from each category), each presented 10 times with the order randomized. This experience potentially provides the learner with important information about the two classes that she may then seek to preserve when exposed to the unlabeled distribution. The learner may notice that members of each category occur about equally frequently during the supervised phase, for example. In the unsupervised phase, she may then select a category boundary that divides the unlabeled items approximately in half, preserving this frequency information. Alternatively, the learner might notice that the two categories both have approximately equal variance, and so might learn category structures that preserve roughly equal variation between members of the category.

Since the unlabeled distribution in the original study was bimodal, symmetrical about the trough with peaks of equal width, either of these strategies would lead the learner to shift the boundary to this trough. Indeed, there are many elements of the unsupervised and supervised distributions that differed in this study, any one of which might account for the observed changes in categorization behavior.

The first hypothesis, then, is that learners are trying to preserve specific parameters of the item and label distribution learned during the initial supervised phase. We refer to this as the *heuristic hypothesis*, since there is no principled reason for choosing to preserve a particular parameter from the labeled distribution. Moreover, note that there are several possible variants of the heuristic hypothesis: participants may try to preserve the relative frequencies of the two categories, their variances, their distance from the boundary, and so on.

The second hypothesis is that human beings are true semi-supervised learners – that is, they learn the category structures most likely to have generated all of the observations, labeled and unlabeled, subject to particular implicit assumptions about the relation between labeled and unlabeled examples. In the semi-supervised mixture model described by Zhu et al. (2007), the assumptions are that (i) items are sampled from a distribution in the feature space that is a mixture of Gaussian components and (ii) items sampled from the same component of the mixture receive the same category label. With these assumptions, it is possible to estimate, from all labeled and unlabeled items, the most likely components of the mixture (and their parameters) and the most likely labels associated with each component. We refer to this as the *SSL hypothesis*.

The remainder of this paper attempts to adjudicate which of these hypotheses best explains category-shifts that occur following exposure to unlabeled examples, as documented in prior work. The effort is nontrivial, insofar as it requires us to design a SSL study under which the different heuristic hypotheses and the SSL hypothesis all make different predictions about how initial category structures should change following unsupervised learning. To achieve this goal, we first formalize the nature of the learning task and describe a series of computational SSL models, each representing one of the hypotheses under consideration. Using simulations with the different models, we next discern a particular combination of supervised and unsupervised learning experiences that are expected to produce quite different learning outcomes under the different hypotheses. Finally, we report the results of behavioral studies with human subjects exposed to these learning experiences, and consider how their behaviors align with predictions of the different learning models. The results of these studies allow us to clearly determine what is causing category-shifts in human SSL.

Cognitive Models and Experimental Design

To address the question posed above we formulate a set of models and then attempt to determine which model or models best fit human behavior on a classification task.

The task we will be using for investigation is a 1D binary classification task (feature values $x \in [0,1]$ with labels $y \in \{0,1\}$). We make the strong, yet common, assumption that humans are making use of a Gaussian Mixture Model (GMM). Formally, we define the parameters of a two-component GMM as $\theta = \{w_0, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2\}$, and let $\Theta = \{\theta\}$, the set of all parametrization of this model. The learner is presented first with a set of labeled items: $L = \{(x_i, y_i)\}$, $i = 1 \ldots n_L$, drawn from a 2-component GMM defined by θ_L, followed by a set of unlabeled items $U = \{(x_j)\}$, $j = n_L + 1 \ldots n_L + n_U$ drawn from another GMM with different parameters θ_U.

We assume that, when training on L, humans find the maximum likelihood estimate (MLE) denoted $\hat{\theta}_{SL} \in \Theta$. The learner is then presented with a new set of unlabeled data U which may be drawn from a different distribution than L. Learning from U amounts to performing a search in Θ for a set of parameters that best fit the observed stimuli. Under the heuristic hypotheses, humans search some subspace of Θ for the new optimum, while under the SSL hypothesis, humans search in the whole of Θ.

We also assume the learner uses some form of expectation-maximization (EM) as the search procedure to find this optimum, the MLE on U, with $\hat{\theta}_{SL}$ as the starting point for the search (Dempster, Laird, & Rubin, 1977; Bishop, 2007). Note that, as an optimization procedure, EM can be applied even when labeled and unlabeled items come from different distributions. Although unusual in machine learning, EM used on non-iid data is plausible as a mechanism for how humans adapt. Under this assumption, participants are not focused on matching or maintaining particular aspects of the labeled distribution, but are trying to find a parametric model

that jointly "explains" the labeled and unlabeled distributions.

For example, humans might be only willing to change the proportion of one class to another ($\hat{w}_0$) leaving the rest of the learned parameters $\left(\hat{\mu}_0, \hat{\mu}_1, \hat{\sigma}_0^2, \hat{\sigma}_1^2\right)$ fixed as they were in $\hat{\theta}_{SL}$. Or, they may update both $\hat{w}_0$ and the peaks of the learned distribution $(\hat{\mu}_0,\ \hat{\mu}_1)$, but remain insensitive to changes in spread, or variance $\left(\hat{\sigma}_0^2, \hat{\sigma}_1^2\right)$, of the data. This behavior might be interpreted as the human learner "hanging on" to some beliefs learned on L.

Formalized Cognitive Models

With this task in mind we describe the cognitive models under consideration as models of human behavior.

unconstrained SL ($\hat{\theta}_{SL}$) : This model is a purely supervised learner defined by the parameters $\hat{\theta}_{SL}$. This model estimates the GMM parameters using the MLE over the labeled set L alone and holds them fixed over the unlabeled test data, in effect ignoring the unlabeled data. It is included as comparison, as we know humans are affected by U. Updates are made using

$$\hat{\mu}_0 = \frac{1}{n_0}\sum_{i=1}^{n_L} \mathbb{1}\{y_i = 0\} x_i \quad (1)$$

$$\hat{\sigma}_0^2 = \frac{1}{n_0}\sum_{i=1}^{n_L} \mathbb{1}\{y_i = 0\} (x_i - \hat{\mu}_0)^2 \quad (2)$$

$$\hat{w}_0 = \frac{n_0}{n_L} \quad (3)$$

with $n_0 = \sum_{i=1}^{n_L} \mathbb{1}\{y_i = 0\}$ ($\hat{\mu}_1$, $\hat{\sigma}_1$ are defined similarly).

unconstrained SSL ($\hat{\theta}_{SSL}$) : We define the SSL model, defined by the parameters $\hat{\theta}_{SSL}$, before the heuristic models as all other models are derived from this unconstrained version. Consideration must be given as to whether to perform EM on the full data set $(L+U)$ or to use $\hat{\theta}_{SL}$, the MLE on L, as initialization and perform EM on U alone. We choose the latter as it more closely approximates the situation faced by human learners in the task: initially exposed to L but with no additional feedback as they classify U. For each M-step of EM, the MLE estimates become

$$\hat{\mu_0} = \frac{\sum_{i=n_L+1}^{n} \gamma_i x_i}{\sum_{i=n_L+1}^{n} \gamma_i} \quad (4)$$

$$\hat{\sigma_0}^2 = \frac{\sum_{i=n_L+1}^{n} \gamma_i (x_i - \hat{\mu_0})^2}{\sum_{i=n_L+1}^{n} \gamma_i} \quad (5)$$

$$\hat{w_0} = \frac{1}{n_U}\sum_{i=n_L+1}^{n} \gamma_i \quad (6)$$

$n = n_L + n_U$, responsibilities γ_i calculated at each E-step as

$$\gamma_i = \frac{\hat{w}_0 \mathcal{N}(x_i; \hat{\mu}_0, \hat{\sigma}_0^2)}{\hat{w}_0 \mathcal{N}(x_i; \hat{\mu}_0, \hat{\sigma}_0^2) + (1-\hat{w}_0)\mathcal{N}(x_i; \hat{\mu}_1, \hat{\sigma}_1^2)} \quad (7)$$

and $\hat{\mu}_1$ and $\hat{\sigma}_1$ calculated similarly using $(1-\gamma_i)$.

All remaining models correspond to our heuristic models. They are all similar to $\hat{\theta}_{SSL}$, but assume that learning is being done by fixing *one* of the GMM parameters to the values learned on L while allowing all others to vary:

constrained means ($\hat{\theta}_{\mu}$) : Means $\hat{\mu_0}$ and $\hat{\mu_1}$ are fixed at the initialization values learned on L using (1). It is as though two prototypes are formed at the modes of the labeled distribution and retained when exposed to U. At each EM iteration t, the values of $\hat{\mu}$ at $t-1$ are simply copied forward. The variances $\hat{\sigma_0}^2, \hat{\sigma_1}^2$, weight $\hat{w_0}$ and responsibilities γ_i are updated using (5), (6) and (7) respectively.

fixed standard deviations ($\hat{\theta}_{\sigma}$) : The standard deviations $\hat{\sigma_0}$ and $\hat{\sigma_1}$ are fixed at the initialization values learned on L using (2). Here, it is the spread of the labeled data that is considered important, and is maintained. Again at each EM iteration the values of $\hat{\sigma_0}$ and $\hat{\sigma_1}$ are simply copied forward. Updates for means, weight and responsibilities are the same as in (4), (6) and (7).

fixed ratio of standard deviations ($\hat{\theta}_r$) : At initialization, the ratio of standard deviations learned on L using (2) is calculated as $r = \hat{\sigma}_0/\hat{\sigma}_1$. Again, the spread is considered most important, but now the spread of each class is allowed to vary only so long as the ratio between the two is maintained. As the parameters $\hat{\sigma}_0$ and $\hat{\sigma}_1$ are now tied, the parameter set becomes $\hat{\theta}_r = \{w_0, \mu_0, \mu_1, \sigma\}$. Reformulating the optimization function and solving for σ we find the new update equations

$$\hat{\sigma}^2 = \frac{1}{n_U}\left[\sum_{i=n_L+1}^{n} \left(\gamma_i (x_i - \mu_0)^2 + r^2 (1-\gamma_i)(x_i - \mu_1)^2\right)\right] \quad (8)$$

$$\gamma_i = \frac{w_0 \mathcal{N}(x_i; \hat{\mu}_0, \hat{\sigma}^2)}{w_0 \mathcal{N}(x_i; \hat{\mu}_0, \hat{\sigma}^2) + (1-w_0)\mathcal{N}(x_i; \hat{\mu}_1, (\hat{\sigma}/r)^2)}. \quad (9)$$

Updates for means and weight are the same as in (4) and (6).

constrained weight ($\hat{\theta}_w$) : The weight parameter $\hat{w}_0$ is fixed at the initialization value learned on L. In this case it is the frequency of each class which is considered most important to retain from the labeled data. All other updates remain unchanged.

The above models each fix one property. We also consider cognitive models which constrain multiple parameters. For example, the model $\hat{\theta}_{\sigma,w}$ has only two parameters which are free to vary: $\{\hat{\mu}_0, \hat{\mu}_1\}$, with $\hat{w}_0$, $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ fixed. This results in 5 additional models: $\{\hat{\theta}_{\sigma,w}, \hat{\theta}_{r,w}, \hat{\theta}_{\mu,w/\sigma}, \hat{\theta}_{\mu,w/r}, \hat{\theta}_{\mu,\sigma}, \hat{\theta}_{\mu,r}\}$. The model $\hat{\theta}_{\mu,w/\sigma}$ is constrained in means and weight while standard deviations are allowed to vary. The model $\hat{\theta}_{\mu,w/r}$ is constrained in means and weight while ratio of standard deviations is allowed to vary.

The final cognitive model we examine (**propL**) is not probabilistic. In this model, the learner simply calculates the proportion of negative to positive items seen in L. When the learner is then presented with U, they attempt to place a

boundary in feature space such that this proportion of negative to positive items is preserved. If the distribution generating unlabeled items is different from that generating the labeled items, the boundary learned on the labeled data will not necessarily be the same one applied to the unlabeled data. This model, $\hat{\theta}_{propL}$ has only a single parameter n_0/n_L, with the boundary $\hat{b}$ induced from this ratio:

$$\hat{b} = x_{(j)} : \frac{j}{n_U} = \frac{n_0}{n_L}, \quad j \in [1, n_U] \tag{10}$$

where $b \in [0,1]$ and $\{x_{(1)}, x_{(2)}, ..., x_{(n_U)}\}$ are the items in U, sorted by feature value. Note that this model is related to the cognitive models which preserve the GMM weight w_0. However, since this is not a GMM and classification is simply performed by a step function placed at the learned boundary b, the resulting behavior may be different.

With these cognitive models in hand we now discuss how to compare their performance to human behavioral data in order to assess which may be a better match.

Human Experiment Design

We design an experiment which attempts to discriminate which of our proposed models is a best fit to human behavior in a 1D classification task. An important aspect of this design is the construction of the dataset.

A dataset must be found which will maximally discriminate predictions made by our various models, so that it is as clear as possible which model most strongly matches human behavior. This step is similar in flavor to the *machine teaching* task proposed in (Zhu, 2013). In that setting, a teacher attempts to design an optimal dataset to teach a (potentially unknown) learner a target hypothesis. The difference here is that we do not have a target we wish our learners to learn, but instead would simply like our learners to differ as much as possible in their resulting predictions. The similarity is in the search over potential datasets.

To find a good dataset, first a labeled set L of $n_L = 50$ labeled pairs were drawn from $\theta_L = \{w_0 = 0.75, \mu_0 = 0.4, \sigma_0 = 0.12, \mu_1 = 0.8, \sigma_1 = 0.06\}$. A heuristic search was then made over a sparse grid of parameter settings θ_U, varying in all parameters. At each setting a potential unlabeled set $\tilde{U}$ of $n_U = 300$ was drawn. All cognitive models were then trained on L and predictions made on that $\tilde{U}$. We heuristically selected the dataset $L + \tilde{U}$ with the aim to produce the largest combined pairwise difference between predictions, and therefore largest discriminative power. Additionally, parameters which produced more than one decision boundary in the target range $x \in [0,1]$ were avoided.

In the end the parameters selected from which U was drawn were $\theta_U = \{w_0 = 0.25, \mu_0 = 0.3, \sigma_0 = 0.05, \mu_1 = 0.6, \sigma_1 = 0.1\}$. Plots of the chosen underlying distributions are shown in Figure 2. Importantly note that the labeled and unlabeled distributions vary in all parameters. Figure 2 also shows the estimated distributions and boundaries resulting from training each of the cognitive models on the selected dataset.

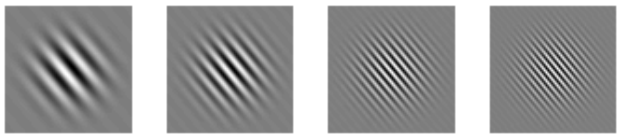

Figure 1: Stimuli at $x = 0,\ 0.25,\ 0.75$ and 1 respectively.

Participants and Procedure

Using this chosen dataset, we performed a human experiment where 49 undergraduate students, participating for partial course credit, were asked to learn a timed classification task. The 1D stimuli used were Gabor patch images varying in only the frequency dimension, with fixed rotation (Figure 1). Each participant was asked to classify the $n_L = 50$ labeled images, each classification followed by feedback indicating whether they were correct or incorrect. The participant was then asked to classify the $n_U = 300$ unlabeled stimuli, with no feedback given. All participants classified the same set of stimuli, each a randomized ordering.

Evaluation Criteria

We call the measurement we use to evaluate our models "agreement". This refers to how well a cognitive model's classification predictions *agree* with observed human behavior. Each participant $k \in \{1, \ldots, K\}$ is asked to classify the set of labeled and unlabeled items in a randomized ordering $(L,U)^{(k)}$. For each participant k we consider the first 50 + 200 items as a training set $(L,U)^{(k)}_{train}$ and the remaining 100 items as a test set $U^{(k)}_{test}$. Though there is certainly no reason to assume that humans will not continue learning on the test set, we do make the assumption that after 200 unlabeled examples, the learned boundary will have stabilized.

Each of our proposed models m is then trained on $(L,U)^{(k)}_{train}$ producing $\hat{\theta}^{(m,k)}$. For the GMM models we use the constrained versions of EM described above while *propL* is calculated directly. We can then calculate the predicted boundary $\hat{b}^{(m,k)}$ for each trained model on each dataset. For each of these model m and dataset k pairs we can then make predictions $\hat{y}^{(m,k)} = \mathbb{1}\left\{x_i^{(k)} \leq \hat{b}^{(m,k)}\right\}, i = 201, \ldots, 300$ and calculate:

$$agreement(m,k) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbb{1}\left\{\hat{y}_i^{(m,k)} = y_i^{(k)}\right\} \tag{11}$$

and total mean-agreement for each model over all K participants:

$$mean\text{-}agreement(m) = \frac{1}{K} \sum_{k=1}^{K} agreement\,(m,k) \tag{12}$$

The mean agreement scores are then used to determine which model is the best fit over all.

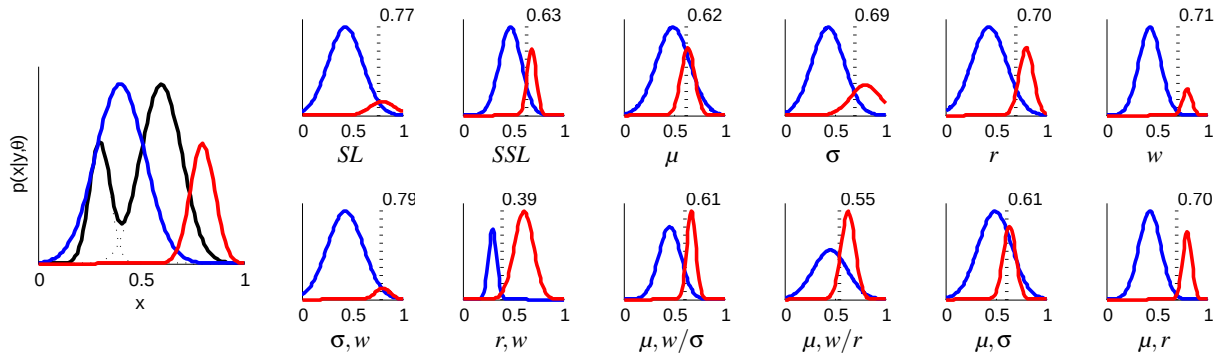

Figure 2: On the left, the ground truth labeled distributions (in blue and red) and unlabeled distribution (in black). On the right the trained models and most central prediction boundary indicated by a dotted line. The boundary for *propL* falls at 0.65.

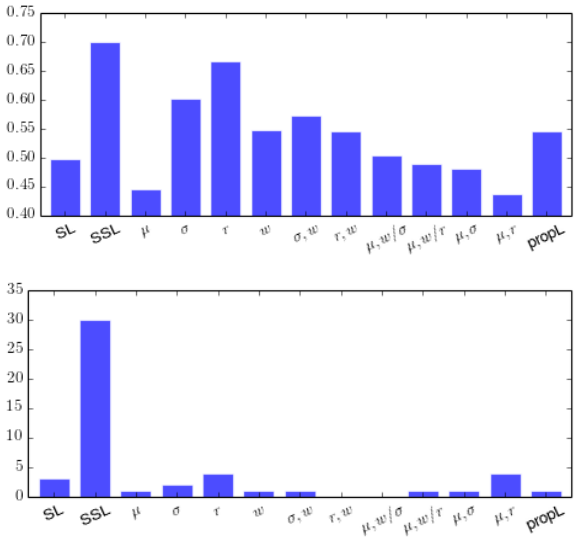

Figure 3: Top: mean agreement scores calculated for each model. Bottom: number of participants for which each model is the best match (highest agreement).

Results

Using the method described above, we found that the maximum mean-agreement score is 0.7 for the completely unconstrained model $\hat{\theta}_{SSL}$, simply standard SSL (Figure 3, top). A repeated measures one-way ANOVA shows significant difference between model agreements per subject, $F(12,624) = 26.68, p = 2 \times 10^{-16}$. Additionally, the unconstrained SSL model, $\hat{\theta}_{SSL}$, is a significantly better fit to human behavior than all other models (post-hoc multiple comparison test with Holm correction, $p \leq 0.05$), save one, SSL constrained by ratio of standard deviations ($\hat{\theta}_r$, $p = 0.11$).

If we look at which model has the best agreement per participant, unconstrained SSL $\hat{\theta}_{SSL}$ is the clear winner, having the highest agreement on 71% of participants (Figure 3, bottom).

Discussion

The question we set out to answer was what causes the category shifts seen in many semi-supervised learning studies? The two hypotheses were 1) *heuristic*: that humans notice and track some properties or set of parameters of the distribution from which labeled items are sampled, and then seek to preserve these properties when integrating information derived from unlabeled items and 2) *SSL*: that humans are true semi-supervised learners, sensitive to all properties.

In this particular categorization task, our results support the latter hypothesis: **humans are sensitive to all parameters and do not constrain their search of the parameter space.** They are sensitive to all changes in the unlabeled data distribution as they try to find the category structure most likely to have generated all observations, labeled and unlabeled.

This result should be of interest to both the cognitive psychology and machine learning communities. From the cognitive psychology perspective we can compare these results to those regarding the distinction between generative and discriminative learning (Hsu & Griffiths, 2010). Recall that to perform categorization, a generative learner attempts to model the full generating distribution $p(x,y)$ while the discriminative learner only attempts to learn a discriminating function $p(y \mid x)$. Several studies have shown that humans are capable of both types of learning (Rips, 1989; Smith & Sloman, 1994; Hsu & Griffiths, 2010). In our task where the underlying generating distribution is important due to its non-*iid* nature, the generative learning model is preferred. Our results argue that humans do in fact use a generative model for this particular task, as the *SSL* model is a better fit than the *propL* model, a discriminative model. It may be that in other tasks, where discrimination between hypothesized models, or models not in the GMM family, is still possible, this result may not be the case. Additional investigation is required to confirm that our conclusion generalizes to other situations.

From the machine learning perspective this result matches the intuition that, for best performance on transfer learning, the learner should not be constrained *a priori* without specific knowledge of the relation between the source domain and the target domain. The learner should be allowed to explore the

full parameters space when attempting to find the best fit approximation.

Finally, though the evidence points to the unconstrained hypothesis dominating over all, no significant difference was found between it and the model constrained by ratio of standard deviations. The difference here is subtle and additional work is necessary to distinguish whether this model is in fact a good approximation of human behavior or just an artifact of the current study.

Acknowledgments This work is supported by National Science Foundation grant IIS-0953219.

References

Bishop, C. M. (2007). *Pattern recognition and machine learning*. Springer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Gibson, B. R., Zhu, X., Rogers, T. T., Kalish, C. W., & Harrison, J. (2010). Humans learn using manifolds, reluctantly. In *Advances in neural information processing systems (NIPS)* (Vol. 24).

Hsu, A. S., & Griffiths, T. E. (2010). Effects of generative and discriminative learning on use of category variability. In *32nd annual conference of the cognitive science society.*

Kalish, C., Kim, S., & Young, A. (2012). How young children learn from examples: Descriptive and inferential problems. *Cognitive Science*, *36*, 1427–1448.

Kalish, C. W., Rogers, T. T., Lang, J., & Zhu, X. (2011). Can semi-supervised learning explain incorrect beliefs about categories? *Cognition*, *120*(1), 106–118.

Kalish, C. W., Zhu, X., & Rogers, T. T. (2014). Drift in children's categories: when experienced distributions conflict with prior learning. *Developmental Science*.

Lake, B. M., & McClelland, J. L. (2011). Estimating the strength of unlabeled information during semi-supervised learning. In *Proceedings of the 33rd annual conference of the cognitive science society.*

McDonnell, J. V., Jew, C. A., & Gureckis, T. M. (2012). Sparse category labels obstruct generalization of category membership. In *Proceedings of the 34th annual conference of the cognitive science society.*

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). New York, NY: Cambridge University Press.

Smith, E., & Sloman, S. (1994). Similarity-versus rule-based categorization. *Memory & Cognition*, *22*(4), 377–86.

Vandist, K., De Schryver, M., & Rosseel, Y. (2009). Semisupervised category learning: The impact of feedback in learning the information-integration task. *Attention, Perception, & Psychophysics*, *71*(2), 328–341.

Zhu, X. (2013). Machine teaching for bayesian learners in the exponential family. In *Advances in neural information processing systems (NIPS).*

Zhu, X., Gibson, B. R., Jun, K., Rogers, T. T., Harrison, J., & Kalish, C. (2010). Cognitive models of test-item effects in human category learning. In *The 27th international conference on machine learning (ICML).*

Zhu, X., & Goldberg, A. B. (2009). *Introduction to semi-supervised learning*. Morgan & Claypool.

Zhu, X., Rogers, T. T., Qian, R., & Kalish, C. (2007). Humans perform semi-supervised classification too. In *Proceedings of the 21st conference on artificial intelligence (AAAI).*