

Original citation:

Martin, J. B., Griffiths, Thomas L. and Sanborn, Adam N.. (2012) Testing the efficiency of Markov Chain Monte Carlo with people using facial affect categories. *Cognitive Science*, 36 (11). pp. 150-162.

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/35996>

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

"This is the peer reviewed version of the following article: Martin, J. B., Griffiths, Thomas L. and Sanborn, Adam N.. (2012) Testing the efficiency of Markov Chain Monte Carlo with people using facial affect categories. *Cognitive Science*, 36 (11). pp. 150-162. which has been published in final form at <http://dx.doi.org/10.1111/j.1551-6709.2011.01204.x> This article may be used for non-commercial purposes in accordance with [Wiley Terms and Conditions for Self-Archiving](#)."

A note on versions:

The version presented here may differ from the published version or, version of record, if you wish to cite this item you are advised to consult the publisher's version. Please see the 'permanent WRAP URL' above for details on accessing the published version and note that access may require a subscription.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

Running head: COMPARING METHODS FOR IDENTIFYING CATEGORIES

Testing the efficiency of Markov chain Monte Carlo with people
using facial affect categories

Jay B. Martin & Thomas L. Griffiths

Department of Psychology

University of California, Berkeley

Berkeley, CA 94720

Adam N. Sanborn

Gatsby Computational Neuroscience Unit

University of College, London

London WC1E 6BT

Keywords: emotion classification, Markov chain Monte Carlo, reverse correlation, classification images

Send all correspondence to:
Jay B. Martin
Dept. of Psychology
New York, New York 10003
Email: jbmartin@nyu.edu

Abstract

Exploring how people represent natural categories is a key step towards developing a better understanding of how people learn, form memories, and make decisions. Much research on categorization has focused on artificial categories that are created in the laboratory, since studying natural categories defined on high-dimensional stimuli such as images is methodologically challenging. Recent work has produced methods for identifying these representations from observed behavior, such as reverse correlation. We compare reverse correlation against an alternative method for inferring the structure of natural categories called Markov chain Monte Carlo with People (MCMCP). Based on an algorithm used in computer science and statistics, MCMCP provides a way to sample from the set of stimuli associated with a natural category. We apply MCMCP and reverse correlation to the problem of recovering natural categories that correspond to two kinds of facial affect (happy and sad) from realistic images of faces. Our results show that MCMCP requires fewer trials to obtain a higher-quality estimate of people's mental representations of these two categories.

Testing the efficiency of Markov chain Monte Carlo with people using facial affect categories

1. Introduction

William James (1890) famously described a baby's experiences as a "blooming, buzzing confusion." People are constantly exposed to a flood of information, which they manage in part by recognizing that objects fall into categories that can be treated similarly. Most research on categorization is based on studies of objects that vary along small numbers of easily identifiable dimensions. However, natural objects can vary along a huge number of dimensions. Thus there is a mismatch between the extremely high-dimensional environment that creates the need to form categories, and the restricted abstract stimuli we used in most experiments in cognitive science. To better understand how people represent natural categories, we must conduct experiments using realistic stimuli with large numbers of dimensions.

Working with complex natural stimuli requires developing sophisticated techniques that can be used to uncover the structure of categories. Recent studies have investigated visual category structures using novel behavioral methods, such as reverse correlation (RC; Mangini & Biederman, 2004; Kontsevich & Tyler, 2002; Gosselin & Schyns 2001; Ahumada & Lovell, 1971) and an alternative method, Markov chain Monte Carlo with People (MCMCP; Sanborn & Griffiths, 2008). Despite the sophistication of these methods, the difficulty of the problem of identifying category structure means that experiments can be costly and time-consuming, requiring large numbers of trials. Consequently, direct comparisons of the efficiency of these approaches are valuable. In this paper, we compare the performance of RC and MCMCP on a particular test case: exploring the visual categories associated with two kinds of facial affect (happy and sad).

The organization of the paper is as follows. We begin with a summary of methods for exploring the structure of natural categories in high-dimensional spaces. We follow that summary with Experiment 1, which uses the two methods, RC and MCMCP, to investigate category structures for facial expressions.

We then present Experiment 2, where we ask naïve participants to judge how well the results from Experiment 1 capture the categories of interest, allowing us to compare the efficiency of the two methods.

2. Methods for Exploring Natural Categories

Experiments investigating the structure of categories based on simple stimuli use a variety of methods, such as asking people to rate the similarity or typicality of different stimuli (Borne 1982; Nosofsky, 1988; Shepard, 1987; Tversky & Gati, 1982) and using categorization judgments to compare computational models (Ashby & Gott, 1988; Nosofsky, 1986; Shepard, Hovland, & Jenkins, 1961). However, these methods scale poorly when applied to large numbers of high-dimensional stimuli, which dramatically increases number of ratings or model parameters required to get a sense for the structure of a category. Reverse correlation and Markov chain Monte Carlo with People are both methods designed to work with naturalistic high-dimensional stimuli.

2.1. Reverse Correlation

RC provides an effective way to determine what aspect of a signal is most relevant to the classification of an object. It is typically used with stimuli that combine a non-informative signal (base), such as an expressionless face, with randomly generated noise. Often, the result of this combination will considerably change the perception of the base signal. For instance, random noise combined with an expressionless face often results in an ostensible change in facial expression. Examining how people categorize the stimuli can thus reveal the features relevant to the categorization decision.

Kontsevich and Tyler (2004) applied this method to facial affect categories by asking participants to assign the emotional expression of multiple copies of Da Vinci's Mona Lisa superimposed with uniformly distributed white noise into one of four categories: sad, slightly sad, slightly happy, and happy. They individually recorded, aggregated, and then averaged the noise patterns of the images assigned to

the happy and sad categories to create classification images, which indicate the features (in this case, pixels) associated with each category. When these classification images were superimposed with the original Mona Lisa, they changed her appearance to match. For example, the classification image from the happy category made Mona Lisa appear happier, while the classification image from the sad category made the Mona Lisa appear sadder. Their results showed that RC is a tool that can be used to uncover the information that people are using in evaluating facial affect categories.

Mangini and Biederman (2004) developed another version of RC called the sinusoidal method. Like the white noise method developed by Kontsevich and Tyler (2004), the sinusoidal method generates noise patterns for determining the relevant information in visual classification. The primary difference between the two methods is the way in which the noise filters are constructed. Sinusoidal filters are created by summing thousands of parameterized wave functions together at various scales and orientations. When combined with a base image, they appear soft and blurry, similar to the appearance of an out-of-focus picture.

Mangini and Biederman conducted several experiments to test the effect of their sinusoidal filters on gender and identity recognition. The procedure was similar the one used by Kontsevich and Tyler (2004). In one experiment, Mangini and Biederman asked participants to classify a series of sinusoidal noise-filtered images as either John Travolta or Tom Cruise. The base image was non-informative, favoring neither actor. Afterwards, when they combined the resulting classification-images to the base image, they found that the superimposed classification-images looked like the respective actors. In another experiment, they asked participants to classify a series of noise-filtered images as either male or female. Similar to the previous experiment, the base-image was non-informative, and the resulting classification-images respectively appeared male and female. Mangini and Biederman showed that the sinusoidal method performed similar to the white-noise method; however, it provided better estimates during early trials.

So far, the RC methods we have discussed use non-informative base-images, but related methods have explored the visual features people use by starting with an informative base-image. Gosselin and

Schyns (2001) proposed a method that uncovers relevant features by testing how well participants can identify category members obscured by a porous or “bubbly” mask, containing several parameterized holes. The method assumes that if a participant is able to correctly categorize an image concealed with a mask containing bubbles, then the bubbles must have exposed the relevant category information.

2.2. Markov chain Monte Carlo with People

Markov chain Monte Carlo is a class of algorithms used in computer science and statistics for generating samples from complex probability distributions (for a detailed introduction, see Gilks, Richardson, & Spiegelhalter, 1996). The basic idea is to construct a sequence of random variables that form a Markov chain, where the value of each variable depends only on the value of the previous variable in the sequence. Artful construction of this Markov chain makes it possible to specify the distribution to which the chain converges, known as the stationary distribution. Over time, the probability that a random variable in the sequence takes on a particular value converges to the probability of that value under the stationary distribution. If we construct a Markov chain that has the distribution from which we want to sample as its stationary distribution, we can generate samples from that distribution simply by simulating the Markov chain and waiting for it to converge.

One of the most popular Markov chain Monte Carlo algorithms is the Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). This algorithm follows a simple procedure to generate samples from a probability distribution $p(x)$ by producing a sequence of values of x . The sequence is initialized with an arbitrary value, x . The next value in the sequence is generated via a process with two steps. First, a candidate for the next value, x' , is chosen by sampling from an arbitrary proposal distribution specified by the designer of the algorithm, $q(x'; x)$. Second, a decision is made as to whether that proposed value will be accepted. This decision is based on the relative probability of x and x' under the target distribution $p(x)$, and tends to move to states that have higher probability. This process is iterated many times, until the Markov chain converges to the target distribution. Convergence is tested

using one of many heuristics (Gelfand, 1990; Brooks, 1998; Cowles, 1996). The early iterations are discarded in order to remove any biases resulting from the value of x used to initialize the chain, a process informally called “burn-in.” After removing the burn-in samples, the remaining samples can be treated as if they were samples from the target distribution, $p(x)$.

The fact that MCMC provides a way to sample from complex probability distributions raises the possibility that it could be used to explore the structure of category distributions. Many psychological models of categorization can be shown to be equivalent to performing Bayesian inference under the assumption that categories are associated with probability distributions over stimuli (Ashby & Alfonso-Reese, 1995). Sanborn, Griffiths, and Shiffrin (2010) proposed a method called Markov chain Monte Carlo with people (MCMCP) that exploits this correspondence, designing a task that would allow people to make the acceptance decisions in the Metropolis algorithm in a way that should yield samples from the probability distributions associated with categories. People are asked to make a series of decisions, each requiring choosing the best category member from two proposed stimuli. The stimuli that are presented in each decision correspond to the values x and x' in the Metropolis algorithm, and the choices that people make determine which proposals are accepted, with the distribution over stimuli associated with the category of interest being $p(x)$. With enough decisions, MCMCP will converge to the region of stimulus space associated with that category, and individual stimuli will be encountered with probability given by $p(x)$.

To illustrate how MCMCP works, imagine a parameter space where each dimension represents a possible facial feature, and these features can be combined to create every possible facial expression. This space is designed so that points, or faces, that are next to each other in the space are similar, but slightly different. Now, suppose two faces are drawn from the space: one at random, call it face A, and one proposed from a Gaussian distribution with the value of face A as its mean (call it face B). A person’s task is then to decide which of the two faces looks happier: face A or face B. Suppose the person selects face B. Her decision causes MCMCP to change its current state from face A to face B, and then MCMCP draws a new proposal face from a Gaussian distribution with the value of face B as its mean (call it face

C). The person has to make a new decision, which of the two faces looks happier: face B or face C.

Suppose the person decides face B. In this case, MCMCP does not change states because the proposal face was rejected. Instead, it redraws another proposal reusing the same Gaussian distribution with its mean at face B (call it face D). Now, the person has to decide between face B and face D. As a person makes these choices, the faces that are selected form a Markov chain. As it continues, this process will converge to the part of the space where happy faces are located, as shown in Fig. 1, providing samples from the probability distribution over happy faces. This process can be easily implemented in any task where the stimuli are parameterized.

MCMCP has been applied to parameterized animals, fruit, and faces (McDuff, 2010; Sanborn, Griffiths, & Shiffrin, 2010) with good results. However, the stimuli used in these experiments had a maximum of nine dimensions, which is not a large enough space to produce a set of realistic images. RC has been used with realistic images, so in the following experiments, we compare the performance of RC and MCMCP on realistic faces with 175 dimensions.

3. Experiment 1: Using Facial Affect Categories to Compare Methods for Identifying Categories

Experiment 1 compared the performance of reverse correlation and Markov chain Monte Carlo with people as methods for identifying the structure of natural categories. Both methods have been applied to facial affect categories in the past (Kontsevich & Tyler, 2004; McDuff, 2010), making this a natural domain in which to conduct a direct comparison. Each method used participants' decisions to estimate the location of visual categories associated with happy and sad faces in a high-dimensional space. We expected MCMCP to search the space more efficiently than RC, because it adapts based on participants' responses while RC relies on randomly generated stimuli.

-----Insert Figure 1 about here -----

3.1 Methods

Participants. We recruited 60 participants from the University of California, Berkeley community. Participants had normal or corrected to normal vision. Participants were randomly assigned to two conditions, with 30 participants using the MCMCP procedure and 30 using RC. Nine additional participants took part in the experiment, but were rejected based on low performance on catch trials, as described below. All participants were compensated either \$10 or 1 course credit per hour of testing.

Stimuli. All experiments were presented two Apple iMacs, and controlled by Matlab 7.4 and PsychToolbox 3.0.8 (Brainard, 1997). Participants sat approximately 44 cm away from the computer monitor. Depending on the randomly assigned condition, MCMCP or RC, participants were shown either two side-by-side 160 x 256 pixel grayscale images or a single 160 x 265 pixel grayscale image, respectively. The images were centered on a black background, and instructions were presented using white onscreen text. After each response, a trial advanced, and a new stimulus was presented.

The stimuli were produced using the California Facial Expression (CAFE) database, a collection of faces comprised of 1280 normalized grayscale portraits, containing 63 individuals (Dailey, Cottrell, & Reilly, 2001), expressing approximately eight distinct “FACS-correct” emotions (Ekman & Friesen, 1978). We applied principal component analysis to this database to generate 175 eigenfaces (Turk & Pentland, 1991), producing a 175-dimensional “face space” that captured 94% of the variance in the original images. Each stimulus was a linear combination of eigenfaces added to the mean of the faces in the CAFE database, corresponding to a point in this high-dimensional space.

Although both methods used eigenfaces to generate stimuli, the particular stimuli used in each condition were sampled differently. In the MCMCP condition, the first stimulus was sampled uniformly from the eigenspace; and then subsequent samples followed the MCMCP procedure, with new proposals being generated by sampling from a proposal distribution that depended on the previous stimulus. For 90% of trials, this proposal was a Gaussian centered on the previous stimulus, with variance 4% long each dimension. Stimuli were limited to those within one standard deviation from the mean of the CAFE database, eliminating artifacts that can result from taking linear combinations of eigenfaces in regions of

the space that are far from the original faces. If a chain from the MCMCP condition stepped outside this boundary, that chain was redirected back onto the opposite side of the space. On the other 10% of trials, the proposed stimulus was sampled uniformly from the set of points within the boundary, making it possible to make large jumps in the stimulus space. In the RC condition, faces were simply sampled uniformly from the set of points within the boundary.

Chains of stimuli in the MCMCP condition were linked within and across participants. Each of the two computers ran four separate chains (two for happy faces, two for sad faces), for a total of eight chains. On each computer, which chain was used to present the stimuli on any given trial was chosen by sampling randomly without replacement from the set of four chains, to minimize repetitions of stimuli across trials. After each participant completed 200 trials in each of the four chains, the last trial of each of the four chains was passed along to the next participant as his or her first trial. That is, the output of one participant was the input of another participant. This iterated process continued for each of the 30 participants in the MCMCP condition, resulting in eight chains, each 6000 trials in length.

Procedure. Participants were randomly assigned to either the RC or MCMCP condition. In the RC task, a participant labeled 880 randomly generated faces as either HAPPY, SAD, or NEITHER. In the MCMCP task, a participant decided which of 880 pairs of faces was either HAPPIER or SADDER, depending on whether that trial corresponded to a chain for happy faces or sad faces. Sample trials for the two conditions are shown in Fig. 2.

-----Insert Figure 2 about here -----

In both conditions, 80 of the trials completed by each participant were “catch trials.” Catch trials used FACS-correct happy and sad faces, for which the task was extremely simple. To create a constant appearance between regular trials and catch trials, catch trials were projected into the same space as the stimuli, and represented as the closest-matching linear combination of our 175 eigenfaces. If participants

misclassified more than 12 (15%) catch trials, we rejected their data and replaced that participant with a new one.

3.2. Results and Discussion

To evaluate how quickly the two methods produced good estimates of the means of the two categories, we calculated the average across all trials from each of the four sub-conditions (RC: sad and happy; MCMCP: sad and happy), and then created classification images. Classification images capture the relevant category information needed to shift the appearance of a base image (in this case the mean face) towards a specified category. Fig. 3 shows these images, the base image, and images corresponding to the estimated means of the two categories. The classification images suggest that MCMCP produced a better characterization of both categories.

-----Insert Figure 3 about here -----

The evaluation presented in Fig. 3 only provides a picture of the final estimates produced by the two methods. To examine how quickly these estimates were produced, we accumulated and averaged the data as a function of trial number. A subset of this data is presented in Fig. 4. From this figure, it qualitatively appears that MCMCP converged faster in both sub-conditions (the happy images are happier, the sad images are sadder). In order to quantify our results, we conducted Experiment 2, where we asked naive participants to evaluate each of the resulting class images.

-----Insert Figure 4 about here -----

4. Experiment 2: Evaluating the Results of the Two Methods

4.1 Methods

Participants. We recruited 20 subjects from the UC Berkeley campus community in the same manner as Experiment 1.

Stimuli. The stimuli for Experiment 2 were cumulative means of the images from Experiment 1. For the RC method, images were averaged across participants. For the MCMCP method, they were averaged across each of the four collapsed chains (two happy chains, two sad chains). These four chains were collapsed again across each of the two sub-conditions (happy and sad). Each of the conditions (RC: happy and sad; MCMCP: happy and sad) was used to generate 49 cumulative averages, resulting in a total of 196 images.

Procedure. Participants evaluated the 196 images, which were randomized across trial number and method for each participant. After the presentation of each image, participants were told that they were going to see 196 face images and that they need to rank each one on a scale from 1 to 9, with 1 being saddest and 9 being happiest.

4.2. Results and Discussion

The mean ratings are shown in Fig. 5, broken down by method and trial number. The quantitative results confirm the qualitative impression from Experiment 1: MCMCP produces more representative estimates of the means of the two categories, and does so more rapidly than RC. This result was confirmed by a two-way repeated-measures ANOVA for each condition, looking for an overall effect of method and trial number. In the happy condition, a significant main effect was obtained for method, $F(1, 39) = 6155.67, p < .001$, and for trial number, $F(48, 1872) = 21.70, p < .001$. In the sad condition, a significant main effect was obtained for method, $F(1, 39) = 1689.03, p < .001$, and for trial number, $F(48, 1872) = 3.96, p < .001$. In the happy condition, MCMCP had significantly higher average rankings ($M = 6.81$) than did RC ($M = 6.20$). In the sad condition, MCMCP had significantly lower average

rankings ($M = 3.51$) than did RC ($M = 4.46$). That is, MCMCP produced, on average, happier faces in the happy condition and sadder faces in the sad condition.

That analysis was followed by a paired samples t-test $p < .001$ for the effect of method at each trial number, providing a Bonferroni correction that kept the overall alpha level below 0.05. Asterisks in Fig. 5 indicate the trial numbers for which method had a significant effect. In the happy condition, 10 cumulative faces were marked significant and in the sad condition, 28 cumulative faces were marked significant.

-----Insert Figure 5 about here -----

5. Conclusion

We set out to compare two methods for inferring the structure of natural categories, reverse correlation and Markov chain Monte Carlo with people. Experiment 1 showed that both RC and MCMCP were able to produce estimates of the features associated with different visual categories. However, the results of Experiment 2 show that MCMCP outperformed RC in both the overall quality of these estimates, as well as the number of trials required to achieve that performance. Thus, as we anticipated, MCMCP reduced the number of trials needed to extract relevant information about the structure of these categories. Our results suggest that MCMCP can be a powerful alternative approach to reverse correlation.

The property of MCMCP that made this improvement in performance possible was using a more sophisticated method for generating the stimuli about which people make judgments. Conventional RC experiments rely on uniform sampling, a process that is blind to decisions made by participants. Uniform distributions generate samples equally over an entire parameter space, disregarding much relevant information. Generating samples in this fashion results in many uninformative trials, especially when the parameter space is large and the category distribution is relatively small. MCMCP overcomes this

problem by generating samples methodically, restricting the scope to only the most informative regions of a parameter space. Consequently, MCMCP seldom visits the less informative areas that provide little or no weight in the overall category distribution. In high-dimensional stimulus spaces, this efficiency can considerably reduce the number of trials required for estimation, and thus the number of participants. Since many natural categories are defined over stimuli that are best represented in high-dimensional spaces, we anticipate that this property of MCMCP will make it a tool that can be used to solve a broad range of problems.

Beyond the insight that these methods give into specific domains such as the perception of facial affect, developing new tools for estimating the structure of natural categories opens up new possibilities for answering questions about categorization that have previously been studied only with simple artificial stimuli. For example, building up a catalogue of natural categories with different kinds of stimuli could provide a novel way to evaluate computational models of categorization, by examining the extent to which natural categories are consistent with the assumptions of those models. Efficient tools for working with complex stimuli and natural categories provide the key to helping us understand how it is that people make sense of the blooming, buzzing world.

References

- Ahumada, A. J., Jr. (1996). Perceptual classification images from Vernier acuity masked by noise. *Perception (Supplement)*, 26, 18.
- Ashby, F. G., & Alfonso-Reese, L. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39, 216-233.
- Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, memory, and Cognition*, 14(1), 33-53.
- Bourne, L. E., Jr. (1982). Typicality effects in logically defined categories. *Memory & Cognition*, 10, 3-9.
- Barker, A. A., (1965). Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, 18, 119-135.
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433-436.
- Brooks, S. P., & Roberts, G. O. (1998). Assessing convergence of Markov chain Monte Carlo. *Statistics and Computing*, 8, 319-335.
- Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883-904.
- Dailey M., Cottrell, G., & Reilly, J. (2001). California Facial Expressions (CAFE). Unpublished digital images: University of California, San Diego.
- Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1, 49-98.
- Ekman, P., & Friesen, W. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto, CA: *Consulting Psychologists Press*.
- Ekman, P. (1999). Basic Emotions. In Dalgleish, T. & Power, M. (eds.). *The Handbook of Cognition and Emotion*. John Wiley & Sons, Ltd, Sussex, U.K.
- Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, 85,

972–985.

- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, Suffolk.
- Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research*, *41*, 2261-2271.
- James, W. (1890). *The Principles of Psychology (Vol. 1)*. New York: Holt.
- Kontsevich, L., & Tyler, C. (2004). What makes Mona Lisa smile? *Vision Research*, *44*, 1493-1498.
- Mackay, D. (2003). *Information Theory, Inference & Learning Algorithms*. Cambridge: Cambridge University Press.
- Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimation information employed for face classifications. *Cognitive Science*, *28*, 209-226.
- McDuff, D. (2010). A human-Markov chain Monte Carlo method for investigating facial expression categorization. In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the 10th International Conference on Cognitive Modeling* (pp. 151-156). Philadelphia, PA: Drexel University.
- Metropolis, N., Rosenbluth, W., Rosenbluth, N., Teller, H., & Teller, E., (1953). *J. Chem. Phys.*, *21*, 1087.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39-57.
- Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 700–708.
- Sanborn, A. N., & Griffiths, T. L. (2008). Markov chain Monte Carlo with people. *Advances in Neural Information Processing Systems* *20*.
- Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*(13, Whole No. 517).
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

Turk, M., & Pentland A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 71-86.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, 89(2), 123-54.

Author Note

This work was supported by a Robert & Colleen Haas grant to JBM, and grant number IIS-0845410 from the National Science Foundation to TLG.

Figure Captions

Figure 1. A simplified illustration of Markov chain Monte Carlo with people (MCMCP).

3.1 Methods

Figure 2. Sample trials from the two methods evaluated in Experiment 1. The panel on the left is a trial from the MCMCP condition, and the panel on the right is a trial from the RC condition.

Figure 3. Results of averaging all trials from Experiment 1 for both the RC and MCMCP conditions. The base images were identical for each task, corresponding to the mean face. The classification images show the features representative of the happy and sad categories, with lighter regions showing pixels that increase in intensity and darker regions showing pixels that decrease in intensity. The addition of these classification images to the base image results in the mean images for the two categories.

Figure 4. Cumulative mean images for the two categories and two estimation methods as a function of trial number.

Figure 5. Results of Experiment 2, showing ratings of happiness and sadness of cumulative averages of images from trials in Experiment 1 broken down by estimation method and trial number. Lines show mean ratings, with error bars indicating one standard error. Asterisks indicate the trial numbers for which method had a significant effect with $p < .001$ by a two samples t-test.









