Learning Objects by Learning Models: Finding Independent Causes and Preferring Simplicity

Gergő Orbán (ogergo@colbud.hu)
Collegium Budapest Institute for Advanced Study
2 Szentháromság utca, Budapest, 1014, Hungary

József Fiser (fiser@brandeis.edu)
Dept of Psychology and Volen Center for Complex Systems, Brandeis University
Waltham, MA 02454, USA

Richard N. Aslin (aslin@rochester.edu)
Dept of Brain and Cognitive Sciences, Center for Visual Science, University of Rochester
Rochester, New York 14627, USA

Máté Lengyel (lmate@gatsby.ucl.ac.uk)
Gatsby Computational Neuroscience Unit, University College London
17 Queen Square, London WC1N 3AR, United Kingdom

Abstract

Humans make optimal perceptual decisions in noisy and ambiguous conditions. Computations underlying such optimal behavior have been shown to rely on Bayesian probabilistic inference. A key element of Bayesian computations is the generative model that determines the statistical properties of sensory experience. The goal of perceptual learning can thus be framed as estimating the generative model from available data. In previous studies, the generative model that subjects had to infer was relatively simple, its structure was also assumed to be known a priori, so that only a few model parameters had to be estimated. We investigated whether humans are capable of inferring more complex generative models from experience. In a completely unsupervised perceptual task subjects learnt subtle statistical properties of visual scenes consisting of 'objects' that could only be identified by their statistical contingencies not by low-level features. We show that human performance in this task can be accounted for by Bayesian learning of model structure and parameters within a class of models that seek to explain observed variables by a minimum number of independent hidden causes.

Introduction

There is a growing number of studies supporting the classical view of perception as probabilistic inference (Helmholtz, 1962; Barlow, 1990). These studies demonstrated that human observers parse sensory scenes by performing optimal estimation of the parameters of the objects involved (Ernst & Banks, 2002; Körding & Wolpert, 2004; Kersten, Mamassian, & Yuille, 2004). A core element of this Bayesian probabilistic framework is an internal model of the world, the generative model. The generative model serves as a basis for inference by specifying how the different sources of currently available sensory evidence are integrated with prior expectations about the external world. Thus, in order to understand the computational principles of perception, it is important to characterize the forms of generative models that are available for perceptual inference.

Most previous studies testing the Bayesian framework in human psychophysical experiments used fundamentally restricted generative models of perception. The generative models considered in these studies consisted of a few observed and hidden variables, and only a limited number of parameters that needed to be adjusted by experience (Ernst & Banks, 2002; Körding & Wolpert, 2004; Kersten et al., 2004; Weiss, Simoncelli, & Adelson, 2002). More importantly, these generative models were tailor-made to the specific psychophysical tasks presented in the experiments. However, in principle, inference can be performed on several levels: the generative model can be used for inferring the values of hidden variables from observed information, and the generative model itself may also be inferred from previous experience (MacKay, 1992). Thus, it remains to be shown whether more flexible, 'open-ended' generative models could be used and learned by humans during perception.

We used an unsupervised visual learning task to show that a general class of generative models (Sigmoid Belief Networks) quantitatively reproduced experimental data, including paradoxical aspects of human behavior, when not only the parameters of these models but also their structure (ie. the number and identity of hidden variables) was subject to learning. Crucially, the applied Bayesian model learning embodied the Automatic Occam's Razor (AOR) effect (MacKay, 1995) that preferred the models that were 'as simple as possible, but no simpler'. This process led to the extraction of independent causes that efficiently and sufficiently accounted for sensory experience, without a pre-specification of the number or complexity of potential causes.

All the presented experimental results were reproduced and had identical roots in our simulations: the model that was most probable based on the training data developed hidden variables corresponding to the real chunks that were originally used to generate the training scenes. These results demonstrate that humans can infer complex models from experience and implicate Bayesian model learning as a powerful computation underlying such basic cognitive phenomena as the decomposition of visual scenes into meaningful chunks.[1]

Experimental Paradigm

Human adult subjects were trained and then tested in four different experiments using the same unsupervised learning paradigm. Subjects saw a sequence of complex visual scenes consisting of 6 of 12 abstract unfamiliar black *shapes* arranged on a 3x3 (Exp 1) or 5x5 (Exps 2-

[1] A previous version of this work has been already presented elsewhere to a rather different audience (Orbán, Fiser, Aslin, & Lengyel, 2006). The main novel aspect of the version presented here is the inclusion of the Gestalt-based model, and a clearer explanation of human performance on embedded combos. This also enabled us to account for the effects of traning length and performance on embedded triplets in Exp. 4.

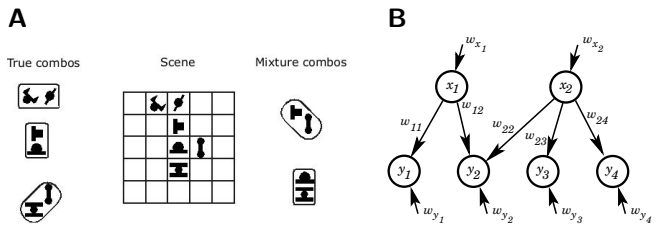

Figure 1: *A,* Experimental design. *B,* Explanation of graphical model parameters.

4) white grid (Fig. 1A). Unbeknownst to subjects, various subsets of the shapes were arranged into fixed spatial combinations *(combos)* (doublets, triplets, quadruplets, depending on the experiment). Whenever a combo appeared on a training scene, its constituent shapes were presented in an invariant spatial arrangement, and in no scenes elements of a combo could appear without all the other elements of the same combo also appearing. Subjects were presented with 100–200 training scenes, each scene was presented for 2 seconds with a 1-second pause between scenes. No specific instructions were given to subjects prior to training, they were only asked to pay attention to the continuous sequence of scenes.

The test phase consisted of 2-alternative forced choice (2AFC) trials, in which two arrangements of shapes were shown sequentially in the same grid that was used during training, and subjects were asked which of the two scenes was more familiar based on the training. One of the presented scenes was either a combo that was actually used for constructing the training set *(true combo)*, or a part of it *(embedded combo)* (e.g., a pair of adjacent shapes from a triplet or quadruplet combo). The other scene consisted of the same number of shapes as the first scene in an arrangement that might or might not have occurred during training, but was in fact a mixture of shapes from different true combos *(mixture combo)*.

Here four experiments are considered that assess various aspects of human observational learning, the full set of experiments are presented elsewhere (Fiser & Aslin, 2001, 2005). Each experiment was run with 20 naïve subjects.

Experiment 1 Our first goal was to establish that humans are sensitive to the statistical structure of visual experience, and use this experience for judging familiarity. In the baseline experiment 6 doublet combos were defined, three of which were presented simultaneously in any given training scene, allowing 144 possible scenes (Fiser & Aslin, 2001). Because the doublets were not marked in any way, subjects saw only a group of random shapes arranged on a grid. The occurrence frequency of doublets and individual elements was equal across the set of scenes, allowing no obvious bias to remember any element more than others. In the test phase a true and a mixture doublet were presented sequentially in each 2AFC trial. The mixture combo was presented in a spatial position that had never appeared before.

Experiment 2 In the previous experiment the elements of mixture doublets occurred together fewer times than elements of real doublets, thus a simple strategy based on tracking co-occurrence frequencies of shape-pairs would be sufficient to distinguish between them. The second, frequency-balanced experiment tested whether humans are sensitive to higher-order statistics (at least cross-correlations, which are co-occurence frequencies normalized by respective invidual occurence frequencies).

The structure of Experiment 1 was changed so that while the same 6 doublet combos were used as before, their appearance frequency became non-uniform introducing *frequent* and *rare combos*. Frequent doublets were presented twice as often as rare ones, so that certain mixture doublets consisting of shapes from frequent doublets appeared just as often as rare doublets. Note, that the frequency of the constituent shapes of these mixture doublets was higher than that of rare doublets. The training session consisted of 212 scenes, each scene being presented twice. In the test phase, the familiarity of both single shapes and doublet combos was tested. In the doublet trials, rare combos with low appearance frequency but high correlations between elements were compared to mixture combos with higher element and equal pair appearance frequency, but lower correlations between elements.

Experiment 3 This experiment tested whether human performance in this paradigm can be fully accounted for by learning cross-correlations. Here, four triplet combos were formed and presented with equal occurrence frequencies. 112 scenes were presented twice to subjects. In the test phase two types of tests were performed. In the first type, the familiarity of a true triplet and a mixture triplet was compared, while in the second type doublets consisting of adjacent shapes embedded in a triplet combo *(embedded doublet)* were tested against mixture doublets.

Experiment 4 This experiment compared directly how humans treat embedded and independent (non-embedded) combos of the same size. Here two quadruplet combos and two doublet combos were defined and presented with equal frequency. Each training scene consisted of six shapes, one quadruplet and one doublet. 120 such scenes were constructed and subjects were either presented with each scene once (half training), or twice (full training). In the test phase four types of 2AFC trials were used: true against mixture quadruplets; embedded against mixture doublets; true against mixture doublets; and embedded against mixture triplets.

Modeling framework

The goal of Bayesian learning is to 'reverse-engineer' the generative model that could have generated the training data. Because of inherent ambiguity and stochasticity assumed by the generative model itself, the objective is to establish a *probability distribution* over possible models. Importantly, because models with parameter spaces of different dimensionality are compared, the marginal likelihood term will prefer the simplest model (in our case, the one with fewest parameters) that can effectively account for (generate) the training data due to the AOR

effect in Bayesian model comparison (MacKay, 1995).

Sigmoid belief networks The class of generative models we consider is that of two-layer sigmoid belief networks (SBNs, Fig. 1B). The same modelling framework has been successfully aplied to configural learning in animal classical conditioning (Courville, Daw, Gordon, & Touretzky, 2004; Courville, Daw, & Touretzky, 2005). The SBN architecture assumes that the state of observed binary variables (y_j, in our case: shapes being present or absent in a training scene) depends through a sigmoidal activation function on the state of a set of hidden binary variables ($\mathbf{x}$), which are not directly observable:

$$P\left(y_j = 1|\mathbf{x}, \mathbf{w}_m, m\right) = \left(1 + \exp\left(-\sum_i w_{ij}x_i - w_{y_j}\right)\right)^{-1} \quad (1)$$

where w_{ij} describes the (real-valued) influence of hidden variable x_i on observed variable y_j, w_{y_j} determines the spontaneous observed activation bias of y_j, m indicates the model structure, including the number of latent variables and identity of the observeds they can influence (the w_{ij} weights that are allowed to have non-zero value), and $\mathbf{w}_m$ stands for all the parameters (w_{ij}, w_{y_j}, w_{x_i} within model structure m).

Observed variables are independent conditioned on the latents (i.e. any correlation between them is assumed to be due to shared causes), and latent variables are marginally independent and have Bernoulli distributions parametrised by their biases, $\mathbf{w}_x$:

$$\begin{aligned} P\left(\mathbf{y}|\mathbf{x}, \mathbf{w}_m, m\right) &= \prod_j P\left(y_j|\mathbf{x}, \mathbf{w}_m, m\right) \quad (2) \\ P\left(\mathbf{x}|\mathbf{w}_m, m\right) &= \prod_i \left(1 + \exp\left((-1)^{x_i} w_{x_i}\right)\right)^{-1} \end{aligned}$$

Finally, training scenes $\mathbf{y}^{(t)}$ are assumed to be *iid* samples from the same generative distribution, and so the probability of the training data ($\mathcal{D}$) given a specific model is:

$$\begin{aligned} P\left(\mathcal{D}|\mathbf{w}_m, m\right) &= \prod_t P\left(\mathbf{y}^{(t)}|\mathbf{w}_m, m\right) \quad (3) \\ &= \prod_t \sum_{\mathbf{x}} P\left(\mathbf{y}^{(t)}|\mathbf{x}, \mathbf{w}_m, m\right) P\left(\mathbf{x}|\mathbf{w}_m, m\right) \end{aligned}$$

The 'true' generative model that was actually used for generating training data in the experiments (Section 2) is closely related to this model, with the combos corresponding to latent variables. The main difference is that here we ignore the spatial aspects of the task, i.e. only the occurrence of a shape matters but not *where* it appears on the grid. Although in general, space is certainly not a negligible factor in vision, human behavior in the present experiments depended on the mere presence or absence of shapes sufficiently strongly so that this simplification did not cause major confounds in our results.

Training Establishing the posterior probability of any given model is straightforward using Bayes' rule:

$$P\left(\mathbf{w}_m, m|\mathcal{D}\right) \propto P\left(\mathcal{D}|\mathbf{w}_m, m\right)\ P\left(\mathbf{w}_m, m\right) \quad (4)$$

where the first term is the likelihood of the model (Eq. 3), and the second term is the prior distribution of models. Prior distributions for the weights were: $P\left(w_{ij}\right) = \text{Exponential}\left(4\right)$, $P\left(w_{x_i}\right) = \text{Laplace}\left(0, 4\right)$, $P\left(w_{y_j}\right) = \text{Laplace}\left(-2, 4\right)$. The prior over model structures preferred simple models and was such that the distributions of the number of latents and of the number of links conditioned on the number of latents were both $\text{Geometric}\left(0.1\right)$. The effect of this preference is 'washed out' with increasing training length as the likelihood term (Eq. 3) sharpens.

Testing When asked to compare the familiarity of two scenes ($\mathbf{y}^A$ and $\mathbf{y}^B$) in the testing phase, the optimal strategy for subjects would be to compute the posterior probability of both scenes based on the training data

$$P\left(\mathbf{y}^Z|\mathcal{D}\right) = \sum_m \int d\mathbf{w}_m \sum_{\mathbf{x}} P\left(\mathbf{y}^Z, \mathbf{x}|\mathbf{w}_m, m\right)\ P\left(\mathbf{w}_m, m|\mathcal{D}\right) \quad (5)$$

and always (ie, with probability one) choose the one with the higher probability. However, as a phenomenological model of all kinds of possible sources of noise (sensory noise, model noise, etc) we chose a soft threshold function for computing choice probability:

$$P\left(\text{choose A}\right) = \left(1 + \exp\left(-\beta\ \log \frac{P\left(\mathbf{y}^A|\mathcal{D}\right)}{P\left(\mathbf{y}^B|\mathcal{D}\right)}\right)\right)^{-1} \quad (6)$$

and used a single β to fit experimental data from all subjects ($\beta = \infty$ corresponds to the optimal strategy, $\beta = 1$ corresponds to probability matching). Here $\log P\left(\mathbf{y}^A|\mathcal{D}\right)/P\left(\mathbf{y}^B|\mathcal{D}\right)$ is the log probability ratio (LPR).

Note that when computing the probability of a test scene, we seek the probability that exactly the given scene was predicted by the learned model. This means that we require not only that all the shapes that are present in the test scene are predicted to be present, but also that all the shapes that are absent from the test scene are predicted to be absent. A different scheme, in which only the presence but not the absence of the shapes need to be matched (i.e. absent observeds are marginalized out just as latents are in Eq. 5) could also be pursued, but the results of the embedding experiments (Exp. 3 and 4, see below) discourage it.

The model posterior in Eq. 4 is analytically intractable, therefore an exchange reversible-jump Markov chain Monte Carlo sampling method (Courville et al., 2004; Green, 1995; Iba, 2001) was applied, that ensured fair sampling from a model space containing subspaces of differring dimensionality, and integration over this posterior in Eq. 5 was approximated by a sum over samples.

Results

In the baseline experiment (Experiment 1) human subjects were trained with six equal-sized doublet combos and were shown to recognize true doublets over mixture doublets (Fig. 2A). When the same training data was used to compute the choice probability in 2AFC

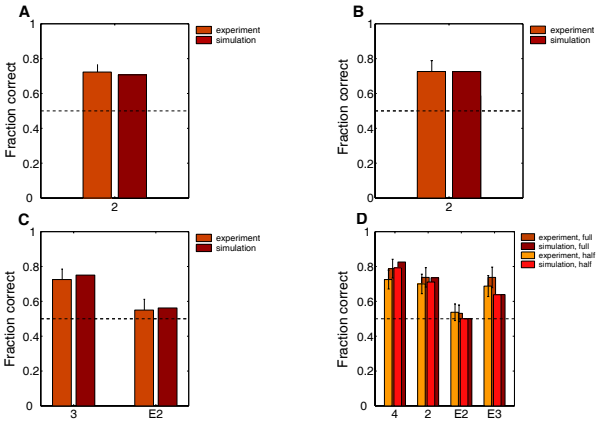

Figure 2: Comparison of human and model performance in four experiments. *A–D,* Results from baseline (Exp. 1), frequency-balanced (Exp. 2), triplet (Exp. 3), and quadruple experiments (Exp. 4), respectively. Bars show fraction of 'correct' responses (choosing a true or embedded combo over a mixture combo) for human experiments (*orange*, average over subjects ±SEM), and 'correct' choice probabilities (Eq. 7) for computer simulations *(brown)*. Labels below the bars denote the type of test trial: *2,* doublet; *E2,* embedded doublet; *3,* triplet; *E3,* embedded triplet; *4,* quadruplet. *D,* results from experiments *(yellow)* and simulations *(red)* at half training length *(half)* and full training length *(full)* are also shown. Dotted lines show 50% correct chance level performance.

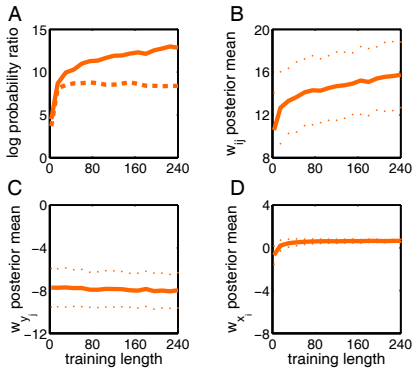

Figure 3: Evolution of model variables with increasing training length in a pilot triplet simulation performed with two triplets. *A,* LPRs for triplet *(solid line)* and embedded doublet *(dashed line)* 2 AFC tests. *B-D,* Mean *(solid line)* ±1s.d. *(dotted lines)* of posterior parameter distributions in the MAP model structure for latent-to-observed weights, w_{ij} *(B)*; observed biases, w_{y_j} *(C)*; and latent biases, w_{x_i} *(D)*.

tests with model learning, true doublets were reliably preferred over mixture doublets. Also, the model with maximal *a posteriori* (MAP) probability showed that the discovered latent variables corresponded to the combos generating the training data (data not shown).

In Experiment 2, we sought to answer the question whether the statistical learning demonstrated in Experiment 1 was solely relying on co-occurrence frequencies, or was using something more sophisticated, such as at least cross-correlations between shapes. Bayesian model learning, as well as humans, could distinguish between rare doublet combos and mixtures from frequent doublets (Fig. 2B) despite their balanced co-occurrence frequencies.

We were interested whether the performance of humans could be fully accounted for by the learning of cross-correlations, or they demonstrated more sophisticated computations. In Experiment 3, training data was composed of triplet combos, and beside testing true triplets against mixture triplets, we also tested embedded doublets (pairs of shapes from the same triplet) against mixture doublets (pairs of shapes from different triplets). If learning only depends on cross-correlations, we expect to see similar performance on these two types of tests. In contrast, human performance was significantly different for triplets (true triplets were preferred) and doublets (embedded and mixture doublets were not distinguished) (Fig. 2C). This may be seen as Gestalt effects being at work: once the 'whole' triplet is learned, its constituent parts (the embedded doublets) loose their significance. Our model reproduced this behavior and provided a straightforward explanation. The main effect of extensive training in the simulations was increasing certainty about the correct model structure (data not shown) and that given that model structure there was a strong causal link between the appearence of a combo and its constituent shapes (shift of w_{ij} weights towards more positive values, Fig. 3B). Given such a confident causal link in the learned model, whenever a combo appeared it could almost only produce triplets, therefore doublets (embedded and mixture alike) could only be created by spontaneous independent activation of individual shapes. In other words, doublets were seen as mere noise that naturally produced embedded and mixture doublets with equal chance. An interesting prediction from this argument is that more training should just further accentuate this affect, that is embedded doublets should become even less preferred Fig. 3A).

The fourth experiment tested explicitly whether embedded combos and equal-sized independent true combos are distinguished and not only size effects prevented the recognition of embedded small structures in the previous experiment. Both human experiments and Bayesian model learning demonstrated that quadruple combos as well as stand-alone doublets were reliably recognized (Fig. 2D), while embedded doublets were not. Moreover, longer training did not help with the recognition of embedded doublets just as predicted before. However, the preference for embedded triplets against mixture triplets was signifcantly above chance, and thus above the level at which embedded doublets were preferred. Our simluations could also account for this effect. The probability of a given combo being produced by the spontaneous activation of its constituent shapes (noise) decreases exponentially with its size. Therefore,

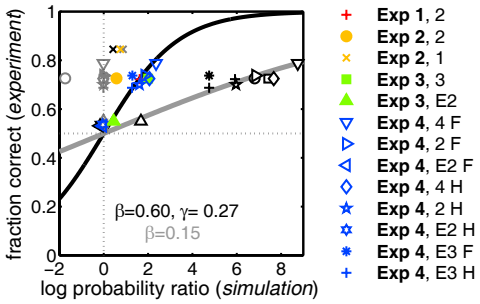

Figure 4: Aggregate plot of the fit of experimental data with simulations. *Black symbols,* LPRs from statistics-based computation; *gray symbols,* LPRs from Gestalt-based computations; *colored symbols,* LPRs from the mixture predictive distribution. *Solid gray line,* sigmoid fit on the 12 black symbols from the data of Fig. 2 (*Exp2, 1* was not included in the fit). *Solid black line,* sigmoid fit on the colored symbols. *Symbol colors*, different experiments: (*red,* baseline (Exp. 1); *yellow,* frequency-balanced (Exp. 2); *green,* triplet (Exp. 3); *blue,* quadruplet (Exp. 4). *Symbol shapes*, different test trial types in a given experiment (see Fig. 2 for an explanation of the legend), F and H in Exp. 4 denote full and half training length, respectively. *Dotted lines* show that fits were constrained to be unbiased, ie. to go across $(0, 0.5)$.

while the appearance of an embedded doublet could be explained away by noise as before, the appearance of an embedded triplet could only be explained by the appearance of a true quadruplet that failed to activate one of its shapes. The only likely mechanism for explaining triplets could thus produce embedded but not mixture triplets, hence the preference for the former.

In order to demonstrate the predictive power of our approach we have replotted all the experimental data (12 data points) against our theoretical predictions from Figure 2 and obtained a strong quantitative match ($r = 0.95$, Fig. 4, *black symbols, solid grey line*). Importantly, we used the same parameter set for modelling all the experiments. Only one parameter was tuned for fitting the data ($\beta = 0.15$, see Eq. 6), while no specific effort was made to optimize the rest of the parameters; indeed changing them in a wide range did not affect the qualitative outcome of our simulations.

In Experiment 2, although rare doublet combos were preferred over frequency-balanced cross-pairs, humans – but not the model – learned about the frequencies of their constituent shapes. Human subjects preferred constituent single shapes of frequent doublets over those of rare doublets (Fig. 5A). Furthermore, experiments showed a slightly greater preference for singlets than for doublets. Since the approach concentrating on the purely statistical aspects of perceptual learning in this task failed to reproduce these findings, we hypothesized that Gestalt cues present in training scenes affect the learning process, therefore cannot be neglected. These Gestalt cues – the clear spatial disjunctness of individual shapes, further accentuated by the grid lines separating them – bias any naïve observer to treat individual shapes as truly independent in spite of conflicting statistical evidence. We modeled this Gestalt-based bias phenomenologically when computing the predictive probabilities of the two test scenes in a 2AFC test trial. The final predictive probability of a scene was a mixture of the predictions of a purely statistics-based model (described before), P_{stats}, and a so-called Gestalt-based model, P_{Gestalt}, which was constrained to have zero latents (ie, the prior distribution on the number of latent variables and correspondingly on the number of links was $\delta(0)$) and thus only learnt about the occurrence frequencies of individual shapes:

$$P_{\text{mixture}}\left(\mathbf{y}^Z|\mathcal{D}\right) \propto P_{\text{stats}}^{\gamma}\left(\mathbf{y}^Z|\mathcal{D}\right) \cdot P_{\text{Gestalt}}^{(1-\gamma)}\left(\mathbf{y}^Z|\mathcal{D}\right) \quad (7)$$

Both P_{stats} and P_{Gestalt} were computed based on Equation 5. The same mixing coefficient γ was fitted to all experimental data (Fig. 4).

We used the previous 12 data points complemented with the data on singlet recognition in Experiment 2 and obtained a good fit when fitting them with the noise parameter β and mixing coefficient γ ($r = 0.76$, $\beta = 0.60$ and $\gamma = 0.27$; Fig. 4, *colored symbols*, *solid black line*). Although the overall fit of the model is somewhat worse than before, recognition of singlets in the model considerably improved when mixing was introduced (Fig. 5A). Also, Bayesian information criterion computed over all data points in the purely statistical model and in the mixture model (-11.6 and -22.6 for the two models, respectively) provided evidence that the inclusion of mixing in the model was justified by the data. Similar to experiments, simulations showed a slightly greater preference for singlets than for doublets (Fig. 4, *orange symbols*). Pilot studies performed with three pairs and similar statistics revealed that Gestalt effects play a significant role in this result (Fig. 5B). While purely statistics-based processing preferred both frequent singlets and rare doublets, its preference for doublets was always markedly greater – simply because the true generative model discovered by statistical learning produced doublets and not singlets. Gestalt-based computations, conversely, only processed singlet frequencies and therefore strongly disfavored rare doublets and preferred frequent singlets. Thus, ameliorating the predictions of statistics-based computations with that of Gestalt-based processing led to a stronger preference for frequent singlets than for rare doublets, just as seen in humans.

Discussion

We demonstrated that humans flexibly yet automatically learn complex generative models in visual perception. Bayesian model learning has been implicated in several domains of high level human cognition, from causal reasoning (Tenenbaum & Griffiths, 2003) to concept learning (Tenenbaum, 1999). Here we showed it being at work already at a pre-verbal stage. Thus such a probabilistic framework might be an adequate unified basis for modeling learning processes from early sensory to complex cognitive levels.

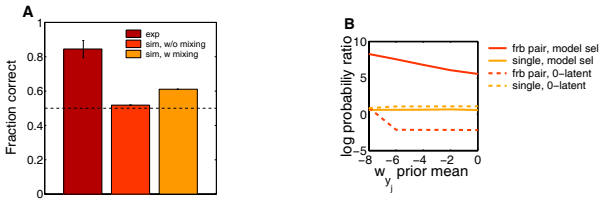

Figure 5: *A*, Recognition of singlets in Experiment 2 in human experiments (*red*), in simulation with purely statistics-based computations (*orange*), and with Gestalt-based computations (*yellow*). *B,* Recognition of frequency balanced *(red)* doublets and frequent singlets *(yellow)* in a series of pilot frequency-balanced simulations. LPRs (Eq. 6) were calculated in simulations with purely statistics-based computations *(solid lines)* and with Gestalt-based computations *(dashed lines)*. Calculations were performed at various observed prior mean values in order to test robustness.

An important finding of the present work was that human data can be explained through an interplay between statistical computations and computations that are governed by Gestalt principles. Since in the experiments presented here only one case (when single shapes were tested in the frequency balanced experiment) necessitated the mixing of statistics- and Gestalt-based predictions, this result should be taken as preliminary. Nevertheless, in that case it seems that the strong Gestalt cues of the experimental paradigm suggesting the independence of individual shapes substantially interfered with statistics-based computations. In line with this account, infants, known to lack some of the Gestalt-based processing capabilities of adults (Kovács, 2000), did not keep track of single shape frequencies in an adapted version of the frequency balanced experiments (Fiser & Aslin, 2002). Such interactions between statistics- and Gestalt-based perceptual systems are the target of future research.

Our approach is very much in the tradition that sees the finding of independent causes behind sensory data as one of the major goals of perception (Barlow, 1990). The results demonstrate that humans can infer complex models from experience and implicate Bayesian model learning as a powerful computation underlying such basic cognitive phenomena as the decomposition of visual scenes into meaningful chunks.

Acknowledgments

This work was supported by IST-FET-1940 program, National Office for Research and Technology under grant no.: NAP 2005/KCKHA005 (GO), NIH research grant HD-37082 (RNA, JF), and the Gatsby Charitable Foundation (ML).

References

Barlow, H. B. (1990). Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Res*, *30*, 1561-71.

Courville, A. C., Daw, N. D., Gordon, G. J., & Touretzky, D. S. (2004). Model uncertainty in classical conditioning. In *NIPS 16.* Cambridge: MIT Press.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2005). Similarity and discrimination in classical conditioning: A latent variable account. In *NIPS 17.* Cambridge, MA: MIT Press.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate information in a statistically optimal fashion. *Nature*, *415*, 429-33.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psych Sci*, *12*, 499–504.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proc Natl Acad Sci USA*, *99*, 15822–6.

Fiser, J., & Aslin, R. N. (2005). Encoding multi-element scenes: Statistical learning of visual feature hierarchies. *J Exp Psychol Gen*, *134*, 521–37.

Green, P. J. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Helmholtz, H. L. F. (1962). *Treatise on physiological optics*. New York: Dover. (original published in 1867)

Iba, Y. (2001). Extended ensemble Monte Carlo. *Int J Mod Phys C*, *12*, 623–56.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu Rev Psychol*, *55*, 271–304.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244-7.

Kovács, I. (2000). Human development of perceptual organization. *Vision Res*, *40*, 1301–10.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, *4*, 415–447.

MacKay, D. J. C. (1995). Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Comput Neural Syst*, *6*, 469–505.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2006). Bayesian model learning in human visual perception. In *NIPS 18.* Cambridge, MA: MIT Press.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In *NIPS 11.* Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. In *NIPS 15.* Cambridge, MA: MIT Press.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nat Neurosci*, *5*, 598–604.