Theoretical Investigations in Visual Short-Term Memory: Structured Probabilistic Representations, Model Mismatch and Neural Population Coding

by

A. Emin Orhan

Submitted in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Supervised by

Professor Robert A. Jacobs

Department of Brain and Cognitive Sciences
Arts, Sciences and Engineering
School of Arts and Sciences

University of Rochester
Rochester, New York

2013
Biographical Sketch

The author was born in 1984 in Manisa, Turkey. He attended Boğaziçi University in Istanbul, Turkey, graduating with a Bachelor of Arts degree in Philosophy in 2005 and later receiving a Master of Arts degree in Cognitive Science in 2008. He began doctoral studies in Brain and Cognitive Sciences at the University of Rochester in 2008 and received the Master of Arts degree from the University of Rochester in 2011. He pursued his research in visual short-term memory under the supervision of Professor Robert Jacobs.

The following publications were a result of work conducted during doctoral study:


Acknowledgments

I would like to thank, first and foremost, my advisor Robbie Jacobs for giving me almost unlimited freedom to work on questions that interested me and for his unfailing generosity with his time and patience whenever I needed help or guidance. Thanks also to my committee members Jiebo Luo and Dave Knill for kindly agreeing to be on my committee and for making many helpful comments. I would especially like to thank Dave Knill for his very helpful comments on the work presented in Chapters 2 and 3. Over the past five years, I have learned a great deal from discussions with a number of very talented postdocs: Ruben Moreno-Bote, Xaq Pitkow and Chris Sims. Ruben and Xaq made helpful comments on the work presented in Chapter 4 and Chris may actually be the reason why I got interested in visual short-term memory in the first place. I would like to thank fellow graduate students in the lab, İlker Yıldırım and Gökler Erdoğan, for their friendship and support. I am grateful to all the undergraduate research assistants, unfortunately too many to name all of them here, who helped me recruit and run subjects in countless experiments over the years. Finally, I would like to thank Kathy Corser, Jennifer Gillis, Chris Dambra and Chris Freemesser for making it extremely easy for me to focus on research only, by so ably taking care of everything else.
Abstract

Visual short-term memory (VSTM) makes a stable and coherent perception of the world possible by maintaining visual information about a scene as it was a moment ago. As VSTM is widely considered to be a severely limited form of memory, it is important to know what kinds of information about a scene can or cannot be held in VSTM. This thesis proposes a probabilistic clustering model that attempts to capture the representations that subjects encode from brief presentations of simple visual scenes typically used in VSTM studies (Chapter 2). The model explains a number of previously observed biases in VSTM and predicts dependencies between representations of different stimuli that should decrease with the difference between their feature values. We show that results from three experiments are qualitatively in agreement with this novel prediction of the model. We further show that this type of dependency can arise as a natural consequence of the simultaneous encoding of multiple stimuli in a population of neurons with correlated responses (Chapter 4). We also investigate the consequences of possible mismatches, of the type demonstrated in Chapter 2, between the internal model that subjects might use to encode simple visual displays and the model actually used by the experimenter to generate the stimuli (Chapter 3). We tentatively suggest that such mismatches can arise due to the long-term adaptation of the visual system to the stimulus statistics in the natural environment which is generally quite different from the stimulus statistics used
in VSTM studies. Remarkably, we show that this type of mismatch alone, without assuming any resource or capacity limitation, can explain some of the main qualitative characteristics of performance limitations observed in VSTM studies typically attributed to capacity limitations. Finally, we take a critical look at the use of impoverished visual stimuli that lack the rich statistical structure displayed by stimuli that are more representative of the natural environment and discuss how the study of VSTM can be extended to ecologically more realistic stimuli, as well as possible implications of such an extension (Chapter 5). Taken together, this thesis proposes novel theoretical ideas at multiple levels of analysis, from computation to neural implementation, for an ecologically more realistic and behaviorally more adequate modeling of representations in VSTM.
Contributors and Funding Sources

This work was supervised by a dissertation committee consisting of Professors Robert Jacobs (advisor) and David Knill of the Department of Brain and Cognitive Sciences and Professor Jiebo Luo of the Department of Computer Science.

The research described in Chapters 2, 3, 4 and 5 was conceived by the student in collaboration with Professor Robert Jacobs and was published, submitted for publication, or will be submitted for publication, in the articles Orhan & Jacobs (2013), Orhan & Jacobs (in preparation), Orhan & Jacobs (2011a) and Orhan & Jacobs (under review) respectively, listed in the Biographical Sketch. The data used in Section 2.6.1 below were provided by Professor Wei Ji Ma of the Department of Neuroscience at Baylor College of Medicine. All other work conducted for dissertation was completed independently by the student.

This work was funded by research grants from the National Science Foundation (DRL-0817250) and the Air Force Office of Scientific Research (FA9550-12-1-0303), both awarded to Professor Robert Jacobs.
# Table of Contents

Biographical Sketch ii

Acknowledgments iv

Abstract v

Contributors and Funding Sources vii

List of Tables xii

List of Figures xiv

1 Introduction 1

2 A Probabilistic Clustering Theory of the Organization of Visual Short-Term Memory 6

2.1 Biases and Dependencies in VSTM ......................... 8

2.2 Hierarchical Encoding of Items in VSTM .................... 17

2.3 Probabilistic Clustering Theory (PCT) ....................... 23
2.4 Relationship between PCT and previous works on hierarchical encoding in human memory .................................................. 26
2.5 Models ................................................................................. 29
2.6 Simulations .......................................................................... 42
2.7 Experiments ......................................................................... 60
2.8 Discussion ............................................................................. 92

3 Consequences of Model Mismatch in Visual Short-Term Memory 98
3.1 General framework ................................................................. 104
3.2 Models for the subject’s prior $q(s)$ ............................................ 110
3.3 Strong hypothesis: Model mismatch alone accounts for performance limitations in VSTM ......................................................... 116
3.4 Weak hypothesis: Model mismatch and capacity limitations together account for performance limitations in VSTM ......................... 143
3.5 Possible changes in the noise distribution due to adaptation to natural stimulus statistics ......................................................... 154
3.6 Discussion ............................................................................. 166

4 A Neural Implementation of Dependencies in Visual Short-Term Memory 174
4.1 An Analytic Expression for the Fisher Information Matrix ............. 175
4.2 A more general model and the dependence of predicted correlations on model parameters ......................................................... 179
4.3 Encoding and decoding multiple stimuli in a network of spiking neurons 190
4.4 Discussion ............................................................................. 193
C Appendix for Chapter 4

C.1 Derivation of Equation 4.4 ........................................ 250
2.1 BIC values of the BFMMs ($K = 2$ and $K = 4$) and the two-level HBM relative to the BIC value of the DPMM on three previous studies that reported biases in VSTM. Negative values indicate better fits than the DPMM, positive values worse fits. Smaller values indicate better fits.
List of Figures
2.1 Schematic illustration of biases and dependencies that may arise in the joint estimates of two items based on their VSTM representations. (A) Biases manifest themselves as shifts of the distribution \( p(\{\hat{\theta}_i\}^N_{i=1}|\{\theta_i\}^N_{i=1}) \) (represented here by a single contour) from the actual feature values of the items. In this example, the mean of the distribution \( p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2) \) (represented by the black dot) is shifted away from the actual feature values of the items, \((\theta_1, \theta_2)\) (represented by the square) and toward the main diagonal, indicating that the estimates of the feature values of both items are biased toward the mean of \(\theta_1\) and \(\theta_2\). Note, however, that the distribution is spherical, hence, assuming a Gaussian distribution for simplicity, there are no dependencies between \(\hat{\theta}_1\) and \(\hat{\theta}_2\). (B) Dependencies manifest themselves as statistical dependencies among \(\{\hat{\theta}_i\}^N_{i=1}\). In this example, representations of the two items, \(\hat{\theta}_1\) and \(\hat{\theta}_2\), are correlated. Note, however, that there are no biases in the representations as \( p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2) \) is centered on \((\theta_1, \theta_2)\). Also note that this example depicts only a simple form of dependence between \(\hat{\theta}_1\) and \(\hat{\theta}_2\), namely second-order correlation. More complex or higher-order dependencies between \(\hat{\theta}_1\) and \(\hat{\theta}_2\) are also possible. (C) A hypothetical example where there are both biases and dependencies in the joint estimates of the two items.

2.2 Single trial of a hypothetical VSTM experiment in which subjects are asked to remember the horizontal locations of 6 colored squares.
2.3 A graphical representation of the models considered in this chapter. All models have a common structure represented by the graphical model shown on the left. We illustrate this graphical model using plate notation, where the nodes inside the plate are meant to be replicated \( N \) items. The shaded node represents the observable variables (i.e., the noisy observations \( x_i \)). The other variables are latent or unobservable. The remaining plots illustrate the generative processes defining the models. Each row illustrates the generation of variables at the corresponding level in the graphical model on the left. The only difference between the models is in the variable \( G \). For the DPMM, \( G \) is a discrete distribution with an infinite number of “atoms”; for the BFMMs, it is a discrete distribution with \( K \) atoms; and for the HBM, it is a single atom. In the example shown here, the DPMM uses three clusters to generate the four items represented by \( \theta_i \)s, the BFMM uses 2 clusters and the HBM uses a single cluster. \( \mu_i \) (mean) and \( \tau_i \) (precision) represent the cluster parameters for item \( i \). The distributions at the bottom two rows illustrate the distributions from which the variables at the corresponding levels were drawn.

2.4 (A) Three noisy observations, \( x_1, x_2 \) and \( x_3 \) (vertical lines) and the three marginal posteriors, \( p(\theta_i|x_1, x_2, x_3) \) for \( i = 1, 2, 3 \) (solid curves), for four different settings of the parameters, \( \alpha_c \) and \( \tau_{\text{obs}} \). (B) Posterior distributions over the number of clusters for each of the corresponding subplots in (A).
2.5 (Leftmost plot) Three items with feature values \( \theta_1, \theta_2 \) and \( \theta_3 \) (vertical lines) and the marginal distributions of the estimates, \( p(\hat{\theta}_i|\theta_1, \theta_2, \theta_3) \) (solid curves), computed over 1000 simulated presentations of the same set of feature values. The remaining plots show each of the three two-dimensional marginals of the estimates, \( p(\hat{\theta}_i, \hat{\theta}_j|\theta_1, \theta_2, \theta_3) \). The means of the marginals are represented by the circles and the actual feature values of the items are represented by the crosses. The numbers inside the plots are the correlation coefficients between the estimates of each pair of items.

2.6 (A) The observed biases (from Wilken and Ma, 2004) and biases predicted by a non-optimized DPMM with \( \alpha_c = 1 \) and \( \tau_{\text{obs}} = 1 \) for four different set sizes. Lighter colors represent larger set sizes. Error bars represent \( \mp 1 \) SEM across subjects. (B) Posterior distributions over the number of clusters inferred by the DPMM averaged over all trials for different set sizes.
2.7 Predictions of a non-optimized DPMM with parameters set to $\alpha_c = 1$ and $\tau_{obs} = 1$. (A) Representative trials from the medium and high homogeneity conditions of Viswanathan et al. (2010). For purposes of illustration, the noisy observations were set to the actual spatial frequencies of the study gratings in the two conditions (i.e., $x_i = \theta_i$). These observations are indicated by solid vertical lines and the frequency of the lure probe at $2\ JND$ is represented by the dashed line. The solid black and gray curves show the combined posterior densities $p(\theta|\{x_i\}_{i=1}^N)$ in the medium and high homogeneity conditions, respectively. Lower panel shows the posterior distributions over the number of clusters in the two conditions. (B) Probabilities of responding ‘old’ in the two conditions as a function of probe frequency. (C) The observed and predicted probabilities of responding ‘old’ to the probe at $2\ JND$ in the two conditions. The model prediction was estimated over 1800 simulated trials.
2.8 Simulation results for the optimized multivariate DPMM with $\kappa$ and $\sigma^2_{obs}$ set to their ML estimates ($\kappa = 6, \sigma^2_{obs} = 225$). (A) Left and right columns show simulation results for representative trials from Experiments 1 and 2, respectively. Top row shows the visual stimuli used in each trial in color-size space. Bottom row shows the posterior distributions over the number of clusters for the corresponding trials. (B) Posterior probabilities over the number of clusters averaged over all trials in the two simulated experiments. (C) Predicted biases for the DPMM in Experiment 1 (gray) and Experiment 2 (black) of Brady and Alvarez (2011) as a function of $\kappa$ and (D) $\sigma^2_{obs}$. Horizontal lines represent the observed mean biases of subjects in these experiments, and gray areas represent $\pm 1SEM$ around the mean biases. The dashed gray and black lines represent the mean predicted biases for the DPMM in Experiment 1 and in Experiment 2 respectively, and the error bars correspond to $\pm 1SEM$ around the mean predicted biases over 25 simulations of both experiments, with 50 simulated trials in each experiment. In (C), $\sigma^2_{obs}$ was fixed at 225 and $\kappa$ was varied, whereas in (D), $\kappa$ was fixed at 6 and $\sigma^2_{obs}$ was varied. . . . . . . . . . 58

2.9 The sequence of events in a single trial of Experiment 1B: (a) a small fixation cross is presented at a random location for 1 s; (b) the target configuration is flashed briefly (100 ms); (c) a delay interval of 1 s follows the target configuration; (d) the probe display (initially containing only the dark horizontal lines) remains on until the subject indicates the horizontal locations of all items using the computer mouse. 63
2.10 (A) Results from a representative subject (RD) in Experiment 1A (left). Crosses represent the 36 stimulus configurations presented to the subject, dots represent the subject’s mean estimates for each configuration and contours represent single-level contours of bivariate Gaussian distributions fitted to the subject’s responses for each configuration (red contours show cases where the subject’s estimates of the horizontal locations of the two presented items were significantly and positively correlated, whereas blue contours show cases where the subject’s estimates were significantly and negatively correlated). The middle and the right panels show the predictions of the best-fitting DPMM and the best-fitting HBM, respectively (colors and line styles same as in subject’s data). (B) Correlations between $\hat{\theta}_1$ and $\hat{\theta}_2$ as a function of the distance $|\theta_1 - \theta_2|$ for subject RD (left) and the corresponding predictions from the best-fitting DPMM (middle) and the best-fitting HBM (right). Error bars represent standard errors of the mean.
2.11 (A) Pairwise correlations between $\hat{\theta}_i$ and $\hat{\theta}_j$ as a function of the distance $|\theta_i - \theta_j|$ between the horizontal locations of pairs of items for combined data from all 8 subjects in Experiment 1A (left) and for combined data from all 11 subjects in Experiment 1B (right). (B) Standard deviations of $\hat{\theta}_i$ as a function of the standard deviation of the actual feature values of items for combined data from all 8 subjects in Experiment 1A (left) and for combined data from all 11 subjects in Experiment 1B (right). Also shown are the predictions of the best-fitting DPMMs and the best-fitting HBMs. Error bars represent standard errors of the mean.

2.12 Posterior distributions over the number of clusters inferred by the DPMM and the BFMM with 2 and 4 components for two representative subjects in Experiment 1A [subject RD] (A) and in Experiment 1B [subject DK] (B) respectively.
2.13 (A-B) Target and lure configurations are represented by open squares and circles, respectively. Lure configurations are defined along two perpendicular axes (parallel and orthogonal axes, represented by the gray and black lines, respectively). The contour lines represent the shapes of the joint distributions \( p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2) \) describing the contents of a subject’s memory of the target configurations. In (A), items in the target configuration have similar feature values (small \( \Delta \)), whereas in (B) they have different feature values (large \( \Delta \)). (C-D) Unnormalized cross-sections of the joint distributions along the parallel (gray line) and orthogonal axes (black line). If the joint distribution is positively correlated for small \( \Delta \) values, but uncorrelated for large \( \Delta \) values, then for small \( \Delta \), lure configurations along the parallel axis should have higher probability under the joint distribution than equidistant lure configurations along the orthogonal axis (C). In contrast, when \( \Delta \) is large, lure configurations along the parallel axis and equidistant configurations along the orthogonal axis should have more or less the same probability under the joint distribution (D).

2.14 The sequence of events in a single trial of Experiment 2: (a) a small fixation cross is presented at the center of the screen for 1 s; (b) the target configuration is flashed briefly; (c) a delay interval of 1 s follows the target configuration; (d) the probe configuration is presented until the subject responds either ‘same’ or ‘different’ (in this example the probe configuration is the same as the target configuration).
2.15 Probability of ‘same’ responses as a function of the position along the parallel/orthogonal axis for the five different target configurations with different ∆ values in Experiment 2 [location change detection] (A) and Experiment 3 [orientation change detection] (B). ‘0’ indicates the target configuration itself (i.e., no-change trials), whereas the other numeric labels represent different lure configurations along the parallel (solid line) or the orthogonal axis (dashed line). Error bars represent standard errors of the mean across subjects. 83

2.16 (A) Results for a representative subject (Subject MG): for the five target configurations used in Experiment 2, the means and shapes of the distributions \( p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2) \) with the parameters of these distributions set to their maximum a posteriori (MAP) estimates. The five target configurations used in Experiment 2 are represented by the black crosses. Red contours represent cases where the 95% credible interval for \( \rho \) excludes 0. (B) The means and the standard errors (across subjects) of the MAP estimates of \( \rho \) for the five target configurations in Experiment 2. (C-D) Similar to (A-B), but results are shown for Experiment 3. In (C), results are shown for a representative subject in Experiment 3 (Subject TH). For better visualization, a different scaling was chosen to create the ellipses in (C) than in (A). 86
2.17 The sequence of events in a single trial of Experiment 3: (a) a small fixation cross is presented at the center of the screen for 1 s; (b) the target display, consisting of two oriented Gabor gratings, is flashed briefly; (c) a delay interval of 1 s follows the target configuration; (d) the probe display is presented until the subject responds either ‘same’ or ‘different’ (in this example the probe display is different from the target display). 89

3.1 Illustration of mismatches that can arise between the experimenter’s prior $p(s)$ (line orientations are independently sampled from uniform distributions) and the subject’s prior $q(s)$ (line orientations are sampled from a joint distribution characterizing orientations in natural visual environments). The configurations of orientations shown in panels A, B and C have equal probability under $p(s)$, but the configurations in B and C are more probable than the configuration in A under $q(s)$ due to dependencies between the orientations of different line segments in natural images. 108

3.2 Under the mixture of Gaussians prior, more homogeneous configurations of orientations are given greater probability. In the examples shown here, the homogeneity of orientations, and hence the probability under $q(s)$, decreases from A to F. See the main text for details of how the homogeneity preference is implemented in the mixture of Gaussians prior. 115
3.3 (A) Recall precision as a function of set size in the Gaussian model with uniform non-negative correlations. Different curves correspond to different correlation values, denoted $\rho_q$. 34 different $\rho_q$ values are used in this example, from 0 to 0.99 with increments of 0.03. Lighter colors represent lower $\rho_q$ values. Following Bays & Husain (2008), precision is defined as 1 over the standard deviation of errors. (B) Recall precision as a function of set size in the Gaussian model with random positive correlations. Different curves correspond to different $\gamma$ values. Six different $\gamma$ values are used in this example, $\gamma = 0.1, 0.15, 0.3, 0.9, 1.8, 3.6$. Lighter colors represent higher $\gamma$ values. (C) Recall precision as a function of set size in the mixture of Gaussians model. Different curves correspond to different weight exponents, $k$. Six different $k$ values are used in this example, $k = 0, 0.2, 1, 7, 13, 19$. Lighter colors represent smaller $k$ values. (D) Recall precision as a function of set size in the Gaussian model with uniform correlations across dimensions. In this example, in addition to the difference in correlation, the experimenter’s prior and the subject’s prior also differ in variances, with the experimenter’s prior having a standard deviation in each dimension that is three times the standard deviation of the subject’s prior ($\sigma_p = 3, \sigma_q = 1$). Different curves correspond to different $\rho_q$ values. As in (A), 34 different $\rho_q$ values are shown, from 0 to 0.99 with increments of 0.03. Lighter colors represent lower $\rho_q$ values. (E) Exponents of power law fits, $1/\sigma \propto N^\xi$, to precision-set size curves shown in (A) [solid line] and (D) [dashed line] as a function of $\rho_q$. . . . . . . . . . . 124
3.4 (A) Recall precision as a function of set size in the Gaussian model with uniform non-negative correlations. In these simulations, \( p(s) \) as well as \( q(s) \) is assumed to be a multivariate Gaussian with uniform correlations. Correlations \( \rho_p \) and \( \rho_q \) denote the uniform correlations between pairs of dimensions in \( p(s) \) and \( q(s) \), respectively. \( \rho_p \) was set to 0.9. Precision-set size curves are shown for 34 different values of \( \rho_q \) from 0 to 0.99 in increments of 0.03. Lighter colors represent lower \( \rho_q \) values. Note that when \( \rho_q \) exceeds \( \rho_p \), the inverse set size effect starts to decrease and ultimately reverses to a set size effect (a decline in precision with set size). (B) Exponents \( \xi \) of power-law fits to precision-set size curves for each pair of \( (\rho_p, \rho_q) \) values. Positive values indicate an increase in precision with set size, whereas negative values indicate a decrease in precision with set size. (C) Results for the mixture of Gaussians model. \( p(s) \) is a multivariate Gaussian with uniform correlations across dimensions, and \( q(s) \) is a mixture of Gaussians with weight exponent set to \( k = 1 \). \( \rho_p \) is varied from 0 to 0.99 in increments of 0.03. Each curve represents the precision-set size curve for a different \( \rho_p \) value. Lighter colors represent lower \( \rho_p \) values. (D) Exponents of power law fits, \( \xi \), to precision-set size curves in the mixture of Gaussians model shown in (C) as a function of \( \rho_p \). . . . . 127
3.5 (A-B) Results for the case of uncorrelated Gaussian $p(s)$ and Gaussian $q(s)$ with random positive correlations ($\gamma = 0.15$): (A) BIC score of the Gaussian fit relative to the BIC score of the $t$-distribution fit for each set size. Lower scores indicate better fits. Positive values indicate better fits for the $t$-distribution than for the Gaussian distribution. (B) For three different set sizes, the estimated Gamma distributions over precision for the best fit $t$-distributions. The vertical lines show the means of the Gamma distributions. (C) Similar to (A), but here the experimenter’s prior contains correlations ($\rho_p = 0.9$). (D-E) Similar to (A-B) except the results are for the case of uncorrelated Gaussian $p(s)$ and the mixture of Gaussians $q(s)$ that favors homogeneity ($k = 1$). (F) Similar to (D), but the experimenter’s prior contains correlations ($\rho_p = 0.9$).
3.6 (A-C) Recall precision as a function of presentation time for different set sizes. Darker colors indicate larger set sizes with set sizes ranging from 1 to 6. Simulations were conducted with the experimenter’s prior \( p(s) \) set to an uncorrelated Gaussian distribution \( (\rho_p = 0) \). Results are shown for the cases of Gaussian \( q(s) \) with uniform correlations \( [\rho_q = 0.96] \) (A), Gaussian \( q(s) \) with random positive correlations \( [\gamma = 0.15] \) (B), and the mixture of Gaussians prior favoring homogeneity \( [k = 1] \) (C). (D-F) Initial encoding rate (blue) and the asymptotic precision (green) as a function of set size. The exponents \( \xi \) of power-law fits to each curve are also provided. Results are again shown for the case of Gaussian \( q(s) \) with uniform correlations (D), Gaussian \( q(s) \) with random positive correlations (E), and the mixture of Gaussians prior favoring homogeneity (F). Due to computational demands, \( 10^4 \) trials per presentation time per set size were simulated for the Gaussian model with random positive correlations. For the other two models, \( 10^5 \) trials per presentation time per set size were simulated. The simulated presentation times varied from 5 ms to 1005 ms in steps of 10 ms. The noise distribution has an asymptotic precision of 2 and \( \tau = 0.1 \) for all models.
3.7 Blown-up views of precision in the first two time steps of the simulation (corresponding to presentation times of 5 ms and 15 ms respectively) for the Gaussian $q(s)$ with uniform correlations ($\rho_q = 0.96$). Different colors correspond to different set sizes with darker colors representing larger set sizes. Solid lines show the actual simulation results and dashed lines show the precisions that would be predicted if the initial encoding rate had the same dependence on set size as the asymptotic precision (taking the precision for $N = 1$ as the baseline). (A) shows the results for the case where the experimenter’s prior is an uncorrelated Gaussian ($\rho_p = 0$) as in Figure 3.6A and (B) shows the results for the case where the experimenter’s prior is a Gaussian with uniform correlations ($\rho_p = 0.9$) as in Figure 3.8A.

3.8 Results when the experimenter’s prior $p(s)$ was a Gaussian distribution with uniform correlations ($\rho_p = 0.9$). The formats of the plots are identical to those in Figure 3.6.
3.9 Biases (A-C) and dependencies (D-F) predicted by the three models for \( q(s) \) considered in this chapter. Biases (A-C) are computed by dividing the target stimulus value \( s_t \) into 8 bins and measuring the mean and the SEM of the estimation errors, \( s_t - \hat{s}_t \), for each bin. The experimenter's prior \( p(s) \) is taken to be an uncorrelated multivariate Gaussian. Dependencies (D-F) are computed by presenting a particular set of stimuli consisting of a variable stimulus \( s_1 \) and \( N - 1 \) zeros over \( 10^5 \) simulated trials and measuring the correlation coefficient between the estimates of \( s_1 \) and \( s_2 \) over all trials (note that \( s_2 \) is always zero). \( \Delta \) denotes the absolute difference between \( s_1 \) and \( s_2 \) (\( \Delta = |s_1 - s_2| \)). For biases, simulation results are shown for 6 different set sizes (from \( N = 1 \) to \( N = 6 \)). For dependencies, simulation results are shown for 5 different set sizes (from \( N = 2 \) to \( N = 6 \), as correlation does not have any meaning for \( N = 1 \)). In each plot, darker lines correspond to larger set sizes. (A) and (D) show the results for a multivariate Gaussian \( q(s) \) with uniform correlations (\( \rho_p = 0.96 \)); (B) and (E) show the results for the case of multivariate Gaussian \( q(s) \) with random positive correlations (\( \gamma = 0.15 \)); (C) and (F) show the results for the mixture of Gaussians prior (\( k = 1 \)).
3.10 (A) Recall precision as a function of set size when \( p(s) \) is an uncorrelated Gaussian (\( \rho_p = 0 \)) and \( q(s) \) is a multivariate Gaussian with uniform non-negative correlations \( \rho_q \). Precision-set size curves are shown for 34 different values of \( \rho_q \) from 0 to 0.99 in increments of 0.03. Lighter colors represent lower \( \rho_q \) values. (B) Similar to A, but \( p(s) \) is a multivariate Gaussian with uniform correlations with \( \rho_p = 0.9 \). (C) Exponents \( \xi \) of power-law fits to precision-set size curves for each pair of \( (\rho_p, \rho_q) \) values. (D) Results for the mixture of Gaussians model. \( p(s) \) is a multivariate Gaussian with uniform non-negative correlations, and \( q(s) \) is a mixture of Gaussians with the same parameter values used in previous simulations (\( k = 1 \)). \( \rho_p \) is varied from 0 to 0.99 in increments of 0.03. Each curve represents the precision-set size curve for a different \( \rho_p \) value. Lighter colors represent lower \( \rho_p \) values.

3.11 Variability in encoding precision when the noise variance increases with set size. (A)-(C) show the results for a multivariate Gaussian \( q(s) \) with random positive correlations (\( \gamma = 0.3 \)) (D)-(F) show the results for the mixture of Gaussians model (\( k = 1 \)). The formats of the plots are identical to those in Figure 3.5.

3.12 Time course of recall precision when the noise variance increases with set size and the experimenter’s prior \( p(s) \) is an uncorrelated Gaussian. The formats of the plots are identical to those in Figure 3.6. (A) and (D) show the results for a multivariate Gaussian \( q(s) \) with uniform correlations (\( \rho_q = 0.81 \)). (B) and (E) show the results for a multivariate Gaussian \( q(s) \) with random positive correlations (\( \gamma = 0.3 \)). (C) and (F) show the results for the mixture of Gaussians prior (\( k = 1 \)).
3.13 Time course of recall precision when the noise variance increases with set size and the experimenter’s prior $p(s)$ is a multivariate Gaussian with uniform correlations ($\rho_p = 0.9$). The formats of the plots are identical to those in Figure 3.12.

3.14 (A) A simple demonstration of stimulus-dependent noise in a one-dimensional example. The noise precision is computed from Equation 3.15 where $q(s)$ is taken to be a Gaussian distribution with zero mean and standard deviation equal to either 1 (solid line), 2 (dashed line) or 3 (dotted line). Parameter values are $\sigma_0^2 = 0.025$ and $\gamma = 0.01$.

(B-D) Empirical distributions of the noise precision over $10^5$ simulated trials for different set sizes from 1 to 6. Hotter colors represent larger set sizes. In (B), both the subject’s prior $q(s)$ and the experimenter’s prior $p(s)$ are uncorrelated Gaussians with standard deviation 2 along each dimension. In (C), $p(s)$ is an uncorrelated Gaussian with standard deviation 2 along each dimension, but $q(s)$ is a Gaussian with uniform correlations across its dimensions ($\rho_q = 0.9$ and $\sigma_q = 2$).

In (D), both $p(s)$ and $q(s)$ are multivariate Gaussians with uniform correlations ($\rho_p = \rho_q = 0.9$ and $\sigma_p = \sigma_q = 2$).
3.15 Simulation results for the case where $p(s)$ is an uncorrelated Gaussian. (A-B) Precision-set size relationships for the SED model where $q(s)$ is either a multivariate Gaussian with uniform non-negative correlations $\rho_q$ (A) or a mixture of Gaussians (B). For the multivariate Gaussian model (A), 9 different $\rho_q$ values were tested ranging from 0 to 0.96 in increments of 0.12. Darker colors correspond larger $\rho_q$ values. For the mixture of Gaussians model (B), 5 different $k$ values were tested: $k = 0, 0.2, 0.5, 1, 1.5$. Darker colors correspond to larger $k$ values. (C-D) Power-law exponents for different combinations encoding and decoding schemes where $q(s)$ is either a multivariate Gaussian with uniform non-negative correlations $\rho_q$ (C) or a mixture of Gaussians (D). SED: suboptimal encoding and decoding, SE: suboptimal encoding, SD: suboptimal decoding, UV: uniform variance.

3.16 Similar to Figure 3.15, but the results are shown for the case where $p(s)$ is a multivariate Gaussian with uniform correlations ($\rho_p = 0.96$).
3.17 (A-B) Results for the case of uncorrelated Gaussian $p(s)$ and the SED model with Gaussian $q(s)$ with uniform non-negative correlations ($\rho_q = 0.96$): (A) BIC score of the Gaussian fit relative to the BIC score of the $t$-distribution fit for each set size. Lower scores indicate better fits. Positive values indicate better fits for the $t$-distribution than for the Gaussian distribution. (B) For three different set sizes, the estimated Gamma distributions over precision for the best fit $t$-distributions. The vertical lines show the means of the Gamma distributions. (C) Similar to (A), but here the experimenter’s prior contains correlations ($\rho_p = 0.96$). (D-E) Similar to (A-B) except the results are for the case of uncorrelated Gaussian $p(s)$ and the SED model with the mixture of Gaussians $q(s)$ ($k = 1$). (F) Similar to (D), but the experimenter’s prior contains correlations ($\rho_p = 0.96$).

4.1 (A) A population of neurons satisfying all the assumptions made in deriving the FIM. For neurons in the upper row $f_k(s) = f_k(s_1)$, and for neurons in the lower row $f_k(s) = f_k(s_2)$. The magnitude of correlations between two neurons is indicated by the thickness of the line connecting them. (B) Correlation coefficients estimated from the inverse of the FIM for all stimuli pairs $s_1, s_2$. (C) Mean correlation coefficients as a function of $|s_1 - s_2|$ (red: model’s prediction; black: combined data from all 8 subjects in Experiment 1A). Parameters: $\alpha = 0.5, g = 50, a = 1$ (these were set to biologically plausible values); other parameters: $K = 1000, \sigma = 9.0, L = 0.015$ (the last two were chosen to provide a good fit to the experimental results).
4.2 The effect of model parameters on the predicted correlations between the estimates of the stimuli. Results are shown for the model with Gaussian tuning functions. The default parameter values are as follows: $c_0 = 0.4$, $L = 0.5$, $w_1 = 0.75$, $\nu = 1$, $g = 40$, $\sigma^2 = 25$, $W = 0.99$, $a = 1$, $K = 1000$ (total number of neurons, with $K/2$ neurons in each group). In each plot, one of the parameters is varied while the remaining parameters are held fixed at their default values. Different lines correspond to simulation results with different parameter values, with darker colors always indicating larger parameter values. (A) $w_1$ is varied from 0.5 to 1.0 in steps of 0.05. (B) $W$ is varied from 0.09 to 0.99 in steps of 0.1. (C) $K$ is varied from 1000 to 5000 in steps of 1000. (D) $L$ is varied from 0.1 to 1000 in 10 logarithmically spaced points. (E) $c_0$ is varied from 0 to 0.5 in steps of 0.1. (F) $\sigma^2$ takes the values 4, 9, 16, 25, 36, 49.
4.3 The effect of model parameters on the predicted correlations between the estimates of the stimuli. Results are shown for the model with von Mises tuning functions. The default parameter values are as follows: $c_0 = 0.4$, $L = 0.5$, $w_1 = 0.75$, $\nu = 1$, $\beta = 19$, $\gamma = 2$ [the last three parameter values are taken from Ecker et al. (2011)], $W = 0.99$, $a = 1$, $K = 1000$ (total number of neurons, with $K/2$ neurons in each group). In each plot, one of the parameters is varied while the remaining parameters are held fixed at their default values. Different lines correspond to simulation results with different parameter values, with darker colors always indicating larger parameter values. (A) $w_1$ is varied from 0.5 to 1.0 in steps of 0.05. (B) W is varied from 0.09 to 0.99 in steps of 0.1. (C) K is varied from 1000 to 5000 in steps of 1000. (D) L is varied from 0.1 to 1000 in 10 logarithmically spaced points. (E) $c_0$ is varied from 0 to 0.5 in steps of 0.1. (F) $\gamma$ is varied from 1 to 6 in steps of 1.

4.4 The effect of noise level, $a$, on the predicted correlations between the estimates of the stimuli. Results are shown for the model with Gaussian (A) and von Mises (B) tuning functions. In each case, $a$ is varied from $10^{-3}$ to $10^3$ in 7 logarithmically spaced points. Darker lines correspond to larger values of $a$. 

187

188
4.5 Predictions from the FIM compared with the performance of the ML estimator. (A) and (B) show the results for the model with von Mises tuning functions; (C) and (D) show the results for the model with Gaussian tuning functions. (A) and (C) show the correlations between the estimates of the two stimuli; (B) and (D) show the variance of the estimate of the first stimulus, $\hat{s}_1$. Results for the MLE were obtained from 1000 simulated trials for each stimulus pair $(s_1, s_2)$. For both models, all parameters were set to their default values. ML estimates were numerically computed from Equations 4.1, 4.6 and 4.11 using a Nelder-Mead simplex method (the `fminsearch` function in Matlab).

4.6 (A) Results for the spiking network model. The actual display configurations $s$ are represented by magenta dots, the estimated means based on the model’s responses are represented by black dots and the estimated covariances are represented by contours (with red contours representing cases for which the two dimensions were significantly correlated at the $p < 0.05$ level). (B) The mean correlation coefficients (and standard errors of the means) as a function of $|s_1 - s_2|$ (red: model prediction; black: combined data from all 8 subjects in Experiment 1A). Model parameters: $g_{in} = 120, \sigma_{in} = 2, \sigma_{out} = 2$. Parameters were chosen to provide a good fit to the experimental data.
5.1 Example natural images highlighting a variety of different types of visual information that we can perceive and remember in the real world. (A) Complex real-world scenes. The first two images are from the SUN database (Xiao, Hays, Ehinger, Oliva & Torralba, 2010); the rightmost image is from http://www.flickr.com/photos/13774569@N07/5923460837/. (B) Textures. Images are from the UIUC texture database (Lazebnik, Schmid & Ponce, 2005). (C) Material properties such as glossiness, rigidity, roughness, being made of leather, etc. Images are from the Flickr Material Database (Sharan, Rosenholtz & Adelson, 2009). (D) Actions. Images are from the SUN database.
5.2 Example probabilistic graphical models that can be used for modeling representations in VSTM. (A) This figure is adapted from Sudderth (2006, p. 250) and shows a hierarchical probabilistic graphical model for representing the structure of natural scenes. Scenes are represented hierarchically in terms of latent variables (unshaded nodes) corresponding to the objects in the scene and their locations, visual parts composing each object and their locations with respect to the object, as well as low-level features for each part and the locations of these features. Plates represent the replication of variables inside them. The largest plate represents the replication of variables for different images ($J$ images in total) and the smaller plate inside it represents the replication of variables for different low-level features in the image ($N_j$ features in total for image $j$). A difference from the model of Sudderth (2006, p. 250) is that we assume that subjects do not have access to the actual values of the low-level features $v_{ji}, w_{ji}$ but only to noisy observations thereof (represented by $t_{ji}$ and $y_{ji}$ respectively). (B) The probabilistic clustering model proposed by Orhan and Jacobs (2013) to capture the internal model that subjects might be employing to encode the simple displays typically used in VSTM experiments. The model implements the assumption that stimuli, whose actual values are denoted by the variables $\theta_i$, are generated in clusters, with cluster parameters $\mu_i$ (cluster mean) and $\tau_i$ (cluster precision). The number of clusters and the assignment of stimuli to clusters are uncertain and have to be inferred probabilistically from the noisy observations of the stimuli, denoted by $x_i$, even if the actual generative process used by the experimenter to generate the stimuli does not involve any clusters. Note that variables with the same names do not necessarily represent the same things in the two models shown in (A) and (B).
1 Introduction

We can maintain visual information across brief disruptions or discontinuities in the visual signals reaching our eyes. For example, we frequently make eye movements and head movements; we move our bodies; we blink; sometimes objects in the world temporarily occlude our view. Yet our perception of the world is surprisingly stable in the face of these disruptions and discontinuities. We do not find ourselves having to make sense of the world from scratch each time we make an eye movement or each time an object temporarily occludes our view. Visual short-term memory (VSTM) makes this possible by maintaining visual information about the scene as it was a moment ago: right before we initiated a saccade (Hollingworth, Richard, & Luck, 2008) or right before a moving object occluded our view.

As in most other areas in cognitive psychology, research on VSTM generally uses simple visual displays consisting of simple objects with easily parametrizable features and with no statistical structure within or between objects, to address questions such as the nature of capacity limits or the units of storage in VSTM (Brady, Konkle & Alvarez, 2011). This choice of stimuli confers several advantages to the design of experiments and the interpretation of results obtained from such experiments. First,
because they have very few degrees of freedom and very little structure, such stimuli are easy to generate and manipulate as desired. Second, it is easy to formulate and test hypotheses within the context of such stimuli. Similarly, the results obtained from experiments using such stimuli are easy to interpret, because one can precisely control the differences between different conditions of an experiment, making it possible to unambiguously attribute an effect to the variable manipulated between the conditions. Third, it might be assumed that because subjects are relatively unfamiliar with such impoverished stimuli in the natural environment, they eliminate any prior knowledge or assumptions subjects might bring to the task, which would again be thought advantageous for the interpretation of results. In this regard, they might be considered to be very different from natural scenes for which people can utilize their extensive prior knowledge to perform various tasks (Biederman, 1972; Bar, 2004; Steyvers & Hemmer, 2012) and it may be extremely difficult to determine precisely what that prior knowledge consists of. A more comprehensive discussion of the relative merits of using natural vs. unnatural stimuli and tasks in VSTM studies is deferred to Chapter 5 below, but it is the validity of this last assumption, the assumption that the use of simple visual displays eliminate any prior assumptions subjects might bring to the task, that we would like to question here.

Recent evidence suggests that even in experiments that use simple items with unnatural stimulus statistics that subjects are presumably relatively unfamiliar with, e.g. oriented Gabor gratings with independently drawn orientations, their internal model of the stimuli might display a rich, hierarchical structure (Brady & Alvarez, 2012; Brady & Tenenbaum, 2013; Orhan & Jacobs, 2013) that does not match the (typically unstructured) model used by the experimenter to generate the stimuli. Subjects might also make inaccurate assumptions about the statistical structure of
the stimuli used in a task. This might happen not just in VSTM tasks, but in other behavioral tasks as well. For example, subjects may implicitly assume dependencies between certain latent variables in the task (Turnham, Braun, & Wolpert, 2011), between different stimuli in a given trial (Jiang, Olson, & Chun, 2000; Orhan & Jacobs, 2013) or between stimuli across trials (Yu & Cohen, 2009; Huang & Sekuler, 2010) when, in fact, there are no such dependencies.

The existence of such discrepancies between the actual stimulus statistics used in an experiment and a subject’s internal model of the stimuli raises several questions: (i) what is the internal model used by subjects in standard VSTM experiments where the stimuli are defined by simple features and where there are no statistical dependencies between the stimuli? (ii) Why do such discrepancies exist in the first place, given that they would incur a performance cost in the task, and what are their consequences? (iii) Because various prominent phenomena in the VSTM literature are explained by appealing to ideas from neural population coding, is it also possible to explain simple forms of such discrepancies in terms of neural population coding? The following three chapters address these three questions respectively.

In Chapter 2, we propose a probabilistic clustering model that attempts to capture the internal model that subjects might be employing to encode the simple visual scenes typically used in VSTM experiments. The model implements the assumption that stimuli are generated in clusters, where the number of clusters and the assignment of stimuli to clusters are uncertain and have to be inferred probabilistically from the scene, even if the actual generative process used by the experimenter generates the stimuli independently and hence does not involve any clusters. The model predicts biases in the estimates of stimuli based on their VSTM representations, as has been observed previously (Brady & Alvarez, 2011; Huang & Sekuler, 2010). It also
predicts dependencies between the estimates of different stimuli that should decrease with the difference between their feature values. We test this novel prediction of the model using both continuous recall (delayed estimation) and change-detection tasks with small set sizes. The results of these experiments are qualitatively in agreement with the prediction of the model.

Chapter 3 more fully investigates the consequences of possible mismatches, of the type considered in Chapter 2, between the internal model that subjects might be using to encode simple visual scenes and the model actually used by the experimenter to generate the stimuli. We show that, surprisingly, this type of mismatch alone, without assuming any resource or capacity limitation, can explain some of the main qualitative characteristics of performance limitations observed in VSTM tasks typically attributed to capacity limitations, including the monotonic decline in memory precision with increasing set size, variability in memory precision across items and trials and different set-size dependencies of initial encoding rate and asymptotic precision when the duration of image presentation is varied. This chapter also tentatively proposes an explanation for the existence of mismatches between the actual stimulus statistics used in an experiment and a subject’s internal model of the same stimuli, namely that such mismatches might arise due to the long-term adaptation of the visual system to the statistics of visual stimuli in the natural environment, which may be difficult to change during the course of an experiment.

Chapter 4 considers the problem of simultaneously encoding multiple stimuli in a population of neurons with correlated responses. We demonstrate that dependencies between representations of different items that are qualitatively similar to the dependencies observed in the experiments described in Chapter 2 can arise as a natural consequence of optimal population coding in a population of neurons with correlated
responses.

The final chapter, Chapter 5, discusses the broader issues of how the study of VSTM can be extended to stimuli and tasks that are ecologically more realistic than those commonly used in standard VSTM tasks and what the implications of such an extension could be for our current view of VSTM.

Taken together, this thesis proposes novel theoretical ideas at multiple levels of analysis (Marr, 1982), from computation to neural implementation, for an ecologically more realistic and behaviorally more adequate modeling of representations in VSTM.
Questions about the capacity and precision of visual short-term memory (VSTM) have attracted much attention in recent years (Luck & Vogel, 1997; Wilken & Ma, 2004; Rouder, Morrey, Cowan, Zwilling, Morey, & Pratte, 2008; Bays & Husain, 2008; Zhang & Luck, 2008). Understanding these properties is important due to their theoretical (Cowan, 2001) and practical implications (Fukuda, Awh & Vogel, 2010). However, there is a more fundamental and often neglected issue that bears directly on memory capacity and precision, namely the content and organization of VSTM (Jiang, Olson, & Chun, 2000; Vidal, Gauchou, Tallon-Baudry, & O’Regan, 2005; Brady, Konkle, & Alvarez, 2011).

When subjects are presented with a display containing multiple items for a brief period of time, what exactly do they encode in VSTM? What would a complete description of the content of their visual memory for the display include and how is this content organized in VSTM? Do subjects only encode information about individual items or do they also encode more global information about the ensemble
of items in the display? Is the information encoded about an item independent of the information encoded about other items? These and other questions about the content and the organization of VSTM are, in a sense, more fundamental than questions about the capacity and precision of VSTM because how much information can be encoded in VSTM (capacity) and how precisely it can be encoded (precision) depend on exactly what information is encoded. For instance, the finding that subjects encode information about ensemble statistics of items in a display (Brady & Alvarez, 2011) could have a significant impact on our estimate of how much information subjects encode about individual items in VSTM.

Here, we introduce a probabilistic modeling approach that attempts to address these questions about the content and the organization of VSTM. Although, as we discuss in Section 2.8, our approach has implications for the nature of capacity limitations in VSTM, it is intended to be a more general theory of the content and organization of VSTM. We call our approach the Probabilistic Clustering Theory (PCT) of the organization of VSTM. PCT states that VSTM infers probability distributions over partitions or clusterings of visual items. Probabilistic clustering of items gives rise to biases in, and dependencies among, VSTM representations. Representations of items belonging to the same cluster share parameters, and thus are dependent. Representations of items belonging to different clusters do not share parameters, and thus are independent. However, VSTM does not infer a single partition. Rather, it infers a probability distribution over all possible partitions. As we discuss below, this property allows it to represent items at multiple granularities or scales.

The chapter is organized as follows. The next section lays out the general framework and reviews experimental evidence for biases and dependencies in VSTM rep-
resentations. The phenomena reviewed in this section are the type of phenomena our theory is primarily intended to explain. Section 2.2 discusses previous attempts at explaining some of these phenomena, focusing, in particular, on hierarchical encoding schemes. Although these schemes have many attractive properties, we argue that they also have important shortcomings. Section 2.3 introduces PCT. We motivate PCT as a natural generalization of hierarchical encoding approaches in VSTM that addresses the shortcomings of these approaches discussed in Section 2.2. We then discuss the relationships between PCT and previous works on hierarchical encoding in human memory. Section 2.5 describes the computational models that will be used in the remainder of the article. As we discuss in this section, these models can all be regarded as specific implementations of PCT with varying degrees of generality. Section 2.6 demonstrates that PCT accounts for a variety of phenomena observed in previous visual short-term recall and recognition experiments. Section 2.7 presents three new experiments designed to directly measure dependencies and biases in subjects’ VSTM representations. These experiments reveal a hitherto unrecognized form of dependence between VSTM representations of different items that is qualitatively predicted by PCT. Finally, Section 2.8 provides a summary, discusses connections with related ideas, and suggests avenues for future research.

2.1 Biases and Dependencies in VSTM

In this section, we first lay out the general framework and the mathematical notation that we use throughout the article and then review experimental evidence for biases and dependencies in VSTM.
2.1.1 Probabilistic encoding in VSTM

Consider an observer that briefly views a display containing $N$ visual items. The observer is asked to remember the feature values of these items and, after a brief delay interval, to report one or more of them. We denote the actual feature values of the items by the random variables $\theta_1, \ldots, \theta_N$. We assume that the observer only has access to noisy internal observations of these features, denoted by the variables $x_1, \ldots, x_N$, that are assumed to be corrupted by both sensory and memory noise. The generation of these noisy observations can be described by a likelihood function $p(x_i^N | \theta_i^N)$, which we assume to be a normal distribution in this chapter. In addition, the observer might have prior assumptions about the feature values of the items. These assumptions can be described by a prior distribution $p(\theta_i^N)$. Given the likelihood and the prior, the observer’s goal is to compute the posterior distribution over the feature values $\theta_1, \ldots, \theta_N$ in accordance with Bayes’ rule:

$$p(\theta_i^N | x_i^N) \propto p(x_i^N | \theta_i^N) \cdot p(\theta_i^N)$$ (2.1)

In a recall task, the observer then makes point estimates of the feature values of the items based on the posterior distribution. We denote the observer’s estimates of the feature values by the random variables $\hat{\theta}_1, \ldots, \hat{\theta}_N$. Note that $\hat{\theta}_1, \ldots, \hat{\theta}_N$ are random variables (i.e., they are stochastic) even when conditioned on a specific $\theta_1, \ldots, \theta_N$, because they depend on the noisy observations $x_1, \ldots, x_N$. In this chapter, we use the posterior mean as the observer’s estimate of the feature values of the items in recall tasks, although we found that the results presented here were robust to the choice of a specific estimator so long as the estimator was reasonable. If, for example, the observer is asked to report the feature value of a single target item $t$,
the marginal posterior corresponding to that item, \( p(\theta_t|x_i^{N}) \), is computed from the joint posterior in Equation 2.1 and the observer’s estimate is taken to be the mean of the marginal posterior: \( \hat{\theta}_t = E[\theta_t|x_i^{N}] \).

Next consider the joint probability distribution over the estimates given the feature values of the visual items:

\[
p(\{\hat{\theta}_i\}_{i=1}^{N} | \{\theta_i\}_{i=1}^{N}) = \int p(\{\hat{\theta}_i\}_{i=1}^{N} | \{x_i\}_{i=1}^{N}) p(\{x_i\}_{i=1}^{N} | \{\theta_i\}_{i=1}^{N}) \, d\{x_i\}_{i=1}^{N} \quad (2.2)
\]

where the noisy internal observations \( \{x_i\}_{i=1}^{N} \) are now integrated out. This joint distribution provides a complete characterization of how the observer represents the specific set of items \( \theta_1, \ldots, \theta_N \) in his/her VSTM. We note that most of the previous works in the VSTM literature were mainly concerned with elucidating the encoding of individual items and how it changes with set size (e.g., how the encoding precision for individual items decreases with the number of displayed items). In our framework, this corresponds to characterizing only the marginals of the full joint distribution (e.g., the precision of the marginals and how it changes with set size). In contrast, we develop experimental and computational methods to characterize the properties of the full joint distribution (thereby focusing on the joint encoding of all items), instead of emphasizing only the marginals (i.e. focusing on the encoding of individual items).

For a given set of feature values \( \theta_1, \ldots, \theta_N \), the joint distribution in Equation 2.2 can be determined empirically by presenting the same set of feature values over a number of trials and recording the observer’s estimates \( \hat{\theta}_1, \ldots, \hat{\theta}_N \) for each presentation. A contribution of this chapter is that we design novel short-term recall and recognition tasks to determine the properties of the joint distribution \( p(\{\hat{\theta}_i\}_{i=1}^{N} | \{\theta_i\}_{i=1}^{N}) \) experimentally. We will say more about how to measure this dis-
tribution experimentally in Section 2.7 below. For now, our discussion of the experimental results reviewed in the current section will limit the range of suitable forms for the joint distribution $p(\{\hat{\theta}_i\}_{i=1}^N | \{\theta_i\}_{i=1}^N)$ by ruling out some simple proposals.

Figure 2.1 schematically illustrates biases and dependencies that may arise in the joint estimates of the feature values of multiple items. The phenomena reviewed in Section 2.1.2 below provide evidence for biases in VSTM, and hence suggest that $p(\{\hat{\theta}_i\}_{i=1}^N | \{\theta_i\}_{i=1}^N)$ should have a form similar to the example shown in Figure 2.1A. The phenomena reviewed in Section 2.1.3 provide evidence for dependencies among VSTM representations of multiple items, and hence suggest that $p(\{\hat{\theta}_i\}_{i=1}^N | \{\theta_i\}_{i=1}^N)$ should have a form similar to the example shown in Figure 2.1B. Together, these biases and dependencies paint a picture of $p(\{\hat{\theta}_i\}_{i=1}^N | \{\theta_i\}_{i=1}^N)$ that has a form similar to the example shown in Figure 2.1C.

### 2.1.2 Biases in VSTM

A simple suggestion for the form of the joint distribution of the estimates is to assume that feature values of different items are represented independently in VSTM, and that estimates of individual items are only affected by the actual feature values of the corresponding items. This corresponds to the assumption that the distribution can be factorized as $p(\{\hat{\theta}_i\}_{i=1}^N | \{\theta_i\}_{i=1}^N) = \prod_{i=1}^N p(\hat{\theta}_i | \theta_i)$. This simple proposal is implicitly assumed in many studies on the capacity and precision of VSTM (Bays & Husain, 2008; Zhang & Luck, 2008). Assuming this factorized form for the joint distribution, one can then model individual distributions $p(\hat{\theta}_i | \theta_i)$ using, for example, univariate Gaussian distributions (Bays & Husain, 2008) or mixtures of Gaussian and uniform distributions (Zhang & Luck, 2008). However, there is extensive evidence against this simple factorized proposal. Here, we briefly review some of the evidence against
Kahana and Sekuler (2002) showed that inter-item similarity between stimuli influences subjects’ performances in an old/new recognition task. In their Experiment 1, subjects were shown a set of study items consisting of a series of sinusoidal gratings with different spatial frequencies. After a blank interval, they were then shown a test grating which, on half of the trials, had the same spatial frequency as one of the study items (old) and, on the other half of the trials, had a novel spatial frequency (new). The task was to decide if the spatial frequency of the test probe was old or new. The authors fit subjects’ data using a simple ‘noisy exemplar’ model that included terms for the effects of both the probe-item similarity between the test probe and each of the study items and the inter-item similarity among the study items. They found that inter-item similarity had a significant effect on subjects’ old/new decisions. In particular, when probe-item similarities were fixed, larger inter-item similarities increased the likelihood of a ‘new’ response. This result suggests that the estimate of each individual item $\hat{\theta}_i$ depends on the inter-item similarity among all study items [contrary to the assumption that memories for individual items depend only on the feature values of their corresponding items; i.e., $p(\hat{\theta}_i|\{\theta_j\}_{j=1}^N) \neq p(\hat{\theta}_i|\theta_i)$]. In a later section (Section 2.6), we show how PCT explains the inter-item similarity effect.

Kahana, Sekuler and colleagues replicated the inter-item similarity effect in later works (Zhou, Kahana & Sekuler, 2004; Kahana, Zhou, Geller & Sekuler, 2007; Viswanathan, Perl, Visscher, Kahana, & Sekuler, 2010), and showed that the same qualitative inter-item similarity effect can be observed in visual short-term memory for realistic-looking synthetic face stimuli (Yotsumoto, Kahana, Wilson & Sekuler, 2007) as well as in auditory short-term memory (Visscher, Kaplan, Kahana & Sekuler,
Huang and Sekuler (2010) showed that visual short-term recall memory for the spatial frequency of a sinusoidal grating is biased toward both the non-target gratings shown on the same trial and the average frequency of the gratings shown on all previous trials. On each trial of their Experiment 2, a subject was successively shown a pair of Gabor stimuli. One of the Gabors was then cued, and the subject reported the spatial frequency of the cued Gabor by adjusting the spatial frequency of a comparison Gabor using a computer mouse. The recall error (the difference between the actual spatial frequency of the target Gabor and the subject’s reproduction) was measured on each trial. Over trials, this yielded an error distribution that reflected the precision of, and potential biases in, a subject’s short-term memory for spatial frequency. It was found that there were two distinct biases influencing subjects’ recall of the spatial frequencies of target Gabors: a bias toward the spatial frequency of the non-target Gabor shown on the same trial, and a bias toward the average spatial frequency of stimuli shown on previous trials in the experiment. A similar bias toward mean spatial frequencies was observed in Experiment 9 of Wilken and Ma (2004). Again, these results indicate that the estimate of a target item \( t \) depends on the non-target items presented on the same trial, as well as on items shown on previous trials, and not solely on the feature value of the target item itself [i.e., \( p(\hat{\theta}_i|\{\theta_i\}_{i=1}^N) \neq p(\hat{\theta}_i|\theta_i) \)].

Specifically (and assuming, for example, \( N = 2 \) items), the biases observed in Huang and Sekuler (2010) and in Wilken and Ma (2004) suggest that \( p(\hat{\theta}_1|\theta_1, \theta_2) \) is biased (or shifted) toward \( \theta_2 \) and, conversely, \( p(\hat{\theta}_2|\theta_1, \theta_2) \) is biased toward \( \theta_1 \) (see Figure 2.1A). In a later section (Section 2.6), we show that our PCT satisfies this property and explains the biases observed in Experiment 9 of Wilken and Ma (2004).
Figure 2.1: Schematic illustration of biases and dependencies that may arise in the joint estimates of two items based on their VSTM representations. (A) Biases manifest themselves as shifts of the distribution $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$ (represented here by a single contour) from the actual feature values of the items. In this example, the mean of the distribution $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2)$ (represented by the black dot) is shifted away from the actual feature values of the items, $(\theta_1, \theta_2)$ (represented by the square) and toward the main diagonal, indicating that the estimates of the feature values of both items are biased toward the mean of $\theta_1$ and $\theta_2$. Note, however, that the distribution is spherical, hence, assuming a Gaussian distribution for simplicity, there are no dependencies between $\hat{\theta}_1$ and $\hat{\theta}_2$. (B) Dependencies manifest themselves as statistical dependencies among $\{\hat{\theta}_i\}_{i=1}^N$. In this example, representations of the two items, $\hat{\theta}_1$ and $\hat{\theta}_2$, are correlated. Note, however, that there are no biases in the representations as $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2)$ is centered on $(\theta_1, \theta_2)$. Also note that this example depicts only a simple form of dependence between $\hat{\theta}_1$ and $\hat{\theta}_2$, namely second-order correlation. More complex or higher-order dependencies between $\hat{\theta}_1$ and $\hat{\theta}_2$ are also possible. (C) A hypothetical example where there are both biases and dependencies in the joint estimates of the two items.
2.1.3 Dependencies in VSTM

The biases in VSTM reviewed in the previous subsection indicate that estimates of individual items depend not just on the actual feature values of their corresponding item, but also on the feature values of other items presented in the display. Consequently, VSTM cannot be characterized using a simple joint probability distribution of the form $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N) = \prod_{i=1}^N p(\hat{\theta}_i|\theta_i)$. However, these biases do not rule out slightly more complex joint probability models of the form $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N) = \prod_{i=1}^N p(\hat{\theta}_i|\{\theta_j\}_{j=1}^N)$. Here, the estimates of individual items $\hat{\theta}_i$ depend on the feature values of all visual items, but these estimates are independent of each other given the feature values of all items. A series of elegant experiments by Jiang, Olson and Chun (2000), however, rule out this latter form of joint probability model as an accurate description of the organization of VSTM.

In each trial of their Experiment 1, Jiang, Olson and Chun (2000) briefly presented a display consisting of colored squares. Following a blank interval, subjects were shown a test display. There were two test conditions. In the single probe condition, only one of the squares (called the target probe) reappeared, either with the same color as in the original display or with a different color. In the minimal color change condition, the target probe (again with the same color or with a different color) reappeared together with distracter squares which always had the same colors as in the original display. In both conditions, the task was to decide whether a color change occurred in the target probe. It was found that subjects’ performances were significantly better in the minimal color change condition than in the single probe condition. This result suggests that the color for the target square was not encoded independently of the colors of the distracter squares because, otherwise, the absence or presence of the distracter squares would not have affected change detection per-
formances for the target. In Experiment 2, the authors observed a similar result for location memory. Location memory for a target was better in the minimal change condition than in the single probe condition or in a maximal change condition in which all distracters were presented but at locations differing from their original locations.

These results are easy to understand in terms of a joint probability model for the estimates $p(\{\hat{\theta}_i\}_{i=1}^N | \{\theta_i\}_{i=1}^N)$ (in what follows, we omit the dependence on $\{\theta_i\}_{i=1}^N$ for brevity of notation, but all distributions should be considered to be implicitly conditioned on $\{\theta_i\}_{i=1}^N$). Intuitively, the single probe condition taps into the marginal probability distribution of a subject’s estimate of the target item $p(\hat{\theta}_t)$ where $t$ indexes the target item, because in the single probe condition distracters are not shown to the subject during test, and thus he/she has to marginalize over his/her uncertainty regarding the feature values of the distracter items. In contrast, the minimal color change condition taps into the conditional probability distribution of the estimate of the target given that the estimates of the distracters are set to the actual feature values of their corresponding items (i.e., $p(\hat{\theta}_t | \hat{\theta}_{-t} = \theta_{-t})$ where $-t$ is the set of indices of the distracter items) because the actual distracters $\theta_{-t}$ are shown to the subject during test. If the target probe has high probability under these distributions, then the subject will be more likely to respond ‘no-change’, whereas if it has low probability, then the subject will be more likely to respond ‘change’. Importantly, if the items are represented independently in VSTM, the marginal and conditional distributions are the same [i.e., $p(\hat{\theta}_t) = p(\hat{\theta}_t | \hat{\theta}_{-t})$]. Because subjects’ performances in the single probe and minimal color change conditions were different, subjects’ marginal and conditional probability distributions must have been different (see Figure 2.1B for a graphical illustration of a simple form of dependence between estimates of different
items). Hence, the results of Jiang, Olson and Chun (2000) provide evidence against the independence assumption.

It is also easy to understand why subjects performed better in the minimal color change condition than in the single probe condition. The conditional distribution $p(\hat{\theta}_t|\hat{\theta}_{-t})$ is, in general, a lower-variance distribution than the marginal distribution $p(\hat{\theta}_t)$. Although this is not exclusively true for the Gaussian distribution, it can be analytically proven in the Gaussian case. If $p(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N)$ is modeled as an $N$-dimensional multivariate Gaussian distribution:

$$\begin{bmatrix} \hat{\theta}_t \\ \hat{\theta}_{-t} \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix})$$

then the conditional distribution $p(\hat{\theta}_t|\hat{\theta}_{-t})$ has mean $a + CB^{-1}(\hat{\theta}_{-t} - b)$ and variance $A - CB^{-1}C^T$, whereas the marginal distribution $p(\hat{\theta}_t)$ has mean $a$ and variance $A$ which is always greater than $A - CB^{-1}C^T$.

Using three novel experiments, Section 2.7 provides additional experimental evidence for dependencies in VSTM and discusses how these dependencies are accounted for by our PCT.

### 2.2 Hierarchical Encoding of Items in VSTM

The phenomena reviewed in the previous section restrict the range of suitable joint probability distributions for characterizing VSTM, but they do not completely determine the exact form of this distribution. In this section, we discuss a proposal for the joint distribution recently put forward by Brady and Alvarez (2011). We first review their experimental results and their hierarchical encoding models. The discussion
of their work will help us motivate PCT as a natural generalization of hierarchical encoding models, as we believe that PCT addresses some of the shortcomings of their hierarchical modeling approach.

2.2.1 Brady and Alvarez (2011)

It has recently been argued that VSTM is organized hierarchically where items are simultaneously encoded at multiple levels of abstraction (Brady & Alvarez, 2011). At a fine scale, each item might be represented individually. Items might also be represented at a coarser scale through summary or ensemble statistics of the feature values of all items in a display. Brady and Alvarez (2011) formalized this idea using a hierarchical modeling approach. We show below that, in our framework, this corresponds to using a specific hierarchical form for the prior distribution over the feature values, \( p(\theta_1, \ldots, \theta_N) \), in Equation 2.1.

Similar to the inter-item similarity effects shown by Kahana, Sekuler and colleagues, Brady and Alvarez (2011) demonstrated that memory for individual items in a display is influenced by ensemble statistics of all presented items. In their Experiment 1, subjects were presented with blue, red and green circles of different sizes for a brief duration. Subjects were explicitly instructed to ignore the green circles but to remember the sizes of the red and blue circles. After a delay interval, a comparison circle appeared at the location of a red or blue circle in the original display. Subjects’ task was to indicate the size of the original circle that was at that location, referred to as the target circle, by using the mouse to resize the comparison. The authors found that the reported size of the target circle was biased toward the average size of the circles having the same color as the target.

Brady and Alvarez (2011) hypothesized that the fact that color was a task-
relevant feature in Experiment 1 (subjects had to remember only the red and blue circles and ignore the green circles) might have increased the salience of this feature, thereby inducing subjects to use a color-based encoding for the items. If so, then the observed bias toward the mean size of the same-colored circles should disappear when performing a task that is similar except that color is task-irrelevant. In their Experiment 2, the authors tested this prediction by removing the green circles from the display and presenting only red and blue circles in each trial. Subjects were asked to remember the sizes of all circles in the display. Therefore, color was no longer a task-relevant feature. Consistent with their hypothesis, the authors found that subjects did not show a bias toward the mean size of the same-colored circles. Instead, subjects’ estimates showed a bias toward the mean size of all circles in a display. The results of Experiments 1 and 2 suggest that subjects employ a flexible strategy, encoding stimuli at different levels of abstraction in VSTM in different task contexts.

To explain the distinctive pattern of biases when color was a salient feature versus when it was not, Brady and Alvarez (2011) used a two-level hierarchical model to account for subjects’ data in Experiment 2 and a three-level hierarchical model to account for data from Experiment 1. The two-level model assumes that subjects encode the items at two different levels of abstraction: the level of individual circles (individual encoding) and the ensemble mean and variance of the feature values (sizes) of all circles in a display. The model also assumes that the feature values of individual items are conditionally independent given the ensemble statistics (though they are still marginally dependent due to the shared ensemble statistics). Using our notation, their two-level model corresponds to choosing the following prior over the
feature values of the items in Equation 2.1:

$$p(\theta_1, \theta_2, \ldots, \theta_N) = \int p(\theta_1, \ldots, \theta_N | \Phi)p(\Phi) d\Phi = \int p(\theta_1 | \Phi) \ldots p(\theta_N | \Phi)p(\Phi) d\Phi$$  

(2.4)

where \( \Phi \) denotes the ensemble statistics for the feature values of all items in a display. Since the color-based grouping of circles is not taken into account in this model, only a bias toward the overall mean size is predicted when estimating the sizes of individual circles (Gelman, Carlin, Stern & Rubin, 2004, p. 117), in accord with the results of Experiment 2.

In the three-level model, the three levels were: (i) the level of individual circles (individual encoding), (ii) the group-level means of the sizes of the red circles and of the blue circles, and (iii) the ensemble mean size of all circles in the display. Similar to Equation 2.4, the three-level model corresponds to using the following prior in Equation 2.1:

$$p(\theta_{r,1}, \ldots, \theta_{r,N_r}, \theta_{b,1}, \ldots, \theta_{b,N_b}) =$$

$$\int \int \int p(\theta_{r,1}, \ldots, \theta_{r,N_r} | \Phi_r)p(\theta_{b,1}, \ldots, \theta_{b,N_b} | \Phi_b)p(\Phi_r, \Phi_b | \Phi)p(\Phi) d\Phi_r d\Phi_b d\Phi$$  

(2.5)

where \( \theta_{r,i} \) and \( \theta_{b,j} \) are the feature values of individual red and blue circles at the finest level, \( \Phi_r \) and \( \Phi_b \) are the summary statistics at the group or color level, \( \Phi \) is the global ensemble statistics of all circles, and \( N_r \) and \( N_b \) are the number of red and blue circles respectively. Since the color-based grouping of circles is explicitly incorporated into the model, biases toward group-level means and toward the global ensemble mean are predicted by this model, mostly consistent with the results of Experiment 1 (whether the authors observe a bias toward the global mean in addition to the group-level bias in Experiment 1 is unclear).
The hierarchical encoding framework of Brady and Alvarez (2011) provides an elegant way of accounting for biases in VSTM by assuming that subjects simultaneously encode items at multiple levels of abstraction in VSTM. As such, their framework furthers our understanding of the organization of VSTM. However, we believe that it also has important disadvantages which are illustrated by considering the specific models proposed by the authors. The two-level and three-level models allow for the representation of items at multiple levels of abstraction, but there are at least two problems with the way they do so. First, the use of hierarchical models with different numbers of levels to account for different patterns of results observed in different experiments is ad hoc. In general, it is not clear what determines the number of levels that should be used for a given experiment or the appropriate ‘grain’ of those levels. Second, the number of groups (the number of relevant colors) and the assignment of each circle to a group was explicitly specified in an a priori manner when formulating the three-level model (see Equation 2.5). Although it is easy to do so for the purposes of modeling Experiment 1, it is generally not clear how to define groups or how to assign items to groups in more naturalistic cases. Grouping of visual items is often highly ambiguous both with respect to the number of groups and the assignment of items to groups. A model that could automatically determine these properties and also take into account the uncertainty about them would provide significant explanatory power.

We illustrate these issues within the context of a hypothetical VSTM experiment. Consider an experiment in which subjects are asked to remember the horizontal locations of a number of briefly presented colored squares. A representative display from a single trial of such an experiment is shown in Figure 2.2 where the vertical locations of the squares are linearly spaced and fixed (represented by the six hori-
Figure 2.2: Single trial of a hypothetical VSTM experiment in which subjects are asked to remember the horizontal locations of 6 colored squares.

Horizontal lines), whereas their horizontal locations are assigned randomly. Each of the squares in the display can be represented at multiple scales (or levels of abstraction) in VSTM. Consider, for instance, the square that lies on the fifth line from the top. At the finest scale, this square can be represented individually. At a slightly coarser scale, it can be encoded together with the square that lies on the sixth line from the top (whose horizontal location is closest to that of the fifth square). At a still coarser scale, it can be encoded together with the first, second and the sixth squares from the top, and at the coarsest scale, all squares can be represented together.

Once the possibility of encoding items at multiple scales in VSTM is established, three important questions arise: (i) how many scales should be used to represent
a set of items; (ii) how are the appropriate scales for the representation of items determined; and (iii) how much weight should be given to representations at different scales? With respect to the example shown in Figure 2.2, the previous paragraph mentioned four possible scales for the representation of the fifth square from the top. But why these scales in particular? Instead of representing this square together with the sixth square at an intermediate scale, why not group it with the fourth square and encode them together, or why not introduce a different scale and represent the fifth square together with, say, the third and the fourth squares? Note that the hierarchical modeling approach of Brady and Alvarez (2011) cannot give satisfactory answers to these questions, as the specific scales and groups for representing an item are explicitly specified in advance, and not inferred by the theoretical model. In contrast, as discussed more extensively below, our proposed PCT provides answers to these questions.

PCT is based on describing a set of items in terms of a probability distribution over all possible partitions where each partition might have a different ‘granularity’. Since PCT considers multiple possible partitions, it can represent an item at multiple scales simultaneously (see the next section). Through standard probabilistic (Bayesian) inference, PCT automatically determines the appropriate partitions for the particular set of items at hand and the probabilities or weights that should be allocated to each partition.

### 2.3 Probabilistic Clustering Theory (PCT)

Just as Brady and Alvarez’s (2011) two- and three-level hierarchical models can be expressed as specific choices for the prior distribution over the feature values of items
(i.e., $p(\theta_1, \ldots, \theta_N)$; Equations 2.4-2.5), PCT also corresponds to a specific choice for the prior distribution. In this section, we describe the properties of the prior assumed by PCT in an informal way. Mathematical details will be provided in a later section (Section 2.5).

Intuitively, the prior assumed by PCT imposes a probabilistic clustering structure on the feature values. According to PCT, an observer’s internal model of the generative process for $\theta_1, \ldots, \theta_N$ assumes that these feature values are generated in clusters, even if the actual generative process does not involve any clusters, as is the case in all experiments considered in this chapter. In other words, the observer assumes that the world has a “clumpy” structure. In estimating the feature values of a set of items based on noisy observations of these feature values, the observer takes into account (or integrates over) his/her uncertainty about the clustering structure of the set of items. This uncertainty might concern both the number of clusters and the assignment of items into clusters.

More specifically, PCT assumes that VSTM automatically infers a probability distribution over all possible partitions of a set of items. Consider a set of three items with feature values denoted by $\theta_1$, $\theta_2$, and $\theta_3$. There are five possible partitions of these items: (i) $\{\theta_1\}, \{\theta_2\}, \{\theta_3\}$ (each item belongs to its own group or cluster); (ii) $\{\theta_1, \theta_2, \theta_3\}$ (all items belong to the same cluster); (iii) $\{\theta_1, \theta_2\}, \{\theta_3\}$; (iv) $\{\theta_1\}, \{\theta_2, \theta_3\}$; and (v) $\{\theta_1, \theta_3\}, \{\theta_2\}$. Based on the similarities among the items’ feature values, VSTM infers a distribution over these five possibilities. If, for example, the items are highly similar, VSTM will tend to assign a large probability to the partition that places all items in the same cluster: $\{\theta_1, \theta_2, \theta_3\}$. If items are highly dissimilar, a large probability will be assigned to the partition that places each item in its own cluster: $\{\theta_1\}, \{\theta_2\}, \{\theta_3\}$. And if $\theta_1$ and $\theta_2$ are somewhat similar (and somewhat
dissimilar) but they are both highly dissimilar from $\theta_3$, then moderate probabilities will be assigned to partitions $\{\theta_1, \theta_2\}, \{\theta_3\}$ and $\{\theta_1\}, \{\theta_2\}, \{\theta_3\}$.

We note that if the number of clusters is constrained to be 1 (i.e., all items are necessarily grouped into a single cluster), then PCT reduces to the two-level hierarchical model of Brady and Alvarez (2011) (Equation 2.4). Thus, the latter model can be seen as a special case of PCT, where the 1-cluster partition, where all items are assigned to the same cluster, is given a probability of 1 and the other partitions are given a probability of 0. However, in the general case, PCT does not set any \textit{a priori} bounds on the number of clusters, but instead determines this automatically from the data. According to PCT, VSTM does not infer a single partition of visual items, rather, it infers a probability distribution over all possible partitions of these items. This enables the representation of items at multiple scales as in the hierarchical modeling approach of Brady and Alvarez (2011). However, unlike the Brady and Alvarez approach, the appropriate scales for the representation of items and their weights are determined automatically from the data (i.e., from the noisy observations of the feature values of items).

Importantly, PCT predicts that observers’ VSTM representations will display biases and dependencies. PCT predicts biases in the estimates of feature values of items toward the cluster means of the clusters that they are assigned to. Because PCT does not infer a single clustering, but a probability distribution over many clusterings, in general, there will be biases at multiple scales (toward the means of all clusters that an item might be assigned to), the net effect of which will depend on the posterior probabilities, or the weights, of different clusterings.

PCT also predicts dependencies between the estimates of feature values of different items assigned to the same cluster. Two items that are never assigned to the
same cluster will not share parameters and hence their estimates will be independent. In contrast, items assigned to the same cluster share parameters and thus their estimates are dependent. PCT predicts that the magnitude of the dependence between the estimates of the feature values of two items should increase with the similarity between the feature values of the items. This is because, according to PCT, the more similar two items are, the more likely they are to be assigned to the same cluster, hence to share parameters. In Section 2.7, we provide experimental evidence supporting this crucial prediction of PCT.

A more detailed discussion of the predictions of PCT regarding biases and dependencies in VSTM will be given in Section 2.5 below.

### 2.4 Relationship between PCT and previous works on hierarchical encoding in human memory

Prior to Brady and Alvarez (2011), several other researchers argued that human memory is organized hierarchically, thereby accounting for categorical effects in memory (Huttenlocher, Hedges & Duncan, 1991; Hemmer & Steyvers, 2009a; Hemmer & Steyvers, 2009b). Consistent with the predictions of a hierarchical Bayesian model, Hemmer and Steyvers (2009b) showed that in an episodic recall task, subjects’ estimates were biased toward both the mean feature value of the presented object category (e.g., mean size of apples) and the mean feature value of the superordinate category (e.g., mean size of all fruits). Hemmer and Steyvers (2009a) extended this result to unfamiliar objects, and developed a hierarchical Bayesian model that could account for the distinct pattern of biases observed for familiar and unfamiliar objects. For familiar objects, the model predicted, in accord with experimental results,
that the recalled size of an object would be biased more toward the mean size of the objects of that kind. For unfamiliar objects, the recalled size was be biased toward the superordinate-level mean (e.g., mean size of fruits).

These biases are similar to the biases observed in VSTM experiments. PCT accounts for these analogous biases in VSTM, not as an effect of pre-existing categories in memory, but rather as an effect of encoding displayed items at multiple scales. To make the similarity to the work of Hemmer and Steyvers (2009b) clearer, one can argue that in VSTM experiments, subjects spontaneously form clusters or ‘categories’ at multiple scales when shown multi-item displays. These categories influence short-term memories for visual items in a way that is similar to the way categories in long-term memory affect episodic recall performance of individual items.

Huttenlocher, Hedges and Duncan (1991) demonstrated that people show systematic biases even in the simple task of estimating the spatial location of a single dot presented within a circle. Subjects showed a bias toward the centers of the four quadrants dividing the circle in their judgments of the angular locations of single dots presented for a brief duration. Huttenlocher, Hedges and Duncan (1991) conceived of the four quadrants as categories, and the centers of the quadrants as prototypical examples of those categories. Their model of how subjects estimated spatial locations was essentially the same as Brady and Alvarez’s (2011) two-level hierarchical Bayesian model reviewed above, with ‘categories’ or quadrants providing the higher-level representations. Similar to the model of Hemmer and Steyvers (2009a), their model also assumed pre-existing knowledge of categories (in this case, quadrants). Huttenlocher, Hedges and Vevea (2000) extended these results by showing that similar biases were evident when subjects estimated other perceptual features of objects belonging to inductively defined categories, and that the observed biases
were influenced by properties of the distributions describing these categories.

Brady and Tenenbaum (2010) developed a discrete slot-based model of VSTM that encodes both high-order structure about a simple display of dot patterns and detailed information about specific dots. They formalized the concept of high-order structure in terms of a correlation parameter, called the ‘gist’ parameter, in a Markov random field such that larger values of this parameter corresponded to images that tended to have similarly colored neighbors, whereas smaller values corresponded to images that tended to have differently colored neighbors. In addition, the model assumed the encoding of detailed information about K specific dots in such a way that ‘outlier’ dots (dots that did not conform to the overall gist of the image) were more likely to be encoded. Brady and Tenenbaum (2010) presented evidence suggesting that the performance of this ‘gist+exception’ model of encoding correlated well with human performance on an image-by-image basis.

Brady and Tenenbaum’s (2010) gist representation applies only to images sampled from a Markov random field, whereas the representations in PCT can be applied more generally. The individual encoding mechanism in their model is biased toward exceptions to the gist. It is interesting to note that PCT implements a similar bias because the further the feature value of an item is from feature values of other items in a display, the more likely that the item will be assigned to a cluster of its own, meaning that the item will be encoded individually.

The hierarchical models described in this section are significant because they advance our understanding of how hierarchical representations can account for biases and categorical effects observed in human memory. A fundamental difference between the previously proposed hierarchical models and PCT is that the hierarchical models assume pre-specified and fixed levels of abstraction to represent items. Hemmer
and Steyvers (2009a, 2009b) assume two levels of abstraction: an object-based level (e.g., apples) and a categorical level (e.g., fruits). Huttenlocher, Hedges and Duncan (1991) also assume two fixed levels of abstraction: a fine-grained representation of the location of a dot and a more global, coarse-grained prototype representation based on pre-specified prior knowledge. Brady and Tenenbaum’s (2010) model represents a display at two fixed levels of abstraction, and Brady and Alvarez’s (2011) hierarchical modeling framework assumes an appropriate pre specification of the levels of abstraction for the representation of items.

In contrast, an innovation of our approach is that it automatically determines multiple scales or levels of abstraction that are appropriate for the representation of an item in VSTM and how to weight the different scales without assuming pre-specified and fixed levels of abstraction. It does so by inferring multiple partitions with different granularities that are appropriate for a given set of items along with the posterior probabilities of those partitions. Our focus is on VSTM but we speculate that studies of other human memory systems would benefit from our approach which emphasizes the use of multi-scale representations without pre-specified, fixed hierarchies.

2.5 Models

This section describes the specific computational models that will be used in the following sections. Three models are described. The first model is a Dirichlet process mixture model (DPMM), also known as an infinite mixture model. The second model is a Bayesian finite mixture model (BFMM). Both models automatically infer posterior distributions over multiple partitions of a set of items. The only difference
between these two models is that they make different assumptions about the maximum number of clusters that the data can be grouped into. The DPMM does not set any *a priori* limit on the maximum number of clusters, whereas a BFMM assumes that the data can be grouped into at most $K$ clusters, for a finite $K$ specified in advance. Because of this difference, the DPMM can infer a probability distribution over all possible clusterings or partitions of a set of items but a BFMM with too few clusters cannot. For this reason, we regard the DPMM as an exact implementation of the PCT, whereas the BFMM can be regarded as an approximation to the DPMM (where the quality of approximation will be determined by $K$). The last model we describe is the two-level hierarchical Bayesian model (HBM) proposed by Brady and Alvarez (2011). As discussed below, the HBM can be regarded as a special case of BFMMs where the number of components $K$ is constrained to be 1. A graphical representation of the models considered in this chapter is shown in Figure 2.3.

### 2.5.1 Dirichlet process mixture models (DPMMs)

DPMMs are commonly used in statistics and machine learning (Neal, 2000; Görür & Rasmussen, 2010). In cognitive science, they are gradually becoming popular in the study of perception and cognition where they have been used as normative models of word segmentation (Goldwater, Griffiths & Johnson, 2009), causal learning (Gershman, Blei, & Niv, 2010) and categorization (Sanborn, Griffiths & Navarro, 2010). Excellent introductions to these models can be found in Goldwater, Griffiths and Johnson (2009) and Navarro, Griffiths, Steyvers and Lee (2006).

Here, we describe the application of the DPMM to the problem of encoding multiple items in VSTM. Consider a single trial of a hypothetical VSTM experiment in which an observer is asked to remember the feature values (e.g., horizontal locations
Figure 2.3: A graphical representation of the models considered in this chapter. All models have a common structure represented by the graphical model shown on the left. We illustrate this graphical model using plate notation, where the nodes inside the plate are meant to be replicated $N$ items. The shaded node represents the observable variables (i.e., the noisy observations $x_i$). The other variables are latent or unobservable. The remaining plots illustrate the generative processes defining the models. Each row illustrates the generation of variables at the corresponding level in the graphical model on the left. The only difference between the models is in the variable $G$. For the DPMM, $G$ is a discrete distribution with an infinite number of “atoms”; for the BFMMs, it is a discrete distribution with $K$ atoms; and for the HBM, it is a single atom. In the example shown here, the DPMM uses three clusters to generate the four items represented by $\theta_i$s, the BFMM uses 2 clusters and the HBM uses a single cluster. $\mu_i$ (mean) and $\tau_i$ (precision) represent the cluster parameters for item $i$. The distributions at the bottom two rows illustrate the distributions from which the variables at the corresponding levels were drawn.
of squares or orientations of Gabor gratings) of $N$ items in a display. For the moment, we consider items defined by a single feature (e.g., position, orientation, color, shape). We denote the actual feature value of item $i$ by $\theta_i$. As laid out in an earlier section, we assume that the observer does not have access to the actual feature values of the items, but to noise-corrupted observations thereof, denoted by $x_i$. In addition, the observer’s internal model of the generative process for $\theta_i$s assumes that these feature values are generated in clusters (even if the actual generative process does not involve any clusters). In estimating the feature values of a set of items based on the corresponding noisy observations, the observer integrates out its uncertainty about the clustering structure of the set of items. Mathematically, the full model can be specified as follows (see Figure 2.3):

$$G \sim DP(G_0, \alpha) \quad (2.6)$$

$$G_0(\mu_i, \tau_i) = U(\mu_i; a, b)G(\tau_i; \alpha_\tau, \beta_\tau) \quad (2.7)$$

$$\mu_i, \tau_i | G \sim G \quad (2.8)$$

$$\theta_i | \mu_i, \tau_i \sim N(\theta_i; \mu_i, \tau_i) \quad (2.9)$$

$$x_i | \theta_i \sim N(x_i; \theta_i, \tau_{obs}) \quad (2.10)$$

Here, $x_i \sim N(x_i; \theta_i, \tau_{obs})$ means that $x_i$ is distributed according to a normal distribution with mean $\theta_i$ and precision $\tau_{obs}$. The precision $\tau_{obs}$ is meant to capture the combined effects of both sensory and memory noise in generating noisy internal observations of the actual feature values. However, sensory noise is likely to be negligible compared with memory noise. We assume that $\tau_{obs}$ may depend on set size, but otherwise it is identical for all items in a given trial and across different trials with the same set size. Recent results suggest that introducing variability in $\tau_{obs}$
across trials and across items in a given trial can lead to better models of capacity limitations in VSTM (van den Berg, Shin, Chou, George & Ma, 2012). We found that variability in $\tau_{\text{obs}}$ was not essential for accounting for the phenomena we consider in this chapter (i.e., biases and dependencies in VSTM), therefore for simplicity, we decided to assume identical $\tau_{\text{obs}}$ across trials and across items. $\mu_i$ and $\tau_i$ represent the mean and the precision of the cluster to which item $i$ belongs, and they are jointly distributed according to a countably infinite discrete distribution denoted by $G$. $G$ is itself distributed according to a Dirichlet process with base distribution $G_0$ and concentration parameter $\alpha$. The base distribution $G_0$ is the product of a uniform distribution for mean $\mu_i$ defined over the interval $[a, b]$ and a gamma distribution for precision $\tau_i$ with scale parameter $\alpha_{\tau}$ and shape parameter $\beta_{\tau}$.

In a given trial, the observer’s goal is to infer the feature values of the items, or the feature value of a single target item, given the noisy observations $\{x_i\}_{i=1}^N$. This problem can be formalized as the inference of the joint posterior distribution $p(\{\theta_i\}_{i=1}^N | \{x_i\}_{i=1}^N)$ if the feature values of all items are to be estimated, or the marginal posterior $p(\theta_t | \{x_i\}_{i=1}^N)$ if only the feature value of a single target item $t$ is to be estimated. In the simulations reported below, these posterior distributions were computed using a Markov chain Monte Carlo (MCMC) sampling algorithm with auxiliary variables (Algorithm 8 of Neal, 2000; see Görür, 2007, and Appendix A.1 for additional details). For recall tasks, we then use the mean of the marginal posterior distribution as the observer’s estimate of the feature value of the target item in that trial. For old/new recognition tasks, on the other hand, the posterior distribution is transformed into probabilities of responding ‘old’ or ‘new’ to a given probe item (see the next section for details).

We now describe the clustering properties of the DPMM, given by Equations
Let $p$ denote a vector or “atom” in $(\mu, \tau)$ space (i.e., $p$ is a possible set of values for mean $\mu$ and precision $\tau$). It can be shown that a single draw, $G$, from a Dirichlet process is a countably infinite discrete distribution over atoms $p$ (Ferguson, 1973). That is, $G$ is a weighted sum of an infinite number of discrete atoms $\{p_k\}_{k=1}^{\infty}$:

$$G(p) = \sum_{k=1}^{\infty} \pi_k \delta(p = p_k)$$  \hspace{1cm} (2.11)

The base distribution of the Dirichlet process, $G_0$, is a prior distribution over the $(\mu, \tau)$ space. $G_0$ determines the locations of the atoms because $\{p_k\}_{k=1}^{\infty}$ are independent samples from $G_0$. The concentration parameter $\alpha$ to the Dirichlet process determines the weights of the atoms $\pi_k$ in Equation 2.11. For small values of $\alpha$, a small number of atoms are given large weights and the rest are assigned very small weights. For large values of $\alpha$, weights are distributed more broadly across atoms.

A clearer understanding of the concentration parameter $\alpha$ emerges when one considers the relationship between $\alpha$ and the clustering properties of the DPMM. Recall that the mean $\mu_i$ and the precision $\tau_i$ are the parameters of the cluster to which item $i$ is assigned. When performing inference, $\mu_i$ and $\tau_i$ for different items $i$ are assigned identical values if these items are assigned to the same component. The concentration parameter $\alpha$ acts as a bias on this assignment process by influencing the probability that two items will be grouped together. Roughly, $\alpha$ controls the observer’s tendency to group items. For small values of this parameter, the model is biased toward a small number of clusters or groups of similar items (“chunks”). For large values, it tends to assign each item to its own cluster.

As indicated above, the base distribution of the Dirichlet process $G_0$ is the product of a uniform distribution over $\mu_i$ and a gamma distribution over $\tau_i$. Since we apply the model to small datasets (typically displays with 2-8 items), it is important to
use a relatively non-informative base distribution for \( \mu_i \). Otherwise its posterior distribution would be strongly affected by the base distribution, creating large biases due to the choice of this distribution alone. Consequently, we use a uniform base distribution over \( \mu_i \) defined over a sufficiently large interval (thereby making the model non-conjugate; see Göğür and Rasmussen, 2010). We set the range of the uniform distribution, \([a, b]\), to a sufficiently large interval that includes the minimum and maximum possible values for the relevant variable in each experiment considered below. For the parameters of the gamma distribution on \( \tau_i \), we put a \( \mathcal{G}(1, 1) \) prior on scale parameter \( \beta_{\tau} \) and set \( \alpha_{\tau} = 1 \). Lastly, we put a \( \mathcal{G}(\alpha_c, 1) \) prior on the Dirichlet process concentration parameter \( \alpha \) and treat \( \alpha_c \) as a free parameter. For a given set size, this reduces the number of free parameters to just two, namely \( \alpha_c \) (a prior parameter for concentration parameter \( \alpha \)) and \( \tau_{obs} \) (memory precision). The same values of \( \alpha_c \) and \( \tau_{obs} \) were used for all trials of a simulated experiment.

We illustrate the working of the model with a simple example in Figure 2.4. Figure 2.4A shows three noisy observations, \( x_1, x_2 \) and \( x_3 \) (represented by the vertical lines) and the three marginal posteriors, \( p(\theta_1|x_1, x_2, x_3) \), \( p(\theta_2|x_1, x_2, x_3) \) and \( p(\theta_3|x_1, x_2, x_3) \) (represented by the solid curves) for four different settings of the parameters. Figure 2.4B shows the posterior distributions over the number of clusters for each setting of the parameters in Figure 2.4A. The marginal posteriors in Figure 2.4A display biases (i.e., their means are not centered on the \( x_i \)s). This is because an item is often grouped with one or both other items, shifting the marginal posteriors toward the \( x_i \)s associated with those items. This is essentially the mechanism by which the DPMM accounts for the biases reviewed in Section 2.1 above (also see Section 2.6 below). Increasing \( \alpha_c \) forces the model to use a larger number of clusters. This has the effect of reducing biases in the marginal posteriors, because each item is
Figure 2.4: (A) Three noisy observations, $x_1$, $x_2$ and $x_3$ (vertical lines) and the three marginal posteriors, $p(\theta_i | x_1, x_2, x_3)$ for $i = 1, 2, 3$ (solid curves), for four different settings of the parameters, $\alpha_c$ and $\tau_{obs}$. (B) Posterior distributions over the number of clusters for each of the corresponding subplots in (A).

now more likely to be assigned to its own cluster. Increasing $\tau_{obs}$, on the other hand, reduces the variance of the marginal posteriors and also reduces the biases, because when $\tau_{obs}$ is high, the model relies more heavily on the observations, $x_i$, in computing the posteriors over $\theta_i$s, and less on the prior over $\theta_i$s induced by the DPMM.

The DPMM also predicts dependencies between the estimates of feature values of different items encoded in VSTM. These dependencies arise in the model when different items are grouped into the same normal component. Specifically, the representations corresponding to these items become dependent due to the shared parameters of the normal component. Importantly, the model further predicts that
the dependency between the memory representations of two items should decrease with the distance between their feature values. Intuitively, this is because the probability that two items will be assigned to the same component decreases with the distance between their feature values. In the extreme case, if two items are highly dissimilar, they will never be assigned to the same component by the model and there should be no dependency between the representations of those items. Conversely, if two items are highly similar and, thus, consistently assigned to the same component, there should be a high degree of dependency between their representations, the exact magnitude of which depends on other factors such as the precision of the component that they are both assigned to. In Section 2.7 below, we present experimental evidence supporting this prediction.

Figure 2.5 illustrates this prediction of the model with a simple example. The leftmost plot in Figure 2.5 shows three items with feature values \( \theta_1, \theta_2 \) and \( \theta_3 \) (vertical lines) and the marginal distributions of the estimates [i.e., \( p(\hat{\theta}_i|\theta_1, \theta_2, \theta_3) \)] (solid curves) computed over 1000 simulated presentations of the same set of feature values. The remaining plots show each of the three two-dimensional marginals of the estimates [i.e., \( p(\hat{\theta}_i, \hat{\theta}_j|\theta_1, \theta_2, \theta_3) \)]. (Note that the distributions shown here are different from the ones shown in Figure 2.4A. The distributions in Figure 2.4A depict \( p(\theta_i|x_1, x_2, x_3) \) for a specific set of noisy observations \( x_1, x_2, x_3 \). For the data plotted in Figure 2.5, noisy observations are integrated out. The relationship between the distributions depicted in Figures 2.4A and 2.5 is as follows: the posterior mean of the distribution depicted in Figure 2.4A would correspond to a single point in Figure 2.5 [also see Equation 2.2].) The two-dimensional marginals of the estimates in Figure 2.5 show that the estimates are correlated and the correlations decrease with the difference between the actual feature values of the items. The biases in
Figure 2.5: (Leftmost plot) Three items with feature values $\theta_1$, $\theta_2$ and $\theta_3$ (vertical lines) and the marginal distributions of the estimates, $p(\hat{\theta}_i|\theta_1, \theta_2, \theta_3)$ (solid curves), computed over 1000 simulated presentations of the same set of feature values. The remaining plots show each of the three two-dimensional marginals of the estimates, $p(\hat{\theta}_i, \hat{\theta}_j|\theta_1, \theta_2, \theta_3)$. The means of the marginals are represented by the circles and the actual feature values of the items are represented by the crosses. The numbers inside the plots are the correlation coefficients between the estimates of each pair of items. The estimates are also apparent in this figure. Note, for example, that the means of the two-dimensional marginals (represented by the circles) are closer to the diagonal than the actual feature values of the items (represented by the crosses), indicating that the estimates are biased toward the mean of the feature values.

**Multivariate extension:** We also consider a multivariate version of the DPMM presented above, where items are now defined not by a single feature dimension, but by multiple feature dimensions. In this case, the univariate normal components are replaced by multivariate normal components. In detail, the multivariate DPMM is
defined by the following equations (see Görür and Rasmussen, 2010):

\[
G \sim DP(G_0, \alpha) \tag{2.12}
\]
\[
G_0(\mu_i, \Sigma_i) = U(\mu_i; a, b)IW(\Sigma_i; \Psi, \kappa) \tag{2.13}
\]
\[
\mu_i, \Sigma_i | G \sim G \tag{2.14}
\]
\[
\theta_i | \mu_i, \Sigma_i \sim N(\theta_i; \mu_i, \Sigma_i) \tag{2.15}
\]
\[
x_i | \theta_i \sim N(x_i; \theta_i, \Sigma_{obs}) \tag{2.16}
\]

where \( \theta_i, \mu_i \) and \( x_i \) are now \( d \)-dimensional vectors, \( N(\theta_i; \mu_i, \Sigma_i) \) is a multivariate normal distribution with mean \( \mu_i \) and covariance matrix \( \Sigma_i \). \( \Sigma_{obs} \) is the common covariance matrix of the noisy observations, \( x_i \). We assume \( \Sigma_{obs} \) to be a diagonal matrix in accordance with recent findings that recall errors are largely independent across different stimulus dimensions in VSTM (Bays, Wu & Husain, 2011; Fougnie & Alvarez, 2011). The uniform base distribution for \( \mu_i \), \( U(\mu_i; a, b) \), is defined over a \( d \)-dimensional hypercube. Similar to the univariate case, we set the region over which the uniform base distribution for \( \mu_i \) is defined to a large volume that includes the minimum and maximum possible values of each component of \( \mu_i \). The base distribution for \( \Sigma_i \) is an inverse-Wishart distribution with inverse scale parameter \( \Psi \) and degrees-of-freedom parameter \( \kappa \). We place a vague inverse-Wishart prior on \( \Psi \) and treat the degrees-of-freedom parameter \( \kappa \) as a free parameter. The concentration parameter \( \alpha \) is given a \( G(1, 1) \) prior. As in the univariate case, posterior inference is performed via an MCMC algorithm with auxiliary variables (Algorithm 8 in Neal, 2000).
2.5.2 Bayesian finite mixture models (BFMMs)

In a finite mixture model, each \( \theta_i \) is assumed to be generated by one of \( K \) Gaussian components, where \( K \) is a fixed, finite positive integer. Formally, a Bayesian finite mixture of Gaussians is very similar to the DPMM introduced above (Equations 2.6-2.10). The only difference between the DPMM and a BFMM comes from the discrete distribution \( G \) over the component parameters (see Figure 2.3). In the DPMM, \( G \) is distributed according to a Dirichlet process with base distribution \( G_0 \) and concentration parameter \( \alpha \) and can be expressed as a weighted sum of an infinite number of discrete atoms, where atoms represent component parameters (see Equation 2.11). In a finite mixture model, on the other hand, \( G \) is a weighted sum of a finite number of atoms only (reflecting the assumption that the data were generated by a fixed, finite number of components):

\[
G(p) = \sum_{k=1}^{K} \pi_k \delta(p = p_k)
\]  

(2.17)

As in the DPMM, the atoms (i.e., the component parameters) are drawn independently from a base distribution \( G_0 \). The component weights \( \pi \), on the other hand, are drawn from a symmetric Dirichlet prior with concentration parameters \( \alpha/K \):

\[
\pi \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)
\]  

(2.18)

whereas the weights \( \pi \) in the DPMM are distributed according to what is known as a GEM (or stick-breaking) process with concentration parameter \( \alpha \) (\( \pi \sim \text{GEM}(\alpha) \)). The close similarity between the DPMM and the BFMM is not accidental. Indeed, it can be shown that the DPMM is mathematically equivalent to a BFMM in the
limit $K \rightarrow \infty$ (Rasmussen, 2000).

We use the same base distribution and hyper-priors for the BFMM as for the DPMM. Specifically, for the base distribution, we use $G_0(\mu_i, \tau_i) = U(\mu_i; a, b)G(\tau_i; \alpha_{\tau}, \beta_{\tau})$ (we put a $G(1, 1)$ prior over $\beta_{\tau}$ and set $\alpha_{\tau} = 1$). We put a $G(\alpha, 1)$ prior over the precision parameter $\alpha$ of the BFMM and treat $\alpha_c$ as a free parameter. Thus, the DPMM and BFMM have the same number of free parameters. As with the DPMM, it is also straightforward to extend the BFMM to multivariate components. In what follows, we consider BFMMs with $K = 2$ and $K = 4$ components.

BFMMs predict qualitatively similar biases and dependencies in VSTM as DPMMs, using essentially the same mechanisms. However, the quantitative details of the biases and dependencies predicted by a BFMM might depend on $K$. In general, for larger and larger $K$, the predictions of a BFMM will be more and more similar to the predictions of a DPMM. Indeed, a popular algorithm for performing efficient approximate inference in the DPMM truncates the infinite sum in Equation 2.11 at a finite but sufficiently large level, thus making the model identical to a BFMM (Ishwaran & James, 2001).

### 2.5.3 Hierarchical Bayesian model (HBM)

In the following sections, we also consider the two-level hierarchical Bayesian model (HBM) used in Brady and Alvarez (2011). The model assumes the following generative process for a single trial of a VSTM experiment (Brady & Alvarez, 2011):

\begin{align*}
\mu, \tau & \sim U(\mu; a, b)G(\tau; \alpha_{\tau}, \beta_{\tau}) \quad (2.19) \\
\theta_i | \mu, \tau & \sim N(\theta_i; \mu, \tau) \quad i = 1, \ldots, N \quad (2.20) \\
x_i | \theta_i & \sim N(x_i; \theta_i, \tau_{obs}) \quad i = 1, \ldots, N \quad (2.21)
\end{align*}
As in the DPMM and BFMMs, \( \{\theta_i\}_{i=1}^N \) (as well as the ensemble statistics \( \mu \) and \( \tau \)) are treated as latent or unobserved variables that the observer does not have access to. Instead, the observer only has access to noisy observations \( \{x_i\}_{i=1}^N \) each generated from a corresponding Gaussian distribution with mean \( \theta_i \) and some constant variance representing the memory noise (or the combined effect of sensory and memory noise). Given these noisy observations \( \{x_i\}_{i=1}^N \), the observer then infers the joint posterior distribution over \( \{\theta_i\}_{i=1}^N \). The group-level mean \( \mu \) is given a uniform prior over a sufficiently large range. To make the two-level HBM truly a special case of BFMM (with \( K = 1 \)), we use the same prior over the group-level precision \( \tau \) as in the BFMM, namely a \( \mathcal{G}(\alpha, \beta) \) prior where \( \beta \) is, in turn, given a \( \mathcal{G}(1, 1) \) prior and \( \alpha \) is set to 1. Brady and Alvarez (2011) use a different prior over \( \tau \) (they use a uniform prior over the group-level standard deviation \( 1/\sqrt{\tau} \)), but we found that this difference did not significantly affect the simulation results reported below. The individual memory precision \( \tau_{obs} \) is treated as the only free parameter of the model.

As the specification of the model in Equations 2.19-2.21 makes clear, the two-level HBM can be regarded as a special case of BFMMs where the number of components \( K \) is constrained to be 1 (see also Figure 2.3).

### 2.6 Simulations

This section studies the DPMM, two versions of the BFMM, and the HBM in the context of three experiments from the visual short-term recall and recognition memory literatures. Our focus will be on the DPMM. Although a BFMM with a sufficiently large \( K \) performs as well as a DPMM, we focus on the DPMM because it is an exact implementation of PCT and, as discussed at the end of this section, has conceptually
appealing properties not shared by BFMMs (e.g., not setting an \textit{a priori} limit on the number of clusters that a set of items can be grouped into). We model experimental results from two short-term recall tasks (Brady & Alvarez, 2011; Wilken & Ma, 2004) and a short-term recognition task (Viswanathan, Perl, Visscher, Kahana & Sekuler, 2010). We also quantitatively compare the fits of the DPMM with those of the BFMMs and the two-level HBM, using the Bayesian information criterion (BIC) measure (Schwartz, 1978). Bayesian model comparison depends on the calculation of marginal log-likelihood of the data under different models. BIC provides only an approximation to the marginal log-likelihood of the data under a given model. We opted for the BIC measure primarily due to computational considerations (it was relatively easy to compute the BIC values given the optimization procedure we adopted in our simulations; see Appendix A.2). Given the similarity of the structures of the models compared in this chapter (see Figure 2.3), it is difficult to see how BIC would unfairly favor one model over the others.

\subsection*{2.6.1 Biases in VSTM: Wilken and Ma (2004)}

As briefly mentioned before, Wilken and Ma (2004) found that subjects displayed systematic biases in their judgments in a VSTM experiment that used spatial frequency as the relevant feature. In each trial of their Experiment 9, subjects briefly viewed a number of Gabor stimuli with different spatial frequencies randomly drawn from 16 frequency values uniformly spaced between 4 and 8 cycles/degree. Different set sizes used in the experiment were $N = 2, 4, 6, 8$. After a delay interval, one of the $N$ Gabors, called the target Gabor, was cued, and subjects adjusted the frequency of a comparison Gabor using the arrow keys to indicate their estimate of the frequency of the target Gabor in the original display. Wilken and Ma (2004) found
that subjects tended to overestimate the spatial frequencies of low frequency Gabors, but tended to underestimate the spatial frequencies of high frequency Gabors (i.e., subjects showed a bias toward the mean spatial frequency in their judgments). These authors also showed that the magnitude of this bias depended on the set size with smaller set sizes leading to smaller biases (see Figure 8 in Wilken and Ma, 2004; also reproduced in Figure 2.6 here).

We sought to determine whether the DPMM could explain the biases observed by Wilken and Ma (2004). We first generated a dataset according to the procedure described above. For each simulated trial, we randomly selected \( N \) spatial frequency values from 16 frequencies uniformly spaced between 4 and 8 cycles/degree. We then generated noisy observations of each of the \( N \) items from Gaussian distributions with mean equal to the true spatial frequency of the item and precision \( \tau_{obs} \). For each set size \( N \), 4000 such trials were simulated. We then ran the univariate version of the DPMM on noisy observations from these simulated trials. In each trial, we used the mean of the marginal posterior over the target spatial frequency as the model's response: \( \hat{\theta}_t = E[\theta_t | \{x_i\}_{i=1}^N] \). We confirmed that using the posterior mode instead of the posterior mean yielded similar results.

To demonstrate that the ability of the DPMM to qualitatively explain the pattern of biases observed by Wilken and Ma (2004) does not critically depend on the optimization of the free parameters, we first arbitrarily set \( \alpha_c = 1 \) and \( \tau_{obs} = 1 \) for all set sizes and did not optimize these free parameters. Figure 2.6 illustrates the behavior of the DPMM with this fixed setting of the parameters. The DPMM was able to capture the two main qualitative patterns in the observed biases: a linear relationship between the bias and frequency of the target Gabor for all set sizes, and an increase in the magnitude of the bias with set size. We emphasize that the
Figure 2.6: (A) The observed biases (from Wilken and Ma, 2004) and biases predicted by a non-optimized DPMM with $\alpha_c = 1$ and $\tau_{obs} = 1$ for four different set sizes. Lighter colors represent larger set sizes. Error bars represent $\pm 1$ SEM across subjects. (B) Posterior distributions over the number of clusters inferred by the DPMM averaged over all trials for different set sizes.

DPMM was able to explain the latter phenomenon without having to use different parameter values for different set sizes.

**Model comparison:** For the purposes of model comparison, we calculated the maximum likelihood (ML) estimates of the free parameters of the models introduced in the previous section: the DPMM, BFMMs with $K = 2$ and $K = 4$ components, and the HBM. An alternative approach would be to put non-informative or vague priors over these parameters and perform Bayesian inference to compute their poste-
riors. Computational infeasibility prevented us from taking this approach. The free parameters of the models were optimized via simple grid searches to find the parameter values that maximized the log-likelihood given the mean observed biases of 15 subjects (details of the model evaluation and optimization procedures are provided in Appendix A.2). When fitting the DPMM and BFMMs to observed biases, $\tau_{obs}$ was allowed to vary across different set sizes, but $\alpha_c$ was fixed across different set sizes. Although fixing both free parameters across different set sizes produced biases that increased with set size (consistent with the biases observed in the experimental data), the differences between biases for different set sizes were less dramatic for the models than in the experimental data (see, for example, Figure 2.6 where $\tau_{obs}$ was fixed at 1 for all set sizes). Allowing $\tau_{obs}$ to vary across different set sizes helped the models achieve better fits to the observed biases. This is consistent with existing hypotheses about relationships between task demands and precision of representations. Wilken and Ma (2004) and Bays and Husain (2008) reported a monotonic decline with set size in the precision with which individual items can be encoded. When we fit the DPMM to data from Wilken and Ma (2004), we found that the best fits were obtained if we allowed $\tau_{obs}$ (which controls the precision of memory noise) to vary across set sizes such that the precision of memory noise monotonically decreased with set size. For the BFMMs and the HBM, the memory precision parameter $\tau_{obs}$ was allowed to vary across different set sizes in a similar manner. Model fits were compared using the BIC measure (see Appendix A.2 for details).

**Results:** Overall, all four models were able to capture the linear relationship between the bias and target frequency for all set sizes and the increase in the magnitude of the bias with set size. Table 2.1 documents the BIC values of the models relative to the BIC value of the DPMM. The two-level HBM was slightly favored over
Table 2.1: BIC values of the BFMMs ($K = 2$ and $K = 4$) and the two-level HBM relative to the BIC value of the DPMM on three previous studies that reported biases in VSTM. Negative values indicate better fits than the DPMM, positive values worse fits. Smaller values indicate better fits.

<table>
<thead>
<tr>
<th></th>
<th>BFMM ($K = 2$)</th>
<th>BFMM ($K = 4$)</th>
<th>HBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilken &amp; Ma (2004)</td>
<td>0.0951</td>
<td>0.0644</td>
<td>-2.8290</td>
</tr>
<tr>
<td>Viswanathan et al. (2010)</td>
<td>61.3027</td>
<td>1.3165</td>
<td>177.3836</td>
</tr>
<tr>
<td>Brady &amp; Alvarez (2011)</td>
<td>0.2810</td>
<td>0.1692</td>
<td>1.7824</td>
</tr>
</tbody>
</table>

the other models due to its smaller number of parameters. However, the differences between the BIC scores of different models were small. For the DPMM (as well as for BFMMs) the posterior distributions were dominated by partitions with small numbers of clusters (typically one or two clusters), suggesting that subjects tended to group items into a small number of clusters.

All four models account for the biases by assuming that subjects spontaneously encode a given display at multiple scales. However, an alternative explanation of the observed biases would be that subjects might simply have a bias toward reporting the overall mean of the range of presented frequencies, and that this bias increases with set size. We believe that this latter explanation is unlikely for two reasons. First, in a similar experiment, Huang and Sekuler (2010) teased apart the contributions to the observed biases of the overall mean of the frequencies presented to the subject on previous trials versus the frequencies presented on the current trial, and found that both make significant contributions to the observed biases, suggesting that the biases cannot be completely attributed to a general bias toward reporting the overall mean frequency. Second, as discussed below, in a carefully controlled experiment, Brady and Alvarez (2011) showed that subjects displayed a specific bias toward the mean size of the same-colored circles presented on the same trial as a target circle which, because of the way their experiment was designed, could not be attributed to
a general bias toward the overall mean size of the circles of a given color.

2.6.2 Inter-item similarity effect: Viswanathan et al. (2010)

As discussed above, Kahana and Sekuler (2002) showed that inter-item similarity between stimuli influences subjects’ judgments in a standard old/new recognition task. Assuming fixed probe-item similarities, they found that a smaller inter-item similarity (i.e., a less homogeneous set of stimuli) increases the likelihood that subjects will judge a probe to be an old or familiar item.

The inter-item similarity effect has been replicated in several other studies. Here, we consider a study by Viswanathan et al. (2010). The design of their experiment was, for our purposes, equivalent to the design of the experiments in Kahana and Sekuler (2002) described in a previous section (Section 2.1). On each trial, a subject viewed three Gabor gratings, referred to as study gratings, followed by a probe grating. The subject then judged whether the spatial frequency of the probe was ‘old’ (the same as the frequency of one of the study gratings) or ‘new’ (a novel frequency).

The experiment used both medium and high homogeneity conditions. Representative trials for these two conditions are schematically depicted in the left and middle plots of Figure 2.7. In this figure, the spatial frequencies (in just-noticeable-difference or JND units) of study gratings are represented by solid vertical lines (at 1, 4 and 8 JND in the medium homogeneity condition, and at 3, 4 and 8 JND in the high homogeneity condition), and the frequencies of probe gratings are represented by dashed vertical lines (at 2 JND in both conditions). In this figure, the probe is a ‘new’ item (or a lure) in both conditions. The experiment was designed so that the individual probe-study item similarities were identical in the two conditions, mean-
ing that the only difference between these conditions was the inter-item similarity of the study items, with the high homogeneity condition having a higher inter-item similarity than the medium homogeneity condition. The inter-item similarity effect refers to the finding that subjects had a significantly higher probability of responding ‘old’ in the medium homogeneity condition than in the high homogeneity condition (mean $P(\text{old}) = 0.69$ vs. mean $P(\text{old}) = 0.57$).

Since the task in Viswanathan et al. (2010) is an old/new recognition task, we cannot use the mean of the marginal posterior of the target item to simulate the model’s responses. Unlike other tasks considered in this section which are recall tasks, there is no single target item in an old/new recognition task. Consequently, we constructed a combined posterior density from the marginal posteriors over each $\theta_i$ by marginalizing over the indices of the items and assuming each item was equally likely: $p(\theta|\{x_i\}_{i=1}^N) \propto \sum_{j=1}^N p(\theta = \theta_j|\{x_i\}_{i=1}^N)$. This density can be roughly thought of as a nonparametric density estimate of the spatial frequencies presented on this trial based on the noisy observations of the frequencies, and it quantifies the posterior density of $\theta$ being the value of any one of the study items.

To model the experimental results, $p(\theta|\{x_i\}_{i=1}^N)$ needs to be transformed into a probability of responding ‘old’. We did this by mapping $p(\theta|\{x_i\}_{i=1}^N)$ between 0.2 and 0.8 so that the minimum probability of responding ‘old’ was 0.2 and the maximum probability of responding ‘old’ was 0.8. We did not map the probabilities between 0 and 1 because subjects tend to have relatively high probabilities of responding ‘old’ even for very dissimilar probes in these types of experiments, and relatively low probabilities of responding ‘old’ even for perfect matches between probes and study items. The specific values of 0.2 and 0.8 were chosen based on a similar previous
study by Kahana et al. (2007). Finally, the model’s response was drawn from a Bernoulli distribution with success probability equal to the probability of responding ‘old’. 1800 trials each of medium and high homogeneity conditions were simulated.

We first demonstrate the behavior of a non-optimized DPMM with the free parameters set to $\alpha_c = 1$ and $\tau_{obs} = 1$. Figure 2.7A shows the combined posterior densities $p(\theta|\{x_i\}_{i=1}^N)$ in a single representative trial of the medium and high homogeneity conditions (solid black and gray curves respectively). For purposes of illustration, the noisy observations were set to the actual feature values of the items in these examples (i.e., $x_i = \theta_i$). The non-optimized DPMM was able to reproduce the inter-item similarity effect without fitting its parameters to the observed data. This can be seen in Figure 2.7A by noting that the black curve, representing the combined posterior density in the medium homogeneity condition intersects the dashed vertical line at a higher point than the gray curve, representing the combined posterior density in the medium homogeneity condition. Figure 2.7B shows the probabilities of responding ‘old’ in the two conditions as a function of probe frequency, which were obtained simply by normalizing the combined posterior densities shown in Figure 2.7A between 0.2 and 0.8. Although the non-optimized DPMM with $\alpha_c = 1$ and $\tau_{obs} = 1$ explained the inter-item similarity effect, it did not provide an excellent quantitative fit to the experimentally observed probabilities of ‘old’ responses in the two conditions (Figure 2.7; $P(old) = 0.68$ vs. $P(old) = 0.49$ in the medium and high homogeneity conditions respectively, compared with the observed mean probabili-

---

1We tried several different ways of transforming $p(\theta|\{x_i\}_{i=1}^N)$ into a probability of responding ‘old’: using different values for the minimum and maximum probabilities of responding ‘old’, normalizing the combined posterior densities separately for the medium and high homogeneity conditions, as well as normalizing them together (i.e., using the same $\max(p(\theta|\{x_i\}_{i=1}^N))$ value in normalizing the combined posterior densities in both cases). Although these manipulations in general affected the model’s quantitative fit, the ability of the model to qualitatively explain the inter-item similarity effect, as well as the relative order of the quantitative fits of different models, were not sensitive to the specific choice of the transformation method.
ties of $P(old) = 0.69$ and $P(old) = 0.57$ for the respective conditions in the actual experiment).

Intuitively, the reason that the DPMM successfully accounts for the inter-item similarity effect is that the posterior distributions over the number of clusters have significant masses at one- and two-cluster partitions for both medium and high homogeneity conditions (see the bottom row in Figure 2.7A). In one-cluster partitions, all items are grouped into a single cluster, and in two-cluster partitions, the leftmost two items are typically grouped into a single cluster and the rightmost item is assigned to its own cluster. Relative to the medium homogeneity condition, the spatial frequencies of items in the high homogeneity condition have a lower variance. Therefore, a single cluster fit to the spatial frequencies of all items in a trial has a lower variance in the high homogeneity condition. Similarly, in two-cluster partitions, the cluster containing the leftmost two items has a lower variance in the high homogeneity condition. This, combined with the fact that the probe at $2 \ JND$ is closer to the means of these clusters in the medium than in the high homogeneity condition, makes the probe more similar to the study items in the medium homogeneity condition.

**Model comparison:** We applied the univariate version of the DPMM, BFMMs and the HBM to simulated trials of the medium and high homogeneity conditions, and computed the ML estimates of the free parameters by searching for the parameter values that maximized the log-likelihood given the experimentally observed probabilities of ‘old’ responses in the two conditions under each model (details of the optimization procedure are provided in Appendix A.2). Models fits were again compared using the BIC measure.

**Results:** BIC values of the models are given in Table 2.1. Qualitatively, all four models were successful at capturing the main inter-item similarity effect. The BIC
Figure 2.7: Predictions of a non-optimized DPMM with parameters set to $\alpha_c = 1$ and $\tau_{obs} = 1$. (A) Representative trials from the medium and high homogeneity conditions of Viswanathan et al. (2010). For purposes of illustration, the noisy observations were set to the actual spatial frequencies of the study gratings in the two conditions (i.e., $x_i = \theta_i$). These observations are indicated by solid vertical lines and the frequency of the lure probe at 2 JND is represented by the dashed line. The solid black and gray curves show the combined posterior densities $p(\theta|\{x_i\}_{i=1}^N)$ in the medium and high homogeneity conditions, respectively. Lower panel shows the posterior distributions over the number of clusters in the two conditions. (B) Probabilities of responding ‘old’ in the two conditions as a function of probe frequency. (C) The observed and predicted probabilities of responding ‘old’ to the probe at 2 JND in the two conditions. The model prediction was estimated over 1800 simulated trials.
values for the DPMM and the BFMM with 4 components were similar. For these models, the posterior distributions over the number of clusters were dominated by three-cluster partitions. However, one- and two-cluster partitions also had significant probabilities. The BIC values for the HBM and the BFMM with 2 components indicated a significantly worse fit for these models.

2.6.3 Encoding at multiple levels of abstraction: Brady and Alvarez (2011)

We now show that the multivariate version of the DPMM accounts for the results of both Experiments 1 and 2 in Brady and Alvarez (2011). Recall from our earlier discussion that in Experiment 1, subjects were presented with blue, red and green circles of different sizes. Subjects were instructed to ignore the green circles. After a brief delay, a comparison circle appeared at the location of a red or blue circle in the original display. Subjects’ task was to indicate the size of the original circle that was at that location, referred to as the target circle, by using the mouse to resize the comparison. It was found that the reported size of the target circle was biased toward the average size of the circles having the same color as the target. Experiment 2 was identical to Experiment 1 except that displays did not include green circles. Brady and Alvarez hypothesized that color is an irrelevant feature in this case, and thus subjects would not use a color-based encoding scheme. Consistent with this hypothesis, subjects did not show a bias toward the mean size of the same-colored circles in their size estimates. Instead, subjects’ estimates showed a bias toward the mean size of all circles in a display.

Brady and Alvarez (2011) accounted for their results using two-level (to account for the results of Experiment 2) and three-level (to account for the results of Experi-
ment 1) hierarchical models (see Equations 2.4-2.5 respectively). Our goal here is to show that the same DPMM explains the results of both experiments in a way that does not require the modeler to stipulate different numbers of levels of abstraction for the encoding of items in the two experiments.

In our simulations, we represented circles as points in a two-dimensional feature space defined by color and size. Following Brady and Alvarez (2011), we assumed that the removal of the green circles in Experiment 2 reduced the salience or weight of the color dimension, thereby shrinking distances along the color dimension in the two-dimensional space. This is illustrated in the top row of Figure 2.8A where blue and red circles are separated along the color dimension for Experiment 1 (color is a relevant feature) but not for Experiment 2 (color is an irrelevant feature). We used arbitrary numerical values for red and blue (red=125, blue=25 in Experiment 1; red=75, blue=75 in Experiment 2). Simulations with different numerical values confirmed that our results did not depend on the choice of specific numerical values for the colors. As long as the salience of the color dimension was sufficiently reduced in Experiment 2, we were always able to get similar results, albeit with different parameter values. Sizes of the blue and red circles in each simulated trial were generated in accordance with the procedures described in Brady and Alvarez (2011). Noisy observations of the sizes and colors of the circles were then generated from multivariate Gaussian distributions with mean equal to the true feature values of the circle and covariance matrix $\Sigma_{obs}$ which we assumed to be a diagonal matrix with equal variances along the diagonal, denoted by $\sigma_{obs}^2$.

The models were then applied to these noisy observations in each simulated trial.² For all models, in each simulated

²To control for potential artifacts, Brady and Alvarez (2011) designed their experiments to use matched pairs of trials. Subjects’ biases were computed based on their responses to these pairs. In our simulations, we followed the same procedures. For brevity, we do not explain these procedures here. The interested reader is referred to Brady and Alvarez (2011).
trial, the mean of the marginal posterior of the target item, \( \hat{\theta}_t = E[\theta_t | \{x_i\}_{i=1}^N] \), was computed and the size dimension of this estimate was taken to be the model’s response in that trial.

**Model comparison:** As in previous examples, we determined the ML estimates of the parameters of the four models using grid searches (see Appendix A.2 for details) and compared the model fits using the BIC measure. The multivariate DPMM and BFMMs have two free parameters, \( \kappa \) and \( \sigma_{obs}^2 \). The two-level HBM has a single free parameter, \( \sigma_{obs}^2 \).

**Results:** BIC values of the models are provided in Table 2.1. The DPMM and BFMMs were able to qualitatively capture the differing pattern of biases observed in Experiment 1 and Experiment 2 of Brady and Alvarez (2011). The DPMM and the BFMM with \( K = 4 \) components yielded similar BIC scores, whereas the BFMM with \( K = 2 \) components provided a slightly worse fit (as evidenced by its higher BIC score). This was because the BFMM with \( K = 2 \) components tended to overestimate the bias toward the mean size of the same-colored circles in Experiment 1, as it almost always favored two-cluster, color-based partitions where the red circles are assigned to one cluster and the blue circles to a separate cluster. The DPMM and the BFMM with 4 components, on the other hand, frequently favored more fine-grained partitions with a larger number of clusters, which had the effect of reducing the magnitude of biases predicted by these models. Predictably, the two-level HBM only generated a bias toward the mean size of all circles in a display and was unable to explain the color-based bias observed in Experiment 1. As discussed earlier, Brady & Alvarez (2011) used a three-level HBM to explain the bias toward the mean size of the same-colored circles observed in Experiment 1 and a two-level HBM to explain the bias toward the mean size of all circles observed in Experiment 2. Our results,
however, demonstrate that it is not necessary to postulate hierarchical models with
different numbers of levels of abstraction to account for the observed pattern of bi-
ases in the two experiments. We applied the same DPMM and BFMMs to both
experiments. Despite changes in experimental conditions across experiments, these
models succeeded at automatically determining the appropriate scales for represent-
ing stimuli in each experiment as well as the weights that should be allocated to each
scale.

The left and right columns of Figure 2.8A show simulation results for the optim-
mized DPMM for a representative trial from Experiment 1 and for the corresponding
trial from Experiment 2 (where the sizes of the circles remained the same and only
the numerical values associated with the colors changed). Also shown are the pos-
terior distributions over the number of clusters for these two trials. For Experiment
1 in which color is a salient dimension, color-based partitions where red circles are
grouped into a single cluster and blue circles into another cluster, are highly proba-
ble (note the high probability of two-cluster partitions in the posterior distribution),
thereby producing a bias toward the mean size of the same-colored circles. For
Experiment 2 in which color is no longer salient, color-based partitions are not prob-
able under the model. Single-cluster partitions where all stimuli are grouped into
a single cluster are highly probable, producing a bias toward the overall mean size
of all circles, consistent with the results from Experiment 2 of Brady and Alvarez
(2011). Although two-cluster partitions are also highly probable, most of the likely
two-cluster partitions are size-based, not color-based, partitions. Figure 2.8B shows
the posterior probabilities over the number of clusters averaged over all trials in the
two simulated experiments. In Experiment 1, as in the representative trial shown
in Figure 2.8A, two-cluster partitions are the most probable, and the overwhelming
majority of these partitions are color-based partitions where blue circles are grouped into one cluster and red circles are grouped into another. In Experiment 2, one-cluster partitions in which all items are grouped together has significant probability (unlike Experiment 1). Again, although two-cluster partitions are also highly probable, most of these two-cluster partitions are size-based. We emphasize that the same values of the parameters $\kappa$ and $\sigma^2_{obs}$ were used for both experiments. Therefore, the models applied to the two simulated experiments were identical.

To demonstrate that the ability of the DPMM to explain the specific pattern of biases observed in Brady and Alvarez (2011) does not critically depend on the optimization of the free parameters $\kappa$ and $\sigma^2_{obs}$ to fit the particular numerical values of these biases observed in their experiments, we varied each one of these two free parameters over a broad range of values while keeping the other parameter fixed and computed the model’s predicted biases in Experiment 1 and Experiment 2. Figure 2.8C-D shows the predicted biases in Experiment 1 and Experiment 2 as a function of $\kappa$ and $\sigma^2_{obs}$, respectively. Note that with the specific bias measure used in Brady and Alvarez (2011), a bias value of 1 indicates that there is no bias toward the mean size of the same-colored circles, whereas a bias value significantly greater than 1 indicates a bias toward the mean size of the same-colored circles. Figure 2.8C demonstrates that the ability of the DPMM to qualitatively explain the pattern of biases observed in Brady and Alvarez (2011) is not very sensitive to the choice of $\kappa$. Except for very large or very small $\kappa$ values, the model was able to capture the experimentally observed pattern of biases. The model preferred a larger number of clusters for larger $\kappa$ values and thus underestimated the bias in Experiment 1. Figure 2.8D shows the predicted biases as a function of $\sigma^2_{obs}$ while $\kappa$ was fixed at 6. For larger values of $\sigma^2_{obs}$, the predicted biases were more variable. For small $\sigma^2_{obs}$ values, on the other hand,
Figure 2.8: Simulation results for the optimized multivariate DPMM with $\kappa$ and $\sigma^2_{\text{obs}}$ set to their ML estimates ($\kappa = 6, \sigma^2_{\text{obs}} = 225$). (A) Left and right columns show simulation results for representative trials from Experiments 1 and 2, respectively. Top row shows the visual stimuli used in each trial in color-size space. Bottom row shows the posterior distributions over the number of clusters for the corresponding trials. (B) Posterior probabilities over the number of clusters averaged over all trials in the two simulated experiments. (C) Predicted biases for the DPMM in Experiment 1 (gray) and Experiment 2 (black) of Brady and Alvarez (2011) as a function of $\kappa$ and (D) $\sigma^2_{\text{obs}}$. Horizontal lines represent the observed mean biases of subjects in these experiments, and gray areas represent $\pm 1SEM$ around the mean biases. The dashed gray and black lines represent the mean predicted biases for the DPMM in Experiment 1 and in Experiment 2 respectively, and the error bars correspond to $\pm 1SEM$ around the mean predicted biases over 25 simulations of both experiments, with 50 simulated trials in each experiment. In (C), $\sigma^2_{\text{obs}}$ was fixed at 225 and $\kappa$ was varied, whereas in (D), $\kappa$ was fixed at 6 and $\sigma^2_{\text{obs}}$ was varied.
the model again tended to underestimate the bias in Experiment 1, as it relied more heavily on the noisy observations and less on the prior induced by the DPMM.

2.6.4 Discussion of simulation results

In summary, the simulation results provide evidence in favor of our proposed Probabilistic Clustering Theory (PCT). The DPMM—the exact implementation of PCT—provided good fits, both qualitatively and quantitatively, to the experimental data of Wilken and Ma (2004), Viswanathan et al. (2010) and Brady and Alvarez (2011). Unsurprisingly, the two-level HBM was unable to explain the main result of Brady and Alvarez (2011), and it also yielded significantly worse quantitative fits than other models to the results from Viswanathan et al. (2010). To account for the different pattern of results in their two experiments, Brady and Alvarez (2011) used HBMs with different numbers of levels for the two experiments (three levels for Experiment 1 and two levels for Experiment 2), where the number of levels for each model was specified in advance. In contrast, the DPMM and the BFMMs succeeded at accounting for the different pattern of results in the two experiments without any changes in the structure or parameters of the models across the experiments and without any specification of the appropriate levels of abstraction in advance.

The BFMM—an approximate implementation of PCT—with 4 components yielded similar results to the DPMM. The BFMM with 2 components yielded a significantly worse fit than the DPMM and the BFMM with 4 components in the simulation of the Viswanathan et al. (2010) study and it overestimated the magnitude of the bias toward the mean size of the same-colored circles in Experiment 1 of Brady and Alvarez (2011). This demonstrates the danger of setting an a priori limit on the number of components, and hence provides indirect support for a more flexible nonparametric
approach such as the DPMM that automatically determines the appropriate number of components in each specific case without assuming an a priori limit on this number. Because of people’s tendency to spontaneously group items in a display (Woodman, Vecera & Luck, 2003; also the studies reviewed in this section) and the relatively small set sizes used in VSTM experiments, a mixture model would typically not need more than 4-5 components to account for the contents of people’s memories for simple multi-item displays, thereby explaining the success of the BFMM with 4 components in modeling the data considered in this section. However, in some cases, subjects might not have a strong tendency to group items in VSTM (which corresponds to a large $\alpha_c$ in the DPMM). Requiring that the items be grouped into 4 components, as the BFMM with 4 components does, might lead to the prediction of biases and dependencies in the estimates of the feature values of items that do not fit such ‘non-grouping’ subjects’ responses well enough. In any case, it is conceptually more appropriate to avoid making unwarranted a priori assumptions about the maximum number of components that would be needed for modeling different subjects’ data in VSTM experiments (or equivalently about the grouping tendency of different subjects). Rather, it is preferable to automatically determine this from the data itself.

2.7 Experiments

As discussed above, our main goal in this chapter is to characterize the organization of VSTM in terms of the joint probability distribution of the estimates of the feature values of multiple items, $p(\hat{\theta}_i^N \mid \theta_i^N)$. This joint distribution allows for the possibility of rich and highly complex dependencies between the feature values
of the visual items and their estimates, as well as among the estimates themselves. It also replaces informal notions of dependence or independence in encoding multiple items that are prevalent in the VSTM literature with well-defined notions of statistical dependence or independence. In earlier sections, we showed how various behavioral phenomena in the VSTM literature constrain the range of appropriate forms for this joint distribution. We also proposed a specific theory of the organization of VSTM, Probabilistic Clustering Theory (PCT), which states that the joint distribution \( p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N) \) should be characterized in terms of a probability distribution over all possible clusterings or partitions of the items (see Section 2.5).

An advantage of this theory is that it postulates that VSTM represents items at multiple granularities or scales without the need for a pre-specified, fixed hierarchy. An implementation of the theory, the DPMM, was shown in the previous section to account for a number of findings in previous research on VSTM.

In this section, we take an empirical approach and try to determine the properties of the joint distribution \( p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N) \) directly. We report the results of three novel experiments specifically designed to uncover potential dependencies between the actual feature values of different items and their estimates, as well as dependencies among the estimates themselves. These novel tasks rely on the idea of probing subjects’ estimates of the feature values of all presented items in each trial, rather than probing their estimates of the feature value of a single target item, as is customary in standard VSTM tasks. We also compare subjects’ experimentally determined joint probability distributions with the joint distributions predicted by the models considered in this chapter (DPMM, BFMMs and the two-level HBM).
2.7.1 Experiment 1

We first designed a VSTM recall experiment where subjects were asked to remember the horizontal locations of a number of briefly presented squares. We then asked subjects to report their estimates of the horizontal locations of all presented squares. This contrasts with previous approaches where only the feature value of a single target item is probed in each trial. Over trials, this procedure allowed us to uncover potential dependencies between joint estimates of the feature values of different items.

Procedure: Subjects were seated 57 cm from a CRT monitor with a screen resolution of 1280×1024 pixels and a refresh rate of 85 Hz. Each trial began with the display of a fixation cross at a random location within an approximately 12° × 16° region of the screen for 1 second. In separate experiments, subjects were then presented with \( N = 2 \) (Experiment 1A) or \( N = 3 \) (Experiment 1B) colored squares (1.4° × 1.4°) on uniformly spaced dark and thin horizontal lines for 100 ms (see Figure 2.9). After a delay interval of 1 second (during which the horizontal lines remained visible, but not the squares), a probe screen was presented. Initially, the probe screen contained only the horizontal lines. Subjects used the computer mouse to indicate their estimate of the horizontal location of each of the colored squares presented on that trial. Unlike other VSTM experiments in which subjects are asked to report their estimate of the feature value of a single probed item from a display, our task required subjects to indicate the feature values of all displayed items. This procedure allowed us to study the dependencies between VSTM representations of all items. Subjects were allowed to adjust their location estimates as many times as they wished. When they were satisfied with their estimates, they proceeded to the next trial by pressing the space bar. Figure 2.9 shows the sequence of events on a single trial of the experiment.
Figure 2.9: The sequence of events in a single trial of Experiment 1B: (a) a small fixation cross is presented at a random location for 1 s; (b) the target configuration is flashed briefly (100 ms); (c) a delay interval of 1 s follows the target configuration; (d) the probe display (initially containing only the dark horizontal lines) remains on until the subject indicates the horizontal locations of all items using the computer mouse.
We used different combinations of horizontal locations: $\Theta = \{\theta_i\}_{i=1}^N$ for the squares. We call each combination $\Theta$ a particular ‘display configuration’. To cover a diverse range of possible display configurations, we first defined a regular grid in the $N$-dimensional configuration space (where each dimension represents the horizontal location of a different square), and then added a small amount of Gaussian jitter ($SD \approx 0.5^\circ$) to each $\theta_i$ in each grid node. In Experiment 1A ($N = 2$), the grid was $6 \times 6$, yielding a total of 36 different configurations. In Experiment 1B ($N = 3$), the grid was $3 \times 3 \times 3$, yielding a total of 27 different display configurations.

To uncover potential dependencies in $p(\{\hat{\theta}_i\}_{i=1}^N | \{\theta_i\}_{i=1}^N)$, we presented the same configuration $\Theta$ multiple times, and collected a subject’s estimates each time. The estimates of the feature values of all items collected in each presentation of a particular display configuration can be thought of as a single sample from the joint distribution $p(\{\hat{\theta}_i\}_{i=1}^N | \{\theta_i\}_{i=1}^N)$ for that particular configuration. In Experiment 1A, each of the 36 display configurations was presented 24 times (yielding a total of 864 trials) with the presentation of different configurations randomly interleaved during the course of the experiment. In Experiment 1B, each of the 27 display configurations was presented 26 times (yielding a total of 702 trials). All subjects participating in the same experiment saw the same set of display configurations. The correlation between $\hat{\theta}_i$ and $\hat{\theta}_j$ for each pair of items $i, j$ in each configuration was estimated by calculating the correlation coefficient between the subject’s estimates of $\theta_i$ and their estimates of $\theta_j$ over all presentations of that particular configuration.

**Participants:** 8 naive subjects participated in Experiment 1A, and 11 naive subjects participated in Experiment 1B. Subjects were undergraduate or graduate students at the University of Rochester. All subjects reported normal or corrected-to-normal vision, and they were compensated at a rate of $10 per hour for their time.
Subjects completed the experiment in two sessions.

**Model comparison:** For each subject, we determined the ML estimates of the free parameters of each of the four models (DPMM, BFMMs with 2 and 4 components, and the two-level HBM) by searching for the parameter values that maximized the subject’s trial-by-trial responses under the distribution $p(\hat{\theta}_i^N|\{\theta_i\}_{i=1}^N)$ induced by the model. In other words, each setting of the free parameters of a model yields a slightly different distribution $p(\hat{\theta}_i^N|\{\theta_i\}_{i=1}^N)$. Since these distribution do not have an analytic form, they were computed by sampling, as in previous cases in this chapter (see Equation 2.2). The likelihood of the subject’s trial-by-trial responses was then computed under these distributions (note that the distributions are dependent on the display configuration $\{\theta_i\}_{i=1}^N$). The parameter values maximizing this likelihood were chosen as the ML estimates of the parameters. The models were then compared using the BIC measure as usual (see Appendix A.2 for additional details).

**Results:** The joint distributions $p(\hat{\theta}_i^N|\{\theta_i\}_{i=1}^N)$ estimated from subjects’ responses displayed notable dependencies between estimates of different items. In particular, there were positive pairwise correlations between $\hat{\theta}_i$ and $\hat{\theta}_j$ for different items $i$ and $j$ and the magnitude of these correlations decreased with the distance $|\theta_i - \theta_j|$ between the actual feature values of the corresponding items.

The leftmost plot in Figure 2.10A shows the results for a representative subject (RD) in Experiment 1A. In this figure, the crosses represent the 36 stimulus configurations $\Theta = \{\theta_1, \theta_2\}$ presented to the subject, the dots represent the subject’s mean estimates for each configuration [i.e., the mean of the joint distribution $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2)$ for each configuration $(\theta_1, \theta_2)$] and the contours represent the shapes of bivariate Gaussian distributions fitted to the subject’s responses for each configuration (i.e., contours of bivariate Gaussian approximations to the joint distributions
\[ p(\hat{\theta}_1, \hat{\theta}_2 | \theta_1, \theta_2) \text{ for all configurations } (\theta_1, \theta_2) \]; red contours show cases where \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) were significantly and positively correlated, whereas blue contours show cases where \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) were significantly and negatively correlated. For stimulus configurations where the two items have similar horizontal locations, the subject’s responses tended to be correlated (note the predominance of red contours near the main diagonal representing \( \theta_1 = \theta_2 \)), whereas for configurations where the two items have dissimilar horizontal locations, this tendency was gradually reduced. This pattern was apparent in most of the subjects. Figure 2.10B (left) shows, for subject RD, the correlations between \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) as a function of the distance \( |\theta_1 - \theta_2| \) between the actual horizontal locations of the two items. In this plot, the 36 stimulus configurations \( (\theta_1, \theta_2) \) presented to the subject were divided into 6 equal-length bins based on \( |\theta_1 - \theta_2| \), and the mean correlation between \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) as well as the standard error of the mean were calculated for each bin. Figure 2.11A (left) shows the correlations as a function of \( |\theta_1 - \theta_2| \) for combined data from all 8 subjects in Experiment 1A. We also performed linear regressions of the correlation between \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) on the distance \( |\theta_1 - \theta_2| \) for each subject separately as well as for combined data from all subjects. For 7 of the 8 subjects in the experiment with \( N = 2 \) (as well as for combined data from all subjects), the linear regression was significant, i.e. the 95% confidence interval for the slope excluded zero \( (p < .05) \), and the slope was negative, suggesting that correlations decreased with the distance \( |\theta_1 - \theta_2| \).

Another notable pattern in subjects’ responses was the bias toward the mean horizontal locations for many of the configurations presented to them (this can be seen in Figure 2.10A by noting that the mean estimates of the subject are closer to the main diagonal than the actual stimulus configuration for many configurations). This bias is consistent with similar biases previously reported in the literature for
Figure 2.10: (A) Results from a representative subject (RD) in Experiment 1A (left). Crosses represent the 36 stimulus configurations presented to the subject, dots represent the subject’s mean estimates for each configuration and contours represent single-level contours of bivariate Gaussian distributions fitted to the subject’s responses for each configuration (red contours show cases where the subject’s estimates of the horizontal locations of the two presented items were significantly and positively correlated, whereas blue contours show cases where the subject’s estimates were significantly and negatively correlated). The middle and the right panels show the predictions of the best-fitting DPMM and the best-fitting HBM, respectively (colors and line styles same as in subject’s data). (B) Correlations between $\hat{\theta}_1$ and $\hat{\theta}_2$ as a function of the distance $|\theta_1 - \theta_2|$ for subject RD (left) and the corresponding predictions from the best-fitting DPMM (middle) and the best-fitting HBM (right). Error bars represent standard errors of the mean.
Figure 2.11: (A) Pairwise correlations between $\hat{\theta}_i$ and $\hat{\theta}_j$ as a function of the distance $|\theta_i - \theta_j|$ between the horizontal locations of pairs of items for combined data from all 8 subjects in Experiment 1A (left) and for combined data from all 11 subjects in Experiment 1B (right). (B) Standard deviations of $\hat{\theta}_i$ as a function of the standard deviation of the actual feature values of items for combined data from all 8 subjects in Experiment 1A (left) and for combined data from all 11 subjects in Experiment 1B (right). Also shown are the predictions of the best-fitting DPMMs and the best-fitting HBMs. Error bars represent standard errors of the mean.
other feature dimensions as reviewed earlier in this chapter (see Section 2.1.2).

Results for Experiment 1B \((N = 3)\) were qualitatively similar. There were again positive pairwise correlations between \(\hat{\theta}_i\) and \(\hat{\theta}_j\) for different items \(i\) and \(j\), and the magnitude of these correlations decreased with the distance between the actual horizontal locations of the corresponding items. Figure 2.11A (right) shows the pairwise correlations between \(\hat{\theta}_i\) and \(\hat{\theta}_j\) as a function of the distance between the actual horizontal locations of the two items for combined data from all 11 subjects in Experiment 1B. In this figure, all \(\{\theta_i, \theta_j\}\) pairs for all stimulus configurations \((27 \times 3 = 81\) pairs in total) were divided into 3 equal-length bins based on the distance \(|\theta_i - \theta_j|\) between the horizontal locations of items \(i\) and \(j\). As in the analysis of data from Experiment 1A, we also performed linear regressions of the correlation between \(\hat{\theta}_i\) and \(\hat{\theta}_j\) on the distance \(|\theta_i - \theta_j|\) for each subject separately as well as for combined data from all subjects. For all 11 subjects in the experiment with \(N = 3\) (as well as for combined data from all subjects), the linear regression was significant \((p < .05)\), and the slope was negative, suggesting that correlations decreased with the distance \(|\theta_i - \theta_j|\).

Overall, subjects exhibited a smaller number of significant \(\{\hat{\theta}_1, \hat{\theta}_3\}\) correlations than \(\{\hat{\theta}_1, \hat{\theta}_2\}\) or \(\{\hat{\theta}_2, \hat{\theta}_3\}\) correlations. This is probably due to the fact that the \(\{\theta_1, \theta_3\}\) pair had a larger vertical distance than the other pairs. We also carried out a similar analysis using the two-dimensional locations of the items (considering both horizontal and vertical locations of the items in computing the Euclidean distance between two items rather than considering only their horizontal locations). This analysis yielded qualitatively similar results. As in Experiment 1A, subjects also displayed a bias toward the mean horizontal locations in their estimates.

We also looked at the standard deviation of subjects' estimates, \(\text{Std}(\hat{\theta}_i)\), as a
function of the standard deviation of the actual feature values of items in a configuration, \( \text{Std}(\{\theta_i\}_{i=1}^N) \) (Figure 2.11B). This function had an inverse U shape both in Experiment 1A and in Experiment 1B. Subjects’ estimates had small standard deviations for configurations with homogeneous or low-variance feature values. \( \text{Std}(\hat{\theta}_i) \) increased for configurations with more heterogeneous or high-variance feature values, but decreased again for the most heterogeneous or highest-variance configurations.

We note that a similar increase in the variance of encoding for individual items with increased heterogeneity of the actual feature values of items has been recently reported in Sims, Jacobs and Knill (2012). The decrease in \( \text{Std}(\hat{\theta}_i) \) for the most heterogeneous configurations is attributable to edge effects, because the most heterogeneous configurations (i.e. configurations for which \( \text{Std}(\{\theta_i\}_{i=1}^N) \) has the highest value) are configurations where one of the items is close to the left or the right edge of the screen and the other item or items are close to the opposite edge. For these configurations, subjects’ estimates had low variance, because their estimates were constrained on one side by the edge of the screen.

Interestingly, the DPMM and, to a lesser extent, the other models were able to qualitatively explain this inverse U-shaped relationship between \( \text{Std}(\{\theta_i\}_{i=1}^N) \) and \( \text{Std}(\hat{\theta}_i) \). Intuitively, this can be understood as follows. By grouping items together, the DPMM reduces the variance of encoding for individual items, because information about an individual item can be gained from other items in the same group. This reduction in the variance of encoding for individual items is largest for the most homogeneous configurations, because the group has the lowest possible variance in such configurations, increasing the amount of information one can get about an individual item from other items in the same group. For more heterogeneous groups, one can get less information about an individual item from other items. For the
most heterogeneous groups, since we used a uniform distribution over the horizontal length of the screen as the base distribution over $\mu$, the edge effect mentioned above acts to reduce the variance in the model’s estimates. We note that the models were only fit to the subjects’ trial-by-trial responses and were not optimized to display this inverse U-shaped relationship between $\text{Std}(\{\theta_i\}_{i=1}^N)$ and $\text{Std}(\hat{\theta}_i)$.

The models overestimated the standard deviations of subjects’ estimates (Figure 2.11B). This is because subjects’ estimates displayed significant biases. To be able to explain these biases, $\tau_{\text{obs}}$ had to be sufficiently small. But such small $\tau_{\text{obs}}$ values led to feature estimates with higher variances than observed in the data. Larger $\tau_{\text{obs}}$ values led to relatively unbiased estimates and did not fit subjects’ estimates well enough. On the other hand, very small $\tau_{\text{obs}}$ values led to estimates with much higher variances than observed in the data and did not fit subjects’ estimates well either.

How did the models do in explaining the biases and dependencies in subjects’ estimates? Qualitatively, the models were able to capture the major aspects of the biases and dependencies observed in the empirical joint distributions $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$ determined from subjects’ responses. In particular, as shown in Figure 2.10A-B for the DPMM (middle column) and the HBM (right column) respectively, the models were able to generate correlated estimates for the horizontal locations of different items when their horizontal locations were very similar. However, a common failure of all four models was that they underestimated the magnitude of pairwise dependencies between estimates of different items. The correlations predicted by the models decayed faster than the actual correlations observed in subjects’ responses. Note that the models were fit to the trial-by-trial responses of the subjects and not directly to the observed pattern of correlations or biases.
The models were, in principle, capable of generating correlations that resembled the correlations observed in subjects’ responses (that is, correlations that decayed more slowly). However, this required using fewer clusters and/or noisier memory observations in the case of the DPMM (a smaller $\alpha_c$ and/or a smaller $\tau_{obs}$) or noisier memory observations in the case of the HBM (a smaller $\tau_{obs}$) than warranted by subjects’ trial-by-trial responses.

In a similar vein, the models also underestimated the magnitude of subjects’ biases. Again, to be able to generate biases that matched subjects’ biases in their magnitude, the models had to use noisier representations than warranted by subjects’ trial-by-trial responses.

In Experiment 1A ($N = 2$), the DPMM had the best BIC score for 6 out of 8 subjects and the HBM had the best BIC score for the remaining 2 subjects. The HBM tended to provide better fits for subjects with broader error distributions. However, the differences between the BIC scores of the four models were not very large and all four models made qualitatively similar predictions. In Experiment 1B ($N = 3$), the DPMM had the best BIC score for 7 out of 11 subjects, the BFMM with 4 components had the best BIC score for 2 subjects, and the HBM had the best BIC score for the remaining 2 subjects. Again, the differences in the BIC scores of these three models were not very large. In contrast, the BFMM with 2 components provided consistently worse fits than the other models. The reason for this is illustrated in Figure 2.12 which shows the posterior distributions over the number of clusters inferred by the DPMM and the BFMM with 2 and 4 components for two representative subjects in Experiments 1A and 1B respectively. The posterior distributions for the DPMM and the BFMM with 4 components give substantial probability to 3-cluster partitions (where each item is represented individually with no grouping). Because the BFMM
with 2 components could use at most 2 clusters, it was not able to account for these ‘non-grouping’ partitions, and consequently could not fit subjects’ responses well. A similar problem arises in the HBM, but in BIC calculations, the HBM benefits from having one less parameter than the other models.

**Discussion:** Our main contribution in Experiment 1 was that we designed a novel experimental task allowing us to empirically determine the form of the joint probability distribution \( p(\{\hat{\theta}_i\}_{i=1}^N | \{\theta_i\}_{i=1}^N) \) characterizing a subject’s estimates of the
feature values of multiple items based on VSTM. The basic idea in our task is to ask subjects to report their estimates of the feature values (horizontal locations) of all presented items in each trial, rather than asking them to report the feature value of a single target item as is customary in more standard VSTM tasks. The estimates of the feature values of multiple items collected in each trial can be thought of as a single sample from the joint distribution \( p(\{\hat{\theta}_i\}_{i=1}^N | \{\theta_i\}_{i=1}^N) \). Over several presentations of the same configuration, this procedure allows us to determine the dependencies between the estimates of the feature values of different items.

Using this novel paradigm with horizontal location as the relevant feature dimension, we found a hitherto unrecognized form of dependence between estimates of the feature values of different items, namely the existence of positive pairwise correlations between \( \hat{\theta}_i \) and \( \hat{\theta}_j \) for different items \( i, j \) that decrease with the distance \( |\theta_i - \theta_j| \) between the feature values of the corresponding items. In addition, consistent with previously reported biases in VSTM for spatial frequency (Wilken & Ma, 2004; Huang & Sekuler, 2010) and size (Brady & Alvarez, 2011), we found biases toward mean horizontal locations in subjects’ estimates.

Among the four models we consider in this chapter, the Bayesian finite mixture model with 2 components (BFMM-2) can be ruled out because it yielded poor fits in Experiment 1B due to its inability to account for 3-cluster partitions where each item is assigned to its own cluster. This shows the importance of not setting an \textit{a priori} limit on the number of components, and hence favors a flexible nonparametric approach such as the DPMM that automatically infers the granularity of groups or clusters that is appropriate for each subject in an experiment. In our experiments, we used at most 3 items, therefore in a sense, one does not need more than 3 components for modeling the representations of the items. Consequently, it could be argued
that the Bayesian finite mixture model with 4 components (BFMM-4) can be safely used in place of the nonparametric DPMM. However, in experiments that use larger set sizes, some subjects might not have a strong tendency to group items in VSTM. In such cases, requiring that the items be grouped into 4 components might lead to the overestimation of biases and/or dependencies in such ‘non-grouping’ subjects’ responses. Therefore, it would be safest to avoid making any a priori assumptions about the maximum number of components that would be needed to characterize different subjects’ behaviors in a VSTM experiment, or equivalently about the grouping tendency of different subjects. This is the approach taken by the nonparametric DPMM.

Although the remaining 3 models (DPMM, BFMM-4 and HBM) were able to generate pairwise correlations between $\hat{\theta}_i$ and $\hat{\theta}_j$ for different items $i$ and $j$ as observed in our subjects’ responses, the correlations predicted by these models decayed significantly faster than the observed correlations. Similarly, these models also underestimated the magnitude of the biases in subjects’ responses. These discrepancies suggest that further modeling efforts will be needed to better capture these aspects of the experimental data.

2.7.2 Experiment 2

A potential concern about Experiment 1 is that subjects were asked to indicate their responses using the computer mouse, and thus the experimental task had a significant motor component. Therefore, the correlations observed between the estimates of the horizontal locations of different items might have been caused by subjects’ hand movements rather than by their VSTM representations of the items. Consider, for instance, a hypothetical trial where two squares have similar horizontal locations.
Suppose that when the probe screen comes up, a subject first indicates his/her estimate $h$ of the horizontal location of the upper square using the computer mouse. To minimize his/her vertical hand movement, the subject might simply move the cursor with a straight downward movement and indicate a horizontal location very similar to $h$ for the lower square. This minimum vertical movement strategy would generate correlated estimates for the horizontal locations of different squares when their actual horizontal locations are similar, exactly as observed in Experiment 1.

To rule out this possibility, we designed a change-detection variant of Experiment 1 that minimizes the motor component of the task. The basic idea behind this change-detection task is schematically illustrated in Figure 2.13. Figure 2.13A shows a target configuration (i.e., a specific setting of the feature values of two items) represented by the black square and a number of ‘lure’ configurations represented by the black circles. The lure configurations are equally spaced along two axes referred to as the parallel axis (because it is parallel to the main diagonal) represented by the gray line, and the orthogonal axis (because it is orthogonal to the main diagonal) represented by the black line. In Figure 2.13A, the two items in the target configuration have similar feature values (the black square representing the target configuration is close to the main diagonal), and therefore the difference $\Delta = \theta_1 - \theta_2$ between the feature values of the two items in the target configuration is small. Figure 2.13B is similar to Figure 2.13A, except that the two items in the target configuration in Figure 2.13B have dissimilar feature values, and therefore $\Delta$ is large. The ellipses in Figure 2.13A-B indicate the predicted shapes of the distributions $p(\hat{\theta}_1, \hat{\theta}_2 | \theta_1, \theta_2)$ characterizing subjects’ estimates of the feature values of the two items in the target configurations (these can be compared to the contours in Figure 2.10). Results from Experiment 1 suggest that $\hat{\theta}_1$ and $\hat{\theta}_2$ are positively correlated when $\Delta = \theta_1 - \theta_2$ is small (as in
Figure 2.13A), and that this correlation decreases with Δ and ultimately vanishes when Δ becomes too large (as in Figure 2.13B).

If this is indeed the case, then when Δ is small, the lure configurations along the parallel axis should be judged to be more ‘similar’ to, and therefore be more confusible with, the target configuration than equidistant (using the Euclidean metric) lure configurations along the orthogonal axis. This is illustrated in Figure 2.13C where the (unnormalized) cross-sections of \( p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2) \) along the parallel and orthogonal axes are shown by gray and black lines, respectively. The lure configurations along the parallel axis have higher probability under \( p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2) \) than equidistant lure configurations along the orthogonal axis. Therefore, in a same/different change-detection task, when the target configuration is presented successively with a lure configuration, the lure configurations along the parallel axis should elicit a higher probability of ‘same’ responses than equidistant lure configurations along the orthogonal axis.

In contrast, when Δ is large, representations \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) should be less correlated, and ultimately uncorrelated as Δ becomes sufficiently large (as in Figure 2.13C). In this case, lure configurations along the parallel axis and equidistant configurations along the orthogonal axis should have more or less the same probability under \( p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2) \) as shown in Figure 2.13D. Thus, they should be equally confusible with the target configuration, and therefore elicit more or less the same proportion of ‘same’ responses when successively presented with the target configuration in a change-detection task. In Experiment 2, we tested these predictions using a simple change-detection task.

**Procedure:** Subjects were seated 57 cm from a CRT monitor with a screen resolution of 1280×1024 pixels and a refresh rate of 85 Hz. Subjects’ heads were
Figure 2.13: (A-B) Target and lure configurations are represented by open squares and circles, respectively. Lure configurations are defined along two perpendicular axes (parallel and orthogonal axes, represented by the gray and black lines, respectively). The contour lines represent the shapes of the joint distributions $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2)$ describing the contents of a subject’s memory of the target configurations. In (A), items in the target configuration have similar feature values (small $\Delta$), whereas in (B) they have different feature values (large $\Delta$). (C-D) Unnormalized cross-sections of the joint distributions along the parallel (gray line) and orthogonal axes (black line). If the joint distribution is positively correlated for small $\Delta$ values, but uncorrelated for large $\Delta$ values, then for small $\Delta$, lure configurations along the parallel axis should have higher probability under the joint distribution than equidistant lure configurations along the orthogonal axis (C). In contrast, when $\Delta$ is large, lure configurations along the parallel axis and equidistant configurations along the orthogonal axis should have more or less the same probability under the joint distribution (D).
stabilized with the help of a chin rest. The sequence of events in a single trial of the experiment is shown in Figure 2.14. In each trial, a small fixation cross was presented at the center of the display for 1 second. As in Experiment 1A, subjects were then presented with a target configuration consisting of two differently colored squares (1.4° × 1.4°) placed on two dark and thin horizontal lines (spaced approximately 10° apart) for 100 ms. Subjects were asked to remember the horizontal locations of the two squares in the target configuration. We used five different target configurations with ∆ = −5.6°, −2.8°, 0°, 2.8°, 5.6°. Target configurations were presented at random on each trial either completely on the right side or completely on the left side of the fixation cross (as in Bays and Husain, 2008). Therefore, the stimuli were always presented in the visual periphery. A small amount of jitter (uniform between −0.7° and 0.7°) was added to the horizontal locations of both squares in each trial so that the subjects never saw the exact same target configuration twice (the same jitter was used for both squares, so that the difference ∆ remained the same).

After a 1-second delay interval (during which only the horizontal lines remained visible), a probe configuration was presented which remained on the display until the subject responded. On half of the trials, the probe configuration was exactly the same as the target configuration. On the other half of the trials, the probe configuration was a lure configuration. The lure configurations were generated as schematically illustrated in Figure 2.13. Specifically, we first defined two orthogonal axes passing through the target configuration in the two-dimensional space of (θ₁, θ₂) where θ₁ is the horizontal location of the first square and θ₂ is the horizontal location of the second square. The parallel axis is defined to be the line passing through the target configuration that is parallel to the main diagonal (the line θ₁ = θ₂) and the orthogonal axis is the line passing through the target configuration that is or-
Figure 2.14: The sequence of events in a single trial of Experiment 2: (a) a small fixation cross is presented at the center of the screen for 1 s; (b) the target configuration is flashed briefly; (c) a delay interval of 1 s follows the target configuration; (d) the probe configuration is presented until the subject responds either ‘same’ or ‘different’ (in this example the probe configuration is the same as the target configuration).
thogonal to the main diagonal. Six lure configurations were generated on each of these two axes (these are represented by the open circles in Figure 2.13A-B and the target configuration is represented by the open square). On the parallel axis, the horizontal locations of the squares in the six lure configurations differed from the horizontal locations of the squares in the target configuration by $(-1.32^\circ, -1.32^\circ)$, $(-0.88^\circ, -0.88^\circ)$, $(-0.44^\circ, -0.44^\circ)$, $(0.44^\circ, 0.44^\circ)$, $(0.88^\circ, 0.88^\circ)$, or $(1.32^\circ, 1.32^\circ)$. As a matter of notation, these six lure configurations along the parallel axis are given the numeric labels -3, -2, -1, 1, 2, 3, respectively (the target configuration was given the label 0). On the orthogonal axis, the horizontal locations of the squares in the six lure configurations differed from the horizontal locations of the squares in the target configuration by $(-1.32^\circ, 1.32^\circ)$, $(-0.88^\circ, 0.88^\circ)$, $(-0.44^\circ, 0.44^\circ)$, $(0.44^\circ, -0.44^\circ)$, $(0.88^\circ, -0.88^\circ)$, or $(1.32^\circ, -1.32^\circ)$. Similarly, these six lure configurations along the orthogonal axis are given the numeric labels -3, -2, -1, 1, 2, 3 so that the lure configurations along the parallel and orthogonal axes with the same numeric label are equidistant to the target configuration.

Subjects were asked to remember the horizontal locations of the squares in the target configuration, and to detect any changes in these horizontal locations when the probe configuration came on. Subjects responded by pressing one of two designated keys: ‘f’ for ‘same’ (or no change) and ‘j’ for ‘different’ (or change). Subjects were given auditory feedback after each trial. In addition, written feedback was presented on the screen after every 30 trials. Each subject completed a total of 600 trials (120 trials for each of the five target configurations; half of these 120 trials were ‘same’ trials and the remaining 60 ‘lure’ trials were equally divided between the 12 lure configurations).

**Participants:** 20 naive subjects participated in the experiment. Subjects were
undergraduate students at the University of Rochester. All subjects reported normal or corrected-to-normal vision, and they were compensated at a rate of $10 per hour for their time. Subjects completed the experiment in a single session.

Results: The average accuracy in Experiment 2 was 68.7% correct. Figure 2.15 shows the probability of ‘same’ responses as a function of the position along the parallel/orthogonal axis [recall that ‘0’ denotes the target configuration itself (i.e., no-change trials), whereas the other numeric labels represent different lure configurations] for the five different target configurations with different $\Delta = \theta_1 - \theta_2$ values (the curves shown in the figure were called ‘mnemometric functions’ in Zhou, Kahana and Sekuler, 2004). When the magnitude of $\Delta$ was small, the proportion of ‘same’ responses to a lure configuration along the parallel axis was higher than the proportion of ‘same’ responses to an equidistant lure configuration along the orthogonal axis. This difference between the proportion of ‘same’ responses to lure configurations along the parallel axis versus the proportion of ‘same’ responses to equidistant lure configurations along the orthogonal axis was largest for the smallest $\Delta$ ($\Delta = 0^\circ$) and became smaller and smaller as the magnitude of $\Delta$ increased. As argued earlier in this subsection, this pattern of results is exactly as one would predict if VSTM representations of the two items in the target configurations were correlated and the correlations decreased with the distance between the horizontal locations of the items (compare, for instance, Figure 2.13 with Figure 2.15A).

To quantify these correlations and to be able to compare the results of Experiment 2 with the results of Experiment 1 more directly, we developed a Bayesian model of subjects’ same/different responses in the change-detection task of Experiment 2. This Bayesian model can be considered to be a two-dimensional generalization of the Bayesian psychometric functions introduced in Kuss, Jäkel and Wichmann.
Figure 2.15: Probability of ‘same’ responses as a function of the position along the parallel/orthogonal axis for the five different target configurations with different $\Delta$ values in Experiment 2 [location change detection] (A) and Experiment 3 [orientation change detection] (B). ‘0’ indicates the target configuration itself (i.e., no-change trials), whereas the other numeric labels represent different lure configurations along the parallel (solid line) or the orthogonal axis (dashed line). Error bars represent standard errors of the mean across subjects.
(2005). The model assumes a bivariate Gaussian distribution with mean \( \mu = [\mu_1, \mu_2] \) and covariance matrix \( \Sigma = [\sigma_1^2, \rho \sigma_1 \sigma_2; \rho \sigma_1 \sigma_2, \sigma_2^2] \) for the shape of the joint probability distribution \( p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2) \) for a given target configuration, \((\theta_1, \theta_2)\). In the one-dimensional case, this is similar to using a Gaussian cumulative distribution function for modeling psychometric functions. When a subject is presented with a probe configuration \((l_1, l_2)\), he/she makes a same/different response based on the probability of the probe configuration under \( p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2) \) [i.e., \( p(\hat{\theta}_1 = l_1, \hat{\theta}_2 = l_2|\theta_1, \theta_2) \)]. Intuitively, probe configurations with high probability under \( p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2) \) are likely to elicit more ‘same’ responses or, in other words, they are more likely to be confused with the target configuration. Mathematically, we handled the mapping from \( p(\hat{\theta}_1 = l_1, \hat{\theta}_2 = l_2|\theta_1, \theta_2) \) to probabilities of same/different responses as follows. We linearly mapped the probabilities of probe configurations relative to the maximum attainable probability (i.e., \( \hat{p}(\hat{\theta}_1 = l_1, \hat{\theta}_2 = l_2|\theta_1, \theta_2) = \frac{p(\hat{\theta}_1 = l_1, \hat{\theta}_2 = l_2|\theta_1, \theta_2)}{\max(p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2))} \); note that for a Gaussian distribution, this is equal to \( \exp(-0.5 \ast (l - \mu)^T \Sigma^{-1} (l - \mu)) \) which is the unnormalized exponential part of the definition of the Gaussian density) to go from a lower bound \( b_l \) which represents the minimum probability of ‘same’ responses to an upper bound \( b_u \) which represents the maximum probability of ‘same’ responses (we performed Bayesian inference over these variables based on individual subjects’ responses):

\[
p_l = (b_u - b_l) \ast \hat{p}(\hat{\theta}_1 = l_1, \hat{\theta}_2 = l_2|\theta_1, \theta_2) + b_l \tag{2.22}
\]

This is a reasonable choice because, as is evident from Figure 2.15, subjects do not respond ‘same’ with a probability of 1 even for the target configuration itself, and do not respond ‘same’ with a probability of 0 even for very dissimilar lure configurations. A ‘same’ response in a given trial is modeled as a Bernoulli distributed variable with a success probability of \( p_l \). Finally, all variables of interest in the model
(μ₁, μ₂, ρ, σ₁, σ₂, b₁, b₂) are given suitable priors and their posteriors were computed via MCMC sampling based on individual subjects’ same/different responses (additional details are provided in Appendix A.3). We note that this Bayesian analysis was performed for each subject separately.

We are specifically interested in the posterior distribution of the variable ρ for different target configurations because it represents the correlation coefficient of the underlying joint probability model \( p(\hat{\theta}_1, \hat{\theta}_2 | \theta_1, \theta_2) \) that characterizes a subject’s estimates of the feature values of the items in the target configuration. Given the posterior distribution of ρ, we computed its maximum a posteriori (MAP) estimate for each target configuration.

Figure 2.16B shows the means and the standard errors (across subjects) of the MAP estimates of ρ for the five target configurations. Confirming the qualitative observations we made from Figure 2.15A, it shows that for target configurations with a small |Δ|, ρ is high and it gradually decreases for target configurations with larger |Δ|. The linear regression of the MAP estimates of ρ on |Δ| was significant and the slope was negative, suggesting that ρ decreased with |Δ| (\( b = -0.08, t(98) = -6.55, p < .01 \)).

Figure 2.16A shows, for a representative subject in Experiment 2 (Subject MG), the means and shapes of the underlying distributions \( p(\hat{\theta}_1, \hat{\theta}_2 | \theta_1, \theta_2) \) for the five target configurations, with the parameters of these distributions set to their MAP estimates. The five target configurations used in Experiment 2 are represented by the black crosses.

Although estimates of the correlations and variances of the joint distributions obtained in Experiment 2 are roughly comparable to estimates of the corresponding variables in Experiment 1, biases toward the mean horizontal locations in Experiment
Figure 2.16: (A) Results for a representative subject (Subject MG): for the five target configurations used in Experiment 2, the means and shapes of the distributions $p(\hat{\theta}_1, \hat{\theta}_2 | \theta_1, \theta_2)$ with the parameters of these distributions set to their maximum a posteriori (MAP) estimates. The five target configurations used in Experiment 2 are represented by the black crosses. Red contours represent cases where the 95% credible interval for $\rho$ excludes 0. (B) The means and the standard errors (across subjects) of the MAP estimates of $\rho$ for the five target configurations in Experiment 2. (C-D) Similar to (A-B), but results are shown for Experiment 3. In (C), results are shown for a representative subject in Experiment 3 (Subject TH). For better visualization, a different scaling was chosen to create the ellipses in (C) than in (A).
2, although consistent with the direction of the biases in Experiment 1, were, in general, weaker than the biases obtained in Experiment 1. In this regard, it is interesting to note that most previous reports of biases toward mean feature values in VSTM used recall tasks rather than change-detection or recognition type tasks (Wilken & Ma, 2004; Huang & Sekuler, 2010; Brady & Alvarez, 2011). Change-detection experiments with more target configurations would be needed to determine whether recall tasks (as in Experiment 1) induce additional biases over and above those observed in change-detection tasks (as in Experiment 2).

Discussion: Using a change-detection task, Experiment 2 replicated the main experimental finding of Experiment 1, namely the existence of pairwise correlations between estimates of the horizontal locations of different items that decrease with the distance between the actual horizontal locations of the corresponding items. Because the change-detection task used in Experiment 2 has a minimal motor component, the results of Experiment 2 rule out possible motor explanations of the observed correlations. Furthermore, because subjects make a single decision in each trial of Experiment 2, these results also rule out possible explanations of the observed correlations in terms of sequential reporting of the estimates of the feature values of items (the recall task used in Experiment 1 involved sequential reporting).

2.7.3 Experiment 3

Experiments 1 and 2 used horizontal location as the relevant feature dimension to be remembered. An interesting question is to what extent the results obtained in these experiments generalize to other feature dimensions. In particular, do the specific pattern of correlations observed between the representations of different items in VSTM extend to feature dimensions other than horizontal location as the feature-
independent nature of our models would suggest? To address this question, Experiment 3 used orientations of Gabor gratings instead of horizontal locations of squares as the stimulus feature to be remembered. Otherwise, the experimental design was the same as in Experiment 2.

**Procedure:** Subjects were seated 57 cm from a CRT monitor with a screen resolution of 1280×1024 pixels and a refresh rate of 85 Hz. Subjects' heads were stabilized with the help of a chin rest. The sequence of events in a single trial of Experiment 3 is shown in Figure 2.17. In each trial, a small fixation cross was presented at the center of the display for 1 second. Subjects were then presented with a target configuration consisting of two oriented Gabor gratings for 100 ms. The standard deviation of the isotropic Gaussian envelope of the gratings was 0.8°. The spatial frequency of the gratings was 0.9 cycles/degree. The two gratings were vertically separated by a distance of about 6° (center-to-center) and their horizontal locations were the same. Subjects were asked to remember the orientations of the gratings in the target configuration. We used five different target configurations with \( \Delta = -15°, -7.5°, 0°, 7.5°, 15° \). Specifically, the target configurations used were (37.5°, 52.5°), (41.25°, 48.75°), (45°, 45°), (48.75°, 41.25°) and (37.5°, 52.5°). As in Experiment 2, a moderate amount of jitter (uniform between \(-30°\) and \(30°\)) was added to the orientations of both gratings in each trial so that the subjects never saw the exact same target configuration twice. Again, the same jitter was used for both gratings, so that the difference \( \Delta \) remained the same. On different trials, target configurations were presented either on the right side of the fixation cross or on the left side of the fixation cross, thus the stimuli were always viewed peripherally. The presentation side was determined randomly in each trial.

After a 1-second delay interval, a probe configuration was presented which re-
Figure 2.17: The sequence of events in a single trial of Experiment 3: (a) a small fixation cross is presented at the center of the screen for 1 s; (b) the target display, consisting of two oriented Gabor gratings, is flashed briefly; (c) a delay interval of 1 s follows the target configuration; (d) the probe display is presented until the subject responds either ‘same’ or ‘different’ (in this example the probe display is different from the target display).
mained on the screen until the subject responded. On half of the trials, the probe configuration was exactly the same as the target configuration. On the other half of the trials, the probe configuration was a lure configuration. The lure configurations were generated as in Experiment 2. For a given target configuration, six lure configurations each were generated on the parallel and orthogonal axes. On the parallel axis, orientations of the gratings in the six lure configurations differed from the orientations of the gratings in the target configuration by $(-18^\circ, -18^\circ)$, $(-12^\circ, -12^\circ)$, $(-6^\circ, -6^\circ)$, $(6^\circ, 6^\circ)$, $(12^\circ, 12^\circ)$, or $(18^\circ, 18^\circ)$. These six lure configurations along the parallel axis were given the numeric labels -3, -2, -1, 1, 2, 3, respectively (the target configuration itself was given the label 0). On the orthogonal axis, orientations of the gratings in the six lure configurations differed from the orientations of the gratings in the target configuration by $(-18^\circ, 18^\circ)$, $(-12^\circ, 12^\circ)$, $(-6^\circ, 6^\circ)$, $(6^\circ, -6^\circ)$, $(12^\circ, -12^\circ)$, or $(18^\circ, -18^\circ)$. Similarly, these six lure configurations along the orthogonal axis were also given the numeric labels -3, -2, -1, 1, 2, 3, respectively, so that the configurations along the parallel and orthogonal axes with the same numeric label are equidistant to the target configuration.

Subjects were asked to remember the orientations of the gratings in the target configuration and detect any changes in these orientations when the probe configuration came on. Subjects responded by pressing one of two designated keys: ‘f’ for ‘same’ (or no change) and ‘j’ for ‘different’ (or change). Subjects were given auditory feedback after each trial. In addition, written feedback was presented on the screen after every 30 trials. Each subject completed a total of 600 trials (120 trials for each of the four target configurations; half of these 120 trials were ‘same’ trials and the remaining 60 ‘lure’ trials were equally divided between the 12 lure configurations).

**Participants:** 18 naive subjects participated in the experiment. Subjects were
undergraduate students at the University of Rochester. All subjects reported normal or corrected-to-normal vision, and they were compensated at a rate of $10 per hour for their time. Subjects completed the experiment in a single session.

**Results:** The average accuracy of the subjects was 68.8% correct in Experiment 3. Analysis methods used here are the same as in Experiment 2 above. Figure 2.15B shows the probability of ‘same’ responses as a function of the position along the parallel/orthogonal axis for the five target configurations used in Experiment 3. Figure 2.16D shows the means and the standard errors (across subjects) of the MAP estimates of $\rho$ for the target configurations. As in Experiment 2, $\rho$ tended to be high for target configurations with a small $|\Delta|$, and gradually decreased for target configurations with larger $|\Delta|$. The linear regression of the MAP estimates of $\rho$ on $|\Delta|$ was significant and the slope was negative, suggesting that $\rho$ decreased with $|\Delta|$ ($b = -0.02$, $t(88) = -5.08$, $p < .01$).

Figure 2.16C shows, for a representative subject in Experiment 3 (Subject TH), the means and shapes of the underlying distributions $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2)$ for the five target configurations (represented by the black crosses), with the parameters of these distributions set to their MAP estimates. We did not find a consistent bias toward mean orientations in subjects’ responses. Note, for instance, for the particular subject shown in Figure 2.16C, the means of the underlying distributions (black and red dots) were closer to the main diagonal than the target configurations (black crosses) in only one out of four cases, the opposite pattern was apparent in the remaining three cases (note that the fifth configuration was on the main diagonal). This could either be due to the limited number of target configurations we tested or due to the particular feature dimension used in this study (i.e., orientation). Further experiments using a larger number of target configurations are needed to resolve this
issue.

**Discussion:** Experiment 3 confirmed the specific pattern of correlations observed in Experiments 1 and 2 for a different feature dimension, namely orientation. Correlations between the estimates of the orientations of different items in a target configuration tended to be high (and credibly different from 0) for small orientation differences, and gradually decreased for larger differences.

### 2.8 Discussion

There is now extensive experimental evidence suggesting that the content of a visual memory for even a simple display encoded in VSTM can be very complex. VSTM uses organizational processes that make the representation of an item dependent on the representations of other items as well as on the actual features of the displayed items (Jiang, Olson & Chun, 2000; Kahana & Sekuler, 2002; Huang & Sekuler, 2010; Brady & Alvarez, 2011; Brady, Konkle & Alvarez, 2011). In other words, the way we remember an individual item might depend not only on the actual properties of that item, but also on the properties of other items simultaneously presented with that item and on how we remember those other items as well.

To account for these dependencies, we proposed Probabilistic Clustering Theory (PCT), a theory of the organization of VSTM. PCT states that VSTM infers probability distributions over partitions or clusterings of visual items. Probabilistic clustering of items gives rise to dependencies in the joint representation of multiple items in VSTM. Representations of items belonging to the same cluster share parameters, and thus are dependent. Representations of items belonging to different clusters do not share parameters, and thus are independent. Importantly, VSTM
does not determine a single partition. Rather, it determines a probability distribution over all possible partitions. This property allows it to represent items at multiple granularities or scales simultaneously. Because PCT hypothesizes that VSTM makes use of multiple scales, it can account for experimental data that has previously motivated hierarchical models of VSTM. In fact, as we showed in the section on Models, a two-level hierarchical model proposed by Brady and Alvarez (2011) can be seen as a special case of PCT with a single cluster (i.e., where all items are always assigned to a single cluster). In the general case, however, PCT does not set any \textit{a priori} bounds on the number of clusters, but rather automatically infers the appropriate distribution over the number of clusters to use and the scales of those clusters from the feature values of a given set of items. Because PCT hypothesizes that VSTM automatically determines which particular scales to use, it overcomes many of the shortcomings of hierarchical models with pre-specified, fixed structures (e.g., see Section 2.2).

We explored several possible implementations of PCT: an exact implementation based on Dirichlet process mixture models (DPMMs), and approximate implementations based on Bayesian finite mixture models (BFMMs) with different numbers of components, including a model using a single component which we have shown to be equivalent to the two-level hierarchical model of Brady and Alvarez (2011). We consider the DPMM to be an exact implementation of PCT in the sense that it does not assume any \textit{a priori} limit on the number of components to be used and considers all possible partitions of the given set of items. In practice, however, DPMMs and BFMMs with a sufficiently large number of clusters make indistinguishable predictions. Because of the small set sizes used in VSTM experiments and subjects’ tendency to group items in a display (Woodman, Vecera & Luck, 2003; and the stud-
ies reviewed in this chapter), ‘sufficiently large’ can be as small as 4-5 components, as demonstrated by the quite good performance of the BFMM with 4 components in fitting the data we modeled in this chapter.

Through computer simulations, we demonstrated that PCT explains a number of biases and dependencies in VSTM representations previously reported in the literature. Through novel experiments, we evaluated a crucial prediction of PCT, namely that dependencies between estimates of the feature values of different items based on their VSTM representations should decrease with the distance between the actual feature values of the corresponding items. This prediction was qualitatively confirmed in a series of experiments specifically designed to measure such dependencies. However, quantitatively, the observed dependencies decayed more slowly than predicted by all models considered in this chapter, suggesting that further improvements in the models or better alternative models are needed to more accurately characterize dependencies in subjects’ estimates.

2.8.1 A novel form of dependence between representations of multiple items in VSTM

Section 2.7 above discussed the results of three novel experiments that we designed to empirically determine properties of the joint distribution, $p(\{\hat{\theta}_i\}_{i=1}^N | \{\theta_i\}_{i=1}^N)$, for small set sizes. This led to the discovery of a previously unrecognized form of dependence between estimates of the feature values of different items in VSTM: there were strong positive pairwise correlations between representations of different items when the feature values of the corresponding items were similar, and the magnitude of these correlations gradually decreased with the distance between their feature values. The existence of these correlations were confirmed for two different feature dimensions:
horizontal location (Experiments 1 and 2) and orientation (Experiment 3). Future studies will need to test for similar correlations using other feature dimensions. It is entirely possible that for some feature dimensions (for example, for more categorical features), correlations might not exist.

In a similar vein, Experiments 1 and 2 used a location change-detection task whereas Experiment 3 used an orientation change-detection task. It would be interesting to examine the interactions between the effects of spatial proximity and feature similarity (e.g., orientation similarity) when measuring response correlations by, for instance, parametrically changing both of these factors. For example, it is conceivable that for very large spatial distances between two gratings, correlations between the estimates of their orientations might be reduced or might vanish altogether.

Another important question concerns the dynamics or time-course of these correlations. Do they arise during encoding or maintenance? What would their temporal profile look like over the course of the delay period? For example, initially in the delay period, the representations of the items might be less correlated and correlations might gradually increase or, alternatively, correlations might originate during the encoding phase and might be more or less stable thereafter. These questions can be addressed by parametrically varying the presentation duration (Bays, Gorgoraptis, Wee, Marshall & Husain, 2011) and the length of the delay interval.

Finally, our experiments used small set sizes and considered only pairwise correlations between representations of different items in VSTM. For larger set sizes, and especially for more naturalistic stimuli, the joint distribution characterizing the estimates of the feature values of multiple items will almost certainly involve more complex, higher-order and more interesting dependencies than just pairwise correlations. The extension of our experimental methods for uncovering possible dependencies in
such cases in a feasible way would be very valuable.

2.8.2 Extension of our probabilistic framework to more naturalistic stimuli

How can our probabilistic framework be extended to modeling the organization of VSTM for more naturalistic stimuli? It may be that characterizing VSTM through the joint distribution $p(\{\hat{\theta}_i\}_{i=1}^N | \{\theta_i\}_{i=1}^N)$ would not be a good starting point when considering naturalistic stimuli because these stimuli often cannot be described in terms of a number of simple feature values. That is, they typically exhibit a much richer structure. In recent years, there has been considerable progress in probabilistic modeling of natural scenes (Fei-Fei & Perona, 2005; Sivic, Russell, Efros, Zisserman & Freeman, 2005; Sudderth, 2006) in computer vision. In particular, nonparametric hierarchical Bayesian models (such as extensions of DPMMs) of the structure of natural scenes have been successfully applied to challenging recognition and classification tasks (Sudderth, 2006). Assuming that these models provide a reasonably good description of the structure of natural scenes, one possible approach to modeling the content of a subject’s visual memory for a natural scene might be to introduce capacity constraints in such models, similar to the capacity constraints in the models discussed in this chapter, such as constraints on the precision of the observable nodes in such models, where the observable nodes might simply be noisy observations of the lowest level features, just as in the models we considered in this chapter (although these lowest level features might be more complex than the ones we used in this chapter). A more detailed discussion of this idea will be presented in Chapter 5. In future work, it would be interesting to compare the performances of suitably-constrained extensions of DPMMs with the performances of humans in
VSTM experiments using natural scenes.
3 Consequences of Model Mismatch in Visual Short-Term Memory

It is often claimed that humans have a limited capacity to encode and maintain information in visual short-term memory (VSTM). Evidence for this claim comes from laboratory experiments where subjects are briefly shown random visual displays containing simple objects and, after a brief delay interval, their memory for one of the objects is probed using a recall or recognition task. When the number of objects is large (e.g., 4 objects or more), subjects perform poorly in these tasks. This poor performance is usually attributed to a general resource limitation in VSTM. In different theoretical frameworks, the relevant “resource” is conceptualized as operating at a cognitive level, such as attentional resources (Bundesen, 1990; Bays & Husain, 2008) or the number of discrete slots holding information about individual items (Luck & Vogel, 1997; Rouder, Morey, Cowan, Zwilling, Morey & Pratte, 2008), or at a neural level, such as the number of neural spikes emitted by a population of neurons (Ma & Huang, 2009).

In many behavioral studies, however, a general resource limitation is not the only,
and often not even the major, factor limiting subjects' performances. Another key factor, much neglected in memory research in general, is what we call “model mismatch”. To illustrate the idea of model mismatch, consider the example of perceptual expertise. Perceptual experts can perform tasks in their domains of expertise more accurately and efficiently than novices (Palmeri, Wong & Gauthier, 2004; Curby, Glazek & Gauthier, 2009; Curby & Gauthier, 2009). The inferior performance of novices is not attributed to a general capacity limitation, because novices can be trained to be perceptual experts with sufficient exposure to stimuli from the relevant domain (Gauthier & Tarr, 1997; Gauthier, Williams, Tarr & Tanaka, 1998; Sowden, Davies & Roling, 2000). Rather, the difference between perceptual experts and novices is that experts have better models or representations of their objects of expertise than novices. That is, experts’ models more closely reflect the true structure of the objects (Gauthier & Tarr, 1997; Palmeri, Wong & Gauthier, 2004). Using less accurate models of the same objects, as novices do, results in worse performance. We call this type of mismatch between the true model of the stimuli used in an experiment and a subject’s (usually imperfect) model of the same stimuli “model mismatch”.

More pertinent to our main interest in this chapter—accounting for performance limitations observed in standard VSTM tasks—model mismatch can also occur when a subject’s mental representations are based on statistical regularities of natural visual environments, but an experimental task uses visual stimuli whose statistical properties deviate significantly from those of natural visual environments. Again, this mismatch means that the subject has a suboptimal, or imperfect, model of the stimuli used in the experimental task (Ma, 2012). Model mismatch causes a subject to perform approximate, as opposed to exact, inference. That is, it leads to suboptimal
statistical inference and thus to suboptimal performance on the experimental task.

The role of model mismatch, and of suboptimal inference in general, in behavioral performance has been emphasized recently from a neuroscientific perspective by Beck, Ma, Pitkow, Latham and Pouget (2012). These authors note the tendency in systems neuroscience to consider observers—either animal or human subjects or ideal observers—as having perfect knowledge of the experimental task (including perfect knowledge of the statistics of the stimuli used in the task), and thus to attribute behavioral imperfections to neural variability. Beck et al. (2012) argue, however, that having a suboptimal model of the task is likely to be the major contributor to behavioral variability, especially for complex tasks, as independent neural variability can always be averaged out (especially when considering a large number of neurons), but the deleterious effects of using a suboptimal model cannot be easily overcome. In fact, as Beck et al. (2012) show, with a suboptimal model of the task it is possible to get poor performance with significant behavioral variability without any neural variability.

Paralleling the arguments of Beck et al. (2012), we claim that cognitive science suffers from the same tendency as systems neuroscience. Neural variability is closely related to the concept of internal noise as used in psychophysics and cognitive science (Ma, 2010), such as in models based on signal detection theory and in ideal observer models of perceptual, motor, and cognitive tasks, including VSTM tasks (Wilken & Ma, 2004; Ma & Huang, 2009). In these models, there is a similar tendency to assume that an observer has perfect knowledge of the experimental task, and thus behavioral variability is due to internal noise. We believe, however, that internal noise is only one factor leading to performance limitations and behavioral variability. The use of suboptimal models is another, and in some cases perhaps more significant, factor.
We consider this general claim in the specific domain of visual short-term memory. The basic idea is that model mismatch occurs in many, perhaps most, VSTM experiments, because subjects have poor models of the stimulus statistics used in these experiments and they have a limited ability to adapt to these stimulus statistics during the course of an experiment. Whereas these experiments typically use random displays with features of different stimuli drawn independently from a broad range, the human visual system is adapted to represent and process very different stimulus statistics, particularly the statistics of natural visual environments with rich spatial and temporal dependencies between features of objects.

The visual system is indeed exquisitely adapted to the statistical regularities of the natural environment it inhabits (Barlow, 1961; Field, 1987; Simoncelli & Olshausen, 2001). In psychophysical tasks, this adaptation manifests itself in superior detection and discrimination of stimuli that are more likely under the statistics of the natural world (Knill, Field & Kersten, 1990; Parragha, Troschianko & Tolhurst, 2000; Stocker & Simoncelli, 2006; Yuille, Fang, Schrater & Kersten, 2004; Girshick, Landy & Simoncelli, 2011). It should be emphasized that adaptation to statistical regularities in the natural environment is not restricted to adaptation to statistical regularities of low-level features. Adaptation to high-level statistical regularities between parts and mid-level features of visual object categories, and between objects in natural scene categories, also show their effects in standard psychophysical tasks (Biederman, 1972; Bar & Ullman, 1996; Bar, 2004). Researchers have argued that neural mechanisms enable the visual cortex to efficiently represent more likely stimuli in the visual input (Dan, Atick & Reid, 1996; Schwartz & Simoncelli, 2001; Zetzsche & Rohrbein, 2001), and to acquire hierarchical representations of its input for the purpose of capturing the statistical regularities in the natural visual environment.
(Riesenhuber & Poggio, 1999). Recent neuroimaging evidence suggests that visual information in VSTM is maintained in the same visual cortical areas where perceptual encoding takes place (Harrison & Tong, 2009; Serences, Ester, Vogel & Awh, 2009). Given that these neural areas are adapted to the statistics of the natural visual input, it would be sensible to expect superior VSTM performance for stimuli more likely under the statistics of the natural world. However, most VSTM studies use ecologically unrealistic stimulus statistics where, for example, a set of simple items with randomly sampled features (e.g., Gabor patches with random orientations), are briefly presented in each trial. Configurations of arbitrary orientations are highly improbable in natural visual environments, and thus it is unrealistic to expect our brains to encode and remember such stimuli efficiently.

It is ultimately an empirical question to what extent performance limitations in VSTM tasks are caused by a general resource limitation versus model mismatch. To avoid prejudging this issue, we consider two hypotheses regarding the potential roles of model mismatch and capacity limitations in VSTM tasks. The stronger hypothesis (elaborated in Section 3.3) states that performance limitations in VSTM tasks can be accounted for in terms of model mismatch alone without having to postulate a general resource or capacity limitation. Surprisingly, the results in Section 3.3 show that model mismatch alone can account for some of the main qualitative characteristics of performance limitations observed in VSTM tasks, including: (i) the decline in memory precision with increasing set size; (ii) variability in memory precision across items and trials; and (iii) different set-size dependencies for initial encoding rate and asymptotic precision when the duration of image presentation is varied. However, we also discuss some possible discrepancies between characteristics of performance limitations observed in VSTM tasks and the results presented in Section 3.3. This
leads us to formulate a weaker hypothesis stating that model mismatch and a general resource or capacity limitation together account for performance limitations in VSTM tasks (Section 3.4).

Before proceeding further, it is worthwhile for us to clearly state our main claim. We do not argue that VSTM does not have capacity limitations. Indeed, as mentioned in the previous paragraph, in a later section (Section 3.4), we review and discuss some evidence against the strong hypothesis claiming that performance limitations in VSTM tasks can be explained in terms of model mismatch alone. Rather, our main goal in this chapter is to highlight the potential importance of a hitherto neglected source of performance limitations in VSTM tasks: model mismatch. Specifically, we argue that characteristics of performance limitations in VSTM tasks generally attributed to a capacity limitation might instead, at least in part, be explained in terms of model mismatch.

Our results have at least two important implications for the study of VSTM. First, most memory researchers assume that the characteristics of performance limitations in VSTM tasks discussed in this chapter reveal important insights into the intrinsic properties of the VSTM system and its capacity limitations. Researchers sometimes propose elaborate theoretical and neural models to account for these characteristics. Our results, however, suggest that inference from properties of performance limitations in VSTM tasks to intrinsic properties of the VSTM system is not straightforward. Properties of performance limitations in VSTM tasks might, at least in part, be natural consequences of using ecologically unrealistic stimulus statistics in VSTM experiments (i.e., consequences of evaluating the VSTM system in stimulus regimes in which it is not adapted to operate efficiently). Second, the results here suggest that these properties critically depend on the stimulus statistics
used in VSTM experiments. Researchers would get very different results if they used different stimulus statistics. For example, our results suggest that the decline in recall precision with set size, or variability in encoding precision, might be much less dramatic or, in the case of precision-set size relationships, might even be reversed, if more natural stimulus statistics were used in VSTM studies.

### 3.1 General framework

This section introduces our probabilistic framework. Later sections use this framework to demonstrate how behavior that is usually interpreted as evidence for a capacity limitation in VSTM can also be explained, at least in part, as a natural consequence of the mismatch between the statistics of visual features in the natural world and the statistics of visual features used in most VSTM experiments.

Consider a single trial of a standard VSTM recall task in which a subject is briefly shown a display containing \( N \) simple items with feature values \( s = [s_1, \ldots, s_N] \). After a delay interval, one of the items, called the target item, is cued and the subject is asked to recall the feature value of the target. In each trial, the feature values of the items, \( s \), are randomly sampled from a distribution \( p(s) \) determined by the experimenter. In standard VSTM experiments, feature values of different items are generally drawn independently from a uniform distribution over a fixed interval. This corresponds to \( p(s) = p(s_1)p(s_2)\cdots p(s_N) \) where each \( p(s_i) \) is a uniform distribution over a given range.

The subject is assumed to have access only to noisy measurements or observations, denoted \( x = [x_1, \ldots, x_N] \), of the actual feature values of the items. This observation process can be characterized by the conditional distribution \( p(x|s) \). The optimal
strategy for the subject is to combine the actual prior \( p(s) \) and the likelihood \( p(x|s) \) to compute the marginal posterior over the feature value of the target, \( s_t \), given the noisy observations:

\[
p(s_t|x) \propto \int p(x|s)p(s)ds_{-t} \tag{3.1}
\]

where \( s_{-t} \) denotes the feature values of all items except the target item. The subject can generate a point estimate of the target’s feature value using this marginal posterior (e.g., use the posterior mean if the mean squared error is to be minimized). However, our framework allows for the possibility that the subject might use a prior, denoted \( q(s) \), that is different than the one used by the experimenter to generate the stimuli, thus making responses based on an approximate posterior, \( q(s_t|x) \), rather than on the true posterior:

\[
q(s_t|x) \propto \int p(x|s)q(s)ds_{-t}. \tag{3.2}
\]

We emphasize that this does not mean that the subject is unaware of the prior used by the experimenter. If, for example, the actual feature values of the items are independently drawn from a uniform distribution, the subject may be aware of this fact. The subject may even be explicitly told by the experimenter how the stimuli are generated. We assume that conscious awareness of stimulus statistics is not a reliable indicator of the particular prior used by the subject in a given experimental task. Rather, the subject’s prior \( q(s) \) is intended to characterize the visual system’s \textit{a priori} expectations about configurations of features [similar to subjects’ prior distributions as discussed in Stocker and Simoncelli (2006), and Girshick, Landy and Simoncelli (2011)]. In general, the probability of a particular \( s \) under \( q(s) \) will reflect the statistical regularities of visual features in the natural environment due to the
adaptation of the visual system to stimulus statistics in the natural environment.

In addition to the prior distribution $q(s)$, the noise distribution $p(x|s)$ will also be adapted to the stimulus statistics in the natural environment. For example, stimuli $s$ that are more likely in the natural environment might be encoded with better precision, i.e. with less noisy measurements $x$, under the noise distribution (Wainwright, 1999; Stocker & Simoncelli, 2006; Ganguli & Simoncelli, 2010; Girshick, Landy & Simoncelli, 2011). In many standard detection or discrimination tasks, subjects perform better when stimulus features are consistent with natural scene statistics (i.e., when $s$ is more likely under $q(s)$; see Girshick, Landy & Simoncelli, 2011; Knill, Field & Kersten, 1990; Parragha, Troschianko & Tolhurst, 2000; Stocker & Simoncelli, 2006; Yuille, Fang, Schrater & Kersten, 2004). This, in general, reflects the adaptation of both the prior and the noise distribution to natural scene statistics. Stocker and Simoncelli (2006) and Girshick, Landy and Simoncelli (2011), for instance, found that regions in the stimulus space that are more likely under the subject’s prior $q(s)$ are also encoded with better precision under the noise distribution. Wei and Stocker (2012) show that this pattern is consistent with the efficient coding hypothesis (Barlow, 1961) under some plausible assumptions.

The adaptation of the noise distribution to natural stimulus statistics can be modeled by assuming a stimulus-dependent variance $\sigma^2(s)$ for the noise distribution, with $s$ that are more likely under $q(s)$ having a smaller noise variance. This possibility will be examined in more detail in Section 3.5 below. However, Sections 3.3 and 3.4, will assume a noise distribution with stimulus-independent covariance, focusing exclusively on the effects of the adaptation of the subject’s prior $q(s)$ to stimulus statistics in the natural environment.

To give a more concrete example of a possible mismatch between the subject’s
prior and the experimenter’s prior, consider, for instance, the case where \( s \) represents the orientations of items. The experimenter’s prior \( p(s) \) might be a product of uniform distributions \( p(s_i) \), giving all combinations of orientations equal probability. However, all combinations of orientations are not equally likely in the natural world. In fact, only a very small subset of all possible combinations of orientations have significant probability under the statistics of the natural world: for example, parallel orientations or combinations of orientations that form a smooth contour (Sigman, Cecchi, Gilbert & Magnasco, 2001; Geisler, Perry, Super & Gallogly, 2001). Other combinations of orientations are exceedingly improbable (see Figure 3.1). Through experience-dependent neural plasticity, a person’s visual system is adapted to the statistics of the natural environment. If specific combinations of orientations or other features are highly probable in the natural environment, the person’s visual cortex will allocate more resources to representing those combinations, thereby becoming better at encoding those combinations, whereas random (and thus less likely) combinations of orientations are encoded with greater difficulty or less efficiency.

Unless explicitly noted otherwise, the following procedure was employed in all simulations reported below. In each simulation, we simulated \( 10^5 \) trials of a hypothetical VSTM recall experiment. In each trial, we first draw \( N \) random stimulus values, \( s = [s_1, \ldots, s_N] \), from the experimenter’s prior distribution \( p(s) \). In different sets of simulations, the experimenter’s prior distribution is assumed to be either a

---

\(^1\)We do not require that \( q(s) \) exactly match the statistics of \( s \) in the natural environment. There may be deviations between \( q(s) \) and the actual statistics of \( s \) in the natural environment. For example, a particular \( s \) may have a high probability under \( q(s) \) for reasons other than being common in the natural environment; for example, accurately detecting that particular \( s \) may be associated with a high reward or survival value. Some researchers have even put forward theoretical arguments against trying to learn an exact model of the environment (Feldman, 2013). In general, however, \( q(s) \) will reflect, more or less faithfully, the statistics of \( s \) in the natural environment, because it is difficult to imagine how an organism would survive with a model of the environment that has little resemblance to the properties of the actual environment (Geisler & Diehl, 2002).
Figure 3.1: Illustration of mismatches that can arise between the experimenter’s prior $p(s)$ (line orientations are independently sampled from uniform distributions) and the subject’s prior $q(s)$ (line orientations are sampled from a joint distribution characterizing orientations in natural visual environments). The configurations of orientations shown in panels A, B and C have equal probability under $p(s)$, but the configurations in B and C are more probable than the configuration in A under $q(s)$ due to dependencies between the orientations of different line segments in natural images.

multivariate Gaussian distribution with a diagonal covariance matrix and equal variance along each dimension (the standard deviation along each dimension was set to 2 in most simulations) or a multivariate Gaussian with uniform correlations, denoted $\rho_p$, between each pair of its dimensions. In each simulated trial, one of the stimuli, $s_t$, is randomly chosen as the target item. Given the actual values of the stimuli, $s$, their noisy measurements, $x$, were generated from the noise distribution $p(x|s)$. In Sections 3.3 and 3.4, the noise distribution is assumed to be a stimulus-independent multivariate Gaussian with a diagonal covariance matrix and equal variance along each dimension. In different sets of simulations, the covariance matrix of the noise distribution was either set size independent in which case the standard deviation along each dimension was set to 0.5 (see Section 3.3), or it was dependent on the set size in which case the standard deviation along each dimension was $\lambda_{\text{min}} \sqrt{N}$, where $\lambda_{\text{min}} = 0.5$ and $N$ is the set size (see Section 3.4). In Section 3.5, on the other hand, we investigate the consequences of adaptation in the noise distribution to natural scene statistics by assuming a stimulus-dependent covariance matrix for $p(x|s)$ with a smaller noise variance $\sigma^2(s)$ for stimuli that are more likely under $q(s)$. 
Given the noisy measurements, the subject is assumed to compute an approximate marginal posterior over the target stimulus value, using an approximate prior $q(s)$ rather than the true prior $p(s)$ (Equation 3.2). We consider different choices for $q(s)$ in the next section, each motivated by statistical properties of visual features in the natural environment. The mean of the approximate marginal posterior is then taken to be the subject’s estimate of the target stimulus value, $\hat{s}_t$, in that trial. In Section 3.5, the use of a stimulus-dependent noise distribution makes the computation of the joint and marginal posteriors in Equation 3.2 intractable, thus, to be able to simulate a sufficiently large number of trials, we employ a Laplace approximation for the joint posterior in this case.

Over trials, this procedure generates an error distribution. The inverse standard deviation ($1/\sigma$) of this error distribution is commonly used as a measure of subjects’ performances in many VSTM studies. Unfortunately, this quantity is sometimes called recall “precision” in the VSTM literature (Bays & Husain, 2008; Bays, Gorgoraptis, Wee, Marshall & Husain, 2011), contrasting with the more standard use of the term “precision” in the statistics literature to refer to inverse variance ($1/\sigma^2$). Some VSTM studies use “precision” in the sense of inverse variance (Van den Berg, Shin, Chou, George & Ma, 2012). In what follows, to be consistent with these earlier studies, we use the term “precision” in both senses, but we are explicit about what sense we are intending. Additional details about the simulations can be found in Appendix B.2.
3.2 Models for the subject’s prior $q(s)$

This section introduces the models we use for the subject’s prior distribution $q(s)$ in our simulations. We consider three choices for this distribution: (i) Gaussian $q(s)$ with uniform, non-negative correlations between stimuli; (ii) Gaussian $q(s)$ with random positive correlations between stimuli; and (iii) a mixture of Gaussians prior favoring homogeneity among the stimuli. These models are all too simple to provide complete characterizations of the subject’s prior $q(s)$ presumed to be implemented in his or her visual cortex, or the relevant environmental statistics of the corresponding visual features. Nonetheless, each contains some crucial properties shared by the environmental statistics of many visual features, hence probably by the prior implemented in visual cortex as well. Our goal in the following sections is to demonstrate that important aspects of VSTM performance limitations arise even within the context of these simple models, and thus model mismatch is a viable explanation for these limitations.

3.2.1 Gaussian $q(s)$ with uniform non-negative correlations between stimuli

We first investigate the consequences of a simple mismatch between the correlation structures of the experimenter’s and the subject’s priors. To this end, we consider the case where the experimenter’s prior, $p(s)$, is a multivariate Gaussian distribution with mean $\mathbf{m}$ (we take $\mathbf{m}$ to be a vector of zeros in all simulations), and a covariance matrix with equal marginal variances, $\sigma^2_p$, along each dimension and a uniform non-negative correlation, denoted $\rho_p$, between dimensions (note that setting $\rho_p = 0$ results in a diagonal covariance matrix). The subject’s prior $q(s)$ is also assumed
to be a multivariate Gaussian with the same mean \( \mathbf{m} \) and a covariance matrix with equal marginal variances, \( \sigma_{q}^2 \), along each dimension and a uniform non-negative correlation, denoted \( \rho_{q} \), between dimensions. Thus the covariance matrices of both the experimenter’s prior and the subject’s prior can be written as:

\[
\Sigma_N = \begin{bmatrix}
\sigma^2 & \rho \sigma^2 \\
\rho \sigma^2 & \sigma^2
\end{bmatrix}.
\]  

(3.3)

To investigate the consequences of a mismatch solely between the correlation structures of the experimenter’s and the subject’s priors, we assume \( \sigma_{p}^2 = \sigma_{q}^2 = 4 \) for the main simulations. However, we also briefly discuss the consequences of an additional mismatch between the variances \( \sigma_{p}^2 \) and \( \sigma_{q}^2 \) (see Section 3.3.1).

For many basic stimulus features, their environmental statistics exhibit rich dependencies between feature values at different locations (Sigman, Cecchi, Gilbert & Magnasco, 2001; Geisler, Perry, Super & Gallogly, 2001; Hyvärinen, Hurri & Hoyer, 2009). Although a Gaussian distribution with covariance matrix given in Equation 3.3 is probably not a realistic model of such dependencies for any stimulus dimension, this model is intended to illustrate how even such a simple form of dependency can give rise to prominent performance limitations in VSTM.

### 3.2.2 Gaussian \( q(s) \) with random positive correlations between stimuli

The second model of the subject’s prior \( q(s) \) that we consider is a Gaussian distribution with random positive correlations between the stimuli. This can be thought

\[
\Sigma_N = \begin{bmatrix}
\sigma^2 & \rho \sigma^2 \\
\rho \sigma^2 & \sigma^2
\end{bmatrix}.
\]
of as a crude approximation to experiments where, in each trial, the stimuli are presented at (pseudo-)random locations on the screen. If it is assumed that correlations between features at different locations decrease as a function of the distance between the locations, as in second-order orientation statistics in natural images (Geisler, Perry, Super & Gallogly, 2001), a Gaussian prior with random positive correlations can be considered as an approximate model of the expected correlations between features at random locations. Of course, the real natural joint statistics of a number of features would be much more complex than can be captured by a simple Gaussian distribution but, again, this model is only intended to provide a demonstration of how important VSTM performance limitations can arise even in very simple models. More complex dependencies as exist in the natural statistics of visual features would, if anything, increase the strength of the effects considered in this chapter (such as the decline in recall precision with set size) because higher-order dependencies would exacerbate the discrepancy or mismatch between the experimenter’s prior, which usually displays no dependencies, and the subject’s prior, which exhibits a rich set of dependencies.

For simulations of the Gaussian model with random positive correlations, we generated random correlation matrices by first randomly generating eigenvalues for the correlation matrix (see Appendix B.2 for details). The eigenvalues are drawn from a symmetric Dirichlet distribution with concentration parameter $\gamma$. Small values of $\gamma$ produce sparse eigenvalues which, in turn, generate correlation matrices with large correlations between dimensions. Larger values of $\gamma$ yield a broader distribution of eigenvalues, which corresponds to smaller correlation values.
3.2.3 Mixture of Gaussians prior favoring homogeneity

Natural visual inputs constitute a very small subset of the set of all possible visual inputs. Consider the set of all possible configurations of orientations for, say, 100 oriented line segments at particular locations in an image. The orientations of these segments can be represented as a vector in a 100-dimensional space. But naturally-occurring configurations of oriented line segments typically reside in a much smaller dimensional subspace of this high-dimensional space of all possible configurations of orientations. Natural configurations will be dominated by relatively homogeneous configurations where many orientations are similar, or by configurations that form smooth contours (Geisler, Perry, Super & Gallogly, 2001). If we consider homogeneous configurations where all orientations are similar, these reside near a one-dimensional manifold in the 100-dimensional space, namely the line \( s_1 = s_2 = \ldots = s_{100} \) where \( s_i \) is the orientation of the \( i \)th segment.

Our third model of the subject’s prior \( q(s) \) is intended to capture the low-dimensional or sparse nature of the natural statistics of visual features. Once again, if we think of the joint statistics of \( N \) visual features in a standard VSTM experiment, the experimenter’s prior \( p(s) \) will typically be a uniform (or at least non-sparse) distribution over the \( N \)-dimensional space of all possible configurations of the visual features. In contrast, the subject’s prior \( q(s) \) will have significant probability only over a much lower-dimensional subspace of the \( N \)-dimensional space. We model this scenario with a mixture of Gaussians prior that places progressively smaller probabilities over higher dimensional subspaces.

The model starts with a mixture component that places its probability mass on (or near) a one-dimensional subspace. If \( s_i \) denotes the \( i \)th stimulus, then this one-dimensional subspace is the line \( s_1 = s_2 = \ldots = s_N \). This mixture component has
a covariance matrix of the form given in Equation 3.3 with a large correlation $\rho$ (we use $\rho = 0.99$). This effectively constrains the component’s probability mass to the line $s_1 = s_2 = \ldots = s_N$, thus implementing a preference for homogeneous stimuli. Using additional mixture components, the model then allocates progressively less probability to higher dimensional subspaces and thus to less homogeneous stimulus configurations. There are, for example, $N$ mixture components that place most of their mass on two-dimensional subspaces with all but one $s_i$ having similar feature values. This is achieved by setting the $i^{\text{th}}$ row and the $i^{\text{th}}$ column of the correlation matrix (except for the entry on the diagonal which is always equal to 1) to a small value (we use 0, indicating zero correlations between $s_i$ and the remaining stimuli) and keeping the remaining correlations high. This process is continued until the $N^{\text{th}}$-dimensional space, reducing the probability mass allocated as the dimensionality is increased (see Figure 3.2). Our mixture of Gaussians prior is reminiscent of the feature statistics in the so-called “dead leaves model” (Pitkow, 2010), which is known to be a very good model of natural image patches (Zoran & Weiss, 2012).

Our mixture of Gaussians model incorporates only one type of ecologically plausible statistical regularity among the stimuli, namely homogeneity. In a more realistic extension of this model, there would be many other high-probability components in the mixture corresponding to other low-dimensional manifolds or subspaces that have a high probability in the natural visual environment, such as manifolds corresponding to smooth contours or regular shapes for the case of orientation. It seems quite plausible that such additional regularities would increase the mismatch between the experimenter’s prior, which is generally a completely factorized and uniform distribution, and the subject’s prior and thus amplify the effects of model mismatch considered in the next section. It would be interesting to investigate in more detail
Figure 3.2: Under the mixture of Gaussians prior, more homogeneous configurations of orientations are given greater probability. In the examples shown here, the homogeneity of orientations, and hence the probability under $q(s)$, decreases from A to F. See the main text for details of how the homogeneity preference is implemented in the mixture of Gaussians prior.

the consequences of incorporating such additional regularities in the model, but this is beyond the scope of the current paper.

Parameter values (i.e., mixture weights of the components) for the mixture of Gaussians model were chosen manually such that (i) components corresponding to less homogeneous configurations were given progressively smaller weights as explained above, and (ii) the resulting model produced significant encoding variability as discussed in Section 3.3.3 below. To investigate the effects of varying the parameters of the model, we also conducted simulations where the individual weights were exponentiated to a common non-negative exponent $k$ (i.e., $w_j^k$, where $w_j$ denotes the default manually-chosen weight of the $j^{th}$ component) and then renormalized to ensure that the weights sum to 1. Note that $k = 1$ corresponds to the default setting of the weights that were manually chosen as described above. Smaller $k$ values correspond to more uniform weights and larger $k$ values correspond to sparser weights with most of the total weight concentrated on the most homogeneous component.
3.3 Strong hypothesis: Model mismatch alone accounts for performance limitations in VSTM

We first consider the strong hypothesis that model mismatch alone (without a general resource or capacity limitation) accounts for performance limitations in VSTM. For the simulations presented in this section, we assume that the noise distribution $p(x|s)$ does not depend on the set size (i.e., the variance of the noise distribution along each dimension is constant across different set sizes). This implements the assumption (made under the strong hypothesis) that a general resource or capacity limitation does not play a significant role in the observed performance limitations in VSTM tasks. As discussed in more detail in Subsection 3.3.1 below, under a general resource limitation, the variance of the noise distribution would increase with set size. The results presented in this section suggest that a mismatch between the experimenter’s prior $p(s)$ and the subject’s prior $q(s)$ alone can explain many of the main qualitative characteristics of performance limitations in VSTM without assuming a general resource limitation, lending support to the strong hypothesis. In Section 3.4, however, we will review some evidence suggesting a role for a general resource limitation in VSTM in addition to model mismatch.

3.3.1 Memory precision-set size relationship

A prominent result cited as evidence for a general resource or capacity limitation in VSTM is the monotonic decline in subjects’ memory precision with set size in
standard VSTM experiments. Most theories explain this decline in memory precision with set size in terms of a limited amount of memory resources. As the number of items increases, the amount of available resources per item decreases, hence the precision with which an individual item can be encoded declines (Palmer, 1990). Different theories disagree on the nature of these resources. One type of theory claims that the resources are a small number of discrete, fixed-precision memory ‘slots’ (Zhang & Luck, 2008), whereas another type of theory claims that the resources are continuous (Wilken & Ma, 2004; Bays & Husain, 2008). In one version of the second type of theory, for instance, it is hypothesized that stimuli are encoded by neural populations, and there is a constraint on the amount of neural resources that can be expended to encode a visual display (Bays & Husain, 2008; Ma & Huang, 2009). For example, there might be a limit on the total number of neural spikes (Ma & Huang, 2009). In this case, larger set sizes lead to less spikes per item. Importantly, under a plausible model of neural variability (called Poisson-like variability), this implies that increases in set size lead to increases in the variance of $p(x|s)$ that scale as $N$ (Ma & Huang, 2009). This predicted increase in variance with set size is slower than that observed in most VSTM experiments which typically report a power-law relationship between variance and set size with exponents around 1.2-1.5 (Bays & Husain, 2008; Bays et al., 2011; Van den Berg et al., 2012).

However, the relevance of these theoretical arguments to VSTM performance limitations remains unclear. Consider, for instance, the argument from a limit on the total number of spikes. Although this argument predicts a decrease in encoding precision with set size, in the limit of a large number of spikes (which is the relevant regime for the brain), differences in encoding precision for different set sizes become very small. Thus, without an additional suboptimality such as a suboptimal readout
of neural responses, this argument cannot account for the magnitude of changes in encoding precision with set size. Once the necessity of an additional suboptimality is conceded, however, as argued by Beck et al. (2012), this additional suboptimality, rather than a general resource limitation such as a limit on the total number of spikes, is likely to explain the bulk of performance limitations in VSTM.

In this section, we show that set size effects can be explained in terms of a mismatch between the experimenter’s prior $p(s)$ and the subject’s prior $q(s)$ without assuming a general resource limitation (i.e., without assuming an increase in observation noise with set size). This implies that it should be possible to recall a target stimulus with the same precision regardless of set size, if the stimuli come from an appropriate distribution. As we show later in this section, if $p(s)$ and $q(s)$ have certain properties (see Section 3.3.1), it is even possible to recall a stimulus with greater precision as the set size increases. That is, it is possible to observe an “inverse set size effect”. We believe that these results warrant a reconsideration of the theoretical arguments regarding capacity limitations because they show that it is possible to explain the memory precision–set size relationship without assuming a general resource limitation.

**Mathematical argument**

Intuitively, the model-mismatch perspective proposed here explains set size effects as follows. Recall that on a typical trial of a VSTM experiment, the subject attempts to recall the feature value of a target item. To understand the subject’s performance, we consider two quantities, namely the true marginal posterior distribution of the target feature value $p(s_t|x)$ which is based on the experimenter’s prior $p(s)$, and the approximate marginal posterior $q(s_t|x)$ which is based on the subject’s prior...
$q(s)$. The main result of this subsection concerns how the average difference between these two distributions [that is, averaged with respect to $p(x)$] can grow with set size. In particular, if the experimenter’s prior over stimulus values $p(s)$ involves no dependencies between different stimuli, but the subject’s prior $q(s)$ does contain such dependencies, then adding new dimensions (i.e., adding new items) to each distribution can make the respective marginal posteriors, on average, less and less similar to each other, thus worsening the subject’s performance. A more formal elaboration of this argument follows.

Computing an incorrect posterior distribution incurs a cost to the subject’s performance. In general, the more similar is the subject’s posterior to the true posterior, the better the subject’s performance will be, where similarity between distributions is measured according to a suitable measure. Since the likelihood is the same in the computation of both the true and the approximate posteriors (Equations 3.1-3.2), the deviation between these two distributions is due to differences in the priors $p(s)$ and $q(s)$. Again, the more similar the two priors, the better the subject’s performance will be. A common way to measure the similarity between two distributions is the Kullback-Leibler (KL) divergence. Here, we use the KL divergence per dimension between the priors $p(s)$ and $q(s)$ (KL divergence between the joint priors divided by the dimensionality) as a heuristic measure of the average dissimilarity between the marginal posteriors. It is important to note that KL divergence per dimension between the joint priors is the relevant measure here, and not the KL divergence between the marginal priors, because the computation of the marginal posteriors (Equations 3.1-3.2) uses the joint priors and not the marginal priors. We show that adding a new dimension to each of the respective priors might increase the KL divergence per dimension between the two distributions if the newly added dimension
is independent of the previous dimensions in $p(s)$, but dependent on the previous dimensions in $q(s)$. The implication of this mathematical result is that increasing the set size (i.e., adding a new dimension) can make the true and approximate marginal posteriors, on average, more dissimilar to each other, thus worsening the subject’s performance.

We assume that all one-dimensional marginals of $p$ are identical: $p(s_i) = p(s_j)$ for all $i$ and $j$. Similarly, all one-dimensional marginals of $q$ are identical, as there is no a priori reason to expect any individual item in a display to be represented differently than any other item. We first consider the KL divergence per dimension between the one-dimensional priors $p(s_1)$ and $q(s_1)$:

$$D_{KL}^1(q||p) = -\int q(s_1) \log p(s_1) ds_1 + \int q(s_1) \log q(s_1) ds_1$$

(3.4)

where $H^1(q, p)$ denotes the cross-entropy per dimension for $n$-dimensional $q$ and $p$ and, similarly, $H^1(q)$ denotes the entropy per dimension for $n$-dimensional $q$. We next add a second dimension, $s_2$, to both distributions and recalculate the KL divergence per dimension between the distributions. We assume that the new dimension is independent of the previous dimensions under the true prior: $p(s_1, s_2) = p(s_1)p(s_2)$ as is common in standard VSTM experiments. However, we allow for the possibility that there may be dependencies between the new dimension and the previous dimensions under the subject’s prior. As discussed in the previous section, such dependencies are common between visual stimuli in the natural environment. Given these assumptions, we recalculate the KL divergence per dimension between the two-dimensional
Thus, $D_{KL}^2(q||p) \geq D_{KL}^1(q||p)$ with equality holding if and only if $q(s_1, s_2) = q(s_1)q(s_2)$.

In general, when increasing the dimensionality from $N$ to $N + 1$, it is easy to see that $H^N(q, p) = H^{N+1}(q, p)$ [because we assume $q(s_i) = q(s_j)$ and $p(s_i) = p(s_j)$ for all $i$ and $j$]. Therefore, $D_{KL}^{N+1}(q||p) \geq D_{KL}^N(q||p)$ if and only if $H^N(q) \geq H^{N+1}(q)$ (i.e., if and only if the new dimension decreases the entropy per dimension of the subject’s prior). This is a non-trivial condition that is satisfied if the new dimension adds more “structure” to the distribution than the average of the previous dimensions. In the next subsection, we show that the condition $H^N(q) \geq H^{N+1}(q)$ is satisfied for the case of Gaussian $q(s)$ with uniform non-negative correlations. In addition, using both this simple model and the two other models introduced in Section 3.2, we illustrate, through simulations, how set size effects can arise without assuming a general resource limitation (i.e., without assuming an increase in the variance of the noise distribution $p(x|s)$ with set size).
Simulations

Section 3.2 described three possible models for the subject’s prior $q(s)$. This subsection presents results for each of these models, starting with a Gaussian prior with uniform non-negative correlations across dimensions.

For the Gaussian $q(s)$ with uniform non-negative correlations (see Section 3.2.1), we first show that $D_{KL}^{N+1}(q||p) \geq D_{KL}^{N}(q||p)$ holds for all $N$. Above, we showed that $D_{KL}^{N+1}(q||p) \geq D_{KL}^{N}(q||p)$ if and only if $H^N(q) \geq H^{N+1}(q)$ [assuming that stimuli are independent under $p(s)$]. Therefore, it is enough to show that $H^N(q) \geq H^{N+1}(q)$. For a Gaussian $q(s)$ with uniform non-negative correlations (Equation 3.3), the entropy per dimension can be written as:

$$H^N(q) = \frac{1}{2N} \log[(2\pi e)^N|\Sigma_N|]$$

$$= \frac{1}{2} \log 2\pi + \frac{1}{2} + \log \sigma_q + \frac{1}{2} \log(1 - \rho_q) + \frac{1}{2N} \log(1 + \frac{N\rho_q}{1 - \rho_q})$$

(3.11)

(3.12)

where $\Sigma_N$ is the covariance matrix of $q(s)$, $\rho_q \geq 0$ is the common correlation coefficient across dimensions, and $\sigma_q$ is the common marginal standard deviation along each dimension. Therefore, $H^N(q) \geq H^{N+1}(q)$ if and only if:

$$\frac{1}{N} \log(1 + \frac{N\rho_q}{1 - \rho_q}) \geq \frac{1}{N+1} \log(1 + \frac{(N+1)\rho_q}{1 - \rho_q})$$

(3.13)

In Appendix B.1, we show that this last inequality is always true with equality holding only when $\rho_q = 0$.

As described in Section 3.1, we simulated a large number of trials of a VSTM recall experiment. On each simulated trial, a display of items is created by randomly sampling from $p(s)$, which is an uncorrelated multivariate Gaussian for the
simulations reported in this subsection, a target item is selected, and the subject’s approximate marginal posterior distribution over the target’s feature value, \( q(s_t|x) \), is computed. The subject’s response is set to the mean of this distribution. The recall precision is measured as the inverse standard deviation of the subject’s error distribution over all trials.

Figure 3.3A shows the recall precision (inverse standard deviation of the error distribution) as a function of set size for the case of Gaussian \( q(s) \) with uniform non-negative correlations. Results are shown for 34 different correlations \( \rho_q \) ranging from 0 to 0.99 with increments of 0.03. Clearly, the recall precision decreases with set size and this set size effect increases monotonically with \( \rho_q \). To quantify the set size effect, we fit power-law functions of the form \( cN^{\xi} \) to these data. Smaller exponents indicate more drastic reductions in precision with set size. Empirically, it has been observed that power-law functions with exponents ranging from \(-0.60\) to \(-0.75\) fit subjects’ data well (Bays & Husain, 2008; Bays et al., 2011; Van den Berg et al., 2012). Figure 3.3E (solid line) shows the exponents of the power-law fits as a function of \( \rho_q \). \( \rho_q = 0.96 \) yields an exponent of \( \xi = -0.56 \) (\( R^2 = 0.89 \)) and \( \rho_q = 0.99 \) yields \( \xi = -0.84 \) (\( R^2 = 0.90 \)). For a given correlation value \( \rho_q \), it is possible to produce smaller exponents—that is, more drastic reductions in precision with set size—if a mismatch between the variances in the experimenter’s prior and the variances in the subject’s prior, in addition to the mismatch in the correlation structure, is introduced. Figure 3.3D shows precision as a function of set size when the experimenter’s prior is assumed to have a standard deviation (in each dimension) that is three times the standard deviation of the subject’s prior (\( \sigma_p = 3, \sigma_q = 1 \)), and Figure 3.3E (dashed line) shows the exponents of the corresponding power-law functions.
Figure 3.3: (A) Recall precision as a function of set size in the Gaussian model with uniform non-negative correlations. Different curves correspond to different correlation values, denoted \( \rho_q \). 34 different \( \rho_q \) values are used in this example, from 0 to 0.99 with increments of 0.03. Lighter colors represent lower \( \rho_q \) values. Following Bays & Husain (2008), precision is defined as 1 over the standard deviation of errors. (B) Recall precision as a function of set size in the Gaussian model with random positive correlations. Different curves correspond to different \( \gamma \) values. Six different \( \gamma \) values are used in this example, \( \gamma = 0.1, 0.15, 0.3, 0.9, 1.8, 3.6 \). Lighter colors represent higher \( \gamma \) values. (C) Recall precision as a function of set size in the mixture of Gaussians model. Different curves correspond to different weight exponents, \( k \). Six different \( k \) values are used in this example, \( k = 0, 0.2, 1, 7, 13, 19 \). Lighter colors represent smaller \( k \) values. (D) Recall precision as a function of set size in the Gaussian model with uniform correlations across dimensions. In this example, in addition to the difference in correlation, the experimenter’s prior and the subject’s prior also differ in variances, with the experimenter’s prior having a standard deviation in each dimension that is three times the standard deviation of the subject’s prior (\( \sigma_p = 3, \sigma_q = 1 \)). Different curves correspond to different \( \rho_q \) values. As in (A), 34 different \( \rho_q \) values are shown, from 0 to 0.99 with increments of 0.03. Lighter colors represent lower \( \rho_q \) values. (E) Exponents of power law fits, \( 1/\sigma \propto N^\xi \), to precision-set size curves shown in (A) [solid line] and (D) [dashed line] as a function of \( \rho_q \).
Next, consider the subject’s prior modeled as a multivariate Gaussian with random positive correlations between stimuli. Figure 3.3B shows the recall precision as a function of set size in this case. Results are shown for 6 different $\gamma$ values: 0.1, 0.15, 0.3, 0.9, 1.8, 3.6. The set size effect decreases monotonically with $\gamma$. Setting $\gamma = 0.15$ yields a power-law exponent of $\xi = -0.67$ ($R^2 = 0.87$). Again, for a given $\gamma$ value, it is possible to produce smaller exponents if a deviation between the variances in $p(s)$ and $q(s)$ is introduced.

Finally, Figure 3.3C shows the results when the subject’s prior is modeled using the mixture of Gaussians model. Results are shown for 6 different $k$ values: 0, 0.2, 1, 7, 13, 19. Recall that $k = 0$ corresponds to the uniform weighting of all components, $k = 1$ corresponds to the default setting of the weights (see Section 3.2.3), and larger $k$ values correspond to sparser weights with most of the total weight concentrated on the most homogeneous component. The set size effect increases monotonically with $k$. Different $k$ values also produce qualitatively different shapes for the precision–set size curves. Setting $k = 1$ yields a power-law exponent of $\xi = -0.40$ ($R^2 = 0.85$).

It is interesting to note that both for the Gaussian model with uniform non-negative correlations and for the Gaussian model with random positive correlations, especially for large correlation values, a large proportion of the total reduction in recall precision is obtained when the set size is increased from $N = 1$ to $N = 2$. For the mixture of Gaussians model, on the other hand, substantial reductions in precision can still be observed beyond $N = 2$. This is presumably due to the existence of higher-order dependencies between stimuli under the mixture of Gaussians prior, whereas the Gaussian models incorporate only second-order dependencies.
The case of correlated $p(s)$

Above, we assumed that the experimenter’s prior $p(s)$ displays no dependencies between different stimuli, which is usually the case in standard VSTM studies. This subsection explores what happens if we introduce dependencies between stimuli in $p(s)$. We report the surprising finding that this situation can lead to an “inverse set size effect” in which recall precision increases with set size.

In these simulations, the experimenter’s prior $p(s)$ is assumed to be a multivariate Gaussian with uniform non-negative correlations $\rho_p$ between each pair of its dimensions. Setting $\rho_p = 0$ results in the case of uncorrelated $p(s)$ discussed in previous subsection.

We first consider the case in which the subject’s prior $q(s)$ is a multivariate Gaussian with uniform non-negative correlations, denoted $\rho_q$. Our main finding is that if both $p(s)$ and $q(s)$ display correlations, an inverse set size effect—an increase in recall precision with set size—is predicted. This can be seen in Figure 3.4A which shows the simulation results for $\rho_p = 0.9$. Moreover, the inverse set size effect becomes more pronounced with larger correlations in the experimenter’s prior. This is illustrated in Figure 3.4B which shows the exponents of power-law fits to precision-set size curves for each pair of $(\rho_p, \rho_q)$ values tested. Positive exponents indicate an increase in precision with set size, whereas negative exponents indicate a decrease in precision with set size. Overall, the predicted decreases in precision with set size are larger in magnitude than the predicted increases in precision with set size (note the maximum and minimum values of the color bar in Figure 3.4B). Results for the case of Gaussian $q(s)$ with random positive correlations were qualitatively similar. For brevity, these results are not shown.

The increase in recall precision with set size can be understood intuitively as fol-
Figure 3.4: (A) Recall precision as a function of set size in the Gaussian model with uniform non-negative correlations. In these simulations, \( p(s) \) as well as \( q(s) \) is assumed to be a multivariate Gaussian with uniform correlations. Correlations \( \rho_p \) and \( \rho_q \) denote the uniform correlations between pairs of dimensions in \( p(s) \) and \( q(s) \), respectively. \( \rho_p \) was set to 0.9. Precision-set size curves are shown for 34 different values of \( \rho_q \) from 0 to 0.99 in increments of 0.03. Lighter colors represent lower \( \rho_q \) values. Note that when \( \rho_q \) exceeds \( \rho_p \), the inverse set size effect starts to decrease and ultimately reverses to a set size effect (a decline in precision with set size). (B) Exponents \( \xi \) of power-law fits to precision-set size curves for each pair of \( (\rho_p, \rho_q) \) values. Positive values indicate an increase in precision with set size, whereas negative values indicate a decrease in precision with set size. (C) Results for the mixture of Gaussians model. \( p(s) \) is a multivariate Gaussian with uniform correlations across dimensions, and \( q(s) \) is a mixture of Gaussians with weight exponent set to \( k = 1 \). \( \rho_p \) is varied from 0 to 0.99 in increments of 0.03. Each curve represents the precision-set size curve for a different \( \rho_p \) value. Lighter colors represent lower \( \rho_p \) values. (D) Exponents of power law fits, \( \xi \), to precision-set size curves in the mixture of Gaussians model shown in (C) as a function of \( \rho_p \).
lows. When there are dependencies between different dimensions in \( p(s) \), it becomes possible to gain information about an individual stimulus, \( s_i \), from the remaining stimuli, and the amount of information gained from the remaining stimuli increases as the number of those stimuli increases. This can also be thought of as a “context effect”—it becomes easier to recall or recognize a stimulus within the context of other stimuli correlated with it, compared to recalling or recognizing it in isolation (Bar, 2004).

The implication of the predicted increase in recall precision with set size when both \( \rho_p \) and \( \rho_q \) are positive is that if the experimenter uses a highly correlated prior (e.g., uses highly homogeneous displays in each trial), one would expect to see an increase in subjects’ recall precision. However, one has to be careful in translating the results of the simulations directly into experimental predictions. Several caveats apply. First, the magnitude of the increases in recall precision with set size is, in general, smaller than the magnitude of the decreases. Therefore, it may be difficult to design experiments with sufficient power to detect increases in recall precision. Second, the magnitude of the increases in precision depends on the degree of match (or mismatch) between \( p(s) \) and \( q(s) \). For models more complex than the ones considered in this chapter, it is possible for both \( p(s) \) and \( q(s) \) to display strong dependencies and yet be sufficiently different from each other for the predicted increases in precision to be small or even non-existent.

Figure 3.4C shows the results for the mixture of Gaussians prior favoring homogeneity. In this case, the experimenter’s prior \( p(s) \) is assumed to be a multivariate Gaussian with uniform correlations \( \rho_p \) across dimensions, and the subject’s prior \( q(s) \) is assumed to be a mixture of Gaussians with \( k = 1 \). We varied \( \rho_p \) from 0 to 0.99 in increments of 0.03. Each curve in Figure 3.4C represents the precision-set size curve
for a different $\rho_p$ value, with lighter colors representing lower $\rho_p$ values. Again, as in the Gaussian models, for sufficiently high $\rho_p$, an increase in recall precision with set size is predicted, although, for a given $\rho_p$ value, the magnitude of the predicted increase in recall precision is, in general, smaller than in the Gaussian models. Figure 3.4D shows the exponents of power law fits to the precision–set size curves in the mixture of Gaussians model as a function of $\rho_p$.

In summary, the results of our simulations on the precision–set size relationship are as follows. First, our results show that model mismatch can qualitatively explain the observed decline in memory precision with increasing set size without assuming a general resource limitation. In general, introducing more mismatch between the subject’s prior and the experimenter’s prior leads to steeper declines in precision with set size. For instance, we found that introducing a mismatch in the variances of these two priors, in addition to a mismatch between their correlation structures, amplifies the set size effect. Second, the precision-set size relationship observed in VSTM tasks might crucially depend on the particular choice for the stimulus distribution $p(s)$ used in an experiment. Our results show that if stimuli are independent under $p(s)$, then it is possible to get a set size effect (Figure 3.3A). However, if stimuli are dependent, it is possible to get an inverse set size effect (Figure 3.4A). We only considered a multivariate Gaussian form for $p(s)$, which can only model second-order dependencies between stimuli. Additional higher-order dependencies in $p(s)$ would amplify the predicted inverse set size effect.

### 3.3.2 Variability in encoding precision

Recent experimental evidence suggests that the precision of encoding individual items in VSTM tasks varies both across trials and across items in a single trial (Van den
Berg et al., 2012; Fougnie, Suchow & Alvarez, 2012). Models that incorporate a mixture of multiple precision components, called scale mixtures, generally fit the distribution of recall errors in such tasks significantly better than models with a single precision component such as a single Gaussian distribution. As in set size effects, this result has been interpreted in terms of the properties of the likelihood or noise distribution $p(x|s)$, that is, in terms of variability in the noise distribution, such as variability in the variances along different dimensions of the noise distribution as well as variability in the variances across trials. The potential source of this variability in encoding precision is unclear. Instability of neural representations (drift or diffusion of neural responses) due to neural noise (Fougnie, Suchow & Alvarez, 2012) or fluctuations in attention across trials and across items in a single trial (Van den Berg et al., 2012) have been speculated as possible mechanisms that might explain encoding variability. In this section, we show that these complicated mechanisms are unnecessary for explaining encoding variability in VSTM. Properties of the subject’s prior, and interactions between the experimenter’s prior and the subject’s prior, can naturally account for encoding variability both across trials and across items within a single trial without hypothesizing any variability in the noise distribution $p(x|s)$ itself.

Intuitively, variability in encoding precision arises, according to the perspective proposed here, because when stimuli $s$ are drawn randomly from the experimenter’s prior $p(s)$, in some trials, just by chance, $s$ will fall near a region where the subject’s prior $q(s)$ has high precision. In those trials, the subject will encode the stimuli with high precision. In other trials, however, $s$ will fall in a region where the subject’s prior $q(s)$ has low precision, in which case the subject encodes the stimuli with low precision. Similarly, in different trials, $s$ might fall in regions that have high precision.
along some dimensions according to $q(s)$, but low precision in other dimensions, producing within-trial variability in encoding precision.

There is no straightforward relationship between the degree of mismatch between $p(s)$ and $q(s)$ and the amount of variability in encoding precision. Variability in encoding precision might arise even when the experimenter’s prior exactly matches the subject’s prior. Imagine, for example, that the subject’s prior is a mixture of Gaussians with two components, one highly correlated, and the other uncorrelated. If the experimenter’s prior exactly matches the subject’s prior, in some trials, stimuli will fall in a region where the correlated component has high probability. In those trials, encoding precision will be higher. In other trials, however, stimuli will fall in a region where the uncorrelated component has high probability, leading to a lower encoding precision. Overall, there will be variability in encoding precision despite the perfect match between the experimenter’s and the subject’s priors. Conversely, variability in encoding precision may be low, or even non-existent, even when there is a mismatch between the experimenter’s and subject’s priors. Again, imagine that the subject’s prior is a mixture of a highly correlated Gaussian and an uncorrelated Gaussian. If samples from the experimenter’s prior consistently come from a region where the correlated component in the subject’s prior has low probability, the stimuli will be attributed to the uncorrelated component and the variability in encoding precision will be low.

For a given set size, encoding variability across trials does not arise in the Gaussian model with uniform non-negative correlations (Section 3.2.1). It is easy to see why this is the case. The same prior with a homogeneous covariance matrix (Equation 3.3) is used in each trial, resulting in a posterior distribution with the same homogeneous covariance matrix in each trial, meaning that individual stimulus fea-
tures have identical marginal variances in each trial. Thus, we only consider the remaining two models in this section.

To show that the other two models generate variability in encoding precision, we simulated a large number of trials and fit the error distributions produced by each of these models with both a single precision distribution (a single Gaussian distribution) and an infinite mixture of Gaussians with different precisions or, more technically, a scale mixture of Gaussians with a Gamma distribution over the precisions (where “precision” is now used in the sense of inverse variance). The latter distribution is also known as the $t$-distribution, and it has heavier tails and a more prominent peak than a Gaussian with the same mean and variance. As mentioned previously, empirical error distributions from VSTM experiments are generally significantly better fit by a scale mixture distribution than by a single precision distribution, indicating the presence of variability in encoding precision. To compare the fits of these two distributions to the error distributions produced by our models, we employed the Bayesian information criterion (BIC) (Schwartz, 1978). The single precision model has two free parameters (mean and variance parameters) and the $t$-distribution has three free parameters (location, scale and degrees-of-freedom parameters). For the $t$-distribution, it is possible to infer the implied Gamma distribution over the precisions from the estimated parameters of the $t$-distribution (see the Appendix for details).

As before, the noise distribution $p(x|s)$ is assumed to be a multivariate Gaussian with an isotropic, diagonal covariance matrix (variance along each dimension is equal to 0.25 as in the previous simulations). Importantly, the noise distribution itself does not contain any variability (i.e., the variances in the noise distribution do not vary across items or trials). For the experimenter’s prior, $p(s)$, we first used the uncorre-
lated Gaussian distribution (with $\rho_p = 0$ and $\sigma_p^2 = 4$). Figures 3.5A and D show the BIC scores of the single precision Gaussian distribution relative to the BIC score of the $t$-distribution for different set sizes. Positive values indicate better fits for the $t$-distribution. For the Gaussian $q(s)$ with random positive correlations (Figure 3.5A), all set sizes except for the set size of 1 indicated better fits for the $t$-distribution, suggesting the presence of multiple precision components (the concentration parameter $\gamma$ was set to 0.15 in these simulations). For the mixture of Gaussians prior favoring homogeneity with $k = 1$ (Figure 3.5D), all set sizes other than 1 and 2 indicated better fits for the $t$-distribution, again suggesting the presence of multiple precision components (all positive BIC values were larger than 150 indicating very strong evidence in favor of the $t$-distribution). Figures 3.5B and E show the inferred Gamma distributions over precision for the best fit $t$-distributions (see the Appendix for details of how these Gamma distributions are estimated). Predictably, larger set sizes lead to smaller mean precisions.

To investigate the effects of correlations in the experimenter’s prior, we next used a multivariate Gaussian $p(s)$ with a uniform correlation of $\rho_p = 0.9$ between each pair of dimensions. Introducing correlations in $p(s)$ significantly reduced variability in encoding precision for the case of Gaussian $q(s)$ with random positive correlations (Figure 3.5C) and completely eliminated variability in encoding precision for the mixture of Gaussians model (Figure 3.5F).

The results for the mixture of Gaussians model (Figure 3.5D-F) were sensitive to the variance of the noise distribution and the correlation coefficient $\rho_p$ of the experimenter’s prior. In particular, when the variance of the noise distribution is increased, variability in encoding precision with the uncorrelated $p(s)$ decreases and variability in encoding precision with the correlated $p(s)$ increases (up to a certain level of noise.
Figure 3.5: (A-B) Results for the case of uncorrelated Gaussian \( p(s) \) and Gaussian \( q(s) \) with random positive correlations \( (\gamma = 0.15) \): (A) BIC score of the Gaussian fit relative to the BIC score of the \( t \)-distribution fit for each set size. Lower scores indicate better fits. Positive values indicate better fits for the \( t \)-distribution than for the Gaussian distribution. (B) For three different set sizes, the estimated Gamma distributions over precision for the best fit \( t \)-distributions. The vertical lines show the means of the Gamma distributions. (C) Similar to (A), but here the experimenter’s prior contains correlations \( (\rho_p = 0.9) \). (D-E) Similar to (A-B) except the results are for the case of uncorrelated Gaussian \( p(s) \) and the mixture of Gaussians \( q(s) \) that favors homogeneity \( (k = 1) \). (F) Similar to (D), but the experimenter’s prior contains correlations \( (\rho_p = 0.9) \).
variance \( [\sigma^2 \approx 2.5] \). When \( p(s) \) is uncorrelated, relatively low noise levels enable the occasional activation of mixture components corresponding to relatively heterogeneous configurations. Increasing the noise variance suppresses the activation of these heterogeneous components by increasing the posterior weights of components corresponding to more homogeneous configurations and consequently reduces the variability in encoding precision. When \( p(s) \) is correlated, on the other hand, an inverse-U shaped relationship is observed between noise variance and variability in encoding precision. For the particular noise variance used in the simulations reported in this section \( (\sigma^2 = 0.25) \), there is no significant variability in encoding precision (see Figure 3.5F). Increasing the noise variance up to a certain value \( (\sigma^2 \approx 2.5) \) gradually increases the variability in encoding precision. However, further increases in the noise variance reduce and ultimately eliminate the variability in encoding precision.

The results presented in this subsection have two important implications. First, they show that it is not necessary to hypothesize any variability in the noise distribution to explain encoding variability. Variability in encoding precision might instead arise if the subject’s prior is sufficiently structured. Consider, for instance, the mixture of Gaussians prior favoring homogeneity. Encoding variability arises in this model, because in some trials, the presented stimuli are more homogeneous and hence, under the subject’s prior, fall near more correlated components, which leads to encoding with better precision.\(^2\) Within–trial encoding variability can be explained similarly. This suggests that the encoding variability observed in VSTM tasks might not be due to attentional fluctuations or fluctuations in neural noise as generally thought. Rather, this variability may be due to variability in the homogeneity level.

\(^2\)When a diagonal likelihood distribution is combined with a correlated Gaussian prior, the resulting posterior has lower marginal variances than a posterior obtained from the combination of the same likelihood with an uncorrelated Gaussian prior that has the same marginal variances as the correlated Gaussian prior.
of the display, or in general variability in the “naturalness” of the display under the subject’s prior—by chance, the display, or parts of the display, will look more or less natural in different trials. This, in turn, results in the encoding of items with more or less precision. There is already some experimental evidence for this account of variability in encoding precision in VSTM tasks (Brady & Alvarez, 2012; Brady & Tenenbaum, 2013; Orhan & Jacobs, in press).

Second, as with precision-set size effects discussed in the previous subsection, the results reported in this subsection emphasize the potential importance of the stimulus distribution \( p(s) \) for determining patterns in subjects’ responses. If stimuli are independent, subjects may show significant variability in encoding precision. If stimuli are dependent, on the other hand, variability in encoding precision may be significantly reduced or even eliminated (see Figure 3.5).

3.3.3 Different set-size dependencies for initial encoding rate and asymptotic precision

Bays et al. (2011) recently investigated the time course of recall precision in VSTM by varying the presentation time of the visual display in a standard VSTM recall experiment. In each trial in Experiment 1 of Bays et al. (2011), a visual display consisting of a number of colored bars was presented, followed by a mask and then a delay period. One of the bars was then cued, and subjects recalled the orientation of the cued bar. Importantly, presentation time of the visual display was varied across trials. Bays et al. (2011) found that recall precision increased with presentation duration. The relationship between recall precision and presentation time was well-fit by a negative exponential function such that an initial rapid increase in precision
is followed by saturation at longer presentation times (Bays et al., 2011):

\[ P(t) = P_{\text{max}}(1 - \exp(-t/\tau)) \]  

(3.14)

where \( P(t) \) is the recall precision (inverse standard deviation of the error distribution) for a presentation duration of \( t \) and \( P_{\text{max}} \) denotes the maximum precision. When data from VSTM recall experiments that systematically vary presentation time for different set sizes are fit with a function of the form given in Equation 3.14, a power-law relationship is observed between set size \( N \) and \( P_{\text{max}} \), with an exponent around \(-0.60\) (Bays et al., 2011). This is the familiar set size effect previously discussed in Section 3.3.1. However, initial encoding rate, \( P_{\text{max}}/\tau \), or the slope of the initial part of the negative exponential function in Equation 3.14, when fit to behavioral data, displays a \( 1/N \) relationship to set size (Bays et al., 2011), suggesting that \( \tau \) should be roughly proportional to \( N^{0.4} \). This implies that increasing the set size affects the initial rate of information accrual into VSTM more than it affects the asymptotic precision. The stronger dependence of initial encoding rate, compared to asymptotic precision, on set size led Bays et al. (2011) to suggest that there is an additional cost for initially encoding information into VSTM above and beyond the cost of simply storing information in VSTM.

In this section, we show that it is unnecessary to assume a set-size dependent noise distribution to explain the behavioral data from Bays et al. (2011). Specifically, we show that the data can be qualitatively explained if it is assumed that the precision of the noise distribution \( p(x|s) \) changes with presentation time \( t \) according to Equation 3.14, but is independent of set size (i.e., \( P_{\text{max}} \) and \( \tau \) do not depend on the set size). Given this assumption, different set-size dependencies for asymptotic precision and initial encoding rate are due to an interaction between two factors, the mismatch
between the experimenter’s and the subject’s priors and the time-varying, but set size-independent, noise distribution. This result has important implications for the interpretation of the experimental data. In particular, it suggests that it may not be necessary to assume an extra cost for encoding information into VSTM in addition to the cost of simply storing information in VSTM as hypothesized by Bays et al. (2011). The apparent extra cost may simply be due to the mismatch between the experimenter’s and the subject’s priors.

We start by considering the situation in which the experimenter’s prior is an uncorrelated Gaussian distribution \( \rho_p = 0 \). Figure 3.6A-C shows recall precision as a function of presentation time for different set sizes. Results are shown for the cases in which the subject’s prior \( q(s) \) is a Gaussian distribution with uniform non-negative correlations (Figure 3.6A), a Gaussian distribution with random positive correlations (Figure 3.6B), and a mixture of Gaussians distribution favoring homogeneity (Figure 3.6C). We fit the curves shown in Figure 3.6A-C with negative exponential functions of the form given in Equation 3.14 with the addition of a constant vertical offset \( b \) (this is necessary because, with the relatively informative priors used in our examples, the encoding precision does not drop to 0 for time 0; instead it drops to some positive value determined by the prior variance). From these fits, the initial encoding rate is calculated as \( P_{\text{max}} / \tau \) and the asymptotic precision is calculated as \( P_{\text{max}} + b \). Figure 3.6D-F plots the estimated initial encoding rates (blue) and the asymptotic precisions (green) as a function of set size for the three different models. Also shown are the exponents \( \xi \) of power-law fits to each curve.

Figure 3.7A shows a blown-up view of Figure 3.6A focusing on the shortest two presentation times (5 ms and 15 ms). Also shown in Figure 3.7A with dashed lines are the precisions that would be predicted if the initial encoding rate had the same
Figure 3.6: (A-C) Recall precision as a function of presentation time for different set sizes. Darker colors indicate larger set sizes with set sizes ranging from 1 to 6. Simulations were conducted with the experimenter’s prior $p(s)$ set to an uncorrelated Gaussian distribution ($\rho_p = 0$). Results are shown for the cases of Gaussian $q(s)$ with uniform correlations ($\rho_q = 0.96$) (A), Gaussian $q(s)$ with random positive correlations ($\gamma = 0.15$) (B), and the mixture of Gaussians prior favoring homogeneity ($k = 1$) (C). (D-F) Initial encoding rate (blue) and the asymptotic precision (green) as a function of set size. The exponents $\xi$ of power-law fits to each curve are also provided. Results are again shown for the case of Gaussian $q(s)$ with uniform correlations (D), Gaussian $q(s)$ with random positive correlations (E), and the mixture of Gaussians prior favoring homogeneity (F). Due to computational demands, $10^4$ trials per presentation time per set size were simulated for the Gaussian model with random positive correlations. For the other two models, $10^5$ trials per presentation time per set size were simulated. The simulated presentation times varied from 5 ms to 1005 ms in steps of 10 ms. The noise distribution has an asymptotic precision of 2 and $\tau = 0.1$ for all models.
Figure 3.7: Blown-up views of precision in the first two time steps of the simulation (corresponding to presentation times of 5 ms and 15 ms respectively) for the Gaussian \( q(s) \) with uniform correlations (\( \rho_q = 0.96 \)). Different colors correspond to different set sizes with darker colors representing larger set sizes. Solid lines show the actual simulation results and dashed lines show the precisions that would be predicted if the initial encoding rate had the same dependence on set size as the asymptotic precision (taking the precision for \( N = 1 \) as the baseline). (A) shows the results for the case where the experimenter’s prior is an uncorrelated Gaussian (\( \rho_p = 0 \)) as in Figure 3.6A and (B) shows the results for the case where the experimenter’s prior is a Gaussian with uniform correlations (\( \rho_p = 0.9 \)) as in Figure 3.8A.

Importantly, all models produced different set size dependencies for initial encoding rate and asymptotic precision (Figure 3.6D-F), and these dependencies are qualitatively consistent with the experimental results of Bays et al. (2011), that is, the initial encoding rate declines faster than the asymptotic precision as a function
of set size. These results were achieved using the reasonable assumption that observation noise declines with increasing image presentation times. They were also achieved without making the assumption that observation noise depends on the set size, an assumption that we have consistently avoided throughout this section. The results suggest that there might not be an inherent additional cost to “writing” information into VSTM beyond the cost of storing that information in VSTM. Rather, these apparent costs may be a by-product of using experimental stimuli that do not match the subject’s prior.

Next, we investigate the effects of correlations in the experimenter’s prior $p(s)$ on initial encoding rate and asymptotic precision. This prior was set to a multivariate Gaussian with uniform correlations ($\rho_p = 0.9$). Figure 3.8 shows the results of these simulations. As demonstrated in Section 3.3.1 above, the asymptotic precision-set size curves become shallower when $p(s)$ is correlated. For all models, an inverse set size effect is observed for the initial encoding rate, meaning that encoding rate increases with set size. Surprisingly, the power-law exponents for the initial encoding rate are greater than the power-law exponents for the asymptotic precision for all three models (recall that for uncorrelated $p(s)$, the opposite is true—the initial encoding rate declines more steeply with set size than the asymptotic precision). This can also be seen in Figure 3.7B which shows a blown-up view of Figure 3.8A focusing on the shortest two presentation times. The slopes of the solid lines increase faster than would be predicted if the initial encoding rate and the asymptotic precision had the same dependence on set size. Thus, when a correlated $p(s)$ is used, relative to just maintaining items in VSTM, there is an extra benefit to initially encoding them into VSTM rather than an extra cost!

In summary, our results suggest that the observed declines in asymptotic precision
Figure 3.8: Results when the experimenter’s prior $p(s)$ was a Gaussian distribution with uniform correlations ($\rho_p = 0.9$). The formats of the plots are identical to those in Figure 3.6.
and initial encoding rate with set size should be reduced or even reversed if more ecologically realistic priors (e.g., priors with dependencies among stimuli) are used in VSTM studies. Similarly, our framework predicts that, when more ecologically realistic stimulus distributions are used in VSTM studies, the apparent extra cost to initially encoding items in VSTM compared to just storing them in VSTM should also be either reduced or even reversed.

3.4 Weak hypothesis: Model mismatch and capacity limitations together account for performance limitations in VSTM

In the previous section, we considered the strong hypothesis that performance limitations in VSTM can be explained solely by model mismatch. Our results indicate that many of the main qualitative characteristics of performance limitations in VSTM can be explained in this way. Nonetheless, in our simulation results, there is evidence suggesting that model mismatch alone might be inadequate for explaining performance limitations in VSTM tasks. In this section, we first review this evidence. Through simulations, we then investigate the implications of a weak hypothesis stating that performance limitations can be explained through a combination of a general resource limitation and model mismatch.

3.4.1 Is model mismatch alone enough?

Although model mismatch alone can account for the main qualitative features of performance limitations in VSTM studies, we note here a few potential discrepancies
between our simulation results and properties of performance limitations observed in VSTM tasks. It should be emphasized that these discrepancies relate specifically to the simple models for $q(s)$ considered in this chapter or similar such models, and thus they cannot rule out the viability of the strong hypothesis if more complex and realistic models for $q(s)$ are considered.

First, there seem to be some deviations from a power law in the precision–set size relationships predicted by our models (Figure 3.3). The coefficient of determination, or $R^2$, values for the three models ranged from 0.85 to 0.90 when their parameters were set to values that produced power-law exponents in the empirically observed range of $-0.60$ to $-0.75$. In contrast, power law functions tend to provide better quantitative fits to empirical precision–set size relationships. Bays et al. (2011), for example, report an $R^2$ value of 0.96 for a power-law fit to the precision–set size relationship. An increase in the noise variance that is proportional to the set size $N$ due to a general resource limitation (Section 3.3.1) would readily explain the success of power-law functions in fitting empirical precision–set size relationships, although as discussed in Section 3.3.1, it predicts a shallower decline in precision than observed empirically. A possible explanation of this discrepancy that will be pursued in this section is that the additional decline, not explained by a general resource limitation, might be due to model mismatch.

Second, the models we have considered for $q(s)$ require significant dependencies between stimuli to produce sufficiently large declines in precision with set size (Section 3.3.1). This would be expected to lead to large biases in subjects’ estimates of the actual stimuli and large dependencies between their estimates for different stimuli. Although such biases and dependencies have been reported previously in the VSTM literature (Wilken & Ma, 2004; Huang & Sekuler, 2010; Brady & Alvarez,
2011; Orhan & Jacobs, 2013), it is unclear whether these biases and dependencies are as large as those that would be predicted by the models considered in this chapter or by similar models.

Figure 3.9 shows the biases and dependencies predicted by the three models for \( q(s) \) considered in this chapter. The simulations shown here use a set size independent noise distribution \( p(x|s) \) (see the figure caption for more details about the simulations). For all models, biases increase with set size (Figure 3.9A-C), consistent with the results reported in a previous study (Wilken & Ma, 2004). Correlations between the estimates of different stimuli, on the other hand, decrease with set size (Figure 3.9D-F). The large negative correlations observed in the multivariate Gaussian model with random positive correlations (and to a lesser degree in the mixture of Gaussians model) are due to the fact that in some trials the dependence between two stimuli under \( q(s) \) is small and therefore produces a small bias in the estimates toward the prior mean, whereas in other trials the dependence between two stimuli under \( q(s) \) is large and produces a larger bias toward the prior mean. Over many trials, this variability in biases generates an overall negative correlation.

We note that the magnitude of biases shown in Figure 3.9A-C is not inconsistent with those reported in Wilken and Ma (2004) [see Figure 8 in Wilken and Ma (2004)]. Recently, Orhan and Jacobs (2013) also reported large biases in a VSTM recall experiment for spatial location. However, not all studies find consistently large biases in subjects’ responses. For example, Orhan and Jacobs (2013) found substantially smaller biases in a change detection task for spatial location. In addition, the constant correlations predicted by the Gaussian model with uniform correlations (Figure 3.9D) and the negative correlations predicted by the Gaussian model with random positive correlations (Figure 3.9E) and the mixture of Gaussians model (Fig-
Figure 3.9: Biases (A-C) and dependencies (D-F) predicted by the three models for $q(s)$ considered in this chapter. Biases (A-C) are computed by dividing the target stimulus value $s_t$ into 8 bins and measuring the mean and the SEM of the estimation errors, $s_t - \hat{s}_t$, for each bin. The experimenter’s prior $p(s)$ is taken to be an uncorrelated multivariate Gaussian. Dependencies (D-F) are computed by presenting a particular set of stimuli consisting of a variable stimulus $s_1$ and $N - 1$ zeros over $10^5$ simulated trials and measuring the correlation coefficient between the estimates of $s_1$ and $s_2$ over all trials (note that $s_2$ is always zero). $\Delta$ denotes the absolute difference between $s_1$ and $s_2$ ($\Delta = |s_1 - s_2|$). For biases, simulation results are shown for 6 different set sizes (from $N = 1$ to $N = 6$). For dependencies, simulation results are shown for 5 different set sizes (from $N = 2$ to $N = 6$, as correlation does not have any meaning for $N = 1$). In each plot, darker lines correspond to larger set sizes. (A) and (D) show the results for a multivariate Gaussian $q(s)$ with uniform correlations ($\rho_p = 0.96$); (B) and (E) show the results for the case of multivariate Gaussian $q(s)$ with random positive correlations ($\gamma = 0.15$); (C) and (F) show the results for the mixture of Gaussians prior ($k = 1$).
ure 3.9F) appear to be qualitatively inconsistent with the correlations reported in Orhan and Jacobs (2013).

Attributing part of the observed performance limitations (e.g., declines in precision with set size) to a general resource limitation might reduce the amount of dependence in the subject’s prior that would be needed to explain the empirical performance limitations, which might, in turn, lead to smaller biases and dependencies in the modeled responses.

Third, we recently modeled the biases reported in Wilken & Ma (2004) using probabilistic clustering models (Orhan & Jacobs, 2013). Although the models used for \( q(s) \) in that paper were different from the models considered in this chapter, we found that allowing the noise variance to vary across set sizes led to better fits to data than using a noise variance that was constant across different set sizes.

A general resource limitation is also an \emph{a priori} plausible source of performance limitations in VSTM. As discussed above, the extent to which a general resource limitation and model mismatch contribute to performance limitations in VSTM tasks is an empirical question. In this section, we thus investigate the consequences of combining the effects of model mismatch and a general resource limitation on performance limitations in VSTM. To this end, we repeat the simulations reported in the previous section, but now using a set size dependent noise distribution \( p(x|s) \). Specifically, we assume that the noise distribution is multivariate Gaussian with an isotropic, diagonal covariance matrix \( \lambda^2 I \) where \( I \) is the \( N \times N \) identity matrix and \( \lambda^2 = \lambda_{\text{min}}^2 N \) (\( \lambda_{\text{min}}^2 \) is the minimum achievable variance; we take \( \lambda_{\text{min}}^2 = 0.25 \)) consistent with the theoretical arguments discussed in Section 3.3.1 about the increase in noise variance with set size due to a general resource limitation. Other details of the simulations reported in this section are the same as the corresponding simulations.
in the previous section.

3.4.2 Memory precision-set size relationship

How is the precision–set size relationship discussed in Section 3.3.1 affected if the noise variance increases with set size? Figure 3.10A shows the recall precision as a function of set size when the experimenter’s prior is an uncorrelated multivariate Gaussian and the subject’s prior is a multivariate Gaussian with uniform correlations, where the correlation coefficient $\rho_q$ is varied between 0 and 0.99 in increments of 0.03. Figure 3.10A should be compared to Figure 3.3A. Predictably, the increase in noise variance with set size leads to steeper declines in recall precision. Without any model mismatch (i.e., $\rho_p = \rho_q = 0$), recall precision decreases as a power-law function of $N$ with an exponent of $-0.44$ (the slightly greater than $-0.5$ exponent is due to the use of a relatively informative prior in our simulations). Model mismatch decreases this exponent further, with larger mismatches leading to smaller exponents. A correlation coefficient of $\rho_q = 0.81$ leads to $\xi = -0.61$ ($R^2 = 0.99$) and $\rho_q = 0.84$ produces $\xi = -0.65$ ($R^2 = 0.99$). This result suggests a possible explanation of the steeper than predicted declines in precision observed in VSTM studies. Recall that the theoretical arguments discussed in Section 3.3.1 predict the noise variance to scale as $N$, which in turn corresponds to a precision (inverse standard deviation) that scales as $N^{-0.5}$ (assuming a relatively non-informative prior). But, most VSTM studies report power-law relationships between recall precision and set size with exponents ranging from $-0.60$ to $-0.75$ (Bays & Husain, 2008; Bays et al., 2011; Van den Berg et al., 2012). A possible explanation of this discrepancy is that the additional decline in precision, not explained by the increase in noise variance, might be caused by model mismatch.
Figure 3.10: (A) Recall precision as a function of set size when $p(s)$ is an uncorrelated Gaussian ($\rho_p = 0$) and $q(s)$ is a multivariate Gaussian with uniform non-negative correlations $\rho_q$. Precision-set size curves are shown for 34 different values of $\rho_q$ from 0 to 0.99 in increments of 0.03. Lighter colors represent lower $\rho_q$ values. (B) Similar to A, but $p(s)$ is a multivariate Gaussian with uniform correlations with $\rho_p = 0.9$. (C) Exponents $\xi$ of power-law fits to precision-set size curves for each pair of $(\rho_p, \rho_q)$ values. (D) Results for the mixture of Gaussians model. $p(s)$ is a multivariate Gaussian with uniform non-negative correlations, and $q(s)$ is a mixture of Gaussians with the same parameter values used in previous simulations ($k = 1$). $\rho_p$ is varied from 0 to 0.99 in increments of 0.03. Each curve represents the precision-set size curve for a different $\rho_p$ value. Lighter colors represent lower $\rho_p$ values.
Introducing correlations in the experimenter’s prior ($\rho_p = 0.9$) leads to overall increases in recall precision and consequent increases in $\xi$ (see Figure 3.10B). For instance, $\rho_q = 0$ yields an exponent of $\xi = -0.44$ and $\rho_q = 0.9$ yields $\xi = -0.22$. However, in no cases do we find an inverse set size effect observed in the previous section where the noise distribution was set size independent. This is because the dependencies in the experimenter’s prior are not strong enough to overcome the effect of increasing noise with set size. Figure 3.10C shows the power-law exponents $\xi$ for all pairs of $(\rho_p, \rho_q)$ values tested. Qualitatively similar results were obtained when the subject’s prior was a multivariate Gaussian with random positive correlations (not shown). Figure 3.10D shows the simulation results for the mixture of Gaussians prior for $q(s)$ with $k = 1$. For the mixture of Gaussians prior, simulations with $\rho_p = 0$ yield $\xi = -0.60$ ($R^2 = 0.99$) and simulations with $\rho_p = 0.9$ yield $\xi = -0.25$ ($R^2 = 0.97$).

For the remaining simulations in this section, we use parameter values for the models that, when combined with the set size dependent noise distribution used in this section, produce power-law exponents for the precision–set size relationship that are in the empirically observed range of $-0.60$ to $-0.75$. In particular, we use $\rho_q = 0.81$ for the multivariate Gaussian model with uniform correlations, $\gamma = 0.3$ for the multivariate Gaussian model with random positive correlations and $k = 1$ for the mixture of Gaussians model.

### 3.4.3 Variability in encoding precision

Figure 3.11 shows the simulation results for variability in encoding precision when the noise variance increases with set size. For the multivariate Gaussian $q(s)$ with random positive correlations ($\gamma = 0.3$), the results are qualitatively very similar to
Figure 3.11: Variability in encoding precision when the noise variance increases with set size. (A)-(C) show the results for a multivariate Gaussian $q(s)$ with random positive correlations ($\gamma = 0.3$), (D)-(F) show the results for the mixture of Gaussians model ($k = 1$). The formats of the plots are identical to those in Figure 3.5.

the results we obtained in the previous section with a set size independent noise distribution. When the experimenter’s prior is an uncorrelated Gaussian, there is significant variability in encoding precision for all set sizes except $N = 1$ (see Figure 3.11A-B). This variability is significantly reduced when the experimenter’s prior displays strong correlations ($\rho_p = 0.9$; Figure 3.11C).

For the mixture of Gaussians prior, there is significant variability in encoding precision only for $N = 4$ when the experimenter’s prior is an uncorrelated Gaussian
(Figure 3.11D-E), and for all set sizes except \( N = 1 \) when the experimenter’s prior has strong correlations (Figure 3.11F). In contrast to the results reported in Section 3.3.2 with a set-size independent noise distribution, overall variability in encoding precision is significantly larger for \( \rho_p = 0.9 \) than for \( \rho_p = 0 \). As discussed in Section 3.3.2, this is due to the relatively larger noise variances used in the simulations reported in this section. Through simulations, we confirmed that using a larger but set size independent noise variance produces results that are qualitatively quite similar to those shown in Figure 3.11D-F.

It is important to emphasize that for any given set size, as in Section 3.3.2, it is still the properties of the subject’s prior and its interaction with the experimenter’s prior that primarily generate variability in encoding precision. The set size dependence of the noise distribution by itself cannot explain this variability. The magnitude of the noise variance only modifies the amount of variability in encoding precision.

3.4.4 Different set-size dependencies for initial encoding rate and asymptotic precision

Figures 3.12 and 3.13 show the simulation results on the time course of recall precision when the noise variance increases with set size. Figure 3.12 shows the results for the case where the experimenter’s prior \( p(s) \) is an uncorrelated Gaussian and Figure 3.13 shows the results for the case where \( p(s) \) is a multivariate Gaussian with uniform correlations (with \( \rho_p = 0.9 \)).

For the simulations reported in this subsection, we assume that the precision of the noise distribution changes with presentation time according to Equation 3.14 as in the simulations reported in the previous section. In addition, however, we also assume that \( P_{\text{max}} \) changes with set size according to \( P_{\text{max}} = \kappa N^{-0.5} \), where \( \kappa \) is the
Figure 3.12: Time course of recall precision when the noise variance increases with set size and the experimenter’s prior $p(s)$ is an uncorrelated Gaussian. The formats of the plots are identical to those in Figure 3.6. (A) and (D) show the results for a multivariate Gaussian $q(s)$ with uniform correlations ($\rho_q = 0.81$). (B) and (E) show the results for a multivariate Gaussian $q(s)$ with random positive correlations ($\gamma = 0.3$). (C) and (F) show the results for the mixture of Gaussians prior ($k = 1$).
maximum achievable precision (obtained for \( N = 1 \)). We set \( \kappa = 2 \). Importantly, however, we assume that \( \tau \) in Equation 3.14 does not depend on the set size (we set \( \tau = 0.1 \)).

The results are qualitatively similar to the results obtained with a set size independent noise distribution (see Section 3.3.3). Different set-size dependencies of the initial encoding rate and the asymptotic precision are successfully captured by all three models (Figure 3.12). The models also yield power-law exponents that are similar to the values reported in Bays et al. (2011): \(-1.01 \pm 0.14\) for the initial encoding rate and \(-0.60 \pm 0.12\) for the asymptotic precision.

Introducing strong correlations in the experimenter’s prior (\( \rho_p = 0.9 \)) also produces qualitatively similar effects to those found in the previous section with a set size independent noise distribution (see Section 3.3.3). In particular, correlations in the experimenter’s prior reduce the difference between the power-law exponents for the initial encoding rate and the asymptotic precision (Figure 3.13). Unlike in the previous section, however, only the multivariate Gaussian \( q(s) \) with uniform correlations displays a crossover between the exponents for the initial encoding rate and asymptotic precision (Figure 3.13D). With the particular parameter values used in these simulations, the other two models predict only a reduction in the difference between the exponents without a change in their ordinal relationship.

### 3.5 Possible changes in the noise distribution due to adaptation to natural stimulus statistics

So far, we have assumed that consequences of adaptation to natural scene statistics are confined to the prior distribution or, in other words, that adaptation to natural
Figure 3.13: Time course of recall precision when the noise variance increases with set size and the experimenter’s prior $p(s)$ is a multivariate Gaussian with uniform correlations ($\rho_p = 0.9$). The formats of the plots are identical to those in Figure 3.12.
scene statistics only changes the prior distribution. However, adaptation can change the noise distribution as well. The essential idea is that if a stimulus is common in the natural visual environment, it should be encoded with better precision under the noise distribution at the expense of encoding less likely stimuli with worse precision. This corresponds to a stimulus-dependent expression for the variance of the noise distribution, $\sigma^2(s)$, with smaller $\sigma^2(s)$ for more likely stimuli $s$ and larger $\sigma^2(s)$ for less likely stimuli. This is similar to the stimulus-dependent noise distributions found in Stocker and Simoncelli (2006) and in Girshick, Landy and Simoncelli (2011). These previous works show that when psychophysical data from simple 2AFC discrimination experiments are fit with models assuming flexible, semi-parametric forms for the noise and prior distributions, the noise variance is found to be smaller in regions where the prior is high. This, in turn, explains the higher sensitivity in the stimulus space around regions where the prior probability is high. Wei and Stocker (2012) note that a smaller noise variance for more likely stimuli is consistent with the efficient coding hypothesis as it maximizes the mutual information between the actual stimuli and their noisy measurements subject to a resource limitation. Ganguli and Simoncelli (2010) give a neural implementation of this idea (for a one-dimensional stimulus space) by computing the optimal allocation of gains and densities for tuning functions in a neural population encoding a single one-dimensional stimulus.

Therefore, adaptation to natural scene statistics can manifest its effects both in the prior distribution (by increasing the prior probability of more likely stimuli $s$) and in the noise distribution (by decreasing the noise variance $\sigma^2(s)$ for more likely stimuli). Stocker and Simoncelli (2005) argue that a stimulus-dependent noise variance and its adaptation to stimulus statistics are necessary to account for certain repulsive biases (i.e. biases away from the adapted stimulus) observed in several
adaptation studies. Motivated by these studies that suggest the potential importance of the effects of adaptation in the noise distribution, in this section, we review and discuss the consequences of such adaptation in the noise distribution.

Specifically, we assume an isotropic diagonal noise distribution $p(x|s)$ with variance along each dimension given by:

$$\sigma^2(s) = \frac{f(N)\sigma_0^2}{\gamma + q(s)^{1/N}} \tag{3.15}$$

where $q(s)$ is the probability density function of the subject’s prior, $\gamma$ is a positive real number that prevents the denominator from becoming too small and hence determines the largest possible variance, and $f(N)$ is a function of $N$ that determines how the noise variance depends on the set size. The exponent $1/N$ in $q(s)^{1/N}$ is necessary to counteract the automatic normalization that occurs in the density when the dimensionality of $s$, i.e. the set size $N$, changes. Intuitively, the volume of the effective stimulus space increases as $l^N$ (where $l$ is the length of the effective stimulus range in the one-dimensional case) and the densities become correspondingly smaller. The exponent $1/N$ approximately cancels this dependence on $N$.\(^3\) This is desirable to make $f(N)$ the only set size dependent term in Equation 3.15. Different choices for $f(N)$ correspond to different assumptions about the relationship between the noise variance and set size. $f(N) = N$ corresponds to a linear increase in the noise variance with set size and implements the assumption of a fixed amount of memory resources (as in Section 3.4). In this section, we focus on the case $f(N) = 1$ which corresponds to a set-size independent noise variance (as in Section 3.3) and implements the assumption that the amount of memory resources increases.

\(^3\)The denominator becomes independent of $N$ only when $q(s)$ does not have any dependencies and is only approximately independent of $N$ otherwise, because the volume of the effective stimulus space increases slower than $l^N$ when there are dependencies between dimensions.
approximately as \( N \). Note that the implicit set-size dependence of \( q(s) \) is canceled by the exponent \( 1/N \) as explained above.

It is easy to see that Equation 3.15 works as desired:\(^4\) for stimuli that have high probability under the subject’s prior, the denominator is large, hence the noise variance, \( \sigma^2(s) \), is small; conversely, for stimuli that have low probability under the subject’s prior, the denominator is small, leading to a large noise variance (also see Figure 3.14A).

If the experimenter’s prior is \( p(s) \), ideally the subject should adapt their noise distribution (Equation 3.15) to \( p(s) \): they should encode stimuli that are more likely under \( p(s) \) with better precision. If, for example, \( p(s) \) is uniform over a given range, \( \sigma^2(s) \) should also be uniform. However, following the reasoning in Section 3.1, we assume that subjects cannot always achieve this optimal allocation of resources. Instead their noise distribution might reflect adaptation to a different distribution \( q(s) \) that may, for example, correspond to the statistics of \( s \) in the natural environment. This leads to suboptimal allocation of resources (within the context of the experiment), because intuitively, the subject is “wasting” their resources by encoding \( s \) that have high probability under \( q(s) \) (e.g. homogeneous configurations or configurations that correspond to smooth contours) with high precision at the expense of the remaining \( s \) and yet these configurations might have very small probability under \( p(s) \).

We refer to this situation as suboptimal encoding. Note that this type of suboptimal encoding can occur even if the resources increase with \( N \). Indeed, in this section, we consider the scenario where \( f(N) = 1 \) which corresponds to an approximately linear

\(^4\)Although Equation 3.15 qualitatively works as desired, the specific form chosen in that equation is ad hoc. Ideally, one would want to find out the optimal expression for the noise variance using, for example, population coding ideas as in Ganguli and Simoncelli (2010). Similarly, Equation 3.15 entails a diagonal noise distribution, but one can easily imagine generalizations with correlated noise. These are directions for future work.
Figure 3.14: (A) A simple demonstration of stimulus-dependent noise in a one-dimensional example. The noise precision is computed from Equation 3.15 where $q(s)$ is taken to be a Gaussian distribution with zero mean and standard deviation equal to either 1 (solid line), 2 (dashed line) or 3 (dotted line). Parameter values are $\sigma_0^2 = 0.025$ and $\gamma = 0.01$. (B-D) Empirical distributions of the noise precision over $10^5$ simulated trials for different set sizes from 1 to 6. Hotter colors represent larger set sizes. In (B), both the subject’s prior $q(s)$ and the experimenter’s prior $p(s)$ are uncorrelated Gaussians with standard deviation 2 along each dimension. In (C), $p(s)$ is an uncorrelated Gaussian with standard deviation 2 along each dimension, but $q(s)$ is a Gaussian with uniform correlations across its dimensions ($\rho_q = 0.9$ and $\sigma_q = 2$). In (D), both $p(s)$ and $q(s)$ are multivariate Gaussians with uniform correlations ($\rho_p = \rho_q = 0.9$ and $\sigma_p = \sigma_q = 2$).
increase of resources with $N$. As in Section 3.3 above, we show below that despite this increase in resources, the subject’s performance can decrease with set size as a result of the suboptimal use of these resources.

We also have to choose the prior distribution to be combined with the noise distribution described in Equation 3.15. As in the rest of this chapter, we can use the subject’s prior $q(s)$ for this purpose. Alternatively, the experimenter’s prior $p(s)$ might be used as the prior to be combined with the noise distribution (the noise distribution itself, however, still uses the subject’s prior as described in Equation 3.15).

### 3.5.1 Memory precision-set size relationship

To investigate the consequences of adaptation in the noise distribution for the precision-set size relationship, we compare four models: (i) a model with both suboptimal encoding and decoding (SED), where both the prior and the noise distribution (Equation 3.15) use the subject’s prior $q(s)$; (ii) a suboptimal encoding (SE) model where the prior is identical to the experimenter’s prior $p(s)$, but the noise distribution uses the subject’s prior $q(s)$; (iii) a suboptimal decoding (SD) model where the prior is identical to the subject’s prior, but the noise distribution uses the experimenter’s prior and finally (iv) a uniform variance (UV) model where the prior is identical to the subject’s prior and the noise variance is stimulus-independent. The UV model is identical to the models considered in Section 3.3 above.

For the subject’s prior, $q(s)$, either the Gaussian with uniform non-negative correlations or the mixture of Gaussians model is used. For the models with stimulus-dependent noise variance, the parameters $\sigma_0^2$ and $\gamma$ were chosen such that (i) the performance of these models for set size 1 was approximately equal to the performance of the models with uniform variance considered in Section 3.3 and (ii) the
stimulus-dependence of variance had an appreciable effect on model performance (large values of $\gamma$ diminish the effect of stimulus-dependence). Additional simulation details are provided in Appendix B.2.

Figure 3.15 shows the precision-set size relationships predicted by the models when the experimenter’s prior $p(s)$ is an uncorrelated Gaussian. Figure 3.15A-B shows the precision-set size curves for the SED model where $q(s)$ is a multivariate Gaussian with uniform non-negative correlations $\rho_q$ (Figure 3.15A) and a mixture of Gaussians (Figure 3.15B) respectively. Figure 3.15C-D shows the power-law exponents for different combinations of encoding and decoding schemes with either the multivariate Gaussian model (Figure 3.15C) or the mixture of Gaussians model (Figure 3.15D).

Two prominent features of the results shown in Figure 3.15 are worth highlighting. First, the models with suboptimal encoding (SE and SED) show the steepest declines in performance with set size (Figure 3.15C-D). This suggests that the models considered in Section 3.3 might yield substantially larger set size effects (smaller $\xi$) when supplemented with a suboptimal encoding scheme without having to postulate a resource limitation. For example, in the case of multivariate Gaussian $q(s)$ (Figure 3.15C), the UV model produces a power-law exponent of $\xi = -0.16$ ($R^2 = 0.89$) for $\rho_q = 0.84$, whereas the SED model (which is obtained by replacing the stimulus-independent noise variance in the UV model with a stimulus-dependent noise variance as in Equation 3.15) has $\xi = -0.39$ ($R^2 = 0.83$) for the same $\rho_q$ value. It can be argued that this bolsters the case for the strong hypothesis (Section 3.3), because now a smaller correlation in $q(s)$ is required to explain the steep declines in precision with set size.

It is important to emphasize that the models all assume, as in Section 3.3, that
the resources increase linearly with set size \( f(N) = 1 \) and therefore the set size effects are caused entirely by suboptimal encoding and/or suboptimal decoding. This can also be verified by considering what happens when there is no model mismatch in the prior or in the noise distribution, that is, when \( \rho_q = 0 \) in the SED model. In this case both the prior and the noise distribution use the experimenter’s prior \( p(s) \). The simulation results for this case are shown by the lightest line in Figure 3.15A. The precision-set size curve is approximately flat (the power-law exponent \( \xi = 0.04 \)) suggesting that the model makes efficient use of the increase in resources with set size. The empirical distribution of precisions [distribution of \( \frac{1}{\sigma(s)} \) computed from Equation 3.15 where \( s \) is drawn from \( p(s) \)] over \( 10^5 \) trials for each set size is shown in Figure 3.14B. The distributions for different set sizes look slightly different, but they are mostly overlapping. The net effect of the slight difference in the shape of the distributions over precisions is a very slight increase in precision with set size \( (\xi = 0.04) \). Figure 3.14C shows the empirical distribution of precisions over \( 10^5 \) simulated trials when \( p(s) \) is an uncorrelated Gaussian, but \( q(s) \) is a multivariate Gaussian with uniform correlations \( (\rho_q = 0.9) \). Note how the distributions shift to the left for larger set sizes.

The second prominent feature of the results shown in Figure 3.15 is the remarkably good power-law fits obtained with the suboptimal encoding schemes (SE and SED) using the mixture of Gaussians model with \( k = 1 \) or similar for \( q(s) \) (Figure 3.15B). The SED model with \( k = 1 \) yields \( \xi = -0.71 \) \( (R^2 = 0.99) \). This is a surprising result, because unlike in standard continuous resource models where the power-law relationship between precision and set size is explicitly built into the noise distribution, there is nothing in Equation 3.15 that, on the face of it, suggests a power-law relationship between precision and set size. Indeed, the power-law fits are
Figure 3.15: Simulation results for the case where $p(s)$ is an uncorrelated Gaussian. (A-B) Precision-set size relationships for the SED model where $q(s)$ is either a multivariate Gaussian with uniform non-negative correlations $\rho_q$ (A) or a mixture of Gaussians (B). For the multivariate Gaussian model (A), 9 different $\rho_q$ values were tested ranging from 0 to 0.96 in increments of 0.12. Darker colors correspond larger $\rho_q$ values. For the mixture of Gaussians model (B), 5 different $k$ values were tested: $k = 0, 0.2, 0.5, 1, 1.5$. Darker colors correspond to larger $k$ values. (C-D) Power-law exponents for different combinations encoding and decoding schemes where $q(s)$ is either a multivariate Gaussian with uniform non-negative correlations $\rho_q$ (C) or a mixture of Gaussians (D). SED: suboptimal encoding and decoding, SE: suboptimal encoding, SD: suboptimal decoding, UV: uniform variance.
significantly worse when we use the multivariate Gaussian model with non-negative correlations for \( q(s) \) (as can be visually verified by looking at the curves in Figure 3.15A). Thus, this appears to be a property of the mixture of Gaussians model (or similar models) specifically.

Figure 3.16 shows the precision-set size relationships predicted by the models when the experimenter’s prior \( p(s) \) is a correlated Gaussian with \( \rho_p = 0.96 \). Inverse set size effects can be observed for all models. The SED models, which differ from the UV models only in the stimulus-dependence of the noise distribution, produce larger inverse set-size effects than the UV models. Figure 3.14D shows the empirical distribution of precisions over \( 10^5 \) simulated trials when both \( p(s) \) and \( q(s) \) are correlated Gaussians (with \( \rho_q = \rho_p = 0.9 \)). Note how the distributions this time shift to the right for larger set sizes.

3.5.2 Variability in encoding precision

An immediate consequence of using a stimulus-dependent noise variance as in Equation 3.15 is variability in encoding precision. With a stimulus-dependent noise variance, variability in encoding precision arises even when we use a multivariate Gaussian model for \( q(s) \). In fact, using any non-uniform distribution for \( q(s) \) in Equation 3.15 would lead to some variability in encoding precision. Note that this is not the case if the noise variance is uniform. For example, if the noise variance is uniform, a multivariate Gaussian \( q(s) \) does not lead to variability in encoding precision as explained in Section 3.3.2 above. Figure 3.17 shows the variability in encoding precision predicted by the SED model using a multivariate Gaussian distribution with \( \rho_q = 0.96 \) for \( q(s) \) (Figure 3.17A-C) or using the mixture of Gaussians model with \( k = 1 \) (Figure 3.17D-F). Compared to the uniform noise variance scenario con-
Figure 3.16: Similar to Figure 3.15, but the results are shown for the case where $p(s)$ is a multivariate Gaussian with uniform correlations ($\rho_p = 0.96$).
sidered in Section 3.3.2, three differences are apparent in this figure. First, variability in encoding precision arises even when we use a multivariate Gaussian model for $q(s)$ (Figure 3.17A-C), which was not the case with a uniform noise variance. Second, for both models, variability in encoding precision arises even with set size 1. Again, none of the models with uniform noise variance yielded variability in encoding precision for set size 1. A previous study found evidence for variability in encoding precision even for set size 1 (Fougnie, Suchow & Brady, 2012), which suggests that the models with uniform noise variance considered in Section 3.3.2 might not be able to explain the full range of variability in encoding precision. Third, unlike in Section 3.3.2, introducing correlations in the experimenter’s prior did not have a drastic effect on variability in encoding precision for the SED models (Figure 3.17C and F).

### 3.6 Discussion

In this chapter, we presented a novel account of performance limitations in VSTM based on the idea that many of the main characteristics of these limitations can be understood, at least in part, as the result of a mismatch between the statistics of stimuli used in most VSTM experiments and the statistics of the same stimuli in the natural visual input that the subject’s visual system is adapted to. We formulated two hypotheses regarding the possible role of model mismatch vis-à-vis that of a general resource limitation in generating the observed performance limitations in VSTM tasks: a strong hypothesis claiming that performance limitations in VSTM tasks can be accounted for entirely by model mismatch and a weaker hypothesis claiming that model mismatch and a general resource or capacity limitation both contribute to the observed performance limitations in VSTM tasks. It is ultimately
Figure 3.17: (A-B) Results for the case of uncorrelated Gaussian \( p(s) \) and the SED model with Gaussian \( q(s) \) with uniform non-negative correlations \( (\rho_q = 0.96) \): (A) BIC score of the Gaussian fit relative to the BIC score of the \( t \)-distribution fit for each set size. Lower scores indicate better fits. Positive values indicate better fits for the \( t \)-distribution than for the Gaussian distribution. (B) For three different set sizes, the estimated Gamma distributions over precision for the best fit \( t \)-distributions. The vertical lines show the means of the Gamma distributions. (C) Similar to (A), but here the experimenter’s prior contains correlations \( (\rho_p = 0.96) \). (D-E) Similar to (A-B) except the results are for the case of uncorrelated Gaussian \( p(s) \) and the SED model with the mixture of Gaussians \( q(s) \) \( (k = 1) \). (F) Similar to (D), but the experimenter’s prior contains correlations \( (\rho_p = 0.96) \).
an empirical question to what extent these two factors contribute to performance limitations in standard VSTM studies, but that model mismatch plays at least some role in many, perhaps most, VSTM studies seems undeniable, because subjects’ responses display prominent biases (Wilken & Ma, 2004; Huang & Sekuler, 2010; Brady & Alvarez, 2011), and dependencies (Jiang, Olson & Chun, 2000; Orhan & Jacobs, 2013) and, on a trial by trial basis, their responses tend to be more accurate when, by chance, the presented visual display contains regularities (e.g., homogeneity) in a given trial even when the stimuli are drawn independently and uniformly across trials (Brady & Alvarez, 2012; Brady & Tenenbaum, 2013; Orhan & Jacobs, 2013). These results would be difficult to explain if subjects were using a prior identical to the prior used by the experimenter to generate the stimuli. To the extent that model mismatch plays some role in VSTM tasks, we believe that the role of capacity limitations in VSTM has been overstated by researchers.

Our framework makes a number of interesting predictions that can be tested experimentally. A crucial prediction of our framework is that all three effects discussed in this chapter (precision–set size relationship, variability in encoding precision and the different set size dependence of initial encoding rate and the asymptotic precision) should critically depend on the stimulus statistics used in the experiment. For the precision–set size relationship (Sections 3.3.1 and 3.4.2) and the different set size dependence of initial encoding rate and the asymptotic precision (Sections 3.3.3 and 3.4.4), we found that these effects should be reduced or even reversed if stimuli are drawn from a prior closer to the subject’s prior \( q(s) \) than the independent uniform or non-sparse priors used in most VSTM studies. As discussed in Section 3.3.2, with respect to variability in encoding precision, the relationship between the similarity between \( p(s) \) and \( q(s) \), on the one hand, and the amount of variability in encoding
precision, on the other hand, is less straightforward. In models with uniform noise variance, our simulations suggest that under low noise conditions, using stimulus distributions closer to the subject’s prior might reduce variability in encoding precision as well (see the simulation results for the mixture of Gaussians model in Section 3.4.3 for a caveat under high noise conditions). But, in models with stimulus-dependent noise variance, differences in variability are found to be less drastic (Section 3.5.2).

Consistent with the prediction of our framework regarding the precision–set size relationship, Mazyar, van den Berg and Ma (2012) recently suggested that set size effects might be caused by the use of heterogeneous displays rather than by intrinsic properties of the hypothesized capacity limitations in VSTM. These authors found that, even in tasks that had a significant VSTM component, there was no evidence for set size effects when homogeneous displays were used and, conversely, set size effects were observed in tasks that did not have a VSTM component but involved heterogeneous stimuli. As we showed in Sections 3.3.1 and 3.5.1, this is exactly as one would expect if the subject’s prior favors homogeneity, but the experimenter’s prior is more uniform over the stimulus space. Using homogeneous displays makes the experimenter’s prior more similar to the subject’s prior and is thus expected to reduce or even reverse the decline in precision with set size. Similarly, consistent with the predictions of our framework, it has been observed that subjects can hold more items in VSTM and with better precision if the stimuli are more homogeneous (Lin & Luck, 2009; Sims, Jacobs & Knill, 2012; Brady & Tenenbaum, 2013). But our results suggest that homogeneity is not the only property that would reduce, eliminate or reverse the set size effects. In general, any manipulation that would bring the stimulus statistics used in the experiment closer to the environmental statistics, or to the subject’s prior, would be expected to boost the subject’s performance and
to reduce or even reverse the set size effects. For example, using oriented patches that form smooth contours as stimuli rather than random configurations of orientations would also be predicted to reduce these effects. Our framework also crucially predicts these same effects on a trial-by-trial basis—even when the experimenter’s prior is an independent and uniform distribution, by chance, stimuli will be more or less natural in different trials according to the subject’s prior. Subjects are predicted to perform better in more natural trials than in less natural trials as has already been demonstrated experimentally (Brady & Alvarez, 2012; Brady & Tenenbaum, 2013; Orhan & Jacobs, 2013).

In theoretical neuroscience, the idea that a memory system’s storage and retrieval performance should depend on the stimulus statistics is not a new insight. It is known, for instance, that systems that appear to have restrictive memory capacity limitations for a certain type of stimulus statistic can operate in an entirely different and much less restrictive regime if a different kind of stimulus statistic is assumed, such as correlated vs. uncorrelated (Gardner, 1988) or sparse vs. non-sparse stimuli (Ganguli, Huh, Sompolinsky, 2008; Ganguli & Sompolinsky, 2010).

It could be argued that our framework is unjustifiably restrictive in assuming that subjects cannot override their natural priors and successfully adapt to stimulus statistics used in an experiment. It is important to emphasize that we do not claim that subjects cannot adapt to any aspect of the experimenter’s prior. Some aspects of the experimenter’s prior, such as the stimulus range or the mean, might be easy for subjects to learn during the course of an experiment. However, it is widely documented that subjects cannot always learn a statistical regularity, or the lack thereof, in an experimental task (Newport & Aslin, 2004; Creel, Newport & Aslin, 2004; Fiser & Aslin, 2005; Michel & Jacobs, 2007; Seydell, Knill, & Trommershäuser, 2009).
general, subjects tend to learn statistical regularities with greater ease and success when these regularities are more “natural” (i.e., when they are consistent with the statistical regularities exhibited in the natural environment). Learning “unnatural” regularities requires a much longer training period (Schwarzkopf & Kourtzi, 2008) and tends to be highly stimulus-specific (Zhang & Kourtzi, 2010). Similarly, subjects sometimes make inaccurate assumptions about the statistical structure of the stimuli used in a task. For example, they may assume dependencies between certain latent variables in the task (Turnham, Braun & Wolpert, 2011), between stimuli in a given trial (Orhan & Jacobs, 2013) or between stimuli across trials (Yu & Cohen, 2009; Huang & Sekuler, 2010) when, in fact, there are no such dependencies. These results suggest that in many, perhaps most, behavioral experiments, there is some discrepancy between the actual stimulus statistics used in the experiment and the subject’s prior. The extent of this discrepancy depends on the particular behavioral task as well as the actual stimulus statistics used in the experiment. Most of the simulations reported in this chapter used a minimal form of discrepancy between the experimenter’s prior and the subject’s prior. In particular, we assumed only a mismatch between the correlation structures of the respective priors and did not assume any mismatches between the means or the marginal variances of the two distributions.

Our framework can be extended to performance limitations in other domains. In multiple-object tracking, performance limitations similar to those in VSTM have been observed (Pylyshyn & Storm, 1988). As in VSTM, these limitations have been explained in terms of a limited amount of attentional resources making subjects’ representations noisier as the number of trajectories to be tracked is increased (Alvarez & Franconeri, 2007; Vul, Frank, Alvarez & Tenenbaum, 2009). Again, we hypothe-
size that at least part of these limitations can be explained in terms of a mismatch between the statistics of motion patterns generally used in multiple-object tracking experiments and the environmental statistics of the same variables that the subjects’ visual system would be expected to be adapted to. Indeed, it is known that, for a given set size (i.e., number of trajectories to be tracked), motion trajectories that display certain non-trivial properties are easier to track for subjects than trajectories that do not have those properties (Yantis, 1992). According to our framework, these types of trajectories can be considered as more likely trajectories under the subjects’ prior, which may, in turn, have a basis in the environmental statistics of motion trajectories.

Special populations display significant performance differences in VSTM tasks from normal adults. For example, children perform significantly worse than adults in VSTM tasks even when factors unrelated to VSTM are controlled for (Gathercole, Pickering, Ambridge, & Wearing, 2004; Burnett-Heyes, Zokaei, van der Staaij, Bays & Husain, 2012). Performance in VSTM tasks tends to be impaired in schizophrenia and some other mental illnesses (Farmer, O’Donnell, Niznikiewicz, Voglmaier, McCarley & Shenton, 2000; Fuller, Luck, McMahon & Gold, 2005). Moreover, there are significant individual differences in VSTM performance among the general population (Fukuda, Vogel, Mayr & Awh, 2010). These differences are generally explained in terms of differences in capacity (i.e., differences in the amount of attentional, neural or cognitive resources). Our framework offers a complementary account of performance differences in VSTM tasks. Our results show that differences in subjects’ prior models for the experimental stimuli will manifest themselves as differences in their performance in these VSTM tasks. Thus, for example, individual differences in VSTM performance among the general population may be at least partly due to each
individual’s slightly different prior model for the experimental stimuli or to differences in their ability to adapt these priors to different stimulus statistics, in addition to any differences in attentional, neural or cognitive resources that individuals may have. Similarly, individuals with schizophrenia may be relying on prior models substantially different from the priors used by normal adults in VSTM tasks. Consistent with this view, there is evidence suggesting that perceptual organization abilities are significantly impaired in individuals with schizophrenia (Silverstein, Kovács, Corry & Valone, 2000).

Our results raise the possibility that many phenomena in the VSTM literature attributed to intrinsic properties of the VSTM system may in fact be explained as straightforward consequences of the prior expectations of the visual system due to adaptation to the regularities in the natural visual environment and the mismatch between those prior expectations and the stimulus statistics used in many VSTM tasks. We demonstrated that in many cases it is unnecessary to postulate bottlenecks in visual attention or neural variability or other sophisticated mechanisms to account for characteristics of subjects’ performances in VSTM tasks. These characteristics fall out naturally from very general principles governing an organism’s adaptation to its natural visual environment, and probing that organism’s visual system with unnatural stimulus statistics that it is not adapted to. Almost four decades ago, Ulric Neisser advocated an ecological approach to the study of memory and warned cognitive psychologists against drawing broad conclusions about memory in general from laboratory studies using extremely artificial and unnatural stimuli (Neisser, 1976). Our results suggest that it may be time for researchers of VSTM to heed his call.
4 A Neural Implementation of Dependencies in Visual Short-Term Memory

As discussed in a previous section (Section 3.3.1), arguments from neural population coding are sometimes invoked to explain characteristics of performance limitations in VSTM. For example, it has been suggested that the decline in memory precision with set size can be explained by a constraint on the total activity of neural populations encoding multiple items, which in turn might be implemented by a divisive normalization mechanism (Ma & Huang, 2009; Wei, Wang & Wang, 2012). In this chapter, we ask the question: could the dependencies between representations of different items in VSTM observed in the experiments described in Section 2.7 above arise in simple neural populations encoding multiple items? We show that dependencies that are qualitatively similar to those we observed in these experiments indeed naturally arise as a consequence of encoding multiple items in a population of neurons with correlated responses. To show this, we take a three-pronged approach in this chapter. In Section 4.1, we consider the problem of encoding multiple stimuli with an idealized, correlated neural population and analytically derive an expression
for the Fisher information matrix (FIM) in this model, which provides an upper bound on the performance of any unbiased estimator of the stimuli. We show that this analytical expression for the FIM predicts dependencies of the type we observed in our experiments. In Section 4.2, we soften some of the restrictive assumptions about the encoding model needed for deriving an analytic expression for the FIM and adopt a more general form for our neural population. With this more general form, we use a semi-analytic expression for the FIM, as well as the maximum likelihood estimator (MLE) of the stimuli, to investigate the sensitivity of the predicted dependencies between the estimates of the stimuli to various model parameters. In Section 4.3, we simulate a more detailed and realistic network of spiking neurons and consider the encoding of multiple stimuli in this network. Using a simple but suboptimal decoding scheme, we show that this more realistic network model also predicts dependencies of the type we observed in our experiments. Together, these results suggest that dependencies between representations of different items could arise naturally from the simultaneous encoding of multiple items in a population of neurons with correlated responses.

4.1 An Analytic Expression for the Fisher Information Matrix

We consider the problem of encoding $N$ stimuli ($\mathbf{s} = [s_1, \ldots, s_N]$) in a correlated population of $K$ neurons with Gaussian noise:

$$p(\mathbf{r}|\mathbf{s}) = \frac{1}{\sqrt{(2\pi)^K \det Q(\mathbf{s})}} \exp[-\frac{1}{2}(\mathbf{r} - \mathbf{f}(\mathbf{s}))^T Q^{-1}(\mathbf{s})(\mathbf{r} - \mathbf{f}(\mathbf{s}))] \quad (4.1)$$
where \( \mathbf{r} \) is a vector containing the firing rates of the neurons in the population, \( \mathbf{f}(\mathbf{s}) \) represents the tuning functions of the neurons and \( Q \) represents the specific covariance structure chosen. More specifically, we assume a ‘limited range correlation structure’ for \( Q \) that has been analytically studied several times in the literature (Abbott & Dayan, 1999; Wilke & Eurich, 2001; Ecker, Berens, Tolias & Bethge, 2011). In a neural population with limited range correlations, the covariance between the firing rates of the \( k \)-th and \( l \)-th neurons (the \( kl \)-th cell of the covariance matrix) is assumed to be a monotonically decreasing function of the distance between their preferred stimuli (Wilke & Eurich, 2001, pp. 165-166):

\[
Q_{kl}(\mathbf{s}) = a f_k(\mathbf{s})^\alpha f_l(\mathbf{s})^\alpha [\delta_{kl} + (1 - \delta_{kl}) \exp(-\frac{|\phi_k - \phi_l|}{L})] \quad (4.2)
\]

where \( \phi_k \) and \( \phi_l \) are the preferred stimuli of the neurons, \( L \) is a length scale parameter determining the spatial extent of the correlations, and \( \delta_{kl} \) is the Kronecker delta function which evaluates to 1 if \( k = l \), and 0 otherwise. There is extensive experimental evidence for this type of correlation structure in the brain (Zohary, Shadlen & Newsome, 1994; Smith & Kohn, 2008; Cohen & Kohn, 2011). For instance, Zohary et al. (1994) showed that correlations between motion direction selective MT neurons decrease with the difference in their preferred directions. This ‘limited-range’ assumption about the covariances between the firing rates of neurons will be crucial in explaining our experimental results in terms of the FIM of a correlated neural population encoding multiple stimuli (see Section 4.2).

We are interested in deriving the FIM, \( J(\mathbf{s}) \), for our correlated neural population encoding the stimuli \( \mathbf{s} \). The significance of the FIM is that the inverse of the FIM provides a lower bound on the covariance matrix of any unbiased estimator of \( \mathbf{s} \) and also expresses the asymptotic covariance matrix of the maximum-likelihood estimate.
of \( s \) in the limit of large \( K \). The \( ij \)-th cell of the FIM is defined as:

\[
J_{ij}(s) = -E\left[ \frac{\partial^2}{\partial s_i \partial s_j} \log p(r|s) \right] \tag{4.3}
\]

Our derivation of \( J(s) \) closely follows that of Wilke and Eurich (2001). To derive an analytical expression for \( J(s) \), we make a number of assumptions: (i) all neurons encode the same feature dimension (e.g. horizontal location); (ii) the neurons can be divided into \( N \) groups where in each group the tuning functions are a function of the feature value of only one of the stimuli, i.e. \( f_k(s) = f_k(s_n) \) for neurons in group \( n \), so that the effects of other stimuli on the mean firing rates of neurons in group \( n \) are negligible; (iii) indices of the neurons can be assigned such that neurons with adjacent indices have the closest tuning function centers (regardless of the group they belong to); (iv) the centers of the tuning functions of neurons are linearly spaced with density \( \eta \). The last two assumptions imply that the covariance between neurons with indices \( k \) and \( l \) (Equation 4.2) can be expressed as \( Q_{kl} = \rho^{k-l}g_k^\alpha f_l^\alpha \) (we omitted the \( s \)-dependence of \( Q \) and \( f \) for brevity) with \( \rho = \exp(-1/(L\eta)) \). A population of neurons satisfying the assumptions above is schematically illustrated in Figure 4.1A (for \( N = 2 \)). We consider Gaussian tuning functions of the form:

\[
f_k(s) = g \exp\left(-\frac{(s - \phi_k)^2}{\sigma^2}\right),
\]

with \( \phi_k \) uniformly spaced between \(-12^\circ\) and \(12^\circ\) and \( g \) and \( \sigma^2 \) are assumed to be the same for all neurons. With these assumptions, it can be shown that (see Appendix C.1 for the derivation):

\[
J_{ij}(s) = \frac{1 + \rho^2}{a(1 - \rho^2)} \sum_{k=1}^{K} h_k^{(i)} h_k^{(j)} - \frac{2\rho}{a(1 - \rho^2)} \sum_{k=1}^{K-1} h_k^{(i)} h_{k+1}^{(j)} + \frac{2\alpha^2}{1 - \rho^2} \sum_{k=1}^{K} g_k^{(i)} g_k^{(j)} - \frac{2\alpha^2 \rho^2}{1 - \rho^2} \sum_{k=1}^{K-1} g_k^{(i)} g_{k+1}^{(j)}. \tag{4.4}
\]

\( J^{-1}(s) \) provides a lower bound on the covariance matrix of any unbiased estimator of \( s \) in the matrix sense (where \( A \geq B \) means \( A - B \) is positive semi-definite).
where \( h_k^{(i)} = \frac{1}{f_k} \frac{\partial f_k}{\partial s_i} \) and \( g_k^{(i)} = \frac{1}{f_k} \frac{\partial f_k}{\partial s_i} \).

We take the inverse of \( J(s) \), which provides a lower bound on the covariance matrix of any unbiased estimator of \( s \), and calculate correlation coefficients based on \( J^{-1}(s) \) for each \( s \). For \( N = 2 \), for instance, we do this by calculating:

\[
J_{12}^{-1}(s)/\sqrt{J_{11}^{-1}(s)J_{22}^{-1}(s)}
\]

In Figure 4.1B, we plot this measure for all \( s_1, s_2 \) pairs between \(-10^\circ\) and \(10^\circ\). We see that the inverse of the FIM predicts correlations between the estimates of \( s_1 \) and \( s_2 \) and these correlations decrease with \(|s_1 - s_2|\), just as we observed in Experiments 1A above (see Figure 2.11A; replotted in black in Figure 4.1C). The best fits to data from Experiment 1A were obtained with fairly broad tuning functions (see Figure 4.1 caption). The inverse of the FIM also predicts negative correlations when \(|s_1 - s_2|\) is large. We do not see any evidence for negative correlations in our data (see Section 4.4 for possible explanations of this discrepancy).

Intuitively, the correlations predicted by the model can be understood as follows. Consider the hypothetical neural population shown in Figure 4.1A encoding the pair \( s_1, s_2 \). In this population, it is assumed that \( f_k(s) = f_k(s_1) \) for neurons in the upper row, and \( f_k(s) = f_k(s_2) \) for neurons in the lower row. Suppose that in the upper row, the \( k \)-th neuron has the best-matching tuning function for a given \( s_1 \). Therefore, on average, the \( k \)-th neuron has the highest firing rate in response to \( s_1 \). However, since the responses of the neurons are stochastic, on some trials, neurons to the left (right) of the \( k \)-th neuron will have the highest firing rate in response to \( s_1 \). When this happens, neurons in the lower row with similar preferences will be more likely to get activated, due to the limited-range correlations between the neurons. This, in turn, will introduce correlations in an estimator of \( s \) based on \( r \) that are strongest...
Figure 4.1: (A) A population of neurons satisfying all the assumptions made in deriving the FIM. For neurons in the upper row $f_k(s) = f_k(s_1)$, and for neurons in the lower row $f_k(s) = f_k(s_2)$. The magnitude of correlations between two neurons is indicated by the thickness of the line connecting them. (B) Correlation coefficients estimated from the inverse of the FIM for all stimuli pairs $s_1, s_2$. (C) Mean correlation coefficients as a function of $|s_1 - s_2|$ (red: model’s prediction; black: combined data from all 8 subjects in Experiment 1A). Parameters: $\alpha = 0.5, g = 50, a = 1$ (these were set to biologically plausible values); other parameters: $K = 1000, \sigma = 9.0, L = 0.015$ (the last two were chosen to provide a good fit to the experimental results).

when the absolute difference between $s_1$ and $s_2$ is small.

### 4.2 A more general model and the dependence of predicted correlations on model parameters

In this section, we consider a more general population coding model and, using a semi-analytic expression for the FIM, investigate the effects of different model parameters on the predicted correlations between the estimates of the stimuli. The results presented in this section are all for $N = 2$ stimuli. Generalization to more than two stimuli is straightforward. There are two main differences between the model presented in the last section and the model considered in this section. The first difference is that in the model considered in this section, the responses of neurons
are allowed to depend on both stimuli, $s_1$ and $s_2$. The mean firing rate of the neuron $k$ in response to the stimuli $s = [s_1, s_2]$ is thus given by:

$$f_k(s_1, s_2) = w_1 f_k(s_1) + (1 - w_1) f_k(s_2)$$  \hspace{1cm} (4.6)$$

where $0 \leq w_1 \leq 1$ is a parameter that determines the influence of the first stimulus, $s_1$, on the response of the neuron. The weight for the second stimulus $s_2$ is $w_2 = 1 - w_1$, thus the mean firing rate is appropriately normalized. We assume that the neurons can be divided into two groups (or $N$ groups in the general case) such that the responses of neurons in one group are more strongly affected by $s_1$ and the weights are reversed for the neurons in the other group: for example, if $w_1 = 0.75$ and $w_2 = 0.25$ for the first group, $w_1 = 0.25$ and $w_2 = 0.75$ for the second group. Note that the model in the last section is a special case of the model in this section with $w_1 = 1.0$. Simulations with models that had a less strict separation between the two groups (for example, models where the weights for individual neurons in the two groups are randomly drawn from two distributions with different means) yielded qualitatively similar results.

The second difference from the model presented in the last section is that neural correlations within the same group and those between the groups are now allowed to differ. For a pair of neurons, $k$ and $l$, within the same group, neural correlations are given by:

$$R_{kl} = c_0 \exp(-\frac{|\phi_k - \phi_l|}{L}) + (1 - c_0) \delta_{kl}$$  \hspace{1cm} (4.7)$$

where $\phi_k$ and $\phi_l$ are the preferred stimuli of the neurons, $c_0$ is the maximum correlation between non-identical neurons and $L$ is the scale length of the correlations. Thus, $R_{kk} = 1$ for all neurons $k$; for $k \neq l$, $R_{kl}$ attains a maximum value of $c_0$ and
exponentially decays as a function of the absolute difference between the preferred stimuli of the neurons. For neurons across the two groups, on the other hand, neural correlations are given by:

\[ R_{kl} = Wc_0 \exp\left( -\frac{|\phi_k - \phi_l|}{L} \right) \]  

(4.8)

where \( 0 \leq W \leq 1 \) is a scaling parameter that determines the reduction in neural correlations across groups. Again, the model presented in the last section is a special case of the model presented in this section with \( W = 1 \). In this section, we also consider neural populations encoding circular variables. For circular variables, differences, \( \phi_k - \phi_l \), in Equations 4.7-4.8 are replaced by angular differences and thus neural correlations within and across groups are given by the following equations respectively (Ecker et al., 2011):

\[ R_{kl} = c_0 \exp\left( -\frac{\text{arg exp}[i(\phi_k - \phi_l)]}{L} \right) + (1 - c_0)\delta_{kl} \]  

(4.9)

\[ R_{kl} = Wc_0 \exp\left( -\frac{\text{arg exp}[i(\phi_k - \phi_l)]}{L} \right) \]  

(4.10)

Given these correlations between neurons within and across groups, covariances are determined according to:

\[ Q_{kl} = af_k(s_1, s_2)\alpha f_l(s_1, s_2)\alpha R_{kl} \]  

(4.11)

As in the previous section, we assume a Gaussian noise model for the responses of neurons, as given in Equation 4.1. We will assume \( \alpha = 0.5 \) throughout this section which corresponds to Poisson-like noise where the noise variance scales linearly with the mean response.
For the Gaussian noise model given in Equation 4.1, it can be shown that the $ij$-th cell of the FIM is given by (Abbott & Dayan, 1999):

$$J_{ij}(s) = \frac{\partial f^T}{\partial s_i} Q^{-1} \frac{\partial f}{\partial s_j} + \frac{1}{2} \text{Tr\left[} \frac{\partial Q}{\partial s_i} Q^{-1} \frac{\partial Q}{\partial s_j} Q^{-1} \text{\right]}$$ (4.12)

where $f(s_1, s_2)$ is a vector containing the mean firing rates of all neurons in response to the stimuli $s = [s_1, s_2]$. We use Equation 4.12 to compute the predicted correlation between the estimates of the two stimuli as in the previous section (Equation 4.5), but without deriving an analytic expression for the matrix inverses in Equation 4.12 as we did before (i.e. in our simulations below, we directly invert the matrix $Q$ in Equation 4.12).

For the neural tuning functions $f_k$ in Equation 4.6, we either use Gaussian functions in which case $f_k$ is given by:

$$f_k(s_i) = g \exp\left(-\frac{(s_i - \phi_k)^2}{\sigma^2}\right) + \nu$$ (4.13)

or von Mises tuning functions in which case $f_k$ is given by (Ecker et al., 2011):

$$f_k(s_i) = \beta \exp(\gamma[\cos(s_i - \phi_k) - 1]) + \nu$$ (4.14)

where $\nu$ is the baseline firing rate and $\phi_k$ is the preferred stimulus of the neuron. Von Mises functions are frequently used for modeling the tuning functions for circular variables such as orientation or motion direction.

To investigate the dependence of predicted correlations on different model parameters, we first choose a biologically plausible default setting of the parameters and then for each parameter of interest we vary the value of that parameter while
holding the remaining parameters fixed at their default values. For the model with Gaussian tuning functions, we consider a scenario where the preferred stimulus values of the neurons in each group are uniformly spaced between $-10$ and $10$. The stimuli presented to the model are pairs, $(s_1, s_2)$, of the form $(-i, i)$ where $i$ ranges from 0 to 10 in steps of 0.5. Thus, a total of 21 pairs are presented to the model. The absolute difference between the two stimuli, $|s_1 - s_2|$, ranges from 0 (for $s_1 = s_2 = 0$) to 20 (for $s_1 = -10, s_2 = 10$). The default parameter values for this case are listed in the caption of Figure 4.2.

For the model with von Mises tuning functions, the preferred stimulus values of the neurons in each group are uniformly distributed around the circle and the stimuli presented to the model are 16 pairs, $(s_1, s_2)$, of the form $(-i, i)$ where $i$ ranges uniformly between 0 and $\pi/2$. The default parameter values for this case are listed in the caption of Figure 4.3.

Figure 4.2 shows the dependence of predicted correlations on various model parameters for the model with Gaussian tuning functions. Figure 4.2A shows that for stimulus weights, $w_1$, smaller than about 0.6, the model predicts negative correlations for small $|s_1 - s_2|$ and positive correlations for large $|s_1 - s_2|$. For larger $w_1$ values, this pattern is reversed and the predicted correlations look more similar to the correlations observed in our experiments (except for the predicted negative correlations which do not exist in our data). The predicted negative correlations can be understood as follows. When doing inference over, or estimating, the stimuli $s_1$ and $s_2$ given the neural responses $r$, Equation 4.6 negatively couples the variables $s_1$ and $s_2$. This is because, for a given neural response $r_k$ (considering just a single neuron for simplicity), if $s_1$ is slightly shifted away from the neuron’s preferred stimulus value $\phi_k$, $s_2$ should be shifted toward $\phi_k$ in order to keep the response $r_k$ the same.
Figure 4.2: The effect of model parameters on the predicted correlations between the estimates of the stimuli. Results are shown for the model with Gaussian tuning functions. The default parameter values are as follows: $c_0 = 0.4$, $L = 0.5$, $w_1 = 0.75$, $\nu = 1$, $g = 40$, $\sigma^2 = 25$, $W = 0.99$, $a = 1$, $K = 1000$ (total number of neurons, with $K/2$ neurons in each group). In each plot, one of the parameters is varied while the remaining parameters are held fixed at their default values. Different lines correspond to simulation results with different parameter values, with darker colors always indicating larger parameter values. (A) $w_1$ is varied from 0.5 to 1.0 in steps of 0.05. (B) $W$ is varied from 0.09 to 0.99 in steps of 0.1. (C) $K$ is varied from 1000 to 5000 in steps of 1000. (D) $L$ is varied from 0.1 to 1000 in 10 logarithmically spaced points. (E) $c_0$ is varied from 0 to 0.5 in steps of 0.1. (F) $\sigma^2$ takes the values 4, 9, 16, 25, 36, 49.
Figure 4.2B shows that $W$ has to be sufficiently large (larger than about 0.8), that is, the reduction in neural correlations across groups should be relatively small to produce positive correlations for small $|s_1 - s_2|$ values. If the reduction in neural correlations across is large, the neural correlations across groups (which tend to produce positive correlations between the estimates of the stimuli) cannot overcome the negative correlations introduced through the coupling of $s_1$ and $s_2$ in Equation 4.6, as explained above.

Figure 4.2C shows that the results do not depend on the number of neurons, $K$, significantly, if $K$ is large enough. Figure 4.2D shows that the limited-range assumption regarding the neural correlations, i.e. that the neural correlations decrease as a function of the distance between the preferred stimulus values of the neurons, is crucial to obtain positive correlations for small $|s_1 - s_2|$ values. This can be seen by looking at what happens when we set $L$ to a very large value (e.g. $L = 1000$). In this case, neural correlations become practically uniform, that is, they do not depend on the distance between the preferred stimulus values. The predicted correlations between the estimates (for $L = 1000$) are shown by the darkest line in Figure 4.2D. The predicted correlations display a complex inverse-U shape that is quite different from the predicted correlations for limited-range neural correlations (e.g. when $L = 1$).

Figure 4.2E shows that neural correlations are necessary to obtain positive correlations for small $|s_1 - s_2|$ values. When there are no neural correlations ($c_0 = 0$; the lightest line in Figure 4.2E), the negative coupling between $s_1$ and $s_2$ through Equation 4.6 dominates the predicted correlations between the estimates. Interestingly, however, introducing even a small amount of neural correlations (for example, $c_0 = 0.1$; note that $c_0$ is the maximum correlation between any two non-identical neurons, not the average correlation) can reverse this pattern.
Finally, Figure 4.2F shows that increasing the width of the tuning functions increases the spatial scale of the predicted correlations.

Results for the model with von Mises tuning functions (Figure 4.3) are qualitatively very similar to the results for the model with Gaussian tuning functions. The only difference appears to be the stronger dependence of the predicted correlations on the number of neurons, $K$, in the model with von Mises tuning functions (Figure 4.2C). We observed that the effects of increasing the number of neurons could be compensated by decreasing the length scale, $L$, of neural correlations (not shown). The reason for this discrepancy between the two models is not immediately clear.

Figure 4.4 shows the effect of the noise level, $a$ (see Equation 4.11), on the predicted correlations. In both models, for large noise levels, the negative coupling between $s_1$ and $s_2$ introduced through Equation 4.6 dominates the predicted correlations between the estimates. Therefore, the noise level should be moderate enough to produce positive correlations for small $|s_1 - s_2|$ values.

We also compared the predictions obtained from the FIM with the performance of the maximum-likelihood estimator (MLE). Asymptotically, that is, when $K \rightarrow \infty$, the performance of the MLE converges to the FIM. Figure 4.5A and C show the correlations between the estimates of the two stimuli and Figure 4.5B and D show the variance of the estimate of the first stimulus. The results show that, with $K = 1000$, the fit between the performance of the MLE and the asymptotic predictions obtained from the FIM is quite good.
Figure 4.3: The effect of model parameters on the predicted correlations between the estimates of the stimuli. Results are shown for the model with von Mises tuning functions. The default parameter values are as follows: $c_0 = 0.4$, $L = 0.5$, $w_1 = 0.75$, $\nu = 1$, $\beta = 19$, $\gamma = 2$ [the last three parameter values are taken from Ecker et al. (2011)], $W = 0.99$, $a = 1$, $K = 1000$ (total number of neurons, with $K/2$ neurons in each group). In each plot, one of the parameters is varied while the remaining parameters are held fixed at their default values. Different lines correspond to simulation results with different parameter values, with darker colors always indicating larger parameter values. (A) $w_1$ is varied from 0.5 to 1.0 in steps of 0.05. (B) $W$ is varied from 0.09 to 0.99 in steps of 0.1. (C) $K$ is varied from 1000 to 5000 in steps of 1000. (D) $L$ is varied from 0.1 to 1000 in 10 logarithmically spaced points. (E) $c_0$ is varied from 0 to 0.5 in steps of 0.1. (F) $\gamma$ is varied from 1 to 6 in steps of 1.
Figure 4.4: The effect of noise level, $a$, on the predicted correlations between the estimates of the stimuli. Results are shown for the model with Gaussian (A) and von Mises (B) tuning functions. In each case, $a$ is varied from $10^{-3}$ to $10^{3}$ in 7 logarithmically spaced points. Darker lines correspond to larger values of $a$. 
Figure 4.5: Predictions from the FIM compared with the performance of the ML estimator. (A) and (B) show the results for the model with von Mises tuning functions; (C) and (D) show the results for the model with Gaussian tuning functions. (A) and (C) show the correlations between the estimates of the two stimuli; (B) and (D) show the variance of the estimate of the first stimulus, $\hat{s}_1$. Results for the MLE were obtained from 1000 simulated trials for each stimulus pair $(s_1, s_2)$. For both models, all parameters were set to their default values. ML estimates were numerically computed from Equations 4.1, 4.6 and 4.11 using a Nelder-Mead simplex method (the fminsearch function in Matlab).
4.3 Encoding and decoding multiple stimuli in a network of spiking neurons

The results presented in the previous sections show that correlations between the estimates of different stimuli that are qualitatively similar to those observed in our experiments can arise from the encoding of multiple stimuli in a population of neurons with correlated responses. However, these results were obtained with idealized, firing rate neurons and optimal (or asymptotically optimal) decoding schemes (using either the FIM or the MLE). Therefore, it is not clear if these results would generalize to population coding with spiking neurons and simpler but suboptimal decoding schemes.

To address these questions, we simulated a more detailed and realistic network of spiking neurons and used a simple but suboptimal read-out scheme to decode the neural responses. The network consisted of two layers. In the input layer, there were 169 Poisson neurons arranged in a 13 × 13 grid with linearly spaced receptive field centers between −12° and 12° along both horizontal and vertical directions. Again, we only consider the case \( N = 2 \) in this section. On a given trial, the firing rate of the \( k \)-th input neuron was determined by the following equation:

\[
    r_k = g_{in}[\exp\left(-\frac{||x_1 - \phi_k||}{\sigma_{in}}\right) + \exp\left(-\frac{||x_2 - \phi_k||}{\sigma_{in}}\right)]
\]

Here \( || \cdot || \) denotes the Euclidean norm, \( x_i \) is the vertical and horizontal locations of the \( i \)-th stimulus, \( \phi_k \) is the two-dimensional receptive field center of the input neuron, \( g_{in} \) is a gain parameter and \( \sigma_{in} \) is a scale parameter (both assumed to be the same for all input neurons).
The output layer consisted of simple leaky integrate-and-fire neurons. There were 169 of these neurons arranged in a $13 \times 13$ grid with the receptive field center of each neuron matching the receptive field center of the corresponding neuron in the input layer. We induced limited-range correlations between the output neurons through receptive field overlap, although other ways of introducing limited-range correlations can be considered such as through local lateral connections. Each output neuron had a Gaussian connection weight profile centered at the corresponding input neuron and with a standard deviation of $\sigma_{out}$. The output neurons had a threshold of -55 mV and a reset potential of -70 mV. Each spike of an input neuron $k$ instantaneously increased the voltage of an output neuron $l$ by $10w_{kl}$ mV, where $w_{kl}$ is the connection weight between the two neurons and the voltage decayed with a time constant of 10 ms. We implemented the network in Python using the Brian neural network simulator (Goodman & Brette, 2008).

We simulated this network with the same display configurations presented to our subjects in Experiment 1A described above (Section 2.7). Each of the 36 configurations was presented 96 times to the network, yielding a total of 3456 trials. For each trial, the network was simulated for 100 ms and its estimates of $s_1$ and $s_2$ were read out using a suboptimal decoding strategy. Specifically, to get an estimate of $s_1$, we considered only the row of neurons in the output layer whose preferred vertical locations were closest to the vertical location of the first stimulus and then we fit a Gaussian function (with amplitude, peak location and width parameters) to the activity profile of this row of neurons and considered the estimated peak location as the model’s estimate of $s_1$. We did the same for obtaining an estimate of $s_2$. Alternative suboptimal decoding schemes, such as fitting two-dimensional Gaussian functions to the activity profile of all output neurons and taking the first coordinates
Figure 4.6: (A) Results for the spiking network model. The actual display configurations $s$ are represented by magenta dots, the estimated means based on the model’s responses are represented by black dots and the estimated covariances are represented by contours (with red contours representing cases for which the two dimensions were significantly correlated at the $p < 0.05$ level). (B) The mean correlation coefficients (and standard errors of the means) as a function of $|s_1 - s_2|$ (red: model prediction; black: combined data from all 8 subjects in Experiment 1A). Model parameters: $g_{in} = 120$, $\sigma_{in} = 2$, $\sigma_{out} = 2$. Parameters were chosen to provide a good fit to the experimental data.

of the peak locations as the model’s estimate of $s_1$ and $s_2$, yielded similar results.

Figure 4.6 shows the results for the network model. Similar to our experimental results, the spiking network model predicts correlations between the estimates of $s_1$ and $s_2$ and these correlations decrease with $|s_1 - s_2|$ (correlations differed significantly across different bins as indicated by a one-way ANOVA: $F(5, 30) = 22.9713, p < 10^{-8}$; see Figure 4.6). Interestingly, the model was also able to replicate the biases toward the center of the screen observed in the experimental data. This is due to the fact that output neurons near the center of the display tended to have higher activity levels, since they have more connections with the input neurons compared to the output neurons near the edges of the display.
4.4 Discussion

In this chapter, we showed that correlations qualitatively similar to those we observed in the experiments described in Chapter 2 can arise naturally from the optimal coding of multiple stimuli in a population of neurons with correlated responses and determined the dependence of the pattern of correlations between the estimates of different stimuli on various properties of the neural responses (Section 4.2). In Section 4.3, we showed that similar correlations can be obtained in population coding with a network of spiking neurons using simple but suboptimal decoding schemes.

The optimal population coding models (Sections 4.1 and 4.2) predict negative correlations between the estimates for certain $|s_1 - s_2|$ values. As noted previously, we did not find any evidence for negative correlations in our experimental data. A possible explanation of this discrepancy might be that we have not tested a sufficiently large range of $|s_1 - s_2|$ values to observe negative correlations. Although this may be true for Experiments 2 and 3, the tested range was quite large in Experiment 1A-B. A more plausible explanation for the lack of negative correlations in our data is that subjects may be using suboptimal decoding schemes that reduce or cancel the negative correlations that would have been observed with an optimal decoding scheme (e.g. MLE) without significantly affecting the positive correlations. It would be interesting to determine what kind of suboptimal decoding schemes would satisfy this property.

It is interesting to note that for certain parameter regimes including the default setting of the parameter values, the pattern of correlations predicted by optimal population coding looks remarkably similar to the pattern of correlations predicted by one of the models considered in the last chapter, namely the model with a mixture of Gaussians type prior with a correlated component and an uncorrelated component.
(compare, for example Figure 4.2F for small $\sigma^2$ values with Figure 3.9F). Both models predict Mexican-hat type curves with positive correlations for small $|s_1 - s_2|$ values and prominent negative flanks. An intriguing possibility is that limited-range neural correlations assumed in this chapter might be an adaptation of the neural responses to optimally encode a mixture of Gaussians type prior with a correlated component and an uncorrelated component. As mentioned in the previous chapter, this type of prior provides a good model of feature statistics in natural images with occlusion (Pitkow, 2010).
5 Toward Ecologically Realistic Theories in Visual Short-Term Memory Research

Previous chapters formulated novel theoretical ideas for understanding the encoding of simple visual scenes typically used in standard VSTM studies. This chapter takes a critical look at the use of such impoverished scenes that lack the rich statistical structure displayed by stimuli that are more representative of the natural environment and discusses how the study of VSTM can be extended to ecologically more realistic stimuli and tasks, as well as possible implications of such an extension.

5.1 Introduction

Research on visual short-term memory (VSTM) generally uses artificial visual displays consisting of simple objects with easily parameterizable features and with no statistical structure within or between objects. These displays are used to address questions such as the nature of capacity limits or the units of storage in VSTM (Brady, Konkle, & Alvarez, 2011). This choice of stimuli confers many advantages
to the design of experiments and to the interpretation of results obtained from such experiments. Among the advantages of using artificial and simple stimuli are the ease with which such stimuli can be generated and manipulated, the fact that they make the formulation and testing of hypotheses straightforward, their relative unfamiliarity for subjects, and, lastly, their “bare bones” character, stripped of features that are irrelevant to the hypothesis under question. The last two properties help to minimize the effects of irrelevant prior knowledge or assumptions that subjects might bring to a VSTM task.

Despite these advantages, psychologists have also been aware of the potential problems that are associated with the use of artificial stimuli and tasks in experimental studies (Brunswik, 1943, 1955; Neisser, 1976). The main concern here is the danger that these stimuli and tasks might be too artificial to give an accurate reflection of the problems faced by an observer in the natural world. Given that many aspects of perception and cognition can be profitably thought of as rational solutions or adaptations to problems that observers (and actors) encounter in their natural environments (Anderson, 1990; Geisler, Perry, & Ringach, 2009), unnatural stimuli and tasks might lead to misleading characterizations of perceptual and cognitive processes.

In this chapter, our goal is to take a critical look at VSTM studies that use impoverished, unnatural scenes lacking the rich statistical structure displayed by stimuli that are more representative of the natural environment. Although we acknowledge that experiments using artificial displays with simple statistical structure are often useful first steps in elucidating fundamental perceptual and cognitive processes, we argue that whether and how results obtained from such experiments would generalize to stimuli and tasks that are more representative of the natural environment
should always be considered carefully. If there are any doubts about the generalizability of the results, experimental stimuli and procedures will need to be refined accordingly. In the following sections, we discuss how the study of VSTM can be extended to ecologically more realistic stimuli, as well as possible implications of such an extension.

The chapter is organized as follows. The next two sections emphasize the continuity between visual perception and VSTM by reviewing recent evidence for their common neural and representational substrates (Section 5.2), and by reviewing evidence that knowledge of visual statistics aids both visual perception and VSTM (Section 5.3). We use this continuity to suggest a possible way of modeling the representation in VSTM of natural stimuli with rich statistical structure (Section 5.4). We then discuss why it is important to take the question of the ecological realism of the stimuli and tasks used in VSTM studies seriously (Section 5.5), and end with a discussion of a concrete example from neuroscience of how a good balance between experimental control and ecological realism can be achieved in practice (Section 5.6), which we believe could be relevant for research on VSTM as well.

\section{5.2 Continuity between visual perception and visual memory}

As discussed in Chapters 2 and 3, one of the most popular questions in the VSTM literature addresses the nature of capacity limits in VSTM. For example, from a single fixation, how much information can we encode and then maintain in VSTM across a brief interval and what is the nature of this capacity limit? The notions of information and capacity are derived from information theory and, to be meaningful,
require well-defined stimulus spaces. This, together with reasons mentioned in the Introduction, explains why studies on the nature of capacity limitations in VSTM often use stimulus displays containing simple items with simple features. One consequence of this choice is that theories or models attempting to explain the nature of VSTM capacity limitations are specifically tailored to these types of stimulus displays. For example, these theories put forward rival explanations of how a fixed amount of resources in VSTM gets distributed over some simple “items” in a display. Some theories (so-called high-threshold models) have claimed that only a few items can be selected from the display and encoded with near-perfect resolution, while the remaining items are not encoded at all (Luck & Vogel, 1997; Rouder, Morey, Cowan, Zwilling, Morey, & Pratte, 2008). Other theories posit either a discrete (Zhang & Luck, 2008) or continuous resource (Bays & Husain, 2008) that can be distributed among the items either evenly (Wilken & Ma, 2004) or unevenly (Van den Berg, Shin, Chou, George, & Ma, 2012).

However, even leaving aside the difficulties surrounding the issue of what constitutes an “item” in real-world situations (what constitutes an item in the rightmost image in Figure 5.1A?), it seems evident that these accounts can give, at best, an incomplete picture of the nature of capacity limitations in VSTM for the simple reason that we do not just see and remember “items” in the natural world. This can be appreciated by considering all the different kinds of visual information that can be held in VSTM. Figure 5.1 presents some examples. These examples suggest that we can remember visual information about textures (Figure 5.1B), material properties of objects (Figure 5.1C), complex real-world scenes that are presumably encoded at multiple levels of abstraction (Figure 5.1A) and even actions (Giese & Poggio, 2003).
such as walking or running\(^1\) (Figure 5.1D). Consider, for instance, the leftmost image in Figure 5.1B. If we are briefly shown the texture in this figure, we will remember something about it. Subjectively, it seems sensible to suppose that we will not be able to remember every detail of the texture and that our memory will be more impoverished than our perception of the same texture (see Huang & Sekuler, 2010, for a direct comparison of the precision of visual perception with the precision of VSTM in a simple estimation task). Nonetheless, we will be able to remember something about it. Why is our memory more impoverished than our perception? Why can we not remember every visual detail in this texture and what exactly can we remember about it? Similarly for real-world scenes such as the one shown in the rightmost image in Figure 5.1A. What exactly can we remember about this image from a brief presentation? Existing theories on the nature of capacity limitations in VSTM have little to say about these issues because they are specifically tailored to the types of artificial stimuli typically used in VSTM studies.

The general message from Figure 5.1 seems to be clear. In terms of their representational contents, there is no qualitative difference between visual perception and visual memory—whatever can be visually perceived can also be held in VSTM albeit in a possibly more impoverished, lower resolution form.

The idea that perception and short-term memory in general (and visual perception and VSTM in particular) have common neural and representational substrates has a long history in memory research. For example, a prominent theory of working memory holds that short-term maintenance of information consists of temporary

\(^{1}\) It may be claimed that the last example does not constitute visual information. However, we can perceive and remember rich visual details about, for example, somebody’s manner of walking that cannot be not easily classified as anything other than ‘visual’. It is doubtful whether there is a justifiable sense of ‘visual’ according to which the shape of an object counts as visual information, but the manner of movement does not.
Figure 5.1: Example natural images highlighting a variety of different types of visual information that we can perceive and remember in the real world. (A) Complex real-world scenes. The first two images are from the SUN database (Xiao, Hays, Ehinger, Oliva & Torralba, 2010); the rightmost image is from http://www.flickr.com/photos/13774569@N07/5923460837/. (B) Textures. Images are from the UIUC texture database (Lazebnik, Schmid & Ponce, 2005). (C) Material properties such as glossiness, rigidity, roughness, being made of leather, etc. Images are from the Flickr Material Database (Sharan, Rosenholtz & Adelson, 2009). (D) Actions. Images are from the SUN database.
re-activation of representations in long-term memory thought to be implemented in posterior cortical areas that are also involved in the initial perceptual encoding (Cowan, 1995; Fuster, 1997; Ruchkin, Grafman, Cameron, & Berndt, 2003). Although not uncontroversial (Baddeley, 2003), there is considerable evidence for this theory, not only for the case of short-term maintenance of visual information, but for short-term maintenance of information in other modalities as well (see Postle, 2006, for a review).

Recent results from neuroimaging and psychophysics also provide converging evidence for a similar continuity between visual perception and VSTM. Several studies demonstrated that contents of VSTM during maintenance can be reliably decoded from visual cortical areas that are also involved in the perceptual encoding of the same stimuli (Harrison & Tong, 2009; Serences, Ester, Vogel & Awh, 2009; Christophel, Hebart, & Haynes, 2012; Riggall & Postle, 2012; Emrich, Riggall, LaRocque & Postle, 2013). Kang, Hong, Blake and Woodman (2011) showed that a motion direction maintained in VSTM can interact with a visually perceived motion direction to cause a motion repulsion effect which occurs in the same way as when two visually perceived motion directions interact. Saad and Silvanto (2013) showed that maintenance in VSTM can cause adaptation effects similar to the effects caused by prolonged visual stimulation in the tilt after-effect. Montaser-Kouhsari and Carrasco (2009) showed that discrimination performance in perceptual and VSTM tasks are affected in very similar ways by heterogeneities at isoeccentric locations in the visual field. These studies suggest that VSTM representations and visual perceptual representations share common substrates residing in visual cortical areas.
5.3 Knowledge of stimulus statistics aids visual perception and visual memory

A crucial observation about the visual system is that, through experience-dependent visual development and learning, people acquire sophisticated internal models of the types of visual stimuli they encounter in their natural visual environment. The natural visual environment is very far from random. It exhibits a rich set of statistical regularities ranging from low-level regularities between luminance values at nearby locations, or between orientations or spatial frequencies at different locations, to mid-level regularities between surface properties or mid-level features of natural objects, to high-level regularities between objects in different natural scene categories. The visual system adapts to these statistical regularities, which makes it more efficient at performing perceptual tasks in the natural environment.

Learning a good internal model of the statistical regularities in the natural environment has several benefits. First, in any task, optimal performance can only be achieved if the observer’s internal model matches the actual statistical structure of the stimuli used in the task. For natural tasks involving natural stimuli, this corresponds to the environmental statistics of the relevant stimuli. In many psychophysical tasks, people, in fact, do perform better when visual stimuli are consistent with natural scene statistics than when they are not, suggesting that the visual system is adapted to natural scene statistics (Knill, Field, & Kersten, 1990; Parragha, Troscianko, & Tolhurst, 2000; Yuille, Fang, Schrater, & Kersten, 2004; Stocker & Simoncelli, 2006; Girshick, Landy & Simoncelli, 2011). In general, the closer the observer’s internal model is to natural stimulus statistics, the better the observer’s performance will be in natural tasks involving those stimuli. Deviations
in the observer’s internal model from the natural stimulus statistics (referred to as “model mismatch” in Orhan and Jacobs, in preparation) would lead to suboptimal performance in natural tasks.

Second, related to the first point, learning the statistical regularities in the natural environment enables the visual system to tolerate significant amounts of noise in visual input. This might explain, for example, the reliable extraction of basic-level category information and substantial amounts of visual detail from very brief presentations of natural images (Serre, Oliva & Poggio, 2007; Fei-Fei, Iyer, Koch & Perona, 2007; Sharan, Rosenholtz & Adelson, 2009), even without the benefit of attention (Li, VanRullen, Koch & Perona, 2002).

Third, learning the statistical regularities of the natural environment is necessary for efficient encoding of information under resource constraints—that is, for encoding as much information as possible about the environment subject to inevitable biological constraints on the representational systems of the organism. This idea underlies the influential efficient coding hypothesis according to which the goal of early visual processing is to maximize information transmission by reducing the redundancy inherent in natural visual signals due to the rich variety of statistical regularities they exhibit (Barlow, 1961; Simoncelli & Olshausen, 2001; Geisler, 2008).

The idea of efficient coding of information is closely related to the well-known notion of “chunking” in the short-term memory literature (Miller, 1956). Chunking can be considered as a form of redundancy reduction or compression where one allocates less memory resources to more frequently occurring, or more familiar, stimuli, thereby freeing up resources for the representation of less frequent stimuli. This kind of efficient allocation of available resources might lead to an apparent increase in memory performance, or in the number of objects represented, without any ac-
tual increase in the amount of resources. Two recent studies provided experimental demonstrations of this idea in VSTM. Brady, Konkle and Alvarez (2009) showed that subjects can quickly learn statistical regularities between color pairs to encode more colors in simple visual displays consisting of such color pairs. Sims, Jacobs and Knill (2012) showed that subjects can adapt to a decrease in the variance of the stimulus distribution to increase the precision of their memories. Both studies found that these effects were attributable not to an increase in the available resources but to the efficient allocation of resources due to the adaptation to the stimulus statistics. It is important to note that both studies used artificial stimuli with simple forms of statistical regularities, suggesting that these efficient coding effects are likely to play a more significant role in natural stimuli displaying a much richer set of statistical regularities.

5.4 An approach toward unified models of visual perception and visual memory

The previous sections argued for the continuity of visual perception and VSTM. Section 5.2 reviewed evidence indicating that visual perception and VSTM share common neural and representational substrates. Section 5.3 reviewed experimental and theoretical findings suggesting that knowledge of the statistical regularities of visual stimuli aids both visual perception and visual memory, and that this benefit can be expected to be particularly significant for natural visual stimuli displaying a rich set of statistical regularities.

A challenge for scientists is to develop unified models of visual perception and VSTM. We are particularly interested in models that can account for the variety
of natural visual stimuli that can be represented both in visual perception and in VSTM (Figure 5.1) and for the fact that both visual perception and VSTM benefit from knowledge of statistical regularities inherent in such stimuli. Although this is a long-term challenge, we believe that a sketch of a potentially promising approach can be outlined now.

How can we incorporate into research on VSTM a perspective that unifies visual perception and VSTM and that is ecologically realistic? Is it possible to move beyond simple, artificial stimuli with unnatural stimulus statistics used in most VSTM studies and do at least partial justice to the richness and complexity of visual stimuli that are encoded in VSTM (see Figure 5.1)? We take our clue from the evidence reviewed above for common representational substrates for visual perception and VSTM, and suggest that these goals may be at least partially achieved with the use of probabilistic generative models.

Probabilistic generative models provide a rich, versatile and quantitative format for representing people’s prior information or expectations in different perceptual domains (Knill & Richards, 1996; Kersten & Yuille, 2003). In different cases, this prior information can be regarded either as the product of lifelong adaptation to statistical properties of the natural environment or as shorter-term adaptations acquired by subjects during the course of an experiment (Brady, Konkle & Alvarez, 2009; Sims, Jacobs, & Knill, 2012). For different perceptual problems, people’s prior expectations may be captured by different probabilistic generative models. For example, a generative model that can be used for modeling people’s prior information about visual textures (Portilla & Simoncelli, 2000; Balas, 2006; Balas, Nakano & Rosenholtz, 2009) will be different from a generative model that may be used to capture their prior information about the shapes of objects belonging to different
object categories (Sudderth, 2006) or their prior information about the structure of different natural scene categories (Sudderth, Torralba, Freeman & Willsky, 2008). Crucially, using probabilistic generative models—commonly employed for representing prior information or expectations in visual perception (Knill & Richards, 1996; Kersten & Yuille, 2003)—to capture the representational structure of VSTM would unify the representational formats used in visual perception and VSTM.

Figure 5.2A illustrates more concretely how such an approach might be used for modeling the encoding of natural scenes in VSTM. The details of the model shown in Figure 5.2A are not important for our purposes here—the model is simply intended to illustrate, in broad terms, how visual stimuli with rich statistical structure can be represented for the purposes of both perception and VSTM. The specific model in this figure is adapted from Sudderth (2006, p. 250). According to our approach, both visual perception and encoding in VSTM can be considered as probabilistic inference problems. Given noisy measurements or observations of some basic visual features, called the observable variables (represented by the shaded nodes in Figure 5.2A), the visual system tries to infer certain latent variables. For the case illustrated in Figure 5.2A, these latent variables are hierarchically organized and consist of variables such as visual objects and their locations in the scene, parts comprising each object and their locations with respect to the object, and lower-level visual features belonging to object parts and their locations. A subject’s prior knowledge of natural scenes is captured by scene-level variables (all variables outside the largest plate) that probabilistically describe what kinds of objects tend to occur in which configurations in natural scenes of a given type as well as the appearance of the objects (e.g., their parts and low-level features). Given a new scene, the subject (or the model) capitalizes on this prior knowledge to efficiently infer the latent variables.
in a scene. In different applications, the observable variables of interest might differ. In some cases, as in Figure 5.2A, the observable variables of interest might correspond to some low-level image features and their locations whereas, in other cases, they might correspond to higher-level, more abstract visual features. More realism can be introduced into the model by taking into account the fact that perceptual resolution in the visual periphery, or in other words, the precision of measurements obtained from the periphery, is lower than in the fovea.

Crucially, differences between visual perception and VSTM can be captured in the model entirely by the quality (or precision) of the observable variables. In general, the quality of the observable variables would depend on the presentation duration of the scene (with longer presentation durations leading to more precise measurements), as well as the length of the delay interval during which visual information is maintained in memory (precision of measurements degrading with the length of the delay interval). For visual perception, the scene remains visible throughout, and hence the measurements of the observable variables are higher quality or less noisy. For VSTM, the scene is removed for a certain delay interval after a brief presentation, so the measurements of the observable variables would be noisier. This greater uncertainty in the values of the observable variables “percolates” through the model and can have a large effect on the posterior distributions over the latent variables depending on the strength of the prior (for example, how strong the statistical regularities in natural scenes of this type are), as well as on the noise level (which, as mentioned, depends on the presentation time, the length of the delay interval and possibly other task variables). Other than this difference in the precision of the measurements of observable variables, however, the representation of prior knowledge about the scene and the inference process involved in determining the values of the latent variables
in the scene is identical in visual perception and VSTM. That is, there are no qualitative differences between visual perception and VSTM, just quantitative differences (that may still have a large impact depending on the strength of the prior, as mentioned above). This account is consistent with the results reviewed above suggesting common representational (and neural) substrates for visual perception and VSTM.

It seems clear that if the study of VSTM is to be extended to a richer set of stimuli from different perceptual domains that are more representative of the natural visual environment (Figure 5.1), models like the one shown in Figure 5.2A that can capture the rich statistical regularities in these domains, and people’s prior knowledge of those regularities, will be essential (see Brady, Konkle and Alvarez, 2011, for a similar point).

Recent evidence suggests that, even in experiments using simple items with unnatural stimulus statistics (e.g., oriented Gabors with independently drawn orientations), subjects’ internal models display a rich, hierarchical structure (Brady & Tenenbaum, 2013; Orhan & Jacobs, 2013) that does not match the simple, unstructured models used by the experimenter to generate the stimuli. We recently proposed a probabilistic clustering model that attempts to characterize aspects of the internal model that subjects use to encode simple displays typically used in VSTM experiments (Orhan & Jacobs, 2013). Our model is shown in Figure 5.2B. It implements the assumption that stimuli are generated in clusters, where the number of clusters and the assignment of stimuli to clusters are uncertain and have to be inferred probabilistically from the display, even if the actual generative process used by the experimenter does not involve any clusters. The model predicts biases in memory estimates of stimuli based on their VSTM representations, as has been observed previously in a number of VSTM studies (Brady & Alvarez, 2011; Huang & Sekuler,
Figure 5.2: Example probabilistic graphical models that can be used for modeling representations in VSTM. (A) This figure is adapted from Sudderth (2006, p. 250) and shows a hierarchical probabilistic graphical model for representing the structure of natural scenes. Scenes are represented hierarchically in terms of latent variables (unshaded nodes) corresponding to the objects in the scene and their locations, visual parts composing each object and their locations with respect to the object, as well as low-level features for each part and the locations of these features. Plates represent the replication of variables inside them. The largest plate represents the replication of variables for different images (\( J \) images in total) and the smaller plate inside it represents the replication of variables for different low-level features in the image (\( N_j \) features in total for image \( j \)). A difference from the model of Sudderth (2006, p. 250) is that we assume that subjects do not have access to the actual values of the low-level features \( v_{ji}, w_{ji} \) but only to noisy observations thereof (represented by \( t_{ji} \) and \( y_{ji} \) respectively). (B) The probabilistic clustering model proposed by Orhan and Jacobs (2013) to capture the internal model that subjects might be employing to encode the simple displays typically used in VSTM experiments. The model implements the assumption that stimuli, whose actual values are denoted by the variables \( \theta_i \), are generated in clusters, with cluster parameters \( \mu_i \) (cluster mean) and \( \tau_i \) (cluster precision). The number of clusters and the assignment of stimuli to clusters are uncertain and have to be inferred probabilistically from the noisy observations of the stimuli, denoted by \( x_i \), even if the actual generative process used by the experimenter to generate the stimuli does not involve any clusters. Note that variables with the same names do not necessarily represent the same things in the two models shown in (A) and (B).
Due to the coupling of stimuli through the probabilistic clustering assumption, the model also predicts dependencies between memory estimates of different stimuli that should decrease with the difference between their feature values. We experimentally confirmed this novel prediction of the model using both continuous recall (delayed estimation) and change-detection tasks with small set sizes. Our study also highlighted the importance of using sufficiently powerful experimental paradigms to probe VSTM representations. Unlike in more common paradigms where there is only a single target stimulus, we used experimental tasks where subjects were required to recall, or to detect changes in, the feature values of multiple stimuli in a single trial. This was necessary to uncover potential dependencies between the representations of different stimuli.

Similarly, Brady and Tenenbaum (2013) recently showed that hierarchical models incorporating the encoding of summary statistics (such as the mean and variance) of stimuli in a given display, as well as the perceptual grouping of stimuli, are necessary to account for display-by-display performance in change-detection tasks. This was the case even in experiments where stimulus values are drawn independently in each trial, and hence contain no statistical structure.

5.5 Why is it important to take the question of ecological validity in VSTM research seriously?

As mentioned in the Introduction, although the use of unnatural stimuli and tasks might afford greater experimental control, psychologists have been aware of the problems with generalizing the results of such studies (Brunswik, 1943, 1955; Neisser, 1976). This generalization problem may be more acute in some domains than in
others, but it is one that every cognitive psychologist should take seriously if they do not want to study a problem that exists in a laboratory setting, but that does not exist, or is less important or relevant, for real observers or actors living in natural environments. In VSTM studies specifically, we believe that there are a number of reasons why researchers should take the question of ecological validity seriously. The following four subsections discuss some of these reasons.

5.5.1 Choice of stimuli and tasks influences the questions asked

The choice of stimuli and tasks used by experimenters exerts a significant influence on the questions studied. Many of the questions currently debated in the VSTM literature are either not well-defined or simply do not arise within the context of stimuli shown in Figure 5.1. For example, the question of the distribution of resources among items in simple displays do not arise for stimuli such as those in Figure 5.1B-C, and is poorly-defined for stimuli shown in Figure 5.1A and Figure 5.1D because these stimuli are not merely encoded in terms of separate, individual items, but in more complex, possibly hierarchical ways. Even the “items” or objects in these figures are not encoded as monoliths, but as structured objects composed of many parts put together in specific configurations.

2Another example is the question of the units of storage in VSTM—that is, the question of whether objects are represented as bound units in VSTM or as independent features (Luck & Vogel, 1997; Bays, Wu & Husain, 2011; Fougne & Alvarez, 2011). Again, this question is not meaningful for stimuli that do not involve objects. A more meaningful question for such stimuli is what exactly people can
represent about them in VSTM. For stimuli involving objects, thinking in terms of complex probabilistic representational structures such as the graphical model shown in Figure 5.2A would help researchers formalize, sharpen and extend the question of whether objects are represented as bound units or as independent features in VSTM. In terms of the graphical model shown in Figure 5.2A, the question of whether two features $v_{ji}$ and $v_{jk}$ (corresponding to two noisy observations $t_{ji}$ and $t_{jk}$) belonging to the same object are represented independently can be translated into the question of whether the posterior distributions over these variables are correlated and, if so, how strongly. It can be shown that the answer depends on the prior distribution—in particular, on how strongly these features tend to be correlated in objects of this particular type. One can ask similar questions about object parts and objects themselves (are two objects represented independently in a particular scene?), and formulate reasonable hypotheses from the basic properties of probabilistic inference in hierarchical probabilistic graphical models.

The examples reviewed in this subsection suggest that researchers might ask very different questions if VSTM studies focused on natural stimuli.

## 5.5.2 Generalizability of results

In many cases, it remains unclear to what extent the results from VSTM studies using unnatural stimuli and tasks can be generalized to ecologically realistic conditions. For instance, in many VSTM studies, stimuli are presented solely in the periphery, and eye movements are prevented either by monitoring them or by adopting short presentation times (shorter than about 150 ms). In addition, in most VSTM studies, subjects are not given a task other than to remember the stimuli. In natural vision, in contrast, we sample the world by making frequent eye movements under tight control
of task demands (Land, Mennie & Rusted, 1999; Hayhoe, 2000). Importantly, fixated objects are encoded better than unfixated objects (Hollingworth & Henderson, 2002). Indeed, it has been argued that visual information from successively fixated objects can be integrated into a detailed visual representation of the scene (Hollingworth & Henderson, 2002; Hollingworth, 2004). This suggests that, relative to natural (unconstrained) viewing conditions, the prevention of eye movements in experiments might lead to an underestimation of memory performance.

Another obvious difference between viewing conditions in natural vision and in typical experimental settings is the full-field, 3-D visual stimulation under natural conditions versus limited-field, 2-D projections used in most experiments. It remains to be seen what the consequences of these differences, if any, might be for VSTM studies.

In our daily lives, most of our visual perception involves objects and environments that we are extremely familiar with. How does this familiarity affect the characteristics of visual short-term memories for such objects and environments? Could the results of experiments using unfamiliar stimuli be reliably generalized to extremely familiar stimuli? If not, how should they be modified (e.g., see Hemmer & Steyvers, 2009)? For example, it has been suggested that object features are encoded approximately independently in VSTM (Bays, Wu & Husain, 2011; Fougnie & Alvarez, 2011). Does this result generalize to objects that we are extremely familiar with? If not, how should it be modified? Again, thinking in terms of complex probabilistic representational structures such as the graphical model shown in Figure 5.2A could help researchers formalize such questions and formulate reasonable hypotheses that may be experimentally tested. Familiarity with an object can be modeled as the strength of the prior information about that object. Varying the strength of this
prior information yields qualitatively, and perhaps quantitatively, testable changes in the model’s behavior.

Another example where questions about the generalizability of experimental results to more natural conditions might arise is the recent finding that encoding precision is variable across items and across trials in typical VSTM tasks (Van den Berg et al., 2012; Fougnie, Suchow & Alvarez, 2012). It remains unclear whether this result would generalize to encoding of objects in more realistic scenes. It is also worth pointing out that these studies used peripheral presentation of stimuli, and thus variability here refers to variability in encoding isoeccentric items in visual periphery. However, under more natural viewing conditions, it is possible that any variability in the periphery would be small compared to the differences in encoding precision at different eccentricities (e.g., encoding precision in the fovea versus in the periphery), thus making it a less significant phenomenon.

Finally, some studies reported so-called misbinding errors in VSTM recall tasks (Bays, Catalao & Husain, 2009; Emrich & Ferber, 2012) where subjects, instead of reporting the feature value of the target item, mistakenly report the feature value of a distractor. It is not clear whether these types of errors would happen with significant frequency in more natural conditions, as it seems that they are not even robust to different response modalities in recall tasks (scrolling versus pointing) (Van den Berg et al., 2012).

### 5.5.3 Alternative interpretation of results based on model mismatch

As discussed in Section 5.3, the visual system is adapted to the statistical structure of the natural environment. However, most VSTM studies use stimuli that are un-
structured, such as stimuli in which objects’ feature values are selected independently from a uniform distribution. If a subject assumes that stimuli have the structure of the natural environment (e.g., there are dependencies among feature values and these dependencies have particular forms), but experimental stimuli have a different structure, then this creates a “model mismatch”. Importantly, as discussed in Section 5.3, model mismatch will negatively impact a subject’s performance on a VSTM task.

A consideration of model mismatch might affect the interpretation of results obtained from VSTM studies. For example, the decline in precision with set size found in standard VSTM studies is generally interpreted as a signature of VSTM resource or capacity limitations. However, in a computational study, we have recently demonstrated that the same result can be accounted for as a consequence of model mismatch without assuming resource limitations (Orhan & Jacobs, in preparation). Intuitively, the basic idea is that even if VSTM resources increase linearly with set size (meaning that resources are unbounded), the visual system might be allocating most of these resources to stimulus configurations that are common in the natural environment but not proportionately common in the experiment. This leads to inefficient use of resources (with respect to the stimulus distribution used in the experiment), which may cause a decline in memory recall precision even when the amount of resources increases proportionally to set size. It is thus an open empirical question to what extent set size effects observed in VSTM studies are caused by genuine resource limitations versus inefficient use of resources due to model mismatch.

We have also shown through simulations (Orhan & Jacobs, in preparation) that other prominent experimental findings, such as variability in encoding precision (Vanden Berg et al., 2012; Fougnie, Suchow & Alvarez, 2012) and the different set size dependencies of initial encoding rate and asymptotic precision (Bays et al., 2011),
can be explained from a perspective that takes model mismatch into account. This leads to interpretations of these findings that are different from commonly proposed interpretations (Orhan & Jacobs, in preparation). Finally, through a series of simulations, we have demonstrated that the use of stimulus statistics that share important characteristics of stimulus statistics in the natural environment leads to qualitatively different patterns of results than the ones obtained with unstructured stimulus statistics (e.g., independent feature values sampled from a uniform distribution) typically used in VSTM studies (Orhan & Jacobs, in preparation), again suggesting a need for caution when generalizing results from such studies to ecologically more realistic conditions.

5.5.4 Unnatural stimulus statistics may lead to underestimation of VSTM capacity

Using artificial stimuli and unnatural stimulus statistics might lead researchers to develop a distorted view of the properties of VSTM, including its capacity. In information theory, the capacity of an information channel is defined as the maximum mutual information that can be achieved between an input ensemble $X$ and the output of the channel $Y$, where the maximum is taken with respect to all possible probability distributions over the input ensemble (MacKay, 2003). If the VSTM system is abstractly conceived of as an information channel (Sims, Jacobs & Knill, 2012), then $X$ might correspond to all possible visual stimuli and $Y$ might correspond to all possible responses of neurons or neural populations underlying the VSTM system, presumably the visual cortical areas where the initial perceptual encoding takes place (Section 5.2). If it is assumed, as seems plausible, that the responses of these visual cortical areas are adapted to natural stimulus statistics, then the optimal input
distribution, $P(X)$, will correspond to the statistics of the visual stimuli $X$ in the natural visual environment. Any other input distribution will drive the information rate below the capacity of the channel. However, this is exactly what happens when unnatural stimulus statistics (e.g., feature values of items are drawn independently from a uniform distribution) are used in VSTM studies.

A toy example from information theory illustrates this point. Consider a noisy binary asymmetric information channel with two possible inputs, $X = 0$ and $X = 1$, and two possible outputs, $Y = 0$ and $Y = 1$. Denote the input probabilities by $p_0 = P(X = 0)$ and $p_1 = 1 - p_0 = P(X = 1)$. The channel is defined as follows. If $X = 0$, the signal is flipped ($Y = 1$) with probability $f_0$ due to noise in the channel. If $X = 1$, on the other hand, the signal is flipped ($Y = 0$) with probability $f_1$. Assume that $f_0 < f_1$. This means that the channel transmits the signal $X = 0$ more faithfully (with a smaller flip probability). This can happen, for example, if the channel is optimized (subject to a resource constraint) to operate in an environment in which the signal $X = 0$ is more common. Then the channel capacity is achieved for an input distribution with $p_0 > p_1$. To give a numerical example, if $f_0 = 0.1$ and $f_1 = 0.4$, the channel capacity is achieved with the input distribution $p_0 \approx 0.54$ and $p_1 \approx 0.46$. Any input distribution different than this underutilizes the channel. Using artificial stimuli with unnatural stimulus statistics might yield a similar underutilization of VSTM, leading researchers to underestimate its capacity.

5.6 How to proceed: Example from neuroscience

Researchers interested in VSTM need to strike the right balance between the use of stimuli that are simple and easy to control (but artificial) versus natural stimuli which
are the true targets of scientific interest (but complex and difficult to control). A similar need often arises in the field of visual neuroscience, such as when determining the response properties of neurons in different visual cortical areas. A brief review of how neuroscientists approach this issue could be instructive for the study of VSTM.

Although much of what we know about the response properties of neurons in early visual areas was obtained with unnatural stimulus statistics (e.g., white noise, bars with random orientations or Gabor patches), the use of natural stimuli sometimes reveals surprising, new response properties that are not present in responses to unnatural stimuli, because unnatural stimuli fail to engage neural response mechanisms specifically adapted to the structure of natural stimuli (David, Vinje & Gallant, 2004; Felsen, Touryan, Han & Dan, 2005; Freeman, Ziemba, Heeger, Simoncelli & Movshon, 2013). The use of more natural stimuli becomes even more imperative when studying the response properties of higher visual areas (at least during the initial exploration of stimulus characteristics that elicit strong responses from neurons) because highly artificial stimuli are often ineffective in driving neural responses in these higher visual areas (Olshausen & Field, 2005; Felsen & Dan, 2005). Because unnatural stimuli fail to elicit the full range of neural responses, the information-theoretic argument presented in the previous subsection suggests that these stimuli might lead to an underestimation of capacity of the visual system, if the visual system is conceived of as an information channel.

Even when responses to simple, artificial stimuli are used in characterizing the basic response properties of neurons, neuroscientists often try to validate their models of neural response properties by measuring how well they can predict responses to natural stimuli that are more representative of an animal’s environment (David & Gallant, 2005; Sharpee, Sugihara, Kurgansky, Rebrik, Stryker, & Miller, 2006;
Nishimoto & Gallant, 2011). The models, as well as the stimuli used to probe the models, can be refined in successive cycles by determining exactly where and why the models fail (if they do fail) to predict the responses to more natural stimuli.

In a similar vein, we believe that testing the validity of theories or models in the context of stimuli and tasks that are representative of an observer’s natural environment, and then modifying or refining them if needed, might also be a fruitful strategy in research on VSTM. This is true especially when theories or models are initially developed using simple stimuli and tasks as is often necessary during early stages of experimentation.

### 5.7 Conclusion

Recent findings from neuroimaging and psychophysics suggest common neural and representational substrates for visual perception and VSTM. Visual perception is adapted to a rich set of statistical regularities present in the natural visual environment. The continuity between visual perception and VSTM (Section 5.2) suggests that VSTM is adapted to these statistical regularities too. Thinking about the operation of VSTM in such environments may lead to a re-evaluation of the results obtained from experiments using artificial stimulus statistics and tasks, and may inspire the development of ecologically more realistic, complex models incorporating prior knowledge of natural scene statistics.


Hyvärinen, A., Hurri, J., & Hoyer, P.O. (2009). Natural image statistics: A proba-
bilistic approach to early computational vision. London: Springer Verlag.


Snippe, H.P., & Koenderink, J.J. (1992). Information in channel-coded systems:


A Appendix for Chapter 2

A.1 Posterior inference

For the DPMM, posterior inference was performed using Algorithm 8 of Neal (2000). An excellent description of this algorithm can also be found in Görür (2007). For the BFMMs, we used a Gibbs sampling algorithm for posterior inference (Algorithm 2.1 in Sudderth (2006)). Finally, posterior inference in the HBM was performed in WinBUGS (Lunn, Thomas, Best & Spiegelhalter, 2000) modifying the WinBUGS code provided in the supplementary material of Brady and Alvarez (2011). For all models, in recall tasks, we used the posterior means as the models’ estimates of the feature values of different items. As described in more detail in the main text, the procedure for the simulation of the Viswanathan et al. (2010) was slightly different, because this study involved a recognition memory task, unlike the other studies modeled here, which were all recall tasks. The source code used in the simulations reported in Chapter 2 is available upon request from the author.
A.2 Maximum likelihood estimation of the parameters and model comparison

In our simulations, we used grid searches to find the values of the parameters (for the DPMM and the BFMMs: $\alpha_c$ and $\tau_{obs}$ in the univariate case and $\kappa$ and $\sigma^2_{obs}$ in the multivariate case; for the HBM: $\tau_{obs}$) that maximized an approximation to the likelihood given the observed data. The grid range over which we searched was determined through trial-and-error in each case.

In the simulation of the experiment by Wilken and Ma (2004), the data were the average observed errors in 64 conditions (4 set sizes $\times$ 16 bins; see the top-left plot in Figure 2.6). Each model effectively produces a different likelihood distribution $p_M(biases|FP)$ as a function of its free parameters denoted by $FP$, where $M$ could be $DPMM$, $BFMM - 2$, $BFMM - 4$ or $HBM$ and $biases$ denotes the collection of variables representing the average biases in each of the 64 conditions. These likelihood distributions do not have an analytic form, therefore we computed them by sampling. Specifically, for each model, we simulated the experiment of Wilken and Ma (2004) 25 times and, for each of the 25 runs of the simulated experiment, collected the average biases predicted by the model in each of the 64 conditions. Furthermore, because trials in each simulated experiment are independent of each other, the distribution $p_M(biases|FP)$ can be factorized as $p_M(bias_1|FP)p_M(bias_2|FP)\ldots p_M(bias_{64}|FP)$ with $bias_i$ representing the average bias in the $i$-th condition. We then approximated each $p_M(bias_i|FP)$ using a non-parametric kernel density estimate generated from the 25 collected samples of $bias_i$. The log-likelihood of a specific setting of the parameters $FP$ given the observed data is then calculated as $\sum_{i=1}^{64} \log p_M(bias_i = obs\_bias\_i|FP)$, where $obs\_bias\_i$ denotes
the observed average bias in the $i$-th condition. This procedure was repeated for each setting of the free parameters $FP$ over the grid. The parameter values that maximized the estimated log-likelihood were chosen as the maximum likelihood (ML) estimates of the parameters. For the DPMM and the BFMMs, the parameter $\tau_{obs}$ was allowed to vary across set sizes, whereas $\alpha_c$ was fixed across set sizes. Due to the relatively high dimension of the search space in this particular simulation (5 free parameters for the DPMM and the BFMMs and 4 free parameters for the HBM), grid searches were conducted in a greedy fashion. The uniform base distribution for $\mu$ was defined over the interval $[0, 12]$. The exact values of the endpoints of this interval did not affect the simulation results for this particular experiment, or for other experiments, as long as the interval was large enough to include the minimum and maximum possible values of the relevant feature in an experiment.

In the simulation of the experiment by Viswanathan et al. (2010), the data were the total number of observed ‘old’ (or ‘yes’) responses in 1800 trials each of the medium and high homogeneity conditions. The likelihood in each case was modeled as a binomial distribution with $n = 1800$ (number of trials) and the model’s predicted success probability $p$ which was a function of the parameters of the model. The parameter values that maximized the log-likelihood, $\log(\text{Binomial}(k = obs_{\text{med}}; n, p_{\text{med}}(FP))) + \log(\text{Binomial}(k = obs_{\text{high}}; n, p_{\text{high}}(FP)))$ [where $obs_{\text{med}}$ and $obs_{\text{high}}$ are the total number of observed ‘old’ responses in the medium and high homogeneity conditions, respectively, and $p_{\text{med}}(FP)$ and $p_{\text{high}}(FP)$ are the model’s predicted success probabilities in the two conditions], were chosen as the maximum likelihood (ML) estimates of the parameters. The uniform base distribution for $\mu$ was defined over the interval $[0, 11]$.

In the simulation of the experiments by Brady and Alvarez (2011), the data were
the observed mean biases in Experiment 1 and Experiment 2. As in the simulation of the Wilken and Ma (2004) experiment, each model produces a different distribution over the mean biases in the simulated Experiments 1 and 2, $p_M(bias_1|FP)$ and $p_M(bias_2|FP)$, as a function of the free parameters. Since these distributions do not have an analytic form, we drew 25 samples from these distributions by simulating each experiment 25 times and computing the predicted mean bias in each case. We then approximated $p_M(bias_1|FP)$ and $p_M(bias_2|FP)$ with a nonparametric kernel density estimate using the 25 collected samples. The parameter values that maximized the estimated log-likelihood, $\log(p_M(bias_1 = obs\_mean\_bias_1|FP)) + \log(p_M(bias_2 = obs\_mean\_bias_1|FP))$ where $obs\_mean\_bias_1$ and $obs\_mean\_bias_2$ denote the observed mean biases in Experiment 1 and Experiment 2 of Brady and Alvarez (2011) were chosen as the maximum likelihood (ML) estimates of the parameters. For the multivariate DPMM and BFMMs, the same model was applied to both experiments. As explained in the main text, we placed a vague inverse-Wishart prior on $\Psi$ (the inverse scale parameter of the base distribution for $\Sigma$). Specifically, the inverse scale parameter of the inverse-Wishart prior on $\Psi$ was $3000I$ where $I$ is the identity matrix, and its degrees-of-freedom parameter was 2 to maximize the variability (or vagueness) of the prior. The uniform base distribution for $\mu$ was defined over the interval $[-50, 200]$ for both color and size dimensions.

In the application of the models to data from our own experiments, for each subject, the data were trial-by-trial responses of the subject. In the case of the DPMM and the BFMMs, for each configuration $\{\theta_i\}_{i=1}^N$, the models predicted a joint distribution, $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N, \alpha_c, \tau_{obs})$, over the estimates of the items as a function of the free parameters $\alpha_c$ and $\tau_{obs}$. This distribution was computed by sampling using Equation 2.2. For each configuration, 50 samples were drawn from the joint distribu-
tion of the estimates. The log-likelihood of parameters given the subject’s responses for that particular configuration was then computed by evaluating each response under a bivariate Gaussian approximation to the joint distribution constructed from those 50 samples, taking its logarithm and summing over all responses. This was repeated for all stimulus configurations and the total log-likelihood was computed by summing over all configurations. The parameter values that maximized the estimated total log-likelihood were chosen as the maximum likelihood (ML) estimates of the parameters. The uniform base distribution for $\mu$ was defined over the interval $[-10, 10]$.

We used the Bayesian information criterion (BIC) to compare the model fits (Schwartz, 1978). BIC scores were computed according to:

$$BIC = -2 \log L + k \log n$$  \hspace{1cm} (A.1)$$

where $L$ denotes the maximum likelihood of the model (i.e., the likelihood value achieved when ML estimates of the parameters are used), $k$ is the number of parameters and $n$ is the number of data points. BIC scores reported in the article are relative to the BIC score for the DPMM.

A.3 Details of the Bayesian model of same-different responses in Experiments 2 and 3

In the application of the Bayesian model to experimental data from Experiment 2, the following priors were used: each of $\mu_1$ and $\mu_2$ were given Gaussian priors with means centered on the actual feature values (i.e., horizontal locations) of the two items in
the target configurations and with precision 0.01. The correlation coefficient of the underlying bivariate Gaussian distribution, $\rho$, was given a uniform prior over the interval $[-1,1]$. The standard deviations along each of the two dimensions, $\sigma_1$ and $\sigma_2$, were given uniform priors over the interval $[0.1,1.3]$. The lower bound on the probability of ‘same’ responses, $b_l$, was given a uniform prior over $[0,0.5]$ and the upper bound on the probability of ‘same’ responses, $b_u$, was given a uniform prior over $[0.5,1.0]$.

In the application of the Bayesian model to experimental data from Experiment 3, the following priors were used: each of $\mu_1$ and $\mu_2$ were given Gaussian priors with means centered on the actual feature values (i.e., orientations) of the two items in the target configurations and with precision 0.01. The correlation coefficient of the underlying bivariate Gaussian distribution, $\rho$, was given a uniform prior over the interval $[-1,1]$. The standard deviations along each of the two dimensions, $\sigma_1$ and $\sigma_2$, were given uniform priors over the interval $[1.0,18.0]$. The lower bound on the probability of ‘same’ responses, $b_l$, was given a uniform prior over $[0,0.5]$ and the upper bound on the probability of ‘same’ responses, $b_u$, was given a uniform prior over $[0.5,1.0]$.

We implemented the model in Python using the PyMC package (Patil, Huard & Fonnesbeck, 2010). For each of the two experiments, 11000 samples were drawn from the posterior distributions of the variables. The first 1000 samples were discarded as burn-in and the remaining 10000 samples were ‘thinned’ down to every 10th sample to reduce dependencies between consecutive samples. This reduced the number of samples to 1000. The results reported in this chapter are based on these 1000 samples. Proper convergence and mixing were monitored and confirmed both visually and through a battery of diagnostics provided by the PyMC package.
B  Appendix for Chapter 3

B.1  Proof of equation 3.13

We denote $A = \frac{\rho_q}{1 - \rho_q}$. Note that $A \geq 0$, because we assume $\rho_q \geq 0$. We can then rewrite Equation 3.13 as follows:

$$\frac{1}{N} \log(1 + NA) \geq \frac{1}{N + 1} \log(1 + (N + 1)A) \quad (B.1)$$

$$\Leftrightarrow (1 + NA)^{1/N} \geq (1 + (N + 1)A)^{1/(N+1)} \quad (B.2)$$

$$\Leftrightarrow (1 + NA)^{N+1} \geq (1 + (N + 1)A)^N \quad (B.3)$$

Let $k = 1 + NA$. Then the left-hand side of the last inequality is equal to $k^{N+1}$ (i.e., a product of $N + 1$ terms each of which is $k$). The sum of these $N + 1$ terms is $k(N+1)$. The right-hand side of the last inequality is equal to $(k + A)^N$. Multiplying the right-hand side by 1 does not change the result, so the right-hand side can also be expressed as the product of $N + 1$ terms: $N$ terms that are equal to $k + A$ and an additional term which is equal to 1. The sum of these $N + 1$ terms is $N(k + A) + 1 = kN + NA + 1 = kN + k = k(N + 1)$. Therefore, both the left-hand
side and the right-hand side are the product of $N + 1$ terms whose sum is equal to $k(N + 1)$. But, for a given number of terms whose sum is constant, their product is maximized when all the terms are identical (this can be easily seen by writing the Lagrange function for the product of $p$ terms, $x_1, \ldots, x_p$, under the constraint that their sum be constant: $x_1 + \ldots + x_p = c$ and solving the resulting optimization problem). This is achieved by the left-hand side of the last inequality: all terms are equal to $k$, thus the last inequality in B.3 holds.

B.2 Simulation details

Because the prior mean is always equal to a vector of zeros, for the multivariate Gaussian model with uniform correlations and the multivariate Gaussian model with random positive correlations, the posterior means are found by: $C \Sigma^{-1} \mathbf{x}$, where $C$ is the posterior covariance matrix, $\Sigma$ is the covariance matrix of the noise distribution and $\mathbf{x}$ is the noisy observation vector in a particular trial. The model’s estimate for the target stimulus value is then equal to the corresponding element in the posterior mean vector.

For the multivariate Gaussian model with random, positive correlations, we use the \texttt{gallery} function in Matlab with the ‘\texttt{randcorr}’ option to generate random correlation matrices. Eigenvalues used as seeds for the \texttt{gallery} function are generated by first randomly drawing a sample from a symmetric Dirichlet distribution with concentration parameter $\gamma$ and multiplying it by the number of eigenvalues $N$ (this ensures that the eigenvalues sum to $N$). To prevent numerical instabilities, a small amount of independent positive random noise is added to each eigenvalue (uniform between 0 and $0.0001 \times N$) and then the eigenvalues are re-normalized to sum
to $N$. The gallery function generates correlation matrices with both negative and positive entries. To obtain correlation matrices with positive entries only, we simply take the absolute value of each entry in the correlation matrix. Although this procedure occasionally changes the eigenvalues of the correlation matrix, we confirmed, through simulations, that the resulting changes did not significantly alter the qualitative characteristics of the eigenvalue distribution of the correlation matrices. In particular, small $\gamma$ values still yield sparse eigenvalue distributions, whereas large $\gamma$ values yield broad eigenvalue distributions.

For the mixture of Gaussians model, the joint and marginal posterior means can be calculated analytically. We first derive the joint posterior $p(s|x)$:

$$p(s|x) = \frac{p(x|s)p(s)}{p(x)}$$  \hspace{1cm} (B.4)

$$= \frac{p(x|s)[\sum_{k=1}^{N} w_k p_k(s)]}{p(x)}$$  \hspace{1cm} (B.5)

where $p_k(s)$ is the $k$th normal component in the mixture, with mean $0$, covariance $\Gamma_k$ and weight $w_k$. Then:

$$p(s|x) = \sum_{k=1}^{N} w_k p(x|s)p_k(s)$$  \hspace{1cm} (B.6)

In the last equation, $p(x|s)$ and $p_k(s)$ are both Gaussians. The product of two Gaussians is another Gaussian, but no longer normalized. Therefore, the last equation can be re-written as:

$$p(s|x) = \frac{\sum_{k=1}^{N} w_k Z_k N(s; m_k, C_k)}{p(x)}$$  \hspace{1cm} (B.7)

where $C_k = (\Sigma^{-1} + \Gamma_k^{-1})^{-1}$ is the posterior covariance matrix of the $k$th component ($\Sigma$
being the covariance matrix of the noise distribution), $\mathbf{m}_k = C_k\Sigma^{-1}\mathbf{x}$ is the posterior mean of the $k^{th}$ component (note that the prior means of all components are equal to 0) and $Z_k$ is the normalization constant for the $k^{th}$ component:

$$Z_k = (2\pi)^{-N/2}|C_k|^{1/2}|\Sigma|^{-1/2}|\Gamma_k|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}^T\Sigma^{-1}\mathbf{x} - \mathbf{m}_k^T C_k^{-1} \mathbf{m}_k)\right] \quad (B.8)$$

In this normalization constant, we can ignore factors common to all components, because they cancel out during normalization. Thus, Equation B.7 can be re-written as:

$$p(\mathbf{s}|\mathbf{x}) \propto \sum_{k=1}^{N} \tilde{w}_k \mathcal{N}(\mathbf{s}; \mathbf{m}_k, C_k) \quad (B.9)$$

where $\tilde{w}_k = w_k |C_k|^{1/2}|\Gamma_k|^{-1/2} \exp(\frac{1}{2} \mathbf{m}_k^T C_k^{-1} \mathbf{m}_k)$ is the posterior weight of the $k^{th}$ component. Joint and marginal posterior means can then be easily calculated from Equation B.9 as weighted averages of joint and marginal means of individual components.

The default unnormalized prior mixture weights used in Sections 3.3 and 3.4 are $10^{16}, 5 \times 10^{15}, 10^9, 10^{-5}, 10^{-10}, 10^{-15}$ from the most homogeneous to the least homogeneous components, respectively.

For the models with stimulus-dependent noise variance presented in Section 3.5, posterior means or modes are no longer analytically tractable. For these models, to be able to simulate a large number of trials, we used a Laplace approximation for the joint posteriors and estimated the posterior means using a Nelder-Mead simplex algorithm (the `fminsearch` function in Matlab). We used the marginal posterior mean as the model’s estimate of the target stimulus value in a given trial. 2500 trials per set size were simulated in simulations with the multivariate Gaussian $q(\mathbf{s})$ with uniform non-negative correlations, and 1000 trials per set size in simulations
with the mixture of Gaussians model. For the mixture of Gaussians model, the unnormalized prior mixture weights were $10^{16}$, $10^{15}$, $10^{10}$, $10^{-5}$, $10^{-10}$, $10^{-15}$ from the most homogeneous to the least homogeneous components, respectively.

Matlab code for running all the simulations and generating the figures reported in this chapter is available upon request from the author.

### B.3 Fitting normal and $t$ distributions to the models’ errors

In Sections 3.3.2, 3.4.3, 3.5.2, to test for variability in encoding precision, we fit normal and $t$ distributions to the models’ errors. For this purpose, we use the ‘fitdist’ function in Matlab’s Statistics Toolbox with the ‘normal’ and ‘tlocationscale’ options respectively. BIC scores of the two distributions are then computed according to:  
$$BIC = -2 \log L + k \log n,$$
where $\log L$ is the log-likelihood of the model evaluated at the estimated parameter values, $n$ is the number of data points and $k$ is the number of free parameters. The normal distribution has two free parameters (mean and variance parameters), whereas the $t$ distribution has three free parameters (location $\mu$, scale $\sigma$ and degrees-of-freedom parameter $\nu$). Given the estimated parameters of the $t$ distribution, $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\nu}$, the parameters of the implied Gamma distribution over precisions can be computed according to:  
$$\hat{\alpha} = \frac{\hat{\nu}}{2} \text{ and } \hat{\beta} = \frac{\hat{\sigma}^2 \hat{\nu}}{2}.$$
C Appendix for Chapter 4

C.1 Derivation of Equation 4.4

In a neural population with limited range correlations, the covariance between the firing rates of the $k$-th and $l$-th neurons (the $kl$-th cell of the covariance matrix) is assumed to be a monotonically decreasing function of the distance between their preferred stimuli:

$$Q_{kl}(s) = \delta_{kl} + (1 - \delta_{kl})[a f_k(s)^\alpha f_l(s)^\alpha \exp\left(-\frac{||\phi_k - \phi_l||}{L}\right)]$$  \hspace{1cm} (C.1)

where $\phi_k$ and $\phi_l$ are the tuning function centers of the neurons and $L$ is a length scale parameter determining the spatial extent of the correlations. Under the assumptions we make in Section 4.1 [assumptions (i)-(iv)], Equation C.1 reduces to:

$$Q_{kl} = \rho^{k-l}a f_k^\alpha f_l^\alpha$$ with $\rho = \exp(-1/(L\eta))$ where $\eta$ is the density of the tuning functions.

We are interested in deriving the FIM, $J(s)$, for our correlated neural population
encoding the stimuli \( s \). The \( ij \)-th cell of the FIM is defined as:

\[
J_{ij}(s) = -E\left[\frac{\partial^2}{\partial s_i \partial s_j} \log p(r|s)\right]
\]  

(C.2)

It is easy to show that for a multivariate Gaussian distribution as in Equation 4.1, the \( ij \)-th element of the Fisher information matrix is equal to (Abbott & Dayan, 1999):

\[
J_{ij}(s) = \frac{\partial f^T}{\partial s_i} Q^{-1} \frac{\partial f}{\partial s_j} + \frac{1}{2} \text{Tr} \left[ \frac{\partial Q}{\partial s_i} Q^{-1} \frac{\partial Q}{\partial s_j} Q^{-1} \right]
\]  

(C.3)

Following Ecker et al. (2011), we denote the first term in Equation C.3 as \( J_{\text{mean}} \) and the second term as \( J_{\text{cov}} \) (this is because the first term represents the amount of information encoded in changes in the mean firing rates of the neurons, whereas the second term represents the amount of information encoded in changes in the covariance of the firing rates of the neurons).

In order to derive \( J_{\text{mean}} \) and \( J_{\text{cov}} \), we need to invert \( Q \). For the limited range correlation structure we chose (Equation C.1), it is easy to show that the \( kl \)-th element of \( Q^{-1} \) can be written as (Snippe & Koenderink, 1992):

\[
(Q^{-1})_{kl} = \frac{1}{\alpha f_k \alpha f_l} \frac{1 + \rho^2}{1 - \rho^2} \left[ \delta_{kl} - \frac{\rho}{1 + \rho^2} (\delta_{k+1,l} + \delta_{k-1,l}) \right]
\]  

(C.4)

where \( \delta \) is the delta function. This last equation makes it clear that \( Q^{-1} \) is a banded matrix, which considerably simplifies the calculation of \( J_{\text{mean}} \) and \( J_{\text{cov}} \).
Carrying out the multiplication in the first term of Equation C.3, we get:

\[ J_{\text{mean}} = \frac{\partial f^T}{\partial s_i} Q^{-1} \frac{\partial f}{\partial s_j} \]

\[ = -\sum_{k=2}^{K} \frac{1}{af_k^\alpha f_k^\alpha} \frac{\rho}{1 - \rho^2} \frac{\partial f_{k-1}}{\partial s_i} \frac{\partial f_k}{\partial s_j} + \sum_{k=1}^{K} \frac{1}{af_k^\alpha f_k^\alpha} \frac{1 + \rho^2}{1 - \rho^2} \frac{\partial f_k}{\partial s_i} \frac{\partial f_k}{\partial s_j} \]

\[ - \sum_{k=1}^{K-1} \frac{1}{af_k^\alpha f_k^\alpha} \frac{\rho}{1 - \rho^2} \frac{\partial f_k}{\partial s_i} \frac{\partial f_{k+1}}{\partial s_j}. \]  

(C.5)

\[ = \frac{1}{a} \left( 1 + \frac{\rho^2}{1 - \rho^2} \sum_{k=1}^{K} h^{(i)}_k g^{(i)}_k - \frac{2\rho}{1 - \rho^2} \sum_{k=1}^{K-1} h^{(i)}_k h^{(j)}_{k+1} \right) \]  

(C.6)

where \( h^{(i)}_k = \frac{1}{f_k} \frac{\partial f_k}{\partial s_i} \). In Equation C.5, we used the fact that \( Q^{-1} \) is a banded matrix.

We now move on to the derivation of \( J_{\text{cov}} \). We first derive \( \left( \frac{\partial Q}{\partial s_i} \right)_{kl} \):

\[ \left( \frac{\partial Q}{\partial s_i} \right)_{kl} = a \frac{f_k^\alpha}{f_i^\alpha} \frac{\partial f_k}{\partial s_i} \exp(-||\phi_k - \phi_i||L) + a \frac{f_i^\alpha}{f_k^\alpha} \frac{\partial f_i}{\partial s_i} \exp(-||\phi_k - \phi_i||L) \]  

(C.7)

\[ = \alpha \left( \frac{1}{f_k} \frac{\partial f_k}{\partial s_i} + \frac{1}{f_i} \frac{\partial f_i}{\partial s_i} \right) Q_{kl} \]  

(C.8)

\[ = \alpha (g^{(i)}_k + g^{(i)}_l) Q_{kl} \]  

(C.9)

where \( g^{(i)}_k = \frac{1}{f_k} \frac{\partial f_k}{\partial s_i} \). Using Equation C.9, we get:

\[ J_{\text{cov}} = \frac{1}{2} \text{Tr} \left[ \frac{\partial Q}{\partial s_i} Q^{-1} \frac{\partial Q}{\partial s_j} Q^{-1} \right] \]  

(C.10)

\[ = \frac{\alpha^2}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{m=1}^{K} \sum_{n=1}^{K} Q_{kl}(g^{(i)}_k + g^{(i)}_l)(Q^{-1})_{lm} Q_{mn} (g^{(j)}_m + g^{(j)}_n)(Q^{-1})_{nk} \]  

(C.11)

\[ = \alpha^2 \sum_{k,l=1}^{K} Q_{kl} g^{(i)}_k g^{(j)}_l (Q^{-1})_{lk} + \alpha^2 \sum_{k,l=1}^{K} Q_{kl} g^{(i)}_k g^{(j)}_l (Q^{-1})_{lk} \]  

(C.12)

where in Equation C.11, we used the definition of the trace operator and Equa-
tion C.12 results from the expansion and re-arrangement of Equation C.11. To derive a more explicit expression for Equation C.12, we note that:

\[
Q_{kl}(Q^{-1})_{lk} = a_\rho |k-l| f_k f_l \frac{1 + \rho^2}{1 - \rho^2} \left[ \delta_{lk} - \frac{\rho}{1 + \rho^2} (\delta_{l+1,k} + \delta_{l-1,k}) \right] \quad (C.13)
\]

\[
= \rho |k-l| \frac{1 + \rho^2}{1 - \rho^2} \left[ \delta_{lk} - \frac{\rho}{1 + \rho^2} (\delta_{l+1,k} + \delta_{l-1,k}) \right] \quad (C.14)
\]

If we plug this last expression into Equation C.12, the first term in Equation C.12 becomes \( \alpha^2 \sum_{k=1}^K g_k^{(i)} g_k^{(j)} \), because \( \sum_{k=1}^K Q_{kl}(Q^{-1})_{lk} = 1 \). For the second term of Equation C.12, we have:

\[
= \sum_{k,l=1}^K g_k^{(i)} g_l^{(j)} |k-l| \rho |k-l| \frac{1 + \rho^2}{1 - \rho^2} \left[ \delta_{lk} - \frac{\rho}{1 + \rho^2} (\delta_{l+1,k} + \delta_{l-1,k}) \right] \\
= - \sum_{k=2}^K g_k^{(i)} g_k^{(j)} \frac{\rho^2}{1 - \rho^2} + \sum_{k=1}^K g_k^{(i)} g_k^{(j)} \frac{1 + \rho^2}{1 - \rho^2} - \sum_{k=1}^{K-1} g_k^{(i)} g_{k+1}^{(j)} \frac{\rho^2}{1 - \rho^2} \\
= \alpha^2 \sum_{k=1}^K g_k^{(i)} g_k^{(j)} \frac{1 + \rho^2}{1 - \rho^2} - \alpha^2 \sum_{k=1}^{K-1} g_k^{(i)} g_{k+1}^{(j)} \frac{2 \rho^2}{1 - \rho^2} \quad (C.15)
\]

Combining this with the first term in Equation C.12, we get:

\[
J_{\text{cov}} = \alpha^2 \left( \frac{2}{1 - \rho^2} \sum_{k=1}^K g_k^{(i)} g_k^{(j)} - \frac{2 \rho^2}{1 - \rho^2} \sum_{k=1}^{K-1} g_k^{(i)} g_{k+1}^{(j)} \right) \quad (C.16)
\]

Finally, combining \( J_{\text{mean}} \) (Equation C.6) and \( J_{\text{cov}} \) (Equation C.16) together, the \( ij \)-th
element of the Fisher information matrix is given by:

\[
J_{ij}(s) = J_{\text{mean}} + J_{\text{cov}} = \frac{1}{a} \left( \frac{1}{1 - \rho^2} \sum_{k=1}^{K} h^{(i)}_k h^{(j)}_k - \frac{2 \rho}{1 - \rho^2} \sum_{k=1}^{K-1} h^{(i)}_k h^{(j)}_{k+1} \right) + \alpha^2 \left( \frac{2}{1 - \rho^2} \sum_{k=1}^{K} g^{(i)}_k g^{(j)}_k - \frac{2 \rho^2}{1 - \rho^2} \sum_{k=1}^{K-1} g^{(i)}_k g^{(j)}_{k+1} \right)
\]

(C.17)

where \( h^{(i)}_k = \frac{1}{f_k} \frac{\partial f_k}{\partial s_i} \) and \( g^{(i)}_k = \frac{1}{f_k} \frac{\partial f_k}{\partial s_i} \). Equation C.17 is identical to Equation 4.4 in Chapter 4.