

Matthew T. Harrison — Research Statement

My research has focused primarily on applications related to neuroscience, information theory and computer vision. These topics provide ample opportunities to participate in collaborative, interdisciplinary research — something that I value and enjoy — and they relate to a subject that I find fascinating: the mathematical and computational foundations of learning and intelligence. They also provide an endless supply of interesting mathematical and statistical problems. Here are some recent examples of my research:

- Conditional inference for exploratory data analysis [6][7][11].
- Uniform generation of binary matrices with marginal constraints [9].
- Multiple hypothesis testing for random measures [6][10].
- Estimation of the rate-distortion function [12].
- Large deviations for approximate pattern matching [8].
- A lossy minimum description length (MDL) principle [13][14].

Several of the items in this list are ongoing and will continue to shape my research agenda in the near future.

Motivating Applications

Over the next several years I plan to continue existing collaborations and develop new collaborations with scientists studying biological and machine intelligence. Anticipating the specific research questions that will arise is difficult, but probability, statistics and computing will surely play important roles. I expect neuroscience, information theory and computer vision to remain central to my work.

Neuroscience

Perception, cognition and behavior arise in part from the electrical signals among billions of neurons. The details of this process are still a mystery. Modern neurophysiological techniques permit the simultaneous recording of the electrical activity of hundreds of individual neurons (and this number is quickly growing) or, more coarsely, the global activity of brain regions comprised of hundreds of thousands of neurons. The high dimensionality of the resulting data sets, combined with the peculiarities of neural activity patterns, create challenging statistical issues.

Methods for exploratory statistical analysis are crucial right now in neuroscience, and will probably remain crucial for years to come. This is not to say that statistical modeling is premature — see [15] for a great modeling example — but rather, that models are typically used in an exploratory fashion. The conflux of rapidly changing technology, high-dimensional datasets, demand for exploratory methods, and sustained scientific excitement (and funding) all combine to create a fertile environment for statistical creativity and novel methodology.

Information theory

Information theory provides a rigorous probabilistic setting for understanding the fundamental limits of digital coding and communication. Because of its tight connection to probability theory, many results in probability and statistics can be interpreted from an information theoretic point of view. This new perspective has led to a variety of insights, such as the minimum description length (MDL) approach to model selection and learning [1]. Information theoretic ideas and interpretations are becoming commonplace in several applied areas, such as the biological sciences. Typically these applications borrow ideas, like entropy, that are related to theoretical *lossless* compression.

Theoretical *lossy* compression, on the other hand, is a much less developed field with few existing connections to scientific applications. The rate-distortion function plays the role of the entropy and describes the optimal amount of achievable compression. Furthermore, practical lossy compression lags far behind practical lossless compression in terms of approaching optimal compression performance. Closing this gap will likely require new mathematical and computational insights, as will drawing tighter connections between lossy compression and scientific applications. (I am convinced such connections exist. For example, lossy compression — which attempts to preserve semantic content at the expense of precise detail — seems more analogous to biological learning and perception than does lossless compression — which must preserve every detail.)

Computer vision

The gap between biological and machine vision is substantial. The technological impact of closing this gap and the potential benefits for society are quite profound. One place where this gap is especially apparent is visual learning. A child can accurately learn a novel visual category (like “helicopter”) with only a few examples (maybe only a single example). But a visual category is an incredibly complicated, high-dimensional statistical object (think of the collection of retinal images that contain a helicopter). How is rapid learning possible?

Knowledge sharing (or transfer) among visual categories is often suggested as a possible solution. Inheriting ideas from my PhD advisor, Stuart Geman, I believe that highly structured statistical models (in particular, *compositional* models [2]) are required for efficient sharing and transfer of knowledge. In any event, visual category learning represents an incredibly difficult statistical problem for which we know a solution exists. Even partial progress will push the boundaries of statistical methodology, generate new technology and provide scientific insights into biological vision.

Example problems

For brevity, I will focus on two problems that I am currently studying: random generation of discrete structures and multi-scale multiple hypothesis testing. These problems are broadly applicable, mathematically interesting and will likely remain research interests of mine for several years. Both were initially motivated by statistical applications in neuroscience, and simple versions of each appeared in a recent *Nature Neuroscience* article describing the exploratory analysis of a large dataset from rat prefrontal cortex collected in G. Buzsáki’s lab [6].

Random generation of discrete structures

Binary matrices (two-way zero-one tables) arise in a variety of settings. In neurophysiology they are used to represent multiple discretized point processes. More generally, they are used to represent occurrence matrices, bipartite graphs, directed graphs (for square matrices), undirected graphs (for symmetric matrices) and other such discrete structures. Fast sampling from a specified distribution over a collection of binary matrices is a challenging problem with broad applicability. Sampling can be used for Monte Carlo statistical inference, approximate counting, stochastic search, and so on.

Chen et al. [4] describe a sequential importance sampling (SIS) approach for approximate uniform sampling over the set of $m \times n$ binary matrices with specified row sums $r = (r_1, \dots, r_m)$ and column sums $c = (c_1, \dots, c_n)$,

$$\Omega(r, c) = \left\{ (z_{ij}) \in \{0, 1\}^{m \times n} : \sum_{k=1}^n z_{ik} = r_i, \sum_{k=1}^m z_{kj} = c_j, 1 \leq i \leq m, 1 \leq j \leq n \right\}.$$

I recently observed [9] that a combination of dynamic programming (DP) and improved asymptotic enumeration can be used to simplify and improve the SIS algorithm in [4]. The resulting algorithm gives state of the art performance, scales well to moderately large problems (say, 10^5 nonzero entries) and, most importantly, presents a conceptual framework for generalization. The key observation is that both the structural constraints — in this case, the Gale-Ryser conditions for the existence of a $z \in \Omega(r, c)$ — and the recent asymptotic approximations in [3] factor in a way that permits fast and exact sampling via DP.

The algorithm immediately generalizes to the case of forced zero entries (e.g., zero diagonal) and to the case of $\{0, 1, \dots, k\}$ -valued matrices, including, two-way contingency tables, although in some cases the relevant asymptotic theory is non-existent. The approach of combining asymptotic enumeration theory, SIS and DP also seems promising for a variety of related problems. For example, for symmetric binary matrices (or undirected graphs) the relevant structural constraints are encapsulated by the Erdős-Gallai conditions characterizing graphical degree sequences. This condition does not exactly factor in the manner needed for DP, but it very nearly does. Whether the resulting SIS procedure is practical is an empirical question that I should be able to answer in the coming months.

For another example, consider sampling uniformly from

$$\bigcup_{c \in \Gamma} \Omega(r, c)$$

where Γ enforces certain constraints on the column sums. This problem arises in neuroscience where the constraints take the form

$$\Gamma = \left\{ c \in \{0, \dots, m\}^n : \sum_{j=1}^n \binom{c_j}{\ell} = b_\ell, 1 \leq \ell \leq k \right\}$$

for specified integers b_1, \dots, b_k and some $k \geq 2$. Hierarchical importance sampling is a sensible approach: first sample $c \in \Gamma$ from an appropriate distribution and then sample $z \in \Omega(r, c)$ as before. Of course, sampling from Γ may be difficult. For the neuroscience problem I found that the dimensionality of Γ was small enough that its algebraic-geometric structure (cf. [5]) could be exploited to enable fast and exact sampling. This is interesting because algebraic techniques are

often restricted to rather low-dimensional problems — computing Markov and Groeber bases is computationally demanding — whereas here the hierarchical formulation allows them to be used in the context of a high-dimensional problem. In the near future I plan to investigate the extent to which these techniques can be applied more generally.

Multi-scale multiple hypothesis testing

Random measures are a useful representation for studying point process data and functional data. For a point process, the corresponding random measure is defined by $\mu(A) = \#\{\text{events in } A\}$. For functional data f , the corresponding random (signed) measure is defined by $\mu(A) = \int_A f$. Note that for any fixed set A , $\mu(A)$ is just a real-valued random-variable, but the data allow us to compute the value of μ on infinitely many different sets.

Consider repeated observations of a random measure in two experimental conditions. The goal is to identify whether the experimental condition affects the distribution of the random measure and, if so, to localize where the differences occur. Formally, the observations are (label, random measure)-pairs, $(L_1, \mu_1), \dots, (L_n, \mu_n)$, where $L_k \in \{1, 2\}$ indicates which of the two experimental conditions was used in the k th observation and where μ_k is the random measure representation of the k th observation. The goal is to identify whether the conditional distributions of $\mu|L = 1$ and $\mu|L = 2$ are identical, and, if they differ, to further identify the sets where they differ. This can be viewed as a multiple hypothesis testing problem:

$H_0(A)$: the conditional distributions of $\mu(A)|L = 1$ and $\mu(A)|L = 2$ are identical;

$H_1(A)$: the conditional distributions of $\mu(A)|L = 1$ and $\mu(A)|L = 2$ are different;

where we have a different hypothesis test for each subset A in some collection of subsets \mathcal{A} . For example, in the case of temporal processes, \mathcal{A} might be comprised of subsets of the form $(t - \sigma, t + \sigma)$ where t varies over time and σ varies over scale. I call this *multi-scale multiple hypothesis testing* because it is common in practice to fix the scale a priori (often implicitly via smoothing), whereas here we are explicitly considering multiple scales.

A. Amarasingham and I are developing multiple-testing procedures for this setup [10] that (strongly) control the family-wise error-rate (FWER), which is the probability of at least one false rejection. These procedures are based on permutation tests and the well-known max- t (or min- p) procedures of Westfall & Young [16] which replace the usual critical values for testing $H_0(A)$ with a critical value derived from an appropriate joint null distribution for $(\mu(A) : A \in \mathcal{A})$. The algorithms perform well in neuroscientific applications with very little loss of power compared to the fixed-scale multiple testing situation (i.e., t varies but σ is fixed). Control of the FWER requires a reasonable, but hard to verify, modeling assumption. We have also shown that the procedure is robust to local violations of this modeling assumption. The proofs are interesting in that they combine the probabilistic notions of exchangeability and conditional independence with the algebraic structure of permutation groups.

Future directions and outreach

Within the next year I expect to bring several current projects to completion, although providing distribution quality software and communicating the methodology to neuroscientists and other

practitioners may continue longer. I will also continue researching the two abstract problems of random generation of discrete structures and multi-scale multiple hypothesis testing, both of which are likely to be important and active areas for many years to come. These two problems combine mathematics, statistics and computing, meshing well with my training and interests.

Through existing and new collaborations, I will move forward in developing a long-term research program that contributes to the broad scientific endeavors of understanding human intelligence and creating comparable artificial intelligence. Probability, statistics and computing will undoubtedly play prominent roles.

I believe that research is not just an intellectual endeavor, but a human endeavor. For this reason I will strive to develop a research program that combines intellectual excellence with service and with personal connections to students, colleagues and the broader community. Collaborative research, software development, involving undergraduates in research, and bringing research into the classroom and into local schools are some of the things I have already begun to do.

References

- [1] AR Barron, J. Rissanen, and B. Yu, *The MDL Principle in Modeling and Coding*, IEEE Transactions on Information Theory **50** (1998), 2743–2760.
- [2] E. Bienenstock, S. Geman, and D. Potter, *Compositionality, MDL Priors, and Object Recognition*, Advances In Neural Information Processing Systems (1997), 838–844.
- [3] E.R. Canfield, C. Greenhill, and B.D. McKay, *Asymptotic enumeration of dense 0–1 matrices with specified line sums*, Journal of Combinatorial Theory, Series A **115** (2008), 32–66.
- [4] Y. Chen, P. Diaconis, S.P. Holmes, and J.S. Liu, *Sequential monte carlo methods for statistical analysis of tables*, Journal of the American Statistical Association **100** (2005), 109–120.
- [5] P. Diaconis and B. Sturmfels, *Algebraic algorithms for sampling from conditional distributions*, Annals of Statistics **26** (1998), 363–397.
- [6] S. Fujisawa, A. Amarasingham, M.T. Harrison, and G. Buzsáki, *Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex*, Nature Neuroscience **11** (2008), 823–833.
- [7] S. Geman, A. Amarasingham, M.T. Harrison, and N.G. Hatsopoulos, *The statistical analysis of temporal resolution in the nervous system*, (Submitted).
- [8] M.T. Harrison, *The generalized asymptotic equipartition property: Necessary and sufficient conditions*, IEEE Transactions on Information Theory **54** (2008), 3211–3216.
- [9] ———, *A dynamic programming approach for sequential importance sampling of binary matrices*, (In preparation).
- [10] M.T. Harrison and A. Amarasingham, *Multi-scale multiple hypothesis testing*, (In preparation).
- [11] M.T. Harrison and S. Geman, *A rate and history-preserving resampling algorithm for neural spike trains*, Neural computation (In press).
- [12] M.T. Harrison and I. Kontoyiannis, *Estimation of the rate distortion function*, IEEE Transactions on Information Theory **54** (2008), 3757–3762.
- [13] M.T. Harrison, I. Kontoyiannis, and M. Madiman, *A minimum description length principle in lossy data compression*, (In preparation).

- [14] M. Madiman, M. Harrison, and I. Kontoyiannis, *Minimum Description Length vs. Maximum Likelihood in Lossy Data Compression*, IEEE International Symposium On Information Theory, 2004, pp. 461–461.
- [15] J.W. Pillow, J. Shlens, L. Paninski, A. Sher, A.M. Litke, EJ Chichilnisky, and E.P. Simoncelli, *Spatio-temporal correlations and visual signalling in a complete neuronal population*, Nature **454** (2008), 995–999.
- [16] P.H. Westfall and S.S. Young, *Resampling-based multiple testing*, Wiley, 1993.