Geometric Embedding for Learning Combinatorial Structures

Terran Lane[1], **Ben Yackley, Sergey Plis, Stephen McCracken, Blake Anderson**
University of New Mexico
Albuquerque, NM 87131 USA
{terran,benj,pliz,samcc,blake}@cs.unm.edu
http://www.cs.unm.edu/∼terran/

Introduction We examine the problems of optimization and Bayesian reasoning in combinatorial spaces including graph structures and permutations.

One strategy is to embed combinatorial sets into continuous spaces and work there with powerful analytic tools. This trick has proven quite effective in, for example, continuous relaxations of integer programming problems [2]. Some work in machine learning has used implicit or explicit continuous embeddings of combinatorial problems [1, 6, 4], but the connection to the topology and analytic properties of the embedding space is typically not made explicit, nor fully exploited.

We argue that the embedding approach offers a powerful way to finesse some of the challenges of combinatoric spaces. We outline two of our research directions in this space: rapid Bayesian network structure identification via induced topology, and Bayesian inference on permutations via low-dimensional embedding.

Fast Bayes Net Structure Search A key learning problem for Bayesian networks (BNs) is *structure identification*: finding the "best" structure, corresponding to a set of observed data. The structure identification problem is NP-hard and considerable work has been devoted to efficient algorithms for it.

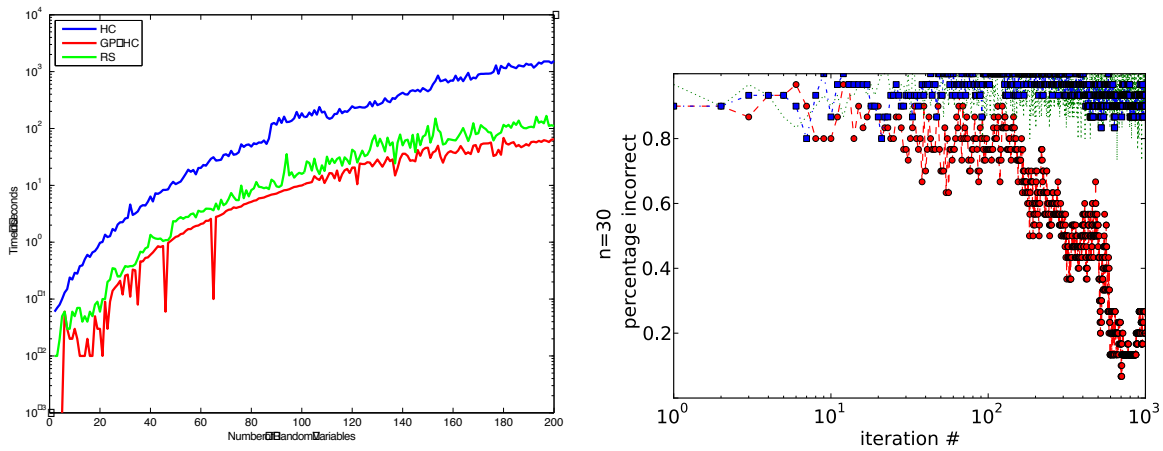
The core optimization problem is $\widehat{G} = \arg\max_{g \in \mathcal{G}} S(g|D)$, where $S()$ is some *score function* (such as BDe or BIC) that measures the goodness of fit between a proposed graph, g, and a set of data, D. The exponential challenge of this optimization problem comes down to the combinatorial structure of the space of DAGs, $\mathcal{G}$, but evaluating the score function can also be a significant computational burden. Even with a decomposable score, the scan incurred in accumulating even per-variable sufficient statistics can be quite expensive for large data sets.

We observe that $S : \mathcal{G} \to \mathbb{R}$ is simply a function mapping, so we can build a function approximator for it. If this approximator, $f() \approx S()$, is sufficiently accurate and is fast to evaluate (say, constant time), then $f()$ can be used as a proxy for $S()$ in any incremental search for $\widehat{G}$.

To build the approximator $f()$, we embed the set of graphs on d nodes into an $O(d^2)$-dimensional Euclidean space by imposing a distance metric between adjacency-matrix representations of the graphs. This imposes a hypercube topology on the space of graphs: the "meta-graph," a graph over the space of graphs. From this topology we construct standard function approximators, such as linear models over the Fourier basis or Gaussian process regressors.

Figure 1 (**a**) demonstrates the speedup that can be achieved by using two versions of our proxy score function, $f()$ (GP-HC and RS), compared to the same search strategy using exact score evaluation, $S()$ (HC). Both GP-HC and RS achieve better than an order-of-magnitude speedup over HC (vertical axis), over a wide range of Bayesian network sizes (horizontal axis). The speedup is particularly pronounced for very large data sets, where HC's time scales with data set size and GP-HC/RS are effectively independent of data set size (data not shown).

[1] To present.



(**a**) (**b**)

Figure 1: (**a**) Structure identification speedup from two variants of the meta-graph approach to BN structure search (GP-HC and RS) over standard hillclimbing search with exact score evaluation (HC). x axis: number of random variables in the target BN; y axis: total time for search, on a log scale. (Down is faster.) (**b**) Error rate (y axis; down is good) of permutation identification from noisy permutation observations ($n = 30$ objects) as a function of number of observations (x axis). Red/circles line: our permutation embedding method; blue/squares line: cumulative averaging of observations; green/plain line: raw observations.

Bayesian Inference on Permutations A key component of object tracking, ranking, and related tasks is *Bayesian inference on permutations*: representing and manipulating probability distributions over the super-exponentially large family of permutations of n objects. An exact representation of a general PDF requires factorially many parameters. Prior approaches have approximated a general PDF with a restricted set of basis functions [3], or have embedded the permutation space only implicitly and worked with a heuristically chosen probability distribution [6].

We start with the $n \times n$ matrix representation of a permutation, and take the metric to be the Euclidean distance between (vectorizations of) these matrices. This metric implies a specific embedding into $\mathbb{R}^d$, in which the permutations fall on the surface of a hypersphere. In turn, this suggests a natural family of probability distributions over permutations: the Fisher-Bingham distributions, of which a common special case is the von Mises-Fisher distribution (vMF) [5].

We can give polynomial (in n) time mappings between permutation space and the Euclidean hypersphere. This allows us to sample data from permutation space, efficiently map to Euclidean space, analytically and quickly manipulate probability distributions there (marginalization and projection steps of Bayesian inference), and then efficiently map resulting inferences back to permutation space.

Coupling the probability representations to the mapping operations bridges the gap between the discrete, combinatorial space of permutations and the continuous, low-dimensional hypersphere. This allows us to lift the large body of results developed for directional statistics [5] directly to permutation inference. Figure 1 (**b**) illustrates that our approach yields accurate reconstructions of permutations from noisy observations, as might arise in object tracking applications.

Acknowledgements This material is based upon work supported by the National Science Foundation under Grant No. 0705681. Dr. Plis's work was supported by the NIH under grant number NCRR 1P20 RR021938.

References

[1] M. A. Fligner and J. S. Verducci. Distance based ranking models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):359–369, 1986.

[2] R. E. Gomory. Outline of an algorithm for integer solutions to linear programs. *Bulletin of the American Mathematical Society*, 64(5):275–278, 1958.

[3] J. Huang, C. Guestrin, and L. Guibas. Fourier theoretic probabilistic inference over permutations. *Journal of Machine Learning Research*, 10:997–1070, May 2009.

[4] T. Jaakkola, D. Sontag, A. Globerson, and M. Meila. Learning bayesian network structure using lp relaxations. In Y. W. Teh and M. Titterington, editors, *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[5] K. V. Mardia and P. E. Jupp. *Directional Statistics*. John Wiley & Sons, Ltd., second edition, 2000.

[6] M. Meila, K. Phadnis, A. Patterson, and J. Bilmes. Consensus ranking under the exponential model. In *Proceedings of the 23rd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.