

Personal Statement

Peter Hoff

1 Research statement

My research involves the development of statistical models that can uncover complicated patterns in data. This began with work as a graduate student on latent variable methods for nonparametric mixture models, and I have since expanded the use of latent variables to construct techniques for cluster analysis and to develop models for relational data. These statistical methods have applications across many disciplines, including the social sciences and cancer genetics.

Much of the excitement of data analysis derives from the uncovering of scientifically meaningful structure in noisy, complex data. However, an overzealous exploration of data may erroneously identify seemingly meaningful patterns that are actually the result of experimental noise. On the other hand, a data analysis approach that is too restrictive runs the risk of mischaracterizing valuable scientific information. My research in statistical inference has tried to address the concerns of both of these approaches, via the development of methods which are based on simple, parsimonious models but have the ability to expand to uncover more complicated structure in the data. Ideally, a researcher using such methods can address basic, pre-specified scientific questions, but can also uncover more complicated patterns in the data if given enough evidence. Some examples of this in my research are outlined below.

1.1 Latent variable methods for relational and social network data

Hoff, P.D., Raftery, A.E., and Handcock, M.S. (2002). Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, Vol. 97, no. 460, 1090-1098.

Hoff, P.D. (2003a). Random Effects Models for Network Data. In *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, 303-312, Ronald Breiger, Kathleen Carley, and Philippa Pattison, eds., Washington, D.C., The National Academies Press.

Ward, M.D., Hoff, P.D., and Lofdahl, C.L. (2003b). Identifying International Networks. In *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, 345-359, Ronald Breiger, Kathleen Carley, and Philippa Pattison, eds., Washington, D.C., The National Academies Press.

Hoff, P.D., and Ward, M.D. (2004). Modeling Dependencies in International Relations Networks. *Political Analysis* Vol. 12, No. 2, 160

Hoff, P.D. (2005). Bilinear Mixed-Effects Models for Dyadic Data. *Journal of the American Statistical Association*, Vol. 100, no. 469, 286-295.

Shortreed, S., Handcock, M.S. and Hoff, P.D. Positional Estimation within the Latent Space Model for Networks. *Methodology*, Vol. 2, no. 1, 24-33.

Ward, M.D. and Hoff, P.D. (2007) Persistent Patterns of International Commerce. *Journal of Peace Research*, Vol. 44, no. 2, 157-175.

Hoff, P.D. (2007) Model averaging and dimension selection for the singular value decomposition *Journal of the American Statistical Association*, Vol. 102, no. 478, 674-685.

Relational data consist of measurements that are made on pairs of objects or under pairs of conditions. This framework has many applications across disciplines including the study of disease transmission, social interactions among individuals, political interactions among nations, and relationships among genes in biological systems. In many cases the relational measurements are all either zero or one, indicating the presence or absence of ties between pairs of objects. Such data are typically referred to as social networks. Two common approaches to analyzing social network data are (a) representation via graphical methods and (b) estimation of parameters in a statistical model. Roughly speaking, the graphical approach is to represent the social network data visually, plotting individuals as points and then drawing lines between those individuals having ties to one another. Although such procedures allow researchers to visualize complicated network structures, these methods are generally descriptive and do not provide statistical inference, such as tests of hypotheses, parameter estimates, confidence intervals or predictions for ties that are unmeasured.

Statistical inference for social network data is generally done in the context of a statistical model. However, models for these data can be difficult to formulate because of the highly interdependent nature of the data. For example, a typical network may include some highly active members as well as some relatively inactive members. This induces a statistical dependence among the relations common to an individual, a dependence that invalidates “standard” statistical models based on independence of the network ties. Many networks also display more complicated dependence patterns: For example, the presence of ties from person i to person j and from j to k often correlates with a tie from i to k . Such a pattern is called transitivity in the social networks literature, and further invalidates the use of models which assume independence of the ties.

Pioneering work in modeling the interdependence in social networks utilized log-linear and exponentially parameterized random graph models. These models are typically very parsimonious, using a small number of parameters to measure global patterns in the network. Parameter estimates in these models are sometimes difficult to obtain and do not capture “local” network structure very well, such as clusters of connected individuals. Additionally, these models can not be easily extended to allow for more general relational data, where the ties take on arbitrary numerical values.

My contribution to this area has been the development of statistical models for the analysis of relational data of various types. My goal has been to develop a methodology which provides the benefits of both the graphical and the model-based approaches, while at the same time allows for feasible parameter estimation. My initial strategy (developed in Hoff, Raftery and Handcock 2002) was to conceptualize the nodes of a network as lying in some unobserved, latent social space. The model is formulated so that the tendency to form ties is increasing as nodes get closer together in this space. This model can represent common network behavior such as reciprocity and

transitivity, and the latent locations themselves can be estimated and plotted to provide a visual representation of the network. Furthermore, the model-based nature of the approach provides confidence intervals for model parameters and latent locations, as well as predictions about potential ties that are unmeasured. In Hoff (2005) I discuss an alternative latent variable representation in which the strength of a tie between two individuals is modeled as the product of latent characteristic vectors specific to the individuals. This formulation has a number of conceptual and computational advantages, and can be extended to a more general class of reduced-rank interaction models, a current research area of mine.

In collaboration with political scientists, I have found these types of models to be extremely useful for uncovering structure in international relations networks. For example, in Ward, Hoff and Lofdahl (2003) and Hoff and Ward (2004), we show that these latent space models are able to outperform more standard logistic regression models in terms of predicting conflicts between nations. In a recent paper (Ward and Hoff, submitted), we show how a latent space model is able to uncover systematic lack-of fit of standard models of international trade, and provide a much more complete picture of international relations.

The analysis of relational data currently dominates my research agenda, and involves a number of collaborators and graduate students as well as two grants on which I am the primary investigator. Specific application areas currently include the evaluation of theories of international trade and conflict and the analysis of protein interaction networks in biological systems. Methodological research that I am focusing on includes inferential methods for dynamic social networks and the forecasting of relational data.

1.2 Subspace clustering

Hoff, P.D. (2004) Discussion of “Clustering objects on subsets of attributes,” by J. Friedman and J. Meulman. *Journal of the Royal Statistical Society*, Vol 66, no. 4, 845.

Hoff, P.D. (2005) Subset clustering of binary sequences, with an application to genomic abnormality data. *Biometrics*, Vol 61, no. 4, 1027-1036.

Hoff, P.D. (2006) Model-based Subspace Clustering. *Bayesian Analysis*, Vol. 1, no.2, 321-344.

Cluster analysis is the categorization of objects into separate groups so that the attributes of objects within a group are similar. For example, consider the case where the objects are tumor cells from different patients, and the attributes are measures of genetic damage at given locations along the genome of a tumor cell. Identifying clusters of tumors could help categorize different forms of cancer and characterize the mechanisms by which normal cells undergo tumorigenesis.

Typical clustering methods compute similarities between objects based on an entire set of selected attributes. However, when the number of measured attributes is large, it may be the case that two given groups differ at only a subset of the measured attributes, and so only a subset of the attributes are “relevant” to the clustering. Furthermore, it may be the case that the attributes that are relevant for differentiating one pair of groups are different from those that differentiate

another pair. In such cases, traditional clustering methods may fail because the differences between any two groups, averaged over all the attributes, is small.

To accommodate such cluster structure a number of techniques have recently been developed called “subspace clustering algorithms”. These algorithms generally operate by searching over a large set of clusterings in order to optimize a heuristic similarity measure. In contrast, in the articles listed above I have formulated a novel model-based approach for subspace clustering and have developed estimation methods which identify the number of clusters, the cluster memberships, and the relevant attributes for each cluster in a unified way. In Hoff (2005) I develop this approach for binary data, and then implement it to arrange a set of renal cell carcinomas into groups based on their patterns of genetic damage. This analysis results in a more refined clustering of these data than had been possible using more restrictive clustering approaches. In Hoff (to appear) I extend the method to the case of continuous data. In this situation, differences between clusters are measured by differences in means and/or variances at subsets of attributes. This model-based approach is shown to outperform an algorithmically-motivated subspace clustering method, and also outperforms model-based methods that form clusters based on differences at all attributes.

My basic approach to subspace clustering is quite general. In ongoing work I am making use of this generality to allow for the measured attributes of an object to be of varying data types, for example, some variables being continuous and others discrete. Additionally I am working on modeling the within-cluster covariance structure of the attributes, as well as developing approximate parameter estimation methods to allow for the analysis of massive datasets.

1.3 Nonparametric inference via mixture models

Hoff, P.D., (2000). Constrained Nonparametric Maximum Likelihood via Mixtures. *Journal of Computational and Graphical Statistics*, Vol. 9, No. 4, 633-641.

Hoff, P.D., Halberg, R.B., Shedlovsky, A., Dove, W.F., and Newton, M.A. (2002). Identifying Carriers of a Genetic Modifier Using Nonparametric Bayes Methods. In *Case Studies in Bayesian Statistics 5*, Springer-Verlag, 327-342.

Hoff, P.D. (2003a). Nonparametric Estimation of Convex Models via Mixtures. *Annals of Statistics*, Vol. 31, No. 1, 174-200.

Hoff, P.D. (2003b). Bayesian Methods for Partial Stochastic Orderings. *Biometrika*, Vol. 90, No. 2, 303-317.

I began research in this area as a graduate student, and was initially motivated by the seemingly simple problem of estimating the mean μ of a population based on samples y_1, \dots, y_n . In this case we typically write $y_i = \mu + \epsilon_i$, where the ϵ_i 's represent mean-zero sampling error or noise. Typical statistical analysis proceeds by presuming the distribution of errors belongs to some known class of distributions, for example, mean-zero normal distributions. I was interested in eliminating such assumptions and allowing the errors to have any mean-zero distribution. My approach to estimation and inference for μ required that I be able to estimate an error distribution restricted to lie in the class of all mean-zero distributions.

This particular restriction is a convex constraint, a type of constraint which arises quite frequently in semiparametric statistical inference. In addition to the estimation of means, convex constraints play a role in the estimation of quantiles, estimation of unimodal distributions and estimation subject to partial stochastic orderings. Nonparametric estimation of a probability distribution constrained to lie in such convex sets can be tricky. In the work listed above I make use of mathematical results from functional analysis to show how this potentially difficult constrained estimation problem can be rewritten in terms of an unconstrained mixture problem, a type of problem that is familiar to statisticians. In Hoff (2000) I show how to use this mixture representation approach to obtain maximum likelihood estimates for constrained distributions. The general mathematical theory for this approach is laid out more formally in Hoff (2003a), in which I also show how to construct prior distributions on convex sets of probability distributions. This last item makes nonparametric Bayesian inference possible for these types of problems.

One of the more useful applications of this technique has been in the estimation of probability distributions subject to a partial stochastic ordering. Roughly speaking, a distribution P_1 is stochastically larger than another distribution P_2 if samples from P_1 tend to be larger than those from P_2 . In Hoff (2002) and Hoff (2003b) I discuss in detail the mixture representation approach in the context of partial stochastic orderings, and give example data analyses showing how this approach can be useful in two different applications to cancer genetics research.

My current research in nonparametric methods makes use of a model averaging procedure to obtain nonparametric confidence intervals for means and other quantities, where the averaging tends to put most of the weight on models that have a small number of parameters unless there is a substantial amount of information in the data suggesting that a more complicated model is appropriate. This approach favors simpler models than the corresponding mixture representation approach to the problem, and the parameter estimates in this new approach are more stable, easier to obtain, and methodology is more transparent.

2 Teaching statement

Statistics 421: Applied regression and experimental design

Statistics 502: Applied regression and experimental design

Statistics-Math 523: Advanced theory of probability

CSSS-Statistics 564: Bayesian statistics for the social sciences

CSSS-Statistics 567: Social network analysis

As is the case for teachers in many service departments, Statistics instructors are often in the position of teaching an extremely heterogeneous group of students whose main academic interests do not include the course material. This was the situation I faced in my first year at the University of Washington, teaching Stat 421, CSSS 564 and CSSS 567. One thing I've learned from teaching these classes is that students are much more open to studying a subject once they understand *why* they should be studying it.

Realization of this somewhat obvious fact has had a great impact on my lecture style. Most of the students in these three classes are in scientific disciplines and will eventually need to gather and analyze their own data. My approach to lecturing in these classes attempts to reflect this. I typically introduce a new statistical concept by first presenting an example scenario consisting of a group of researchers, their data and their scientific questions. We then discuss how we might address the researchers' questions using concepts previously covered in the course, as well as the limitations of these concepts. These limitations then motivate the introduction of a new statistical concept, which we then use to address the scientific questions of the example. With this approach I attempt to emphasize the motivation and main ideas behind the course material during lecture, and then provide opportunities for exploring the finer details via handouts, homework and reading assignments. I believe that for classes like these, such an approach is the most efficient use of lecture time for both the teacher and instructor: The pace at which a group of intelligent, motivated but mathematically heterogeneous students can understand general concepts is much more uniform than the pace at which they can go through mathematical derivations.

My lecture style in Stat-Math 523 is quite a bit different. The students in this class are advanced graduate students who will likely be doing original research in statistical methodology. Such research requires that students are able to understand and prove theorems completely correctly with no errors. To emphasize this, I have used a non-standard homework policy in which students receive no partial credit on homework problems, but can resubmit a problem for grading as many times as they wish. Although this approach generally reduces the total number of problems a student works on during the quarter, even the students who complete only a fraction of the assigned problems have gained the confidence that they can at least prove *some* theorems completely correctly.

I take my teaching very seriously and have put a lot of effort into it. I was delighted to be nominated for a UW Distinguished Teaching award in 2006, and a Distinguished Mentoring award in 2005.