RESEARCH STATEMENT

JYOTISHKA DATTA
DEPARTMENT OF STATISTICS, PURDUE UNIVERSITY

In the recent past, thanks to applications in genomics, finance and astronomy as well as other fields, high-dimensional statistical inference has been a major trend in research and practice. Many large scale applications involve multiple testing, where one simultaneously tests a small proportion of true signals in presence of a large number of noise observations, or sparse regression, where the number of covariates is much larger than the sample size, or covariance matrix estimation where the number of variables is much larger than the number of observations. There has been a tremendous development in terms of theoretical, methodological and computational work that is still going on. With the accelerating growth in the size of datasets, the need for methodological advances as well as the issues of computation and scalability has come to the fore. I have chosen to work in the broad area of model selection, multiple testing and covariance estimation for high dimensional data with a focus on both theory and application to real data problems.

Besides my strong methodological background, I have also gained considerable experience in both applied and computational research, through my collaborative work with other faculty members from the Statistics department at Purdue University as well as interdisciplinary work with researchers from other disciplines. I consider it my strength to recognize challenging problems arising from other disciplines and use my theoretical and computational skills to develop solutions to them.

My doctoral research began with a theoretical study of the Bayes risk of a recently proposed sparse signal-detection procedure, called the Horseshoe estimator due to Carvalho et al. (2010). This paper provided a theoretical footing to the use of a continuous shrinkage prior in the context of a multiple testing problems and provided conditions under which the Bayes risk for the Horseshoe prior attains the Bayes risk for the oracle up to a multiplicative constant. The findings of this research have been published in *Bayesian Analysis*, and I have received the William J. Studden publication award given by the Department of Statistics at Purdue University for an outstanding mathematical statistics paper accepted for publication.

My ongoing methodological research spans model selection and high-dimensional covariance matrix estimation. In model selection, we compare two Bayesian approaches, namely, the “optimal” prior, due to Bayarri et al. (2012), and the continuous shrinkage priors, e.g. the Horseshoe prior due to Carvalho et al. (2010), with the frequentist gold-standard like the LASSO in terms of their variable selection performances as well as estimation accuracies. In covariance matrix estimation, we have developed a novel shrinkage estimator for estimating the inverse covariance matrix that enjoys both computational feasibility and nice theoretical properties.

The applied part of my dissertation deals with different less-studied aspects of multiple testing such as estimation of the false negative rate, effect of cross-validation, and effect of strong dependence on multiple testing. In an article submitted to *Sankhya-B*, I have proposed an Empirical Bayes estimator of the FNR and provided results for the achieved FDR after internally cross-validating the results. In another paper, I have explored a few less studied aspects of Bootstrap, including its applications in many high-dimensional problems, e.g. the estimation of the inverse covariance matrix in a high-dimensional space. This paper has been published in *Statistical Methodology* (see Datta and Ghosh (2013b)).

As part of an ongoing interdisciplinary work with the National Institute of Biomedical Genomics, India, I am studying multiple testing with arbitrarily strong dependence between the test statistics, motivated by the need to locate a disease causing locus on a chromosome through the use of marker loci. To address the combined effect of multiplicity and dependence, we propose a novel test-statistic combining two or more dependent test-statistics. This new test statistic seems to be informative for all the parameters involved.

In the following sections, I describe some of my recent contributions to each of these areas and state my future research goals.

1 Multiple Testing

Asymptotic Properties of Bayes Risk for the Horseshoe Prior

Simultaneous testing of many hypotheses has been one of the most important challenges in the analysis of large-scale datasets arising from microarrays and other sources. A popular model for sparse signal-recovery is the two-groups model (Efron (2008)), with a point mass at zero for noises and a continuous density for the signals. In a series of remarkable papers, Carvalho, Polson, and Scott (2010, 2009); Scott (2011) introduced a continuous “one-group” shrinkage rule based on what they call the horseshoe prior for multiple testing and model selection. The name ‘Horseshoe’ is attributed to the shape of the density of the shrinkage weight, $\kappa$, for each observation. The one-group model needs significantly less computational effort than the two-groups model. Moreover, Carvalho et al. (2010) go on to provide strong numerical evidence that the “one-group” shrinkage rule approximately behaves like the answers from a two-groups model. However, a thorough analysis of the asymptotic properties of the one-group model was still not available.
In my dissertation research I have studied the asymptotic properties of the Bayes risk for the Horseshoe prior in the context of multiple testing. Our paper provides a clear, comprehensive demonstration that the Bayes risk for the Horseshoe prior attains the oracle up to a multiplicative constant, depending on the hyperparameters of the model (Datta and Ghosh (2013a)). Given the (often extreme) computational difficulty associated with the two-groups model, and the wide use of variable-selection priors, this correspondence is of fundamental interest to Bayesians.

To prove the optimality of the natural decision rule induced by the horseshoe prior, we have used the same asymptotic framework as Bogdan, Chakrabarti, Frommlet, and Ghosh (2011) who introduced the Bayes oracle in the context of multiple testing and provided conditions under which the Benjamini-Hochberg and Bonferroni procedures attain the risk of the Bayes oracle. We prove two concentration inequalities for the posterior distribution of the shrinkage coefficient, \( \kappa \), given an observation \( Y \) near zero and near the tails. It turns out that the shrinkage behavior of the Horseshoe rule is governed by two shrinkage parameters: a global shrinkage parameter \( (\tau) \) that adapts to the sparsity level and a local shrinkage parameter \( (\lambda_i) \) that flags the large observations. Using these properties, we have proved the following theorem:

**Theorem 1.1.** If the "global shrinkage parameter" \( \tau \) of the horseshoe prior is chosen to be of the same order as the proportion of signals \( p \), then the natural decision rule induced by the horseshoe prior attains the risk of the Bayes oracle up to \( O(1) \) with a constant close to the constant in the oracle.

We provide numerical evidence that the optimality holds if we use the full Bayes estimate of \( \tau \), which seems to adapt to the unknown sparsity in the data. Our theoretical results provide some insights into the asymptotic behavior of the continuous one-group model in the context of multiple testing. The theoretical as well as the numerical results support the observation of Carvalho et al. (2010) that the one-group model can closely mimic the results from Bayesian methods on the two-groups model when the global shrinkage parameter is suitably tuned to handle the sparsity in the data. For details, see Datta and Ghosh (2013a). The intuitive reason why the horseshoe prior works so well is that the posterior inclusion probability of the two groups model is well captured in the shrinkage weight \( 1 - E(\kappa|Y) \) of the horseshoe prior.

This paper was selected for the William J. Studden publication award given by the Department of Statistics at Purdue University for a paper published in a Mathematical Statistics journal. This award is given at most once a year for an outstanding publication by a Ph.D. candidate in the Department of Statistics.

An Empirical Bayes Approach to False Negative Rate Calculation and Cross-Validation in Large Scale Hypotheses Testing

In the context of multiple testing, controlling the False Discovery Rate (FDR) has become very important. A remarkable result in theory of multiple testing due to Benjamini and Hochberg (1995) asserts that the false discovery rate can be controlled at any pre-assigned given value by using a particular test based on the \( p \)-values. The main focus of most of the research on multiple testing has been on controlling or estimating the FDR and estimating the parameters of the models being used. This theory, due to Benjamini and Hochberg (1995), based on the use of \( p \)-values is quite well-developed.

In this framework, the natural question of estimating the False Negative Rate (FNR) has been somewhat neglected. Another interesting work which deserves further study is by Majumder et al. (2009) where a novel scheme of internal cross-validation using half-samples is proposed. Their cross-validation scheme based on the outcomes of Benjamini-Hochberg rule makes the results of a large-scale testing problem more credible to users.

In this paper, we have examined the estimation of the False Negative Rate in the context of large-scale hypotheses testing using Benjamini-Hochberg’s procedure. We presented an estimator based on Empirical Bayes interpretation of the two-groups model used in multiple testing problems. Our numerical results suggest that for the two-groups normal model, the Empirical Bayes estimator \( \hat{FNR} \) is very accurate in terms of estimating the FNR. We also discuss a class of estimators of FNR based on the estimation of the proportion of true null effects and compare a few popular estimators that exist in literature.

The article also sheds some light on the achieved false discovery rate when the outcomes of the Benjamini-Hochberg procedure is internally cross-validated using the half-sample approach of Majumder et al. (2009). We provide some numerical support to the claim that the final FDR after cross-validation is controlled at \( \alpha^2 \) if Benjamini-Hochberg procedure at significance level \( \alpha \) is applied to both the half-samples.

Multiple Testing Procedures under Dependence

In large-scale association studies in respect of a disease for which a large number of markers are used, it is often found that none of the markers is associated with the disease after the usual FDR correction using Benjamini-Hochberg procedure. Also, the variability of results from one investigator to another in terms of the inferred marker-disease association leads one to believe that there is an intrinsic variability in FDR procedures due to the dependence between the markers. It is believed that because of the high error rates, particularly missed discoveries, caused due to not taking into account the effect of dependence, a lot of effort and public-funds are being wasted.

In this ongoing research project, we propose a new bivariate test for handling the dependence. We combine two highly dependent tests into one test for two parameters, and it seems to be very informative for both parameters. The initial results imply that there is a lot of information in the dependence and getting rid of it can dramatically reduce the information present in the experiment. We also consider generalization of our bivariate tests and its scopes.
2 Model Selection

Optimal Objective Priors for Estimation and Testing

In recent Bayesian studies of linear models, two approaches to optimality have been proposed. One, due to Bayarri, Berger, Forte, and García-Donato (2012), approaches optimality axiomatically by proposing conditions that an optimal prior for model selection should satisfy. They make this plausible by providing a set of definitive properties for such a prior and then producing a prior which indeed has all these properties. The desiderata includes criteria such as invariance under certain groups of transformations (Criterion 7), and consistency in the sense of Information Consistency and the more usual sense of the posterior increasingly putting most of its mass on the true model (Criterion 3, Information Consistency).

A second, more direct, approach towards optimality is based on continuous shrinkage priors for Bayesian inference of linear models. The Horseshoe prior was introduced as a default choice for multiple testing (Carvalho, Polson, and Scott (2010)) and linear model (Polson and Scott (2012)) by Carvalho et al. (2010). Recently, a couple of other innovative priors such as the Generalized Double Pareto and the Three parameter Beta were proposed by Armagan et al. (2013, 2011). These new priors have also been claimed to be optimal by their performance. In particular, Carvalho et al. (2010) provide strong numerical evidence that the inference based on the Horseshoe prior mimics well the answer from the gold standards of Bayesian two-groups model and has good risk properties. In providing a theoretical footing, Datta and Ghosh (2013a) prove a theorem that supports this claim.

Our goal in this paper is to compare these different priors among themselves and also with a frequentist procedure that has become something of a gold-standard for linear models, namely the LASSO and its various modifications. Our numerical results suggest that the dichotomy between the inclusion probabilities for the zero and non-zero components of the \( \beta \) vector is more enhanced for the "robust" prior, leading to a possibly better model selection performance. However, the shrinkage priors lead to better estimation and prediction accuracies across different choices of the design matrix. It also seems that the Lasso is consistently outperformed by the Horseshoe prior and the Robust prior for estimation and prediction accuracies.

3 Bootstrap: An Exploration

Bootstrap was introduced by Efron (1979) to simulate from the given data to estimate bias and variance of a given statistic \( T_n(X_1, \ldots, X_n) \), and also to construct confidence intervals or tests. Efron’s bootstrap began in the late 1970’s initially as “another look at Jackknife”, but, almost immediately in Efron (1982) and several other publications, the subject had an explosive development that is still going on. In this paper, we review a few unusual aspects of Bootstrap and some of the recent theoretical as well as methodological advances in high dimensional statistical inference (vide Datta and Ghosh (2013b)).

We first discuss the handling of non-linearity by Bootstrap through a numerical example. One is that the superiority of bootstrap over delta method for certain statistics, may not be just due to the Edgeworth refinement (vide Singh (1981)) of the normal approximation used in the delta method, the accurate handling of non-linearity by Bootstrap may also play some role in this matter. This may help explain a point made by Efron (1982) and also noted in Ghosh (1992), that the estimate of the standard deviation of the sample correlation coefficient by the delta method is not just inaccurate but also it is ‘badly biased downwards’.

The estimation of the true inverse covariance matrix where we observe only a limited amount of data in a high-dimensional space, remains one of the most important problems. In this paper, we evaluate a few resampling-based approaches of estimating the inverse covariance matrix. We review an approach called the augmented bootstrap proposed by Tyekucheva and Chiaromonte (2008). We also show through numerical examples that augmenting the Bayesian bootstrap improves upon the relative squared error for certain examples.

Several high dimensional examples, such as, Random Forest and its offshoot random survival forest (Ishwaran et al. (2008)) and a Scalable Bootstrap for massive data, introduced by Kleiner et al. (2011), are also discussed. Finally, we discuss some aspects of Bootstrap in the context of hypothesis testing in high-dimension.

4 Covariance Selection using Continuous Shrinkage Priors

Estimation of a sparse precision matrix in the large \( p \), small \( n \) case has recieved a lot of attention from both frequentist and Bayesian statisticians in recent years due to its application in many areas including fMRI, spectroscopy and others. Several frequentist methods such as regularization, penalized likelihood and a few Bayesian methods using the G-Wishart prior and its generalization have gained much popularity in this context. In this project, we consider a Bayesian approach of estimating the precision matrix using shrinkage priors like the Horseshoe prior that has been used for multiple testing and model selection. These priors allow for sparsity by the dint of their unbounded mass near zero and their Cauchy-like tails leave the true non-zero edges unshrunk. Our major contribution is providing a new estimator for the precision matrix through shrinkage of the cholesky factors rather than the original precision matrix. This estimator enjoys both computational feasibility as well as nice theoretical properties.
References


S. Tyekucheva and F. Chiaromonte. Augmenting the bootstrap to analyze high dimensional genomic data. TEST, 17(1):1–18, 2008. ISSN 1133-0686.