A Guided Random Walk Through Some High Dimensional Problems

Junyong Park
University of Maryland, Baltimore County, USA
Jayanta K. Ghosh
Purdue University, USA and Indian Statistical Institute, Kolkata, India

Abstract

We survey some recent work in high dimensional multiple testing estimation and other multivariate inference problems depending on random matrices and graphical problems. Different approaches to these problems are explored featuring classical procedures like the Benjamini-Hochberg multiple tests, empirical Bayes and Bayes tests and estimates and use of shrinkage as a major method in estimation. The choice of topics reflects to some extent our taste and interests.

AMS (2000) *subject classification.* Primary 62H15, 62H39; Secondary 62G05.
Keywords and phrases. High dimensional, multiplicity, sparsity, empirical Bayes, Bayes, shrinkage.

1 Introduction

A striking feature of the current statistical scenario is the preponderance of problems with a high dimensional parameter space, often accompanied by a rather low level of replication, which aggravates the curse of dimension. Most of these problems have led to substantial new research in theory, methodology and computation- so much so that Bickel and Doksum (2007) write in a new edition of their old classic, "As a consequence our second edition, reflecting what we now teach our graduate students, is much enlarged from the first", and of their writing as "an enterprise that at times seemed endless, gratifyingly ended in 1976 but has, with the field, taken on a new life."

We survey some of those problems that have been of some interest to us. The reader may consider our survey as a random walk in a new territory, guided by our interests rather than a desire to represent all aspects of these problems. A representative survey can only be written when this emerging area has stabilized. We will concentrate on two categories of problems on multiple tests and related estimating problems in Section 2, and classification and some other problems in Sections 3 and 4. In Section 5 we comment on some high dimensional multivariate problems.

2 Multiple testing

2.1 Multiple tests and the Benjamini Hochberg. A relatively simple example of many tests being conducted at the same time is a microarray data on gene expression. We begin with a classical parametric formulation, see, for example, Ghosh et al. (2006), Bogdan et al. (2008) for more details.

Formulation I

X_i's, $i = 1, 2, \ldots, m$, are normally distributed, independent, with mean $\mu_i = 0$ under the ith null H_{0i}, and $\mu_i \neq 0$ under the ith alternative H_{Ai}. The variance of X_i is σ^2 same for all i and assumed known. In the context of microarrays, H_{Ai} implies the ith gene is expressed, i.e., it has an effect, for example on a particular kind of tumor which may grow because of this gene. H_{0i} on the other hand asserts the ith gene has no effect on the tumor. One may think of μ_i as the mean effect of the ith gene, while the observation X_i is a summary test statistic based on replicated observations which we do not display.

The assumption of known σ^2 is often made, see, e.g., Abramovich et al. (2006). We believe σ^2 will be sufficiently well estimated from the replicates for most results to go through even if σ^2 is unknown. However, the proofs of some major results may need nontrivial modification.

In typical microarrays m is a few thousands. This is what makes the multiple testing of μ_i's a complex, high dimensional problem. It also permits "borrowing of strength" among the tests, as we see later. The earlier goal of controlling the Family Wise Error rate (FWER) becomes very conservative and hence unacceptable. For a definition of FWER see Ghosh et al. (2006).

A famous classical test is due to Benjamini and Hochberg (1995). Actually as mentioned in the last reference, this test as well as the idea of controlling False Discovery Rate (FDR) goes back to Seeger (1968), Simes (1986), Sorić (1989). But the test and controlling FDR seem to have become popular only after Benjamini and Hochberg (1995) proved a beautiful theorem on control of false discoveries (i.e., Type I errors in some sense). We will refer to the test as BH test or rule.

We first define the test and FDR and then state the theorem of Benjamini and Hochberg (1995).

Let $P_{(1)} < P_{(2)} < \cdots < P_{(m)}$ be the ordered P-values. Fix $0 < \alpha < 1$. Let

$$k = \operatorname{argmax} \left\{ i : P_{(i)} \leq \frac{i}{m}\alpha \right\}. \tag{2.1}$$

Reject the null hypotheses corresponding to $P_{(1)}, P_{(2)}, \ldots, P_{(k)}$. The intuition behind the test seems to be as follows. Each $P_{(i)}$ is being compared with its (approximate) expectation i/m and declared as significantly small if $P_{(i)} < \frac{i}{m}\alpha$, where α is an indicator of how much smaller is $P_{(i)}$. Note that from the definition of $P_{(k)}$,

$$P_{(k)} < \frac{k}{m}\alpha$$

but

$$P_{(k+1)} \geq \frac{k+1}{m}\alpha$$

which together suggest that

$$P_{(k)} \sim \frac{k}{m}\alpha.$$

One then argues if this P-value is significantly small, then so must be all smaller P-values. Benjamini and Hochberg (1995) show that regardless of the values of $(\mu_1, \ldots, \mu_m)$, the above test has

$$FDR \leq \alpha \tag{2.2}$$

where

$FDR = E(\frac{V}{R} I(R > 0))$

$V =$ number of nulls falsely rejected

$R =$ number of nulls rejected

$E =$ expectation under $(\mu_1, \ldots, \mu_m)$ and σ^2.

There have been extensions of this theorem in more recent work by Benjamini and Yekutieli (2001), Sarkar (2002) to mention a few from among many. There has also been an explosion of papers from Stanford on different aspects multiple tests, including alternative formulations, and alternative inference procedures. We list a few of their papers; Abramovich et al. (2006), Donoho and Jin (2004), Efron and Tibshirani (2002), Storey (2002, 2003, 2007), Storey et al. (2004) and Johnstone and Silverman (2004).

The FDR of a multiple test is a useful summary that comes to us naturally. For example the Food and Drug Agency (FDA) of a country can be evaluated by its empirical FDR found from post approval follow up. An increase in empirical FDR would suggest FDA isn't being stringent enough to handle multiplicity.

It is also worth pointing out that in these problems multiplicity caused by the very large values of m, is further aggregated by the rather small number of replicates. Replicates are few because replication is expensive.

Given the above, controlling the FDR is an attractive property of a test as pointed out in Benjamini and Hochberg (1995). Many decision theorists, specially Bayesian decision theorists, have pointed out control of FDR is not easy to justify from a decision theoretic point of view. A decision theorist minimizes the risk, which is an average loss, whereas FDR is an expectation of a ratio under the parameters governing the distribution of the given data. How can they be reconciled? To study this question, Bogdan et al. (2008) introduced a Bayesian Oracle, i.e., a lower bound to a weighted average of the misclassification probability, where the weights for H_{0i} and H_{1i} are $(1-p)$ and p. The lower bound applies to all multiple tests if we formalize the previous description by introducing 0-1 losses, additivity of the losses for the m tests and some more structure on the μ_i's. (This oracle is quite different from other oracles including Sun and Cai, 2007). Through simulations with $m = 200$ and $.05 \leq p \leq .2$ it was shown that the BH does nearly as well as the oracle. This same result has been shown theoretically in Bogdan et al. (2010). To proceed further and describe the oracle as well as other inference procedures, namely, parametric and nonparametric Bayes as well as parametric and nonparametric empirical Bayes, we introduce our second formulation of the multiple test problem as in Ghosh et al. (2006), and examine the usual assumption of sparsity and suitably large "signals" as in Donoho and Jin (2004). The other inference procedures will be studied and compared with the BH test in the next subsection.

Formulation II

A parametric mixture or EB (Empirical Bayes) model introduces the following additional structure to reduce the number of parameters as well as allow "borrowing of strength" in the context of multiple testing. (For more details see Ghosh et al. 2006, Storey, 2007 and Efron, 2008).

Let γ_i's be i.i.d. $B(1,p)$. If $\gamma_i = 1$, H_{Ai} is true. If $\gamma_i = 0$, H_{0i} is true. Moreover, given H_{Ai}, $\mu_i \sim N(0,\tau^2)$ and given H_{0_i}, $\mu_i = 0$. The oracle is obtained as follows. Treat the parameter p as known and use the independence of X_i's and hence that of μ_i's. Also assume the losses are additive. Then the Bayes test is extremely simple, namely,

Reject H_{0i} if $|X_i| > B(\sigma^2,\tau^2,p)$

Accept H_{0i} otherwise.

Here, the threshold $B(\sigma^2,\tau^2,p) = \frac{2(\sigma^2+\tau^2)\sigma^2}{\tau^2}\left[\frac{1}{2}\log\left(\frac{\sigma^2+\tau^2}{\sigma^2}\right) + \log\left(\frac{1-p}{p}\right)\right]$.

A lower bound to the risks of any multiple test is provided by m (risk of the above test, which is same for all i).

The assumptions on m, p and τ^2 will clarify some common assumptions. Of course $m \to \infty$ and, formally, the replication is one per test because $|X_i|$ is the test statistic. Actually, the number of replications per test is a fixed n which remains the same as $m \to \infty$. This is the sort of example that has led to labeling these as "large p and small n problems", where this "p" is the same as m and is the number of unknown original parameters.

The original problem has been considerably simplified in Formulation II, in that we now have only two parameters, p and τ^2. In microarrays, and many other high dimensional examples, the proportion p of signals, i.e., significantly non-zero parameters is supposed to be small. This is the so called "sparsity" assumption. Given the fact that m is large and the proportion of signals is small, the signals can be distinguished from the merely noisy X_i's only if the signal magnitude is large in a manner relative to the proportion p. For example if we take an extremely sparse case, i.e., $mp = 1$. (i.e., on an average only one out of m X_i's may be a signal), the magnitude of the signal must be of the same order of magnitude as $\max_{1\leq i\leq m}|X_i|$ under the global null (i.e., all H_{0i}'s are true). The magnitude of signals is controlled in Formulation II by τ^2. In Bogdan et al. (2008) for very sparse p, τ^2 was taken to be of the order of $\sqrt{2\log m}$, which is the order of $\max_{1\leq i\leq m}|X_i|$ under the global null. Bogdan et al. (2010) suggest that for general p, τ should be of order $\sigma\sqrt{2\log\frac{1}{p}}$. It seems plausible that the above assumptions may have an interesting methodological significance. We conjecture the following.

> Conjecture : If p is small and $\tau = \sqrt{2\log\frac{1}{p}}$, then the optimal inference procedures like PEB (Parametric Empirical Bayes) will have empirical risk
>
> $= \frac{1}{m}\{\#H_{0i}:$ falsely rejected $+ \#H_A:$ falsely rejected $\}$
>
> approximately equal to Bayes risk of a single test.

An earlier paper where the magnitude of signal is related to its sparsity is Donoho and Jin (2004). However, in their case, H_{Ai} postulates $X_i \sim N(\mu_i, 1)$ and it is $|\mu_i|$ that is assumed to be large. See also Bogdan et al. (2010).

Under certain assumptions including Formulation II, $p \to 0$ and τ as explained above, Bogdan et al. (2010) show BH attains the Bayes oracle asymptotically if $\alpha \to 0$. In fact Bogdan et al. (2010) prove the following theorem.

THEOREM 2.1. *Suppose* $(\log m)^r/m \leq p \leq m^{-\beta}$ *and* $\tau = c\sigma\sqrt{2\log p^{-1}}$ *where* $r \leq 1$, $0 < \beta < 1$ *and* $0 < c < \infty$. *Suppose further* $\alpha \to 0$ *such that* $\log\alpha/\log m \to 0$. *Then the BH test with this* α *attains the oracle risk asymptotically as* $m \to \infty$.

(Note α is actually α_m here. We have usually suppressed the dependence on m in the notations. We follow a similar convention about p and τ^2.) Bogdan et al. (2010) conjecture on the basis of their earlier simulations that similar optimality results should hold for PEB and Full Parametric Bayes procedures. Such results also seem plausible for the Nonparametric Empirical Bayes tests.

We end this section with a fully nonparametric formulation due to Efron (2008). In Efron's model, the null is not standard normal and the alternative is nonparametric. Efron (2008) shows several data sets where the null is not standard normal. Efron defines explicit algorithms for estimating them.

It appears there may be a problem of identifiability. An identifiable version of this is due to Martin and Tokdar (2009). Our nonparametric Bayes test in the next section retains a normal null with unknown σ^2 for the density of X_i but allows X_i under the alternative to be nonparametric since μ_i's are assigned nonparametric mixing distribution.

2.2 Full Bayes and empirical Bayes. Under Formulation II, the Bayes multiple test is a threshold test, i.e., one rejects H_{0i} iff

$$X_i^2 > \frac{2(\sigma^2+\tau^2)\sigma^2}{\tau^2}\left[\frac{1}{2}\log\left(\frac{\sigma^2+\tau^2}{\sigma^2}\right)+\log\left(\frac{1-p}{p}\right)\right]$$

same for all i.

The common threshold for all i is denoted by $B(\sigma^2, \tau^2, p)$. A parametric empirical Bayes (PEB) multiple test would estimate τ^2 and p, or only p if τ^2 is known as a function of p, plug in this estimate in $B(\sigma^2, \tau^2(\hat{p}), \hat{p})$ and then reject H_{Ai} iff $X_i^2 > B(\sigma^2, \tau^2(\hat{p}), \hat{p})$. Bogdan et al. (2008) point out that the Type 2 MLE of p doesn't seem satisfactory but alternative estimates including a penalized MLE, do nearly as well as the oracle. The full Bayes approach of Scott and Berger (2006) takes σ^2, τ^2 and p as unknown and puts a prior on all of them. The full Bayes test rejects H_{0i} iff $Pr\{r_i = 1|X_1, \ldots, X_m\} > \frac{1}{2}$. Simulations of Bogdan et al. (2008) show this multiple test also attains the oracle approximately.

Similar findings hold for a Nonparametric Empirical Bayes multiple test based on a nonparametric normal mixture for μ_i under H_{Ai} instead of a simple normal. The test is based on estimating the mixing distribution of μ_i's based on observed X_i's using a recursive algorithm of Newton the convergence of which has been studied in a recent paper of Tokdar et al. (2009).

This method also works very well in comparison with the oracle. Moreover, it seems to be similar to the nonparametric empirical Bayes rule of Storey (2007) and Efron (2008) at least in spirit, but not in details. Bogdan et al. (2008) also use the Full Bayes approach in which μ_i is assumed to be distributed as a mixture of normals with mixing distribution P, which itself is random and has a Dirichlet process distribution. To ensure P can have a point mass at zero with positive probability, the prior mean for the Dirichlet is a distribution that has a positive mass at zero.

Bogdan et al. (2008) were unable to do many simulations for this multiple test. The success of the NPEB approach suggests the Full Bayes approach would also do well.

Scott and Berger (2010) have drawn attention to subtle but important differences between Full Bayes and EB even in the parametric case. Their findings are likely to be valid in the nonparametric case also. There are challenging theoretical issues here of which the resolution would throw more light on possible inadequacies of EB.

We finally return to the BH test. Storey (2003) and Genovese and Wasserman (2004) have pointed out the similarity of the FDR control by the BH test and the control by a suitable threshold test of the so called BFDR (Bayesian FDR) or pFDR(positive FDR) defined as

$$Pr\{\ H_{0i} \text{ is true}|X_i^2 > c\} = \alpha$$

i.e.,

$$\frac{(1-p)P_{H_{0i}}\{X_i^2 > c\}}{(1-p)P_{H_{0i}}\{X_i^2 > c\} + pP_{H_A}\{X_i^2 > c\}} = \alpha$$

where c is the threshold of the BFDR control multiple test.

Genovese and Wasserman (2004), Storey (2003), Efron and Tibshirani (2002) point out that in the numerator of the LHS of the previous equation, one may drop the term $(1-p)$ which is nearly equal to one. Then we need to solve

$$\frac{P_{H_0}\{X_i^2 > c\}}{P_{mixture}\{X_i^2 > c\}} = \alpha.$$

For each given c, we can evaluate the numerator using standard approximations to the normal tail. Moreover we can estimate the denomination by the empirical distribution

$$\frac{\#\{i : X_i^2 > c\}}{m}.$$

Hence we can determine the threshold c for the BFDR control test without having to estimate. It appears heuristically that the multiple test based on this threshold provides a good approximation to the BH test. This has been proved for fixed p by Genovese and Wasserman (2004) and for $p \to 0$ by Bogdan et al. (2010). Thus the BH test seems to be a PEB test but doesn't have to estimate p. This seems to explain why it is optimal, adaptively in p.

2.3 Higher criticism, multiple estimates, estimate of p and beyond . When m is extremely large, a few hundreds of thousands - a situation that Donoho and Jin (2004) take as a plausible model for homeland security - Donoho and Jin use a new principle, which they attribute to Tukey and call Higher Criticism following Tukey. This is a very innovative paper with a complete description of when one can test with both error probabilities tending to zero and when one can also estimate unknown parameters well.

The extremely large value of m is needed for the underlying asymptotic theory, which requires $\log \log m$ is sufficiently large in practice. Given this, Higher Criticism is shown to do better than the BH test in some cases and is never worse than it.

An extension of this paper appears in Jeng (2009). In the same scenario as that of multiple tests, with m a few thousands or even bigger or above, one may wish to estimate the μ_i's well. Johnstone and Silverman (2004) show that an empirical Bayes approach and some thresholding leads to a minimax estimate in the sparse case. The paper contains many new ideas and insights. Even more stunning in the paper by Abramovich et al. (2006) in the same subject. The authors consider three different kinds of sparsity, the first of which is similar to that considered for multiple tests. They also introduce several l_p-losses and ask whether one can get asymptotically minimax estimate of the μ_i's, which are adaptive with respect to both the different definitions of sparsity and the different l_p-losses.

They show that the answer is yes and produce an extremely simple set of estimates, based on a beautiful but long and difficult proof.

The estimates can be described as follows. Fix an $\alpha \leq \frac{1}{2}$ and choose k as in the definition (2.1) of the BH test, i.e.,

$$k = \text{argmax} \left\{ i : P_{(i)} \leq \frac{i}{m} \alpha \right\}.$$

Then estimate $\mu_{(1)}, \dots, \mu_{(k)}$ by $X_{(1)}, \dots, X_{(k)}$ and set the remaining estimates equal to zeros.

Finally, one may also wish to estimate p as well. Meinshausen and Rice (2006) refer to an astronomical example where this is the main goal.

It is clear that the BH test provides an estimate $= k/m$, where k is as defined above. But Meinshausen and Rice (2006) provide estimates that can serve as confidence bounds. They provide several methods for getting such estimates. Cai et al. (2007) explore the problem of providing asymptotically minimax and adaptive estimates.

While many problems have been solved, which originated in microarrays but are meaningful in the much wider context of general multiple tests, many problems remain. A beautiful survey of what is known and a guide to new problems is provided in Efron (2008).

3 Classification in high dimension

Classification is one of the most typical problems in statistics with broad applications to the various areas such as computer science and biological science. One recent characteristic problem in classification is high dimension and low sample size. For example, in the analysis of microarray, dimensionality is often thousands or more, but only tens of samples. In these large p and and small n, there have been many attempts to improve classical classification rules.

Suppose there are n_k many p-dimensional observations $\mathbf{X}_{k1}, \ldots, \mathbf{X}_{kn_k} \sim N_p(\boldsymbol{\mu}_k, \Sigma_k)$ from C_k, $k = 1, 2$ where $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2}, \ldots, \mu_{kp})'$. Let S be the pooled version of sample covariance matrices and S^{-1} be the inverse or Moor-Penrose inverse if S is singular. In such a case, one classical classification rule is Fisher's rule which has the form of

$$\delta_F(\mathbf{X}) = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' S^{-1} \left(\mathbf{X} - \frac{\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2}{2} \right)$$

where $\bar{\mathbf{X}}_k$ is the mean vector of samples from C_k, $k = 1, 2$. Fisher's rule classifies a new observation $\mathbf{X}$ to C_1 if $\delta(X) > 0$ otherwise C_2. Fisher's rule is a plug-in-rule for the Bayes rule under multivariate normal populations, however, it has been widely used even for non-normal populations due to its good performance in many practical problems. Despite its popularity, Fisher's rule has a drawback that it may not work well in high dimension and small sample since S^{-1} is a poor estimate of the inverse of population covariance matrix, so finally leads to the poor prediction of Fisher's rule. To remedy this drawback, there are a couple of methods. One is regularization of sample covariance matrix, for example, Friedman (1989) considered regularized classifiers to avoid ill-posed or poorly-posed inverse problem in S^{-1}. However, the regularization seems to be useful in moderately large dimension or the case when the sample size is nearly close to the dimension p. Thus the regularized classifier in Friedman (1989) may not directly apply to the case of $p >> n$ such as microarray data. Another approach known as the independent rule (IR) ignores all of the off-diagonal terms in S, equivalently all the variables are treated as if they are independent. So S in Fisher's rule is simply replaced by diagonal matrix of S, denoted by $D = diag(S) = diag(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2)$ which results in the IR

$$\hat{\delta}_I(\mathbf{X}) = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' D^{-1} \left(\mathbf{X} - \frac{\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2}{2} \right) = \sum_{j=1}^{p} \hat{\alpha}_j (x_j - \hat{\mu}_j) / \hat{\sigma}_j^2$$

where $\hat{\alpha}_j = \bar{X}_{1j} - \bar{X}_{2j}$ and $\hat{\mu}_j = (\bar{X}_{1j} + \bar{X}_{2j})/2$. The IR is also called the Naive Bayes rule or Idiot's Bayes rule due to its too simplified form (Hand and Yu, 2001), however its performance is not too bad or in many cases, it achieves even better prediction than Fisher's rule especially in high dimension. Although the good performance of the IR was recognized by many practitioners, it had not been well understood theoretically until Bickel and Levina (2004) provided theoretical studies on the performance of the IR. The main result by Bickel and Levina (2004) shows that the IR outperforms Fisher's rule under broad conditions when the number of variables is large compared to the sample size from the point of view of minimaxity. Fisher's rule may obtain 1/2 misclassification probability when $p/n \to \infty$, i.e., almost random guessing while the IR has the misclassification strictly less than 1/2 when $(\log p)/n \to 0$. The IR seems to be successful in high dimension. However, for extremely high dimensional data such as microarray, IR is not satisfactory. From numerical studies in many papers, the IR can achieve as high misclassification error as random guessing even when the two populations can be perfectly classified. The situation may occur when there are too many noisy variables which do not contribute to classification. Fan and Fan (2008) and Greenshtein et al. (2009) showed the poor performance of the IR theoretically. To avoid this accumulation of noisy variables, it is natural to remove such noisy variables and use only a subset of important variables for the improvement of the IR. Tibshirani et al. (2002) proposed nearest shrunken centroid (NSC) incorporating soft threshold to eliminate many of the noisy genes in gene expression data. The number of selected genes are determined through soft shrinkage parameter which is chosen by cross validation in the paper. More recently, Fan and Fan (2008) provides more delicate studies on feature selection in high dimensional classification and proposed Feature Anneal Independence Rule (FAIR). FAIR selects variables by applying hard threshold to the IR, which is for some $b > 0$,

$$\hat{\delta}_{FAIR}(X) = \sum_{j=1}^{p} \hat{\alpha}_j (x_j - \hat{\mu}_j)/\hat{\sigma}_j^2 I_{\{\sqrt{n/(n_1 n_2)}|T_j| > b\}} \tag{3.1}$$

where T_j is the two sample t-statistic and $\hat{\sigma}_j^2$ is pooled sample variance of jth variable. Under some regularity conditions with sparsity, they sorted the features by the absolute values of T_j in the decreasing order and then provided a choice of the number of selected variables as follows; when $\Sigma_1 = \Sigma_2 = I$ is known, for the sorted features by the absolute values of $\hat{\alpha}_j$ in decreasing order, the number of variables, m_0, is estimated by

$$\hat{m}_0 = \text{argmax}_{1 \le m \le p} \frac{\sum_{j=1}^{m} \hat{\alpha}_j^2 + m(n_1 - n_2)/(n_1 n_2)^2}{nm/(n_1 n_2) + \sum_{j=1}^{m} \hat{\alpha}_j^2}.$$

When Σ_1 and Σ_2 are unknown and different from the identity matrix, for the sorted features by the absolute values of T_j in decreasing order, the number of variables, m_1, is estimated by

$$\hat{m}_1 = \text{argmax}_{1 \le m \le p} \frac{1}{\hat{\lambda}_{\max}} \frac{n[\sum_{j=1}^{m} \hat{\alpha}_j^2/\hat{\sigma}_j^2 + m(1/n_2 - 1/n_1)]^2}{nm/(n_1 n_2) + \sum_{j=1}^{m} \hat{\alpha}_j^2/\hat{\sigma}_j^2}$$

where $n = n_1 + n_2$ and $\hat{\lambda}_{\max}$ is the maximum eigenvalue of pooled correlation matrix. The above criterion is purely data dependent procedure for choosing the number of variables (or equivalently b in (3.1)). Since $\hat{\lambda}_1$ diverges with m, $\hat{m}_1$ is usually smaller than $\hat{m}_0$. Greenshtein et al. (2009) reported the simulation studies that $\hat{m}_0$ selects too many features, thus FAIR with $\hat{m}_0$ does not improve the IR significantly.

More recently, Greenshtein et al. (2009) proposed an approach called conditional MLE (CMLE) incorporating Stein's unbiased risk estimator to select a subset of variables and correct the selection bias. For two sample t-test, T_j, they estimate coefficient of linear classifier by CMLE conditioned on $T_j > C$ or $T_j < -C$, say $\delta_C(T_j)$. The tuning parameter C determines the number of selected variables and degree of shrinkage. Greenshtein et al. (2009) provides an criterion for selecting C which maximizes

$$V(C) = \frac{E(\sum_{j=1}^{p} \delta_C(T_j)\alpha_j)}{\|\boldsymbol{\delta}_C\|}$$

where $\|\boldsymbol{\delta}_C\| = \sqrt{\sum_{j=1}^{p} \delta_C(T_j)^2}$, i.e., maximizes the distance between two centroids of the linear classifier for two classes. However, since the numerator in $V(C)$ includes unknown α_j's, it is replaced by Stein's risk unbiased estimator such as $E(\sum_{j=1}^{p} \delta_C(T_j)\alpha_j) = E(\sum_{j=1}^{p} U_j)$ where $U_j = \delta_C(T_j)T_j - \delta_C'(T_j)$ and $\delta_C'$is derivative of δ_C. Thus the following estimator of $V(C)$

$$\mathcal{V}(C) = \frac{\sum_{j=1}^{p} U_j}{\|\boldsymbol{\delta}_C\|}$$

is maximized in terms of C. Greenshtein et al. (2009) compared the performances of the CMLE, FAIR and IR by numerical studies. The numerical study shows that FAIR may not work as well as the CMLE and sometimes FAIR may select too many variables with $\hat{m}_0$ in (2.2).

These rules, namely, NSC, FAIR and CMLE incorporating variable selection are really efficient especially in the case of sparsity when most are noisy variables and only a small number of variables contributes to classification. On the other hand, Greensthein and Park (2009) and Efron (2009) consider Bayesian perspective approach in high dimensional classification, particularly, NPEB estimates for coefficients a_j's in a linear classifier $\sum_{j=1}^{p} a_j X_j + a_0$. Greenshtein and Park (2009) and Efron (2009) introduced the same idea to high dimensional classification independently and almost at the same time. To estimate coefficients in the linear classifier, both considered standardized t-statistic for each variable, say $Z_j = (\bar{X}_{1j} - \bar{X}_{2j})/\sqrt{s_{1j}^2/n_1 + s_{2j}^2/n_2}$, which is approximately normal $\approx N(\Delta_j, 1)$ for some Δ_j. As an estimate of a_j, the posterior mean $E(\Delta_j | Z_j = z) = z + f'(z)/f(z)$ where $f(z) = \int \phi(\Delta - z) dG(\Delta)$ is the marginal density of z_j's and $f'(z)$ is its derivative. To estimate $f(z)$ and $f'(z)$, Efron (2009) used density estimation based on poisson regression while Greenshtein and Park (2009) utilized the kernel density estimation with bandwidth selection $h = 1/\sqrt{\log p}$ which is not the optimal choice in standard density estimation problem. Unlike the previous classifiers, NSC, FAIR and CMLE, the classifier from NPEB estimates for a_j does not select a subset of variables, however $\hat{a}_j$'s are shrunken to 0 if the corresponding variables have little contribution in classification. Greenshtein and Park (2009) concentrated on classification while Efron (2009) showed a variety of applications of NPEB as well as classification. Greenshtein and Park (2009) demonstrated more carefully the case in which NPEB based classifiers work better than threshold based classifiers (FAIR and CMLE). Many classifiers in high dimensions emphasize the case that $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ has the property of sparsity. In Bayesian framework, this sparsity may be expressed by $\alpha_j \sim (1-\epsilon)\delta(0) + \epsilon g(\alpha)$ for small ϵ, delta function δ and some density g. Additionally, Greenshtein and Park (2009) demonstrated other situations such as non-sparse case as well as sparse case. In the paper, three categories are presented depending on the structure of $\boldsymbol{\alpha}$ which are often encountered in high dimensional classification. They first concentrate on the configuration of l_2-norm $\|\boldsymbol{\alpha}\| = o(p)$ as $p \to \infty$ since, otherwise, any reasonable procedure would achieve misclassification probability 0. Three categories are : (i) very few non-zero coordinates of a large/moderate magnitude; (ii) very few coordinates of a large magnitude, mixed with many very small coordinates (i.e., non-sparse case); (iii) Many coordinates of a very small magnitude (i.e., non-sparse vectors).

They provided intensive numerical studies showing that the NPEB classifier outperforms CMLE, FAIR and IR especially for the case of (ii) and (iii). These results coincide with NPEB estimation of normal mean vector in Brown and Greenshtein (2009).

Linear classifiers have played a major role in a variety of classification problems. Although Fisher's rule and the IR are linear classifiers possibly used in non-normal populations, there is another typical example of linear classifier for the case of multivariate binary data. $\mathbf{X}_{ki}$ are multivariate binary data and each of variables is modeled as Bernoulli random variable, say $Bernoulli(p_{ki})$. When all the variables are assumed to be independent, the Bayes rule has the form $\sum_{j=1}^{p} a_j X_j + a_0$ where $a_j = \log\left(\frac{p_{1j}}{p_{2j}} \frac{1-p_{2j}}{1-p_{1j}}\right)$ and $a_0 = \sum_{j=1}^{p} \log(\frac{1-p_{1i}}{1-p_{2i}})$. All parameters are estimated by MLE and they are plugged into the Bayes rule. Park and Ghosh (2007) have studied the performance of the plug-in rule with MLE for multivariate binary data and showed various asymptotic results when the number of variables has the relationship $p = O(n^{\tau})$ for $\tau > 0$. Under the sparsity condition, they showed that identifying a subset of significant variables improves the performance of the plug-in rule. This result is in the same line with most of the previous studies on FAIR and CMLE for normal populations. This also justifies the numerical studies by Wilbur et al. (2002) that variable selection in multivariate binary data improves the IR in DNA fingerprinting data. Park (2009) also investigated the performance of the IR in the aspect of convergence rate of the risk and showed that the IR with a selected variable improves the convergence rate, too.

4 Estimation in high dimension

Estimating simultaneous normal mean vector problem has been considered quite often with many related issues such as admissibility, adaptive nonparametric regression, variable selection, multiple testing and many other areas in statistics. So far, there seem to be three major categories of estimators developed in this area. The first is the James-Stein estimator, shrunk towards zero or mean value which is minimax on the entire space of the unknown mean vector. Efron and Morris (1972, 1973) and Morris (1983) interpreted the James-Stein estimator from the point of view of linear or parametric EB estimator. The second approach is based on some structural assumption on mean vector such as high dimensional and sparse in the sense that many of unknown means are zero or near zero. With the assumption on the sparse mean vectors, threshold methods (vide Section 2 of the paper) such as the universal (Donoho and Johnstone, 1994), soft threshold (Donoho and Johnstone, 1995), FDR (Abramovich et al., 2006 and Benjamini and Hochberg, 1995), the parametric empirical Bayes posterior median (Johnstone and Silverman, 2004) are now known to perform much better than James-Stein type of estimators. The third approach is nonparametric or generalized empirical Bayes estimator which has been investigated deeply by Zhang (1997), Brown and Greenshtein (2009) and Jiang and Zhang (2009). NPEB approach was proposed the earliest among three approaches by Robbins (1951). See Robbins (1956, 1964, 1983) and Zhang (1997, 2003, 2005a). Here, we focus on the last approach since most recent researches have focused on the last approach and it produces excellent performance compared to the others. The oracle estimator under squared loss is the Bayes estimator $E(\mu|Z) = z + f'(z)/f(z)$. This estimator includes unknown $f(z)$ and $f'(z)$. Zhang (1997) estimates f and f' based on a Fourier infinite-order smoothing kernel and Brown and Greenshtein (2009) uses kernel density estimation with normal kernel and provides the optimal choice of bandwidth $1/\sqrt{\log n}$ which is different from those in kernel density estimation. On the other hand, Jiang and Zhang (2009) investigated a way to estimate completely unknown prior G in $f(x) = \int \phi(\mu - x)dG(\mu)$ for a standard normal density $\phi(y)$ only with observations and without smoothing and then plug in the estimator $\hat{G}$ into f and f'. In other words, their estimators are $\hat{f}(z) = \int \phi(\mu - x)d\hat{G}(\mu)$ and $\hat{f}'(z) = \int (\mu - z)\phi(\mu - z)d\hat{G}(\mu)$. Brown and Greenshtein (2009) and Jiang and Zhang (2009) presented simulation studies showing that their estimators outperform all the other estimators especially when the unknown parameters are not of sparse case type. However, Jiang and Zhang (2009) gave some comment at the end of the paper that it is not clear if the kernel method by Brown and Greenshtein (2009) works as well as Jiang and Zhang (2009) in moderate size of samples unless additional theoretical properties are provided.

There is a special variant of the problem related to the above estimation. Under sparsity condition, since most of the variables are noise, sometimes it is interesting to select a few of the variables for a further investigation. Without any further information, the large values of the observations are selected and the total amount of signal for the selected lot, namely, $S_C = \sum_{j=1}^{n} \mu_j I(Z_j > C)$ is of interest to investigate further. Here, for simplicity, we assume $\mu_i \geq 0$. Greenshtein et al. (2008) studied NPEB estimation $\widehat{S}_C = n(\int_C^\infty y f(y)dy - \hat{f}(C))$ where $\hat{f}$ is obtained from kernel density estimation. Under sparsity condition, they showed some form of consistency of their estimator and applied the estimator to local false discovery rate. A similar type of estimation of signals for selected observations goes back to Robbins (1977) who used NPEB estimator for the case of Poisson random variables, $Z_i|\lambda_i \sim Poisson(\lambda_i)$ and $\lambda_i \sim G$ for some prior G. These estimation problems are a special case of the more general problem in Zhang (2005b), namely estimating $\sum_{j=1}^{n} U(Z_j, \theta_j)$ for observed Z_j, $Z_j \sim F_{\theta_j}$ where θ_j is unknown, and a given function U.

5 Some new high dimensional multivariate analysis

The following discussion is based on a special issue of *Ann. Statist.* (2008) devoted to High Dimensional Multivariate Analysis, edited by special editor Peter Bickel. We first try to relate the finding there with some of our earlier observations. In the earlier sections, we have argued that for high dimensional problems with relatively small sample size, some form of sparsity and relatively large signals as measured by signal to noise ratio, seem not only to help but are probably also necessary. A comparison of the graphs of the oracle risk and the risk of common inference procedures are no longer close when the proportion of signals is away from zero (and one), vide Bogdan et al. (2008). It is possible one needs a different oracle in such cases. Similarly, the fact that even the oracle does poorly in the absence of good signal to noise ratio is the major consideration in leaving out $C = \infty$ in Assumption A of Bogdan et al. (2010). That sparsity helps in variable selection was shown earlier in Bogdan et al. (2004) where it is shown an extra penalty for multiplicity of comparisons is needed for BIC to control FWER (Family Wise Error Rate). The modified BIC, which we call mBIC, achieves this. Fan and Fan (2008) (see also our Section 3) discuss the above issue in great generality in the context of classification. They also mention earlier work on the break down of Fisher's classical discriminant function in the high dimensional case as pointed out by Bickel and Levina (2004) and Bai and Saranadasa (1996). Fan and Fan (2008) argue that some form of preliminary screening out of merely noisy variables is needed. The special issue of the *Ann. Statist.* (2008) also features the problems of estimating a high dimensional covariance matrix regularized by banding or thresholding the empirical covariance matrix, vide Anderson and Zeitouni (2008) and Bickel and Levina (2008). Banding means every element of the covariance matrix which is more than a specified distance from the diagonal element in the same row is set equal to zero, i.e., the matrix is being shrunk towards a nearly diagonal matrix. A different kind of regularization is considered by Jeng (2009). In another important paper, Johnstone (2008) derives the Tracy-Wildom limit law for the largest eigenvalue of $(A+B)^{-1}B$ where A and B are central Wishart matrices in p variables with common covariance, and m and n degrees of freedom. This is a well-known test criterion in several problems of classical multivariate analysis based on multivariate normality. This particular criterion was pioneered by S.N. Roy whose birth centenary was celebrated in 2005 by I.S.I. and Calcutta University. The special issue has other interesting papers, including one by Rajaratnam et al. (2008) on Bayesian covariance estimates in graphical Gaussian models, which combines both regularization and shrinkage. Some of the calculations are stunningly beautiful. The other papers in the special issue are also deep and interesting but do not seem as closely related to our problems as the papers discussed above.

Acknowledgement. We thank the editors for inviting us to contribute to this memorial volume for Ashok Maitra. One of us (JKG) remembers him as a friend and outstanding colleague. We are also grateful for being allowed to retain a few references which we think are important in high dimensional inference but which we do not discuss as being somewhat removed from our major concerns in this paper.

References

Abramovich, F., Benjamini, Y., Donoho, D.L. and Johnstone, I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, **34**, 584–653.

Anderson G.W. and Zeitouni, O. (2008). A CLT for regularized sample covariance matrices. *Ann. Statist.*, **36**, 2553–2576.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by example of a two sample problem. *Statist. Sinica*, **6**, 311–329.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate : A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165–1188.

Bickel, P.J. and Doksum, K.A. (2007). *Mathematical Statistics*. Second Edition. Pearson Prentice Hall.

Bickel P.J. and Levina, E. (2004). Some theory for Fisher's Linear Discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. *Bernoulli*, **10** 989–1010.

Bickel, P.J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.*, **36**, 2577–2605.

Bogdan, M., Ghosh, J.K. and Doerge, R.W. (2004). Modifying the Schwartz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics*, **167**, 989–999.

Bogdan, M., Ghosh, J.K. and Tokdar, S. (2008). A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, (N. Balakrishnan, Edsel Pena and Mervyn J. Silvapulle, eds.). Inst. Math. Stat. Collect., **1**. IMS, Beachwood, USA, 211–230.

Bogdan, M., Chakrobarti, A., Frommlet, F. and Ghosh, J.K. (2010). Bayes Oracle and asymptotic optimality for multiple testing procedures under sparsity. Submitted.

Brown, L. and Greenshtein, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.*, **37**, 1685–1704.

Cai, T., Jin, J. and Low, M. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.*, **35**, 2421–2449.

Donoho, D. and Jin, J. (2004). Higher Criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, **32**, 962–994.

Donoho, D. and Johnstone, I.M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, **81**, 425–455.

Donoho, D. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, **90**, 1200–1224.

Du, J., Zhang, H. and Mandrekar, V.S. (2009). Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *Ann. Statist.*, **37**, 3330–3361.

Efron, B. (2008). Microarrays, Empirical Bayes and the Two-Groups Model. *Statist. Sci.*, **23**, 1–22.

Efron, B. (2009). Empirical Bayes Estimates for Large-Scale Prediction Problems. *J. Amer. Statist. Assoc.*, **104**, 1015–1028.

Efron, B. and Morris, C. (1972). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika*, **59**, 335–347.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors-an empirical Bayes approach. *J. Amer. Statist. Assoc.*, **68**, 117–130.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiology*, **23**, 70–86.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.*, **36**, 2605–2637.

Friedman, J.H. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.*, **84**, 165–175.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.*, **32**, 1035–1061.

Ghosh, J.K., Delampady, M. and Samanta, T. (2006). *An Introduction to Bayesian Analysis : Theory and Methods*. Springer Texts in Statistics. Springer, New York.

Greenshtein, E. and Park, J. (2009). Application of nonparametric empirical Bayes estimation to high dimensional classification. *J. Mach. Learn. Res.*, **10**, 1687–1704.

GREENSHTEIN, E., PARK, J. and LEBANON, G. (2009). Regularization through variable selection and conditional MLE with application to classification in high dimensions. *J. Statist. Plann. Inference*, **139**, 385–395.

GREENSHTEIN, E., PARK, J. and RITOV, Y. (2008). Estimating the mean of high valued observation in high dimensions. *J. Statist. Theory Pract.*, **2**, 407–418.

HAND, D.J. and YU, K. (2001). Idiot's Bayes — not so stupid after all? *Internat. Statist. Rev.*, **69**, 385–395.

JENG, X. (2009). Covariance adaptation and regularization in large scale hypothesis testing and high dimensional regression. Ph.D. Thesis, Purdue University.

JIANG, W. and ZHANG, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.*, **37**, 1647–1684.

JOHNSTONE, I.M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy Widom limits and rates of convergence, *Ann. Statist.*, **36**, 2638–2716.

JOHNSTONE, I.M. and SILVERMAN, B. (2004). Needles and hay in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, **32**, 1594–1649.

MARTIN, R. and TOKDAR, S.T. (2009). Kullbak-Leibler projections, recursive estimation of a mixing distribution. Unpublished manuscript.

MEINSHAUSEN, M. and RICE, J. (2006). Estimating the proportion of false null hypotheses among a large number of independent tested hypotheses. *Ann. Statist.*, **34**, 373–393.

MORRIS, C.N. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion). *J. Amer. Statist. Assoc.*, **78**, 47–65.

PARK, J. (2009). Independent rule in classification of multivariate binary data. *J. Multivariate Anal.*, **100**, 2270–2286.

PARK, J. and GHOSH, J.K. (2007). Persistence of the plug-in rule in classification of high dimensional multivariate binary data. *J. Statist. Plann. Inference*, **147**, 3687–3705.

RAJARATNAM, B., MASSAM, H. and CARVALHO, C.M. (2008). Flexible covariance estimation in graphical Gaussian models. *Ann. Statist.*, **36**, 2818–2849.

ROBBINS, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of Second Berkeley Symposium on Mathematical Statistics and Probability*, (J. Neyman, ed.). Univ. California Press, Berkeley, 131–148.

ROBBINS, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of Third Berkeley Symposium on Mathematical Statistics and Probability*, **1**, (J. Neyman, ed.). Univ. California Press, Berkeley, 157–163.

ROBBINS, H. (1964) The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.*, **35**, 1–20.

ROBBINS, H. (1977) Prediction and estimation for the compound poisson distribution. *Proc. Natl. Acad. Sci.*, **74**, 2670–2671.

ROBBINS, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Statist.*, **11**, 713–723.

SARKAR, S.K. (2002). Some results on false discovery rate in stepwise multiple testing procedure. *Ann. Statist.*, **34**, 239–257.

SCOTT, J.G. and BERGER, J.O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plann. Inference*, **136**, 2144–2162.

Scott, J.G. and Berger, J.O. (2010). Bayes and empirical Bayes multiplicity adjustment in the variable selection problem. *Ann. Statist.*, to appear.

Seeger, P. (1968). A note on a method for the analysis of significances en mass. *Technometrics*, **10**, 586–593.

Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

Sorić, B. (1989). Statistical "discoveries" and effect size estimation. *J. Amer. Statist. Assoc.*, **84**, 608–610.

Storey, J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **64**, 479–498.

Storey, J.D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.*, **31**, 2013–2035.

Storey, J.D. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **69**, 347–368.

Storey, J.D., Taylor, J.E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **66**, 187–205.

Sun, W. and Cai, T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.*, **102**, 901–912.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci.*, **99**, 6567–6572.

Tokdar, S.T., Martin, R. and Ghosh, J.K. (2009). Consistency of a recursive estimate of mixing distributions. *Ann. Statist.*, **37**, 2502–2522.

Wilbur, J.D., Ghosh, J.K., Nakatsu, C.H., Brouder, S.M. and Doerge, R.W. (2002). Variable selection in high-dimensional multivariate binary data with application to the analysis of microbial community DNA fingerprints. *Biometrics*, **58**, 378–386.

Zhang, C.-H. (1997). Empirical Bayes and compound estimation of normal means. *Statist. Sinica*, **7**, 181–194.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes method. *Ann. Statist.*, **31**, 379–390.

Zhang, C.-H. (2005a). General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist.*, **33**, 54–100.

Zhang, C.-H. (2005b). Estimation of sums of random variables: Examples and information bounds. *Ann. Statist.*, **33**, 2022–2041.

Junyong Park
Department of Mathematics
and Statistics
University of Maryland
Baltimore County, USA
E-mail: junpark@umbc.edu

Jayanta K. Ghosh
Department of Statistics
Purdue University, USA
E-mail: jayantag1@gmail.com