Phase Transition for Multiple Testing: A Simulation based comprehensive study of Bayesian and Frequentist Methods

Karoliina Toivonen

19.10.2016

Matemaattis-luonnontieteellinen Matematiikan ja tilastotieteen laitos

Karoliina Toivonen

Phase Transition for Multiple Testing

Soveltava matematiikka

Pro gradu -tutkielma Lokakuu 2016 70 s.

Multiple testing is a statistical inference problem, applied widely in the fields of genomic studies, QTL mapping and national security, where large number of hypotheses is being tested simultaneously. However, it is not always straightforward whether multiple testing can successfully be carried out for a specific dataset.

To measure whether multiple testing works as desired, the error rate, defined as $P(\text{Type I error}) + P(\text{Type II error})$, for investigating performance of different frequentist and Bayesian testing methods is considered. In a grid of all possible combinations of p (proportion of signal in the data) and τ^2/σ^2 (variance), a simulation study is conducted, testing a set of hypotheses with Benjamini-Hochberg procedure and its modified versions, as well as with Parametric Empirical and Full Bayes methods.

As a result, a sharp phase transition phenomenon for the error rate of each of the inference

Multiple testing, Phase transition, Variable selection

Kumpulan tiedekirjasto

Acknowledgements

I would like to express my sincere gratitude to my brilliant thesis advisor Dr. Ritabrata Dutta for his continuous support, patience, motivation and understanding. Many thanks to you Rito, for everything!

I would also like to acknowledge Prof. Samuel Kaski for his wise words and counsel and Prof. Jukka Corander for the help with this thesis and my studies.

Finally, I want to thank my husband Oskar for tolerating me when I'm trying to handle my math. I don't know how he does it.

Contents

Chapter 1

Introduction

"Data! Data! Data!" he cried impatiently. "I can't make bricks without clay."

Already in the 19th century the value of data was acknowledged, at least by Sherlock Holmes. Although, for benefiting from data today you don't really need to be solving crimes with your inhumane deduction skills. With the use of new machine learning techniques and statistical inference schemes data has become the new source of wealth and happiness, almost like the gold of modern times. Some of the most successful businesses at this very moment own data and nearly nothing else; their whole logic is based on data and its smart utilization.

However, working with large amounts of data easily becomes expensive, at least from the computational point of view. For example, in statistics, the asymptotic theory guarantees the consistency of inference schemes when the sample size m (i.e. the size of the dataset) grows to infinity. Infinity? Yes, infinity, which could mean any large numbers, emphasis on the word large. Now, imagine your boss giving you a task to make something inferential out of a dataset with $m \to \infty$ by tomorrow morning with your everyday computational resources. This might potentially be the perfect moment to have a nervous breakdown, but what if instead there was a way to know how much data is actually needed for inference procedures to work properly? What if there were scalable inference schemes for big data scenario? To address these sorts of questions, an explosion of research has recently emerged, trying to find optimal circumstances needed for inference procedures to work properly with large scale data (and to save you from the nuisance of karoshi).

In this thesis, the theme is to look into scalable statistical inference from the perspective of multiple testing: a procedure of testing a set, possibly a large one, of hypotheses simultaneously. The problem with multiple testing, *multiplicity*, is that the probabilities of errors

in inference increase as the sample size m and the proportion of signals among samples (p) increase, and to find a way to prevent this from happening (as well as possible) is indeed one of the goals of the following chapters. Finding sharp changes in the behaviour of error rates of different multiple testing methods could help identifying the circumstances (combinations of m, p and other parameters) needed for reducing multiplicity, even when working with big data.

This sort of phenomenon of sharp change is known as phase transition and has been explored in such fields of mathematics and theoretical physics as random graphs, percolation theory and statistical mechanics. Recent works in the field of compressed sensing like Amelunxen, Lotz, McCoy and Tropp (2014) [1], Donoho and Tanner (2009) [8] and in variable selection like Donoho and Stodden (2006) [6] have shown similar sharp phase transition behaviour. Motivated by these existing works, it is only natural to assume that phase transition could indeed be occurring in the context of multiple testing as well. Actually, in Jin, Zheng and Wang (2015) [17], sharp changes in the success rates of frequentist multiple comparison methods were reported. Still, all the empirical Bayes and full Bayes procedures remain free of investigation from the phase transition point of view but hopefully not anymore after this thesis.

Finding phase transition for multiple testing would not only help finding the optimal parameters needed for the multiple testing rules to be feasible. Because of the well known fact of multiple testing being a form of model selection, the findings could also work as preliminary results for Bayesian variable selection, for which no phase transition behaviour has yet been reported. Also, the boundaries for feasible multiple testing procedures could help understanding and giving new solutions to the peculiar discrepancy between empirical and full Bayes variable selection approaches that Scott and Berger reported in their paper [19].

In this thesis, the initial approach to the investigation is done via a simulation study. With a generated sparse data set from a distribution $N(\ ; \ ^2)$ where either represents a signal ($\neq 0$) with a distribution $N(0; \ ^2)$ or not a signal ($= 0$), the simulation includes inference in the range of 100, 200 and 400 (m) hypotheses of the form $H_0 : \ = 0$ vs. $H_A : \ \neq 0$. Testing all these hypotheses is done with six different methods, starting with Benjamini-Hochberg procedure (BH), continuing with Full Bayes approach, also known as Scott and Berger procedure (SB), then with Parametric Empirical Bayes methods from decision theory (PEB1 and PEB2) and finally modifying BH with the help of PEB1 and PEB2 (resulting in BH1 and BH2). The behaviour of false discovery rate (FDR), Bayes false discovery rate (BFDR), power and misclassification error of each of these testing procedures are looked into but the main focus is put on the behaviour of the overall rate

$OE = P(\text{Type I error}) + P(\text{Type II error}) = BFDR + (1 - Power)$. Traces of phase transition in the behaviour of this error rate are hoped to be found in the domain of the measure of sparsity and the measure of undersampling.

Some restrictions in the computational resources complicated the study to some extent but the results indeed started to yield promising boundary like patterns, indicating the existence of phase transition in multiple testing. Comparing the result to those of the Bayes oracle it is observable that with a sufficiently large value of m, the boundaries might very likely be in complete agreement with each other. Thus the boundary of the oracle can be interpreted as the estimate of the universal phase boundary.

The structure of the thesis is the following:
In Chapter 1, phase transition as a phenomenon is generally explained and previous results in different fields of science are presented. The meaning of a phase diagram and a phase boundary are explained. In Chapter 2, multiple testing is thoroughly explained. Notations such as the model and the error rates are introduced. In Section 3.3 the steps of the testing procedures BH, FB, PEB1, PEB2, BH1 and BH2 are described. Then, Chapter 4 merges phase transition and multiple testing presenting the simulation study and the intuitive expectations of the results. Finally the actual results of the OE rate for each of the methods are revealed with figures and explained in more detail. Some estimations of the phase boundaries are given. The thesis ends with the conclusions in Chapter 5.

Chapter 2

Phase Transition

Phase transition is a term most commonly used to describe sharp changes between solid, liquid and gaseous states of matter such as water. *A phase diagram* in Figure 2.1 shows the conditions, combinations of pressure and temperature, at which these different phases of water occur and *the phase boundaries* distinguishing them from each other.

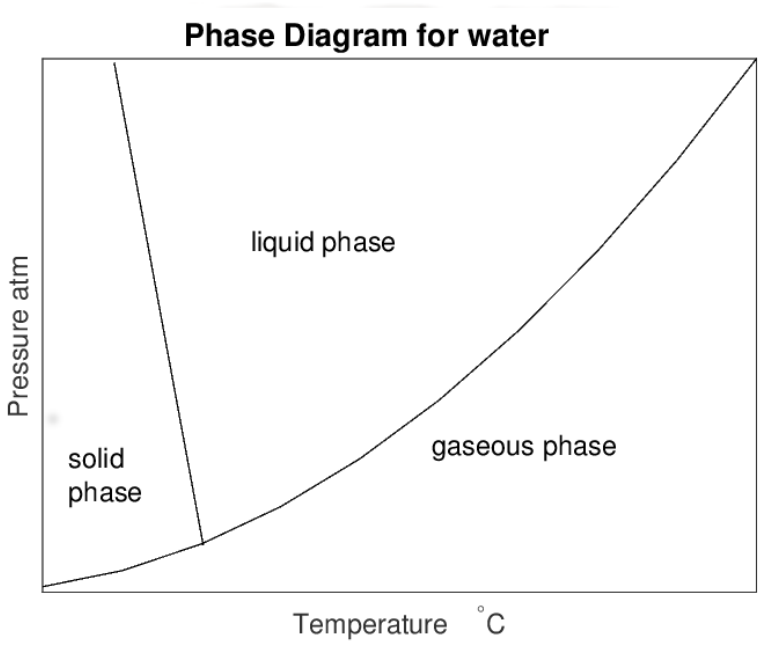


Figure 2.1

There are a couple of well known areas in mathematics and theoretical physics known to exhibit phase transition: random graphs, percolation theory and statistical mechanics. In Section 2.1, these phenomena are shortly introduced to give the reader an idea on the usefulness of phase transitions.

Furthermore, a few statistical processes in machine learning are also known to show phase transition behaviour. These include compressed sensing and variable selection. In statistics, phase transitions occur as abrupt changes in behaviour of statistical or computational procedures when their performance rates are studied in a grid of variables characteristic to the problem in question. These variables can be anything from the properties of the data (e.g. sample size) or the properties of the parameters (e.g. their latency or prior distributions) to the properties of the task (e.g. computing time, storage cost). Finding phase boundaries, thresholds for these changes, would be of assist when identifying the conditions allowing the algorithm to perform satisfyingly. Visual information of these sharp limitations of the method can be given by a phase diagram.

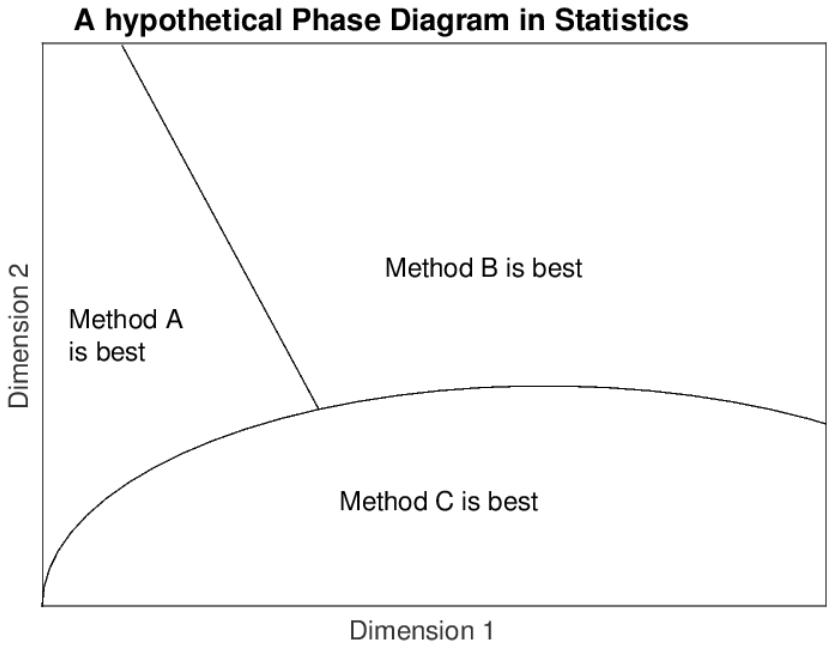


Figure 2.2

In Figure 2.2, Dimension 1 and Dimension 2 represent two variables such as the sample size and the number of parameters to be optimized, whereas the curves dividing the plane in three parts are the phase boundaries showing clearly when (from the perspective of

Dimensions 1 and 2) one of the three hypothetical methods is performing better than the other two.

In Mathematics & Theoretical Physics

Random graphs

For many monotone-increasing properties, such as graph connectedness, of random graphs, it is very unlikely for graphs of a size slightly less than a certain threshold to have the property, whereas graphs with a few more graph edges are almost certain to have it. When such a threshold exists, it is said that a phase transition occurs at that threshold.

Example 2.1.1. The property in question being graph connectedness, a random graph $G(n;p)$ experiences a sharp phase transition at the edge probability $p = \frac{\ln n}{n}$. For any > 0, if $p < \frac{(1-)\ln n}{n}$, then the graph $G(n;p)$ will with high probability contain isolated vertices and thus be disconnected. Instead, if $p > \frac{(1+)\ln n}{n}$, then with high probability the graph will be connected.

Percolation theory

Site percolation theory studies clusters that are defined as sets of occupied sites that can be traversed by jumping from neighbour to occupied neighbour. At random, a site from a large lattice of empty sites could be occupied with probability p or unoccupied with probability $1-p$. Two sites may also be attached with a bond with probability b or unattached with probability $1-b$. Study of such clusters is called bond percolation theory. In this case a cluster is defined as a set of points that can be traversed only by travelling across occupied bonds. Third case of percolation theory, site-bond percolation, has both sites and bonds, filled at random, with bonds only permitted to be between occupied sites.

It is clear that the larger the probabilities p and b are, the larger the average cluster is. An infinite cluster forms at a certain threshold pc. Above pc, the infinite cluster gathers an increasingly greater share of the lattice sites whereas the remaining finite clusters shrink. If $p = 1$, the infinite cluster contains all of the sites. Thus it is said that phase transition occurs at the threshold pc.

Statistical mechanics

Mathematically, phase transition in the field of statistical mechanics is a point in parameter space where free energy (log marginal likelihood) becomes a non-analytic function of

one of its parameters in the thermodynamic limit.

Types of phase transition:

- 1st Order: Free energy is continuous but a first derivative is discontinuous.
- 2nd Order: These are characterized by a divergence in one of the higher order derivatives of the free energy.

In Statistics

Compressed sensing

Let $x \in R^n$ be the signal of interest with very few non-zero coefficients, i.e. x is *sparse*. Further, let A be a $m \times n$ measurement matrix, $m < n$. Now *the compressed sensing problem* is defined as follows:
Recover x from

$$y = Ax + \quad \text{(with error} \quad \in \mathbb{R}^n\text{)}: \tag{2.1}$$

Because m is smaller than n, this inverse problem can not be solved without taking advantage of the *sparsity* of x.

The method of l_1 minimization is a well-established approach to the compressed sensing problem. This procedure searches for the sparse x by solving the convex problem

$$\min ||x||_1 \text{ subject to } y = Ax: \tag{2.2}$$

This technique is sensible because the l_1 norm of a vector can serve as a proxy for the sparsity. It is said that the convex problem 2.2 succeeds at solving the compressed sensing problem when it has a unique solution $\hat{x}$ which equals the true unknown x; otherwise it fails.

Phase transition in Compressed Sensing

The phase transition for compressed sensing is quantified with (;) where $= s/n$ is the proportion of non-zero coefficients with respect to the sample size, i.e. the measure of sparsity, and $= n/m$ is the measure of undersampling. Plotting on the x-axis and on the y-axis means transforming the plane into a phase diagram, and the performance of the recovery procedure can be assessed by evaluating a measure of recovery quality at

each point of the plane.

Vattikuti, Lee, Chang, Hsu and Chow (2014) [25] summarize two lines of research on which their results on compressed sensing rely as follows:

Proposition 1 Suppose that the entries of the measurement matrix A are i.i.d. and $= 0$ (noiseless case). Now the -plane is partitioned by a curve $= {}_{l_1}(\ )$ ($l_1 l_2$ equivalence curve) into two phases. Below this curve the solution to the convex problem (2.2) leads to $\hat{x} = x$ with probability converging to 1 as $n; m; s \to \infty$ in such way that and remain constant. Above the curve, 2.2 leads to $\hat{x} \neq x$ with similarly high probability.

Figure 2.3: Phase boundary from [25]

Proposition 2 Suppose that given the measure matrix A, $E(A'A) = I$ and that there is a smallest number such that for each row a of A, $\max_{1 \leq t \leq p} |a_t|^2 \leq$. If $n > C\ s \log m$ for a constant C then the solution of the problem

$$\min_{\hat{x}} \left(||y - A\hat{x}||_2^2 + \ ||\hat{x}||_1 \right)$$

with a suitable choice of obeys

$$||\hat{x} - x||_1^2 \leq \frac{^2}{n} s \text{ poly} \log(m)$$

where ^{2} is the variance of the residuals in .

In Vattikuti et al. [25] it is mentioned that it has also been shown by Donoho and Stodden [6] and Donoho, Maleki and Montanari (2011) [10] for a curve similar to that in Proposition 1 to have phase transition behaviour in the noisy case $\neq 0$. However, with large residual noise, the transition is to progressive improvement with n rather than to abrupt recovery of the model.

Variable selection

Suppose that Y, a variable of interest, and $\{X_1; :::; X_p\}$, a set of potential variables or predictors, are vectors of n observations. Variable selection is a procedure for selecting the best subset of relevant features in $\{X_1; :::; X_p\}$. Such a situation is particularly interesting when p is large and the set $\{X_1; :::; X_p\}$ is thought to contain many redundant or irrelevant variables.

The problem of variable selection is best known in the context of linear regression. Let index the subsets and let q be the size of the th subset. Then the problem is to select and fit a model of the form

$$Y = X \quad + \; ; \tag{2.3}$$

where X is a $n \times q$ matrix with columns corresponding to the th subset. is a $q \times 1$ vector of regression coefficients and $\sim N_n(0; \ ^2 I)$. More generally, variable selection is a special case of model selection where each model under consideration corresponds to a distinct subset of $\{X_1; :::; X_p\}$.

Frequentist methods

Subset selection in linear regression problems is a discrete process in which variables are either retained or discarded. It often suffers from high variance and thus doesn't reduce the prediction error of the full model. **Shrinkage methods** are more continuous and they don't exhibit as much high variability. The **LASSO** method tries to minimize the

least square error of the regression with an upper bound on the l_1 norm of the parameter vector. The estimate is defined by

$$\hat{}^{lasso} = \min ||y - X \quad ||_2^2, \text{ subject to } || \quad ||_1 \leq t: \tag{2.4}$$

When there are many correlated variables in a linear regression model, their coefficients can become poorly determined and suffer from high variability. A large positive coefficient on one variable can be cancelled by a similarly large negative coefficient on its correlated counterpart. By imposing a chosen size constraint t on the coefficients as in 2.4, this problem is avoided.

Forward stepwise selection starts with no variables in the model and stepwisely adds the variables with the lowest p-value less than a cut-off threshold $_{crit}$. After each addition the model is refitted. **Least angle regression (LARS)** provides a forward stepwise approximation to LASSO. It uses a similar strategy to that of forward stepwise procedure, but only enters as much of a predictor as it deserves.

If there are p potential predictors then there are 2^p possible models. In **criterion based selection**, first all of these models are fitted and then the best one is chosen according to some criterion, for example the **Akaike Information Criterion (AIC)**

$$\text{AIC} = \log L - p; \tag{2.5}$$

where L is the maximum value of the likelihood function for the model. The preferred model is the one with the minimum AIC value.

Bayesian methods

Given a variable selection problem in which a choice needs to be made between two models, say M_1 and M_2, parameterized by model parameter vectors $_1$ and $_2$, is assessed by the **Bayes factor**:

$$B(x) = \frac{(M_1|x)}{(M_2|x)} \times \frac{p(M_2)}{p(M_1)} = \frac{(M_1|x) = \ (M_2|x)}{p(M_1) = p(M_2)}; \tag{2.6}$$

where the posterior odds in favor of Model 1 versus Model 2 is

$$\frac{(M_1|x)}{(M_2|x)} = \left(\int_1 \frac{p(M_1) f_1(x|\Theta_1) p_1(\ _1) d\ _1}{p(x)} \right) = \left(\int_2 \frac{p(M_2) f_2(x|\ _2) p_2(\ _2) d\ _2}{p(x)} \right):$$

It is clear that when $B(x) >> 1$, model M_1 is more supported by the observations concerned than M_2.

If the likelihood corresponding to the maximum likelihood estimate of the parameter for each model is used instead of the Bayes factor integral, then the test becomes a classical **likelihood-ratio test**.

Next, in comparison to AIC, **Bayes Information Criterion (BIC)**

$$\text{BIC} = \log L - \frac{1}{2} p \log n \tag{2.7}$$

tends to prefer smaller models penalizing larger models more heavily. As with AIC, the preferred model is the one with the minimum BIC value. It is known (Bogdan, Ghosh and Doerge (2004) [4]) that under sparsity the phenomenon of overestimating arises using BIC. As a remedy, Bogdan et al. [4] introduced a modification of BIC as

$$\text{mBIC} = \log L - \frac{1}{2} p \log n - p \log(w); \tag{2.8}$$

where w can be interpreted as a probability of a particular covariate being relevant.

Phase transition in Variable Selection

For an appropriate choice of t in 2.4, the LASSO estimate describes the same problem as the equation 2.2 in compressed sensing (Donoho and Stodden [6]). This suggests that it might be possible to benefit from the equivalence of these equations in the model selection setting and observe a threshold in behaviour such that for sufficiently sparse models traditional variable selection works while for more complex models the algorithm's ability to recover the model breaks down. Following this intuition Donoho and Stodden [6] documented the existence of a well-defined breakdown point for linear regression algorithms in the $p > n$ case. They observed that when the true model is sufficiently sparse (less than the threshold point), forward stepwise selection, LASSO and LARS can all recover a good model, and when p is close to n, variable selection methods do not work as well.

No phase transition behaviour has been reported for the Bayesian methods in variable selection.

Chapter 3

Multiple Testing

Errors in statistical inference are likely to occur when one considers a set of inferences simultaneously. For hypothesis testing (see for more in Appendix A), this usually means rejecting the true null hypotheses or failing to reject false null hypotheses and is called *the multiple testing problem*. Several statistical procedures, multiple testing methods, have been developed for controlling this effect, allowing significance levels for single and multiple testing to be directly compared.

Traditionally multiple testing methods have been focusing on the analysis of variance with only a modest number of hypotheses considered. However, in recent years multiple testing has also gained interest for its applicability in understanding large data sets. Many techniques have been developed for large scale testing and are widely applied for example in bioinformatics and national security.

In the simulation study (Chapter 4) of this thesis a group of testing methods is used, both frequentist and Bayesian, and also a set of error rates is defined to describe the problems in multiple testing.

The model

Consider a data vector X with the length m consisting of test statistics $X_i \sim N(\ _i;\ ^2);\ i = 1;:::;m;$ with unknown variance $\ ^2$ and mean $\ _i$ from a distribution $N(0;\ ^2)$, where variance $\ ^2$ is unknown as well. When $\ _i \neq 0$, it's called *a signal* and X_i is representing this signal, its distribution reformulated as $N(0;\ ^2 + \ ^2)$ (the non-null distribution).

For simplifying the forthcoming denotions, a random indicator $_i;\ i = 1;:::;m;$ is defined as

$$_i = \begin{cases} 0; & \text{if } \ _i = 0 \\ 1; & \text{if } \ _i \neq 0: \end{cases}$$

Now it can be assumed that $(X_i;\ _i); 1 \leq i \leq m$, are i.i.d. random vectors with X_i from a mixture distribution,

$$X_i \sim pN(0;\ ^2 + \ ^2) + (1-p)N(0;\ ^2); \tag{3.1}$$

where p is the probability of $_i$ being a signal, i.e. $p = P(\ _i = 1)$. In this thesis, denotations $p_0 = 1 - p$ and $p_A = p$ are used.

For each $i = 1;:::;m$ the aim is to test whether X_i has the null distribution or not. Thus the hypotheses are defined as

$H_{0i} :\ _i = 0$ vs. $H_{Ai} :\ _i = 1:$

Each of these tests has a significance level of , i.e. for any one test, the chance of rejecting the true null hypothesis is . Now, the probability of at least one true null hypothesis being rejected among all the m tests is much higher, making the traditional procedures designed for testing only a single hypothesis useless. Multiple comparison methods aim to control this effect and the two types of errors defined to describe it.

Error rates

There are two types of *errors* in multiple testing:

- Type 1 error: Rejecting H_{0i} when H_{0i} is true.
- Type 2 error: Failing to reject H_{0i} when H_{Ai} is true.

	Fail to reject H_{0i}	Reject H_{0i}
H_{0i} true	Correct	Type 1 error
H_{Ai} true	Type 2 error	Correct

Table 3.1: Errors in multiple testing

The frequencies of these errors can be represented in different manners using variables

describing counts of possible outcomes of multiple testing procedures from Table 3.2, and variables from matrix of losses for making the wrong decision (Table 3.3).

	Fail to reject H_0	Reject H_0	Total
H_0 true	U	V	m_0
H_A true	T	S	m_1
Total	W	R	m

Table 3.2: Counts of possible outcomes of m hypothesis tests (Bogdan, Ghosh and Tokdar (2008) [5]).

	Accept H_{0i}	Reject H_{0i}
H_{0i} true	0	$_0$
H_{Ai} true	$_1$	0

Table 3.3: Matrix of losses from Bogdan et al. [5].

Benjamini and Hochberg (1995) [2] formally defined **the false discovery rate** as the expected value of the ratio of rejected true null hypotheses among all rejected hypotheses, $FDR = E(\frac{V}{R})$ where $\frac{V}{R} = 0$ if $R = 0$. Furthermore, **the positive false discovery rate** for a case where there is at least one rejected hypothesis was defined as $pFDR = E(\frac{V}{R}|R > 0) = \frac{FDR}{P(R>0)}$ in Storey (2003) [24], and **the Bayesian false discovery rate** as $BFDR = P(H_0 \text{ is true}|H_0 \text{ is rejected})$ in Efron and Tibshirani (2002) [12]. In Storey [24] it is shown that in case such as ours where a two-component mixture model is used to create the individual test statistics, $pFDR = BFDR$.

Then, considering multiple testing from the decision theory approach, denotations (as in Bogdan et al. [5])
$t_1 = P(\text{Type 1 error in a single test})$ and $t_2 = P(\text{Type 2 error in a single test})$ are used. **The Bayes risk** related to the matrix of losses 3.3 is defined as $BR_{0;\ A} = {}_0p_0t_1 + {}_Ap_At_2$. The test that minimizes this risk is called **the Bayes oracle** and it rejects H_{0i} if

$$\frac{f_A(X_i)}{f_0(X_i)} > \frac{(1-p_A)\ {}_0}{p_A\ {}_1};$$

or equivalently if

$$p_{Ai} = P(H_{Ai}|X_i) > \frac{^0}{_0 + \ _A}:$$

Corresponding to $0-1$ loss, $BR_{1;1}$ measures the accuracy of the testing procedure and can also be defined as **the misclassification probability** $MP = \frac{E(V+T)}{m}$. In our parametric settings (3.1) the Bayes oracle minimizing $BR_{1;1} = MP$ rejects the null hypothesis if

$$\sum_{j=1}^{n} x_{ij}^2 > \frac{2(\ ^2 + \ ^2)\ ^2}{^2}\left[\log\frac{p_0}{p_A} + \frac{n}{2}\log\frac{^2 + \ ^2}{^2}\right]: \tag{3.2}$$

Other tests are usually compared to this oracle.

The power of a multiple testing procedure is represented as $Power = E\left(\frac{S}{R}\right)$.

In classical sense, as is stated in Frommlet, Chakrabarti, Murawska and Bogdan (2011) [15], a multiple testing procedure is considered to be optimal if it maximizes the amount of true discoveries while controlling one of the Type 1 error functions at a certain fixed level, or when it minimizes the Bayes risk.

However, when working with large data sets, it is a common opinion that a good multiple testing rule should be able to minimize both of the error types. The following estimation of an error rate OE (**the overall error**) simply computes the sum of the risks of individual tests, thus combining the probabilities of both Type 1 and Type 2 errors in one function:

$$\begin{aligned} OE &= P(\text{Type I error}) + P(\text{Type II error}) \\ &\equiv P(\text{Reject } H_0|H_0\text{True}) + P(\text{Accept } H_0|H_A\text{True}) \\ &\equiv E(V=(V+S))\frac{1}{P(V+S>0)} + E(T=(T+S))\frac{1}{P(T+S>0)} \\ &\equiv BFDR + (1 - Power): \end{aligned}$$

In the simulation study of this thesis the focus is mainly on investigating this rate, especially trying to find out if it can be controlled with parametic empirical Bayes and Full Bayes procedures since no such results have been obtained yet in earlier researches. For frequentist testing methods, succesful comparable results have been achieved in Jin et al. [17]

Inference

A group of multiple testing methods were chosen for the simulation, including the main frequentist procedure BH with its modified versions BH1 and BH2, the main Bayesian procedure FB as well as two parametric empirical Bayes classifiers PEB1 and PEB2.

Benjamini-Hochberg procedure (BH)

Benjamini and Hochberg [2] proved that the stepwise multiple testing procedure of Seeger (1968) [20] and Simes (1986) [21] controls FDR at a desired level when the test statistics are independent. This test is currently known as Benjamini-Hochberg procedure (BH) and even though many methods have been developed after it, it still remains the flagship of frequentist selection.

Algorithm 1: BH

1: Order the unadjusted p-values: $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$.

2: Find the test with the highest rank j, for which the p-value is

$$p_{(j)} \leq \frac{j}{m}:$$

3: Reject tests H_{0i} when $i = 1; \dots; j$.

Full Bayes procedure (FB)

Bayesian approach to multiple testing has been acquiring more attention lately, motivated by the need to analyse DNA microarray data. Among possibly a really large set of genes, the goal might be for example to locate the particular ones activated by a specific stimuli.

The issue of Bayesian specification, i.e. choosing prior distribution, computing key posterior quantities and finding useful ways to display them, is studied in Scott and Berger (2006) [18]. The paper introduces the full Bayes approach for minimizing the posterior $BR_{1;1}$ using importance sampling and for that, the following quantities are defined:

- Prior on p_A :

$(p_A) = \ (1 - p_A)^{\ -1}$ with adjustable parameter $\ :$

- Prior on $\ ^2$ and $\ ^2$:
$(\ ^2;\ ^2) = (\ ^2 +\ ^2)^{-2}$:

- The marginal posterior distribution of $(\ ^2;\ ^2; p_A)$:

$$(\ ^2;\ ^2; p_A|x) = C_1^{-1} \sum_{j=1}^{m} \left[\frac{p_0}{\sqrt{2\ \ ^2}} \exp\left(\frac{-x_j^2}{2\ ^2}\right) + \frac{p_A}{\sqrt{2\ (\ ^2 +\ ^2)}} \exp\left(\frac{-x_j^2}{2(\ ^2 +\ ^2)}\right) \right]$$

$\times\ \ (\ ^2;\ ^2) \times\ \ (p_A)$;
where C_1 is the normalization constant:

As mentioned in Scott and Berger [18], p_A can be expressed as an expectation of a known function $h(\ ^2;\ ^2; p_A)$ with respect to the posterior $\ (\ ^2;\ ^2; p_A|x)$. Evaluating p_A while dealing with a large sample size (m), importance sampling is the most efficient way to compute these kinds of posterior expectations because the same multivariate importance sample can be used for all of the computations. The idea is to generate a sample from an approximation of the posterior and use it to compute an importance sampling estimate of p_A. With a large m the distribution for $(\ ^2;\ ^2; p_A)$ is most likely close to normal, so one suitable proposal for the approximation is Student's t importance density.

The following procedure FB follows the guidelines described in Scott and Berger [18].

Algorithm 2: FB

1: Eliminate domain restrictions by transforming to the parameters

$$= \log(\ ^2); \quad = \log(\ ^2); \quad = \log(\frac{p_0}{p_A}):$$

2: Compute the negative logarithm of the transformed posterior $(\ ^2; \ ^2; p_A|x)$:

$$\begin{aligned} & - \log \ (e \ ; e \ ; (1 + e \ )^{-1}|x) \\ & = - \sum \log \left[(1 - (e \ + 1)^{-1}) f_0(x_i|e \ ) + (e \ + 1)^{-1} f_A(x_i|e \ ; e \ ) \right] \\ & - 2 \log(e \ + e \ ) + (\ - 1)(\log \ + \log(1 - (e \ + 1)^{-1})): \end{aligned}$$

3: Find the mode $(\hat{}; \hat{}; \hat{})$ of the negative logarithm from Step 2.
4: Compute the Hessian matrix of the negative logarithm of posterior at the mode.
5: Define the importance function as Student's t importance density

$t_3(\ ; \ ; \ |(\hat{}; \hat{}; \hat{}); aH^{-1})$; where aH^{-1} is the covariance matrix, $a = 5$:

6: Draw a sample $(\ _i; \ _i; \ _i)$ from $t_3(\ ; \ ; \ |(\hat{}; \hat{}; \hat{}); aH^{-1})$.
7: Define the importance sampling weights

$$w_i = \exp \log(\ (e \ ; e \ ; \frac{e}{1+e}|x)) - \log(t_3(\ _i; \ _i; \ _i|(\hat{}; \hat{}; \hat{}); aH^{-1})):$$

8: Reject H_{0i} if

$$P(\ _i = 0|x) \equiv \frac{\sum_i h(e^{\ _i}; e^{\ _i}; \frac{e^\lambda}{1+e^\lambda}) w_i}{\sum_i w_i} < 0.5;$$

where

$$\begin{aligned} & h(e^{\ _i}; e^{\ _i}; \frac{e^{\ _i}}{1 + e^{\ _1}}) \\ & = \left[(1 + \frac{1 + e^{\ _i}}{e^{\ _i}} - 1) \sqrt{\frac{e^{\ _i}}{e^{\ _i} + e^{\ _i}}} \exp\left(\frac{e^{\ _i} x_i^2}{2e^{\ _i}(e^{\ _i} + e^{\ _i})} \right) \right]^{-1}: \end{aligned}$$

Parametric empirical Bayes procedures (PEB1&PEB2)

The most commonly used approach to parametric empirical Bayes, PEB1, includes estimation of the unknown variables of 3.1 by maximizing the likelihood function

$$L(X_1,\ldots,X_m|p_A,\ ,\ ) = \prod_{i=1}^{m}(p_A f_A(X_i) + (1-p_A)f_0(X_i)). \tag{3.3}$$

These estimates are then plugged into the Bayes oracle.

Algorithm 3: PEB1

1: Fix p_A and find the estimates $\hat{}(p_A)$ and $\hat{}(p_A)$ using **the expectation-maximization algorithm** for the likelihood function 3.3.
2: Estimate p_A by maximizing

$$L(X_1,\ldots,X_m|p_A,\hat{}(p_A),\hat{}(p_A)).$$

3: Plug the estimates into the Bayes oracle (3.2).

Since PEB1 procedure has a large FDR when the data is sparse, i.e. p_A is very small, (Bogdan et al. [5]), the following procedure called PEB2 was created to stabilize the MLE by using prior information on p_A, as in Scott and Berger [18] and Bogdan et al. [5], where the prior density

$$f(p_A) = \ (1-p_A)^{\ -1} \tag{3.4}$$

is used. In this study $\ = 5.58$ is used as proposed in Scott and Berger [18].

Furthermore, instead of the EM algorithm, PEB2 uses the method of moments in the estimation of and . In Bogdan et al. [5] it is stated that using the fourth moment makes the method sensitive to the change in the tail of the mixture model and therefore gives good results in a very sparse case.

Algorithm 4: PEB2

1: Fix p_A and find the estimates $\hat{}(p_A)$ and $\hat{}(p_A)$ using **the method of moments** for the second and fourth moments of the mixture model 3.1.
2: Estimate p_A by maximizing

$$\log L(X_1; :::; X_m | p_A; \hat{}(p_A); \hat{}(p_A)) - \log(f(p_A)): \tag{3.5}$$

3: Plug the estimates into the Bayes oracle (3.2).

Modified Benjamini-Hochberg procedures (BH1&BH2)

In Bogdan et al. [5], the parametric empirical Bayes procedures are used to enhance the Benjamini-Hochberg method. Basically, the estimation of the variables is done with the help of PEB1 and PEB2 after which the steps of BH procedure are executed accordingly.

Algorithm 5: BH1

1: Find the estimates $\hat{}(p_A)$ and $\hat{p}_A$ following Step 1 and Step 2 in PEB1.
2: Follow Step 1, Step 2 and Step 3 of BH, using a threshold

$$p_{(j)} \leq \frac{j}{m(1-\hat{p}_A)}: \tag{3.6}$$

Algorithm 6: BH2

1: Find the estimates $\hat{}(p)$ and $\hat{p}_A$ following Step 1 and Step 2 in PEB2.
2: Follow Step 2 in BH1.

Chapter 4

Phase Transition in Multiple Testing

As mentioned earlier in the Chapter 2, the existence of a sharp phase boundary for frequentist model selection procedures in the domain of the variables 'availability of signal' and 'strength of signal' has been proved. In Chapter 3 the same notions of variables were defined for multiple testing, p being the fraction of signals and $^2 = \ ^2$ the strength of signal.

Furthermore, since the early literature on the subject, see for example Hodges and Lehmann (1956) [16], it has been known that multiple testing can be reformulated as model selection, making it only natural to assume that phase transition might be occurring in the context of multiple testing as well, within these variables. In Jin et al. [17], a sharp phase transition was indeed found when investigating overall error (OE) of the frequentist testing methods, which also motivated the focus of this thesis on this error rate and trying to find traces of phase behaviour in it using the fully Bayes and parametric empirical Bayes approaches as well.

Finding a universal phase boundary for multiple testing procedures would be helpful when figuring out whether the specific inference scheme is feasible or not for the dataset in question. Also, these findings would actually work as a preliminary study for connecting this Bayes boundary with variable selection. Yet, these are not the only approaches to benefit from the boundary. In Scott and Berger [19] a peculiar discrepancy was revealed between full Bayes and empirical Bayes variable selection which doesn't seem to arise from the failure to account for uncertainty in the empirical Bayes estimate, the usual issues in this type of problems. The possibility for a serious difference between the two Bayesian approaches remains even when the empirical estimate converges asymptotically to the true hyperparameter value. This phenomenon and the conflicts originated from it are yet to be fully corrected so in this thesis the matter is investigated from the phase boundary point of view.

The search for phase transition is natural to start with an empirical simulation study based on the intuitions and expectations of the matter. If the phase boundary was real, the results from the simulation would most likely contain clear visualizations of sharp phase transitions and from them the estimation of the universal boundary would be possible.

Intuition

Before starting to work with the simulation, certain expectations were had of the results. Based on earlier research it was known that BH (and PEB1) would not work well under sparsity, i.e. when p_A is very small. Also, when the variance ^{2} is small, the null and not null distributions from the model 3.1 would not differ that much from each other, making the estimation of the variables difficult and resulting in problems in the inference. Thus it was expected to come upon a situation visualized in Figure 4.1: under the red curve, where the strength of the signal ^{2} is small and the proportion of signal p_A very small, the testing methods would not perform well. This red curve would then be the phase boundary for the performance of the testing method in question.

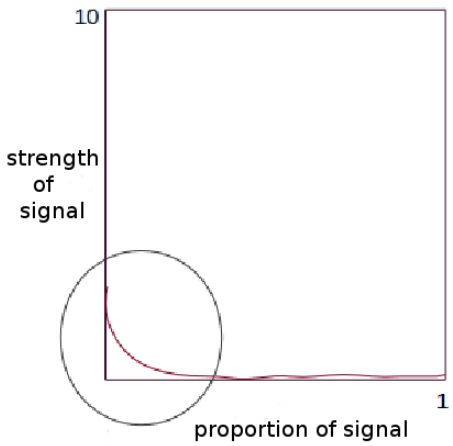


Figure 4.1: Intuition

Simulation study

For the simulation, the data is generated from the distribution 3.1 using 50 varying values of p_A from interval $[0; 1]$ and 50 values of ^{2} from $[0; 10]$, the variance being $^2 = 1$ and the significance level for BH $= 0{:}05$. The length m of the data vector X consisting of the test statistics X_i is first 100, then 200 and finally 400. The experiment would have benefited from much larger volumes of data but unfortunately due to computational costs and difficulties this wasn't viable. However, with the most plentiful data vectors possible, a grid is created to run the inference on, each grid point presenting a test statistic generated with a unique combination of $p;\ ^2$ and m.

For each of the grid points the hypothesis is defined: is this particular data vector generated from the null distribution or not. All the five methods, BH, FB, PEB1, PEB2, BH1 and BH2 are implemented (details in Appendix B) and used to test simultaneously first the data generated with $m = 100$, then $m = 200$ and finally $m = 400$.

After the inference part is finished, the performance results of each of the methods are estimated using the false discovery rate, the Bayes false discovery rate, the Bayes risk, the power and the overall error, main focus being on OE which is estimated as

$$OE \equiv \sum_{i=1}^{m} (V_i = (V_i + S_i)) \frac{1}{\sum_{i=1}^{m} (V_i + S_i > 0)} + \sum_{i=1}^{m} (T_i = (T_i + S_i)) \frac{1}{\sum_{i=1}^{m} (T_i + S_i > 0)}; \qquad (4.1)$$

where the values of $V_i; S_i; T_i; i = 1; :::; m$ are from the table of the counts of the possible outcomes (Table 3.2).

Results

Going through the results, the focus is mostly put on the overall rate OE. The results of *FDR, pFDR, BFDR, BR* and *Power* are presented in Appendix 4.3.

The overall errors of the Bayes oracle are presented in Figure 4.2 with a clear hints of phase transition, visualized as the red boundaries closely resembling the intuitive bound in Figure 4.1. These are 'the best case scenario' results for multiple testing, since the oracle uses the real values of the parameters instead of estimates.

Overall errors of BO

(a) OE of BO, m=100 (b) OE of BO, m=200 (c) OE of BO, m=400

Figure 4.2

The other boundary-like phenomenon in results given by the oracle when p_A is large should not be misconstrued as a phase transition. This is just a numerical boundary due to the right-hand side of the Bayes oracle 3.2 giving only negative values:

$$BO = \frac{2(\ ^2 + \ ^2)\ ^2}{^2}\left[\log\frac{p_0}{p_A} + \frac{n}{2}\log\frac{\ ^2 + \ ^2}{^2}\right] = \underbrace{\frac{2(1 + \ ^2)1}{^2}}_{\geq 0\ \forall\ ^2\in(0;10]}\left[\log\frac{1-p_A}{p_A} + \underbrace{\frac{1}{2}\log\frac{1+\ ^2}{1}}_{\geq 0\ \forall\ ^2\in(0;10]}\right] < 0$$

$$\Leftrightarrow \log\frac{1-p_A}{p_A} < 0 \text{ and } -\log\frac{1-p_A}{p_A} > \frac{1}{2}\log(1 + \ ^2):$$

Now, $\log\frac{1-p_A}{p_A} < 0 \Leftrightarrow 0 < \frac{1-p_A}{p_A} < 1 \Leftrightarrow p_A > \frac{1}{2}$:

Then, $-\log\frac{1-p_A}{p_A} > \frac{1}{2}\log(1 + \ ^2) \Leftrightarrow \log\left(\frac{1-p_A}{p_A}\right)^{-1} > \log\sqrt{1 + \ ^2}$

$$\Leftrightarrow \frac{p_A}{1-p_A} > \sqrt{1 + \ ^2} \Leftrightarrow \frac{p_A^2}{(1-p_A)^2} > 1 + \ ^2 \Leftrightarrow \ ^2 < \frac{p_A^2}{(1-p_A)^2} - 1:$$

So, the Bayes oracle gives only negative values and therefore rejects every null hypothesis if $p_A > 1=2$ and $^2 < p_A^2=(1-p_A)^2 - 1$. This boundary is presented in Figure 4.3 and it closely reminds the ones from Figure 4.2. Obviously this conduct leads to the elimination

of Type 2 errors since no null hypothesis goes unrejected. Also, the probability of Type 1 error shrinks down due to the lack of true null hypotheses. Thus OE gives small values and makes it seem like the oracle makes no mistakes.

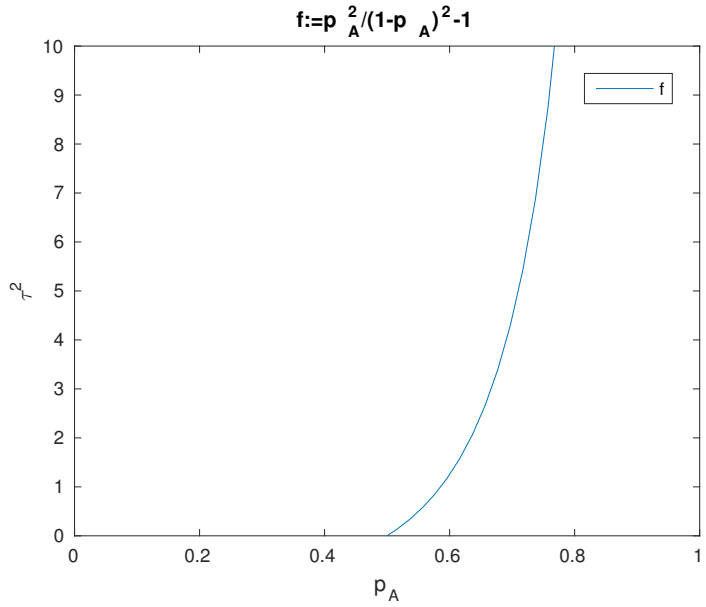
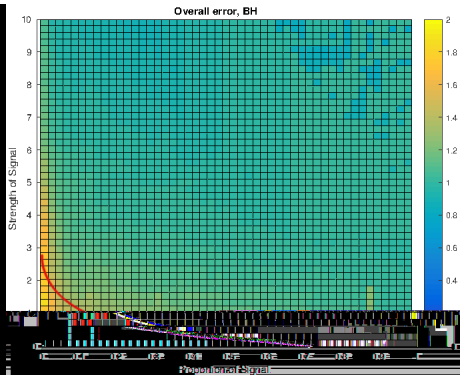


Figure 4.3

The OE functions 4.1 for each of the inference methods with $m = 100; 200; 400$ are plotted below in Figures 4.4-4.9. It is noticeable that all of them are resembling the intuitive expectations of the performance results, presented in Figure 4.1. The color yellow represents high failure rate and it is visible that each of the methods is failing when both the proportion of signal and the strength of signal are small, just like anticipated.

The red curves plotted within the results are the estimated phase boundaries of the oracle from Figure 4.2, with respective values of m.

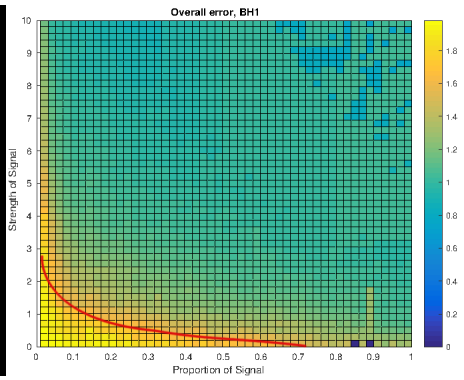
Overall errors of BH



OE of BH, m=100 OE of BH, m=200 OE of BH, m=400

Figure 4.4

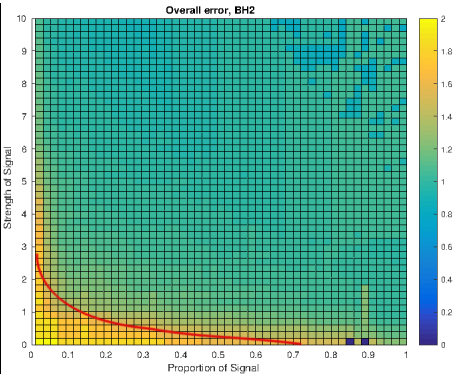
Overall errors of BH1



OE of BH1, m=100 OE of BH1, m=200 OE of BH1, m=400

Figure 4.5

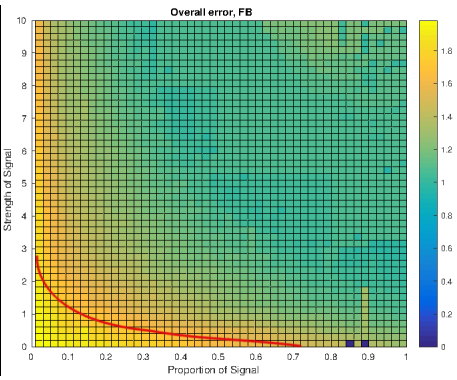
Overall errors of BH2



OE of BH2, m=100 OE of BH2, m=200 OE of BH2, m=400

Figure 4.6

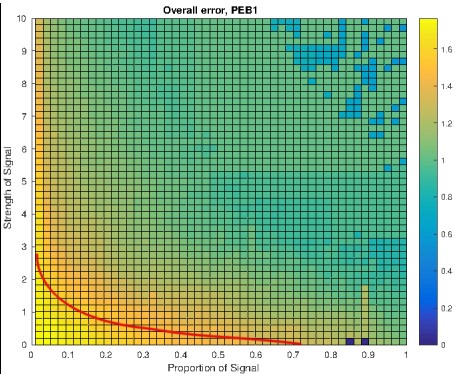
Overall errors of FB



OE of FB, m=100 OE of FB, m=200 OE of FB, m=400

Figure 4.7

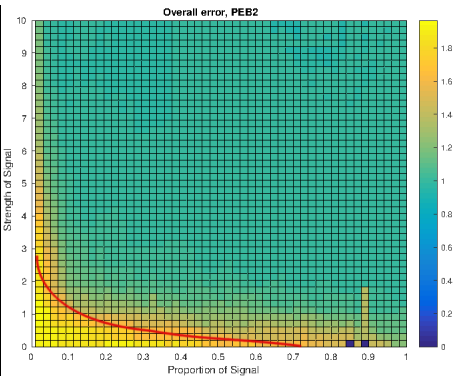
Overall errors of PEB1



OE of PEB1, m=100 OE of PEB1, m=200 OE of PEB1, m=400

Figure 4.8

Overall errors of PEB2



OE of PEB2, m=100 OE of PEB2, m=200 OE of PEB2, m=400

Figure 4.9

It is visible that the estimated phase boundaries of the oracle do fit quite appropriately into the results of each of the methods. The clearest resemblance can be seen in the results of BH2 and PEB2, the methods that were indeed supposed to work well with sparse data. Furthermore, even the most indistinct performances, that being the behaviour of FB and PEB1, are clearly showing similarities with the desired boundary of the oracle.

With increasing values of m, the oracle's boundary fits better and better into the performance rates of the methods. Thus the figures are indeed showing traces of phase transition with boundaries that could be agreeing with a sufficiently large m, since they all bear a close resemblance to the one of the Bayes oracle.

Bayes vs Empirical Bayes: Phase boundary

In Scott and Berger (2010), considerable differences between full Bayes and empirical Bayes approaches in variable selection were found and proved. The paper showed that the failure to account for hyperparameter uncertainty in the empirical Bayes estimate is not the cause for the discrepancy, since even while the estimate converges asymptotically to the true value of the hyperparameter, the potential for a major difference still remains. The motivation of this was to bring to light the conflicts arising from the phenomenon because empirical Bayes approach is sometimes used as an approximation to full Bayes analysis. In the paper Scott and Berger also suggest considering for example some alternative non-Bayesian ways instead of marginal maximum likelihood when estimating p.

However, in the context of this specific multiple testing study, the likeness between the overall errors of PEB1 and FB is obvious in Figures 4.7-4.8. Visually observed from the overall error point of view, there doesn't seem to be any valid reason not to use PEB1 as an approximation to full Bayes analysis, despite PEB1 using the MLE of p_A.

On the other hand, comparing Figures 4.7 and 4.9 it is apparent that FB is not yielding as good results as PEB2 even though PEB2 is utilizing the same prior density of p_A as FB when stabilizing the MLE. The correspondence between the two approaches begins and ends with the shared feature of high overall error rate when both proportion and strength of signal are small. This clearly reveals the lack of competence in the policy of using PEB2 as an estimate of FB universally (even though, in these specific settings of the model, choosing to work with PEB2 instead of FB would of course only be smart).

To make the differences between the three methods more clear, in Figure 4.10 the approximate phase boundaries for the overall error rates when $m = 400$ are visualized as the purple, black and green curves. The red curve represents the boundary of the oracle in Figure 4.2 (c). Then, all three boundaries are combined in Figure 4.11.

FB PEB1 PEB2

Figure 4.10: Estimated boundaries when $m = 400$

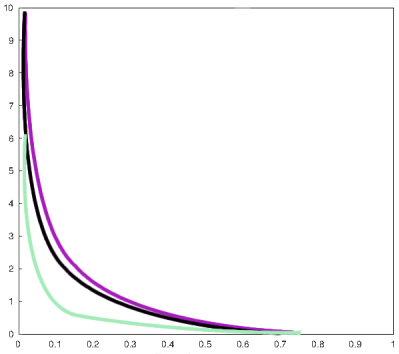


Figure 4.11: Boundaries when $m = 400$

From the phase boundary point of view, as was discussed in the previous section, with a sufficiently large value of m, when all the phase boundaries of the testing methods would agree, using both of the PEB approaches as a computational simplification of FB analysis would indeed be reasonable. Thus defining the universal phase boundary properly would be helping Bayesian scientists to avoid the conflicts while planning to utilize the empirical Bayes.

Chapter 5

Conclusion

The aim of this thesis was to find evidence of phase transition in multiple testing procedures via a simulation study. This was a success; in Chapter 4 a numerous set of figures indicating sharp changes in the overall error $OE = P(\text{Type I error}) + P(\text{Type II error})$ were presented and even proof of a universal phase boundary was discovered.

This boundary when properly defined isn't just about identifying regions of p and $^2 = {}^2$ for which multiple testing is feasible and for which it isn't. Indeed, Bayesian methods in variable selection are expected to show similar results due to multiple testing actually being a form of model selection. Now that the existence of the boundary has already been established the focus can be wholly put on defining the boundary theoretically.

In Chapter 4 also the earlier research on the differences between full Bayes and empirical Bayes was mentioned and discussed from the phase transition point of view: the universal phase boundary for variable selection could shed light on circumstances needed for empirical Bayes selection to actually work as a approximation of full Bayes analysis.

Furthermore, the results of this thesis could actually work as a preliminary results of the search for evidence of the following proposition:

Proposition 5.1. *There exists a sharp phase transition phenomenon in the performance of a statistical inference model in the domain of the variables 'availability of signal' and 'strength of signal', for* ***any*** *statistical model where those variables can be defined.*

Finding the proof for this proposition could be a game changer in practical applications of statistical inference, such as optimizing car performance, cancer treatment and national

security, where large amounts of data are being handled every day with computational costs and different kinds of errors in inference.

On the other hand, until now the asymptotic theory has been considered as the only credible way to prove consistency of statistical methods, and the comparison between procedures has been carried out solely by comparing their asymptotic convergence rates. Working with phase transition could lead to a completely new theoretical and practical perspective on statistical inference methods by finding and analytically defining a sharp phase transition boundary for them. Thus continuing the research of this thesis by analytically defining the universal phase boundary for multiple testing procedures would naturally be a good next step towards establishing this new theory.

Bibliography

[1] Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference* (2014), iau005.

[2] Benjamini, Y., and Hoc berg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* (1995), 289–300.

[3] Bogdan, M., C akrabarti, A., Frommlet, F., and G os , J. K. Asymptotic bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics* (2011), 1551–1579.

[4] Bogdan, M., G os , J. K., and Doerge, R. Modifying the schwarz bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics 167*, 2 (2004), 989–999.

[5] Bogdan, M., G os , J. K., Tokdar, S. T., et al. 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*. Institute of Mathematical Statistics, 2008, pp. 211–230.

[6] Dono o, D., and Stodden, V. Breakdown point of model selection when the number of variables exceeds the number of observations. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings* (2006), IEEE, pp. 1916–1921.

[7] Dono o, D., and Tanner, J. Counting faces of randomly projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society 22*, 1 (2009), 1–53.

[8] Dono o, D., and Tanner, J. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions A: Mathematical, Physical and Engineering Sciences 367*, 1906 (11 2009), 4273–4293.

[9] Dono o, D. L. High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete & Computational Geometry 35*, 4 (2006), 617–652.

[10] Dono o, D. L., Maleki, A., and Montanari, A. The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory 57*, 10 (2011), 6920–6941.

[11] Dutta, R., Bogdan, M., and G os , J. K. Model selection and multiple testing-a bayesian and empirical bayes overview and some new results. *arXiv preprint arXiv:1510.00547* (2015).

[12] Efron, B., and Tibs irani, R. Empirical bayes methods and false discovery rates for microarrays. *Genetic epidemiology 23*, 1 (2002), 70–86.

[13] Erd6s, P., and Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci 5* (1960), 17–61.

[14] Flory, P. J. Molecular size distribution in three dimensional polymers. i. gelation1. *Journal of the American Chemical Society 63*, 11 (1941), 3083–3090.

[15] Frommlet, F., C akrabarti, A., Murawska, M., and Bogdan, M. Asymptotic bayes optimality under sparsity for generally distributed effect sizes under the alternative. *arXiv preprint arXiv:1005.4753* (2010).

[16] Hodges Jr, J. L., and Le mann, E. L. The efficiency of some nonparametric competitors of the t-test. *The Annals of Mathematical Statistics* (1956), 324–335.

[17] Jin, J., Ke, Z. T., and Wang, W. Phase transitions for high dimensional clustering and related problems. *arXiv preprint arXiv:1502.06952* (2015).

[18] Scott, J. G., and Berger, J. O. An exploration of aspects of bayesian multiple testing. *Journal of Statistical Planning and Inference 136*, 7 (2006), 2144–2162.

[19] Scott, J. G., Berger, J. O., et al. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics 38*, 5 (2010), 2587–2619.

[20] Seeger, P. A note on a method for the analysis of significances en masse. *Technometrics 10*, 3 (1968), 586–593.

[21] Simes, R. J. An improved bonferroni procedure for multiple tests of significance. *Biometrika 73*, 3 (1986), 751–754.

[22] Stockmayer, W. H. Theory of molecular size distribution and gel formation in branched polymers ii. general cross linking. *The Journal of Chemical Physics 12*, 4 (1944), 125–131.

[23] STOJNIC, M. Various thresholds for l1-optimization in compressed sensing.

[24] Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*, 3 (2002), 479–498.

[25] Vattikuti, S., Lee, J. J., C ang, C. C., Hsu, S. D., and C ow, C. C. Applying compressed sensing to genome-wide association studies. *GigaScience 3*, 1 (2014), 1.

Appendix A

Hypothesis testing

A statistical hypothesis is a statement about the distribution of a random variable $\mathbf{X} = (X_1; X_2; :::; X_n)$, where $X_i \in S$ is the vector of measurements for the ith object.

In hypothesis testing, the aim is to find out if there is enough evidence to reject a so-called null hypothesis, denoted H_0, in favor of a conjectured alternative hypothesis, usually denoted H_A. An hypothesis test is a statistical decision: H_0 is either rejected or failed to reject. The decision must be based on the observed value x of the data $\mathbf{X}$. An appropriate subset $R \subset S$, a rejection region, needs to be found and H_0 is rejected if and only if $x \in R$.

There are two types of *errors* in hypothesis testing.

- Type 1 Error: Rejecting H_0 when H_0 is true.
- Type 2 Error: Failing to reject H_0 when H_A is true.

	Fail to reject H_0	Reject H_0
H_0 true	Correct	Type 1 Error
H_A true	Type 2 Error	Correct

The maximum probability of a Type 1 Error, ie the probability that a true null hypothesis is rejected, is called the *significance level* of the test and denoted . The most usual significance levels are = 0:05, = 0:01, = 0:001.

The p-value of the test is the smallest significance level for which H_0 can be rejected.

If the distribution of $\mathbf{X}$ depends on an unknown parameter $\theta \in \Theta$, Θ being a parameter space, the hypotheses are of the form

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_A : \theta \notin \Theta_0;$$

where $\Theta_0 \subset \Theta$ is prescribed.

Now, suppose that θ is a real parameter and $\theta_0 \in \Theta$ a specified value. Then three special cases of hypothesis tests can be described:

i) A two-sided test: $H_0 : \theta = \theta_0$ vs. $H_A : \theta \neq \theta_0$

ii) A left-tailed test: $H_0 : \theta \geq \theta_0$ vs. $H_A : \theta < \theta_0$

iii) A right-tailed test: $H_0 : \theta \leq \theta_0$ vs. $H_A : \theta > \theta_0$

Tests Based on a Student's t-Statistic

Let's now assume that the data vector $\mathbf{X} = (X_1, \ldots, X_n)$ is normally distributed with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma \in (0, \infty)$. The sample mean of the data vector $\mathbf{X}$ is

$$M = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

A Student's t-test Statistic will lead to good tests of μ without requiring the knowledge of σ:

$$T(a) = \frac{M - a}{S/\sqrt{(n)}}; \text{ where } a \in \mathbb{R}.$$

$T(\mu)$ has the Student's t-distribution with $n - 1$ degrees of freedom.
For the t-distribution with $k > 0$ degrees of freedom we let f_k denote the probability density function and Φ_k the distribution function.
For $p \in (0, 1)$, let $t_k(p)$ denote the quantile of order p so that $t_k(p) = \Phi_k^{-1}(p)$. For specified values of k and p, the values of $t_k(p)$ can be obtained.

For every $\mu \in (0, 1)$ and $p \in (0, 1)$ the following tests have the significance level α:

i) Reject $H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq \mu_0$ iff $T(\mu_0) \leq t_{n-1}(\alpha - p)$ or $T(\mu_0) \geq t_{n-1}(1-p)$.

ii) Reject $H_0 : \mu \geq \mu_0$ vs. $H_A : \mu < \mu_0$ iff $T(\mu_0) \leq t_{n-1}(\alpha) = -t_{n-1}(1 - \alpha)$.

iii) Reject $H_0 : \quad \leq \quad_0$ vs. $H_A : \quad > \quad_0$ iff $T(\quad_0) \geq t_{n-1}(1 - \quad)$:

The p-values of the tests above are

i) $2(1 - \Phi_{n-1}(|T(\quad_0)|))$

ii) $1 - \Phi_{n-1}(T(\quad_0))$

iii) $\Phi_{n-1}(T(\quad_0))$

Appendix B

Implementation of the methods

Details

Implementation of EM algorithm

Expectation step:

$$Q((\ ^2;\ ^2)|(\ _g^2;\ _g^2))$$
$$= \sum_{i=1}^{m} \log f_A(x_i|\ ^2;\ ^2)p(I = A|x_i;\ _g^2;\ _g^2) + \sum_{i=1}^{m} \log f_0(x_i|\ ^2)p(I = 0|x_i;\ _g^2)$$

$$= \sum_{i=1}^{m}(\sum_{j=1}^{n} \log f_A(x_i|\ ^2;\ ^2))c_{1i} + \sum_{i=1}^{m}(\sum_{j=1}^{n} \log f_0(x_i|\ ^2))c_{2i}$$

$$= \sum_{i=1}^{m}(\sum_{j=1}^{n} -x_{ij}^2 2(\ ^2 + \ ^2) - \log\sqrt{2\ (\ ^2 + \ ^2)})c_{1i} + \sum_{i=1}^{m}(\sum_{j=1}^{n} \frac{-x_{ij}^2}{2\ ^2} - \log\sqrt{2\ \ ^2})c_{2i}$$

$$:= g(\ ;\ );$$

where

$$c_{1i} = p(I = A|x_i;\ _g^2;\ _g^2) = \frac{P(x_i;\ _g^2;\ _g^2; I = A)}{P(x_i;\ _g^2;\ _g^2)}$$

$$= \frac{P(x_i;\ {}^2_g;\ {}^2_g; I = A)}{P(I = A)P(x_i;\ {}^2_g;\ {}^2_g; I = A) + P(I = 0)P(x_i;\ {}^2_g;\ {}^2_g; I = 0)}$$

$$= \frac{p_A f_A(x_i;\ {}^2_g;\ {}^2_g)}{p_A f_A(x_i;\ {}^2_g;\ {}^2_g) + p_0 f_0(x_i;\ {}^2_g)};$$

$$c_{2i} = p(I = 0|x_i;\ {}^2_g;\ {}^2_g) = \frac{P(x_i;\ {}^2_g;\ {}^2_g; I = 0)}{P(x_i;\ {}^2_g;\ {}^2_g)}$$

$$= \frac{P(x_i;\ {}^2_g;\ {}^2_g; I = 0)}{P(I = A)P(x_i;\ {}^2_g;\ {}^2_g; I = A) + P(I = 0)P(x_i;\ {}^2_g;\ {}^2_g; I = 0)}$$

$$= \frac{p_0 f_0(x_i;\ {}^2_g)}{p_A f_A(x_i;\ {}^2_g;\ {}^2_g) + p_0 f_0(x_i;\ {}^2_g)}:$$

Maximization step:

$$\text{(i)}\ \frac{@g}{@} = \sum_{i=1}^{m}(\sum_{j=1}^{n}\frac{x_{ij}^2}{(\ {}^2 +\ {}^2)^2} - \frac{}{\ {}^2 +\ {}^2})c_{1i} = \sum_{i=1}^{m}\sum_{j=1}^{n}(\frac{c_{1i}\ x_{ij}^2}{(\ {}^2 +\ {}^2)^2} - \frac{c_{1i}}{\ {}^2 +\ {}^2}):$$

$$\text{(ii)}\ \frac{@g}{@} = \sum_{i=1}^{m}(\sum_{j=1}^{n}\frac{x_{ij}^2}{(\ {}^2 +\ {}^2)^2} - \frac{}{\ {}^2 +\ {}^2})c_{1i} + \sum_{i=1}^{m}(\sum_{j=1}^{n}\frac{x_{ij}^2}{3} - \frac{1}{})c_{2i}$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{n}c_{1i}x_{ij}^2\frac{}{(\ {}^2 +\ {}^2)^2} - n\sum_{i=1}^{m}c_{1i}\frac{}{\ {}^2 +\ {}^2} + \sum_{i=1}^{m}\sum_{j=1}^{n}c_{2i}x_{ij}^2\frac{1}{3} - n\sum_{i=1}^{m}c_{2i}\frac{1}{}:$$

$$\frac{@g}{@} = 0$$

$$\Leftrightarrow \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{c_{1i}x_{ij}^2}{(\ ^2 + \ ^2)^2} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{c_{1i}}{^2 + \ ^2}$$

$$\Leftrightarrow \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{1i}x_{ij}^2}{(\ ^2 + \ ^2)^2} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{1i}}{^2 + \ ^2}$$

$$\Leftrightarrow \sum_{i=1}^{m}\sum_{j=1}^{n} c_{1i}x_{ij}^2 = (\ ^2 + \ ^2) \sum_{i=1}^{m}\sum_{j=1}^{m} c_{1i}$$

$$\Leftrightarrow \ ^2 + \ ^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{1i}x_{ij}^2}{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{1i}}$$

$$\Leftrightarrow \ ^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{1i}x_{ij}^2}{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{1i}} - \ ^2:$$

$$\frac{@g}{@} = 0$$

$$\Leftrightarrow \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{1i}x_{ij}^2}_{=:A} \frac{}{(\ ^2 + \ ^2)^2} - \underbrace{n\sum_{i=1}^{m} c_{1i}}_{=:B} \frac{}{^2 + \ ^2} + \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{2i}x_{ij}^2}_{=:C} \frac{1}{\ ^3} - \underbrace{n\sum_{i=1}^{m} c_{2i}}_{=:D} \frac{1}{}$$

$$= 0$$

$$\Leftrightarrow \ -\frac{B}{A}\ (\ ^2 + \ ^2) + \frac{C}{A}\frac{(\ ^2 + \ ^2)^2}{^3} - \frac{D}{A}\frac{(\ ^2 + \ ^2)^2}{} = 0$$

$$\Leftrightarrow \ ^4 - \frac{B}{A}(\ ^6 + \ ^4\ ^2) + \frac{C}{A}(\ ^4 + \ ^4 + 2\ ^2\ ^2) - \frac{D}{A}(\ ^6 + \ ^2\ ^4 + 2\ ^4\ ^2) = 0$$

$$\Leftrightarrow \ ^4 - \frac{B}{A}(\ ^6 + \ ^4(\frac{A}{B} - \ ^2)) + \frac{C}{A}(\ ^4 + (\frac{A}{B} - \ ^2))^2 + 2\ ^2(\frac{A}{B} - \ ^2))$$

$$-\frac{D}{A}(\ ^6 + \ ^2(\frac{A}{B} - \ ^2)^2 + 2\ ^4(\frac{A}{B} - \ ^2)) = 0$$

$$\Leftrightarrow \quad {}^{6}(-\frac{B}{A}+\frac{B}{A}-\frac{D}{A}-\frac{D}{A}+2\frac{D}{A})$$
$$+ \quad {}^{4}(1-1+\frac{C}{A}+\frac{C}{A}-2\frac{C}{A}+2\frac{D}{B}-2\frac{D}{B})$$
$$+ \quad {}^{2}(-2\frac{C}{B}+2\frac{C}{B}-\frac{AD}{B^2})$$
$$+\frac{AC}{B^2}=0$$

$$\Leftrightarrow \quad {}^{2}(-\frac{AD}{B^2})+\frac{AC}{B^2}=0$$

$$\Leftrightarrow \quad {}^{2}=\frac{C}{D}:$$

So the estimates are:

$$\hat{}(p_A)=\sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}c_{2i}x_{ij}^2}{n\sum_{i=1}^{m}c_{2i}}}$$

and

$$\hat{}(p_A)=\sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}c_{1i}x_{ij}^2}{n\sum_{i=1}^{m}c_{1i}}-\hat{}(p_A)^2};$$

where

$$\sum_{i=1}^{m}\sum_{j=1}^{n} c_{2i}x_{ij}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{p_0 x_{ij}^2 \prod_{k=1}^{n} f_0(x_{ik}|\ {}^2_g)}{p_A \prod_{k=1}^{n} f_A(x_{ik}|\ {}^2_g;\ {}^2_g) + p_0 \prod_{k=1}^{n} f_0(x_{ik}|\ {}^2_g)};$$

$$n\sum_{i=1}^{m} c_{2i} = n\sum_{i=1}^{m} \frac{p_0 \prod_{k=1}^{n} f_0(x_{ik}|\ {}^2_g)}{p_A \prod_{k=1}^{n} f_A(x_{ik}|\ {}^2_g;\ {}^2_g) + p_0 \prod_{k=1}^{n} f_0(x_{ik}|\ {}^2_g)};$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} c_{1i}x_{ij}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{p_A x_{ij}^2 \prod_{k=1}^{n} f_A(x_{ik}|\ {}^2_g;\ {}^2_g)}{p_A \prod_{k=1}^{n} f_A(x_{ik}|\ {}^2_g;\ {}^2_g) + p_0 \prod_{k=1}^{n} f_0(x_{ik}|\ {}^2_g)} \text{ and}$$

$$n\sum_{i=1}^{m} c_{1i} = n\sum_{i=1}^{m} \frac{p_A \prod_{k=1}^{n} f_A(x_{ik}|\ {}^2_g;\ {}^2_g)}{p_A \prod_{k=1}^{n} f_A(x_{ik}|\ {}^2_g;\ {}^2_g) + (p_0 \prod_{k=1}^{n} f_0(x_{ik}|\ {}^2_g)}:$$

These E-step and M-step are repeated for several times, using $\hat{}(p_A)$ and $\hat{}(p_A)$ of the previous iteration as the guessed values of ^{2} and ^{2} in the following iteration.

Implementation of Method of Moments

$c_4 = \frac{1}{n}\sum_{i=1}^{n} x_i^4$ (fourth moment);

$c_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ (second moment):

$X_i \sim p_A \underbrace{N(0;\ ^2 + \ ^2)}_{=:W_1} + p_0 \underbrace{N(0;\ ^2)}_{=:W_2}$;

$E(X_i^4) = E(E(x_i^4|I_i))$

$= p_A E(W_1^4) + p_0 E(W_2^4)$

$= p_A(\sqrt{\ ^2 + \ ^2})^4 3!!) + p_0((\sqrt{\ ^2})^4 1!!)$

$= 3(p_A(\ ^2 + \ ^2)^2 + p_0\ ^4) = c_4$:

$E(X_i^2) = E(E(x_i^2|I_i))$

$= p_A E(W_1^2) + p_0 E(W_2^2)$

$= p_A((\sqrt{\ ^2 + \ ^2})^2 1!!) + p_0((\sqrt{\ ^2})^2 1!!)$

$= p_A(\ ^2 + \ ^2) + p_0\ ^2 = c_2$:

These imply

$$(\ ^2 + \ ^2)^2 = (\frac{c_2 - p_0\ ^2}{p_A})^2$$

$$\Leftrightarrow 3p_A(\frac{c_2 - p_0\ ^2}{p_A})^2 + 3p_0\ ^4 = c_4$$

$$\Leftrightarrow \frac{3c_2^2}{p_A} + \frac{3p_0^2}{p_A}\ ^4 - \frac{6p_Ap_0c_2\ ^2}{p_A^2} + 3p_0\ ^4 = c_4$$

$$\Leftrightarrow \ ^4(\frac{3p_0^2}{p_A} + 3p_0) - (6\frac{p_0}{p_A}c_2)\ ^2 + \frac{3c_2^2}{p_A} - c_4 = 0$$

$$\Leftrightarrow \ ^2 = \frac{-(-6c_2\frac{p_0}{p_A}) - \sqrt{(-6c_2\frac{p_0}{p_A})^2 - 4(\frac{3p_0^2}{p_A} + 3p_0)(\frac{3c_2^2}{p_A} - c_4)}}{2(\frac{3p_0^2}{p_A} + 3p_0)}$$

and also

$$^2 + \ ^2 = \frac{c_2}{p_A} - \frac{p_0}{p_A}\ ^2$$

$$\Leftrightarrow \ ^2 = \frac{c_2}{p_A} - \frac{1}{p_A}\ ^2.$$

So the estimates are:

$$\hat{}(p_A) = \sqrt{\frac{-(-6c_2\frac{p_0}{p_A}) - \sqrt{(-6c_2\frac{p_0}{p_A})^2 - 4(\frac{3p_0^2}{p_A} + 3p_0)(\frac{3c_2^2}{p_A} - c_4)}}{2(\frac{3p_0^2}{p_A} + 3p_0)}}$$

and

$$\hat{}(p_A) = \sqrt{\frac{c_2}{p_A} - \frac{1}{p_A}\hat{}(p_A)^2}:$$

Code

```
function [Data,index] = datagenerate(m,n,p,sigma_sq,tau_sq)
%Datagenerate takes in dimension of data m, size of data n and
    fraction of signals p, known variance sigma_sq and signal
    variance tau_sq. It returns the data and a vector of
    dimension mx1 with values 1 (if the corresponding data vector
     is created with the non-null distribution) and 0 (if data
    vector from null distribution).
index = repmat(binornd(1,p,m,1),1,n);
mu = repmat(normrnd(0,sqrt(tau_sq),m,1),1,n);
%The model (3.1)
Data = index.*normrnd(mu,sqrt(sigma_sq))+(1-index).*normrnd(0,
    sqrt(sigma_sq),m,n);
index = index(:,1);
```

```
function [d_index] = BO(Data,p,sigma_sq,tau_sq)
%BO takes in the data of size mxn, p, sigma_sq and tau_sq and
    returns the optimal decision made by the procedure as d_index
    . d_index is a vector of dimension mx1 with values 1 (if null
     is accepted), 0 (if alternative is accepted).
oracle = 2*(sigma_sq+tau_sq)*sigma_sq/tau_sq*(size(Data,2)/2*log
    ((sigma_sq+tau_sq)/sigma_sq)+log((1-p)/p));
if(size(Data,2)>1)
  x_sq = sum(Data'.^2);
else
```

```
  x_sq = Data.^2;
end
d_index = (x_sq>oracle);
```

```
function [d_index] = BH(Data,alpha,est_p)
%BH takes in the data of size mxn and returns the decision made by the procedure as d_index. d_index is a vector of dimension mx1 with values 1 (if null is accepted), 0 (if alternative is accepted).
%Step 1
[v] = ttest_my(Data);
comp = alpha/(size(Data,1)*(1-est_p));
b = 1:size(Data,1);
pvalues = [v' b'];
sortp = sortrows(pvalues, 1);
%Step 2 and Step 3
d_index = zeros(1,size(Data,1));
d_index(sortp(:,2)) = (sortp(:,1)<=(pvalues(:,2).*(comp)));
%%%%%%%%%%%%%%%%%%%%%%%%%%
function [v] = ttest_my(Data)
if (size(Data,2)==1)
    v = 2*normcdf(abs(Data),0,1,'upper')';
else
    for ind = 1:size(Data,1),
        [~,v(ind)] = ttest(Data(ind,:));
    end
end
```

```
function [d_index] = FB(Data)
%FB takes in the data of size mxn and returns the decision made by the procedure as d_index. d_index is a vector of dimension mx1 with values 1 (if null is accepted), 0 (if alternative is accepted).
addpath(genpath('DERIVESTsuite'))
%Step 1 and Step 2
neglogfunpost = @(args)neglogposterior_var_trans(args,Data);
%Step 3
initial_point = [0,0,log(9)];
[mode,~] = fminsearch(neglogfunpost,initial_point);
%Step 4
```

```
[H] = hessian(neglogfunpost,mode);
%Step 5
df = 3; a=5;
N = 10000; %Number of samples
C = a*inv(H);
%Step 6 and Step 7
y = mvtrnd(C,df,N)+repmat(mode,N,1); %y is in ksi,eta and lambda
weights = zeros(N,1);
for ind = 1:N
    weights(ind,1) = exp(-neglogposterior_var_trans(y(ind,:),Data
        )-log(mvtpdf(y(ind,:)-mode,C,df)));
end
%Step 8
y_t = [exp(y(:,1)),exp(y(:,2)),1./(1+exp(-y(:,3)))];
numerator = 0; denominator = 0;
for ind = 1:N
    numerator = numerator + h_SB(y_t(ind,:),Data)*weights(ind);
    denominator = denominator + weights(ind);
end
estimate_prob = numerator/denominator;
c = 3;
[cutoff] = threshold(Data,y_t,weights,c);
d_index = (estimate_prob<cutoff');
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [val] = neglogposterior_var_trans(args,Data)
ksi = args(1); eta = args(2); lambda = args(3);
%Define prior Hyperparameters
alpha = 5.58;
%Definition of posterior as in Eqn. 8 of Scott and Berger 2008
logposterior = @(t,s,p) sum(log(p*normpdf(Data,0,sqrt(s))+(1-p)*
    normpdf(Data,0,sqrt(s+t))))-2*log(s+t)+alpha*log(p);
logJacobian = @(y) sum(y)-2*log(1+exp(y(3)));
val = - logposterior(exp(ksi),exp(eta),1/(1+exp(-lambda))) -
    logJacobian([ksi,eta,lambda]);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [cutoff] = threshold(Data,y_t,weights,c)
%The grid for integration of |mu|
N = 1000;
mu = linspace(-10,10,N);
```

```
%tau_sq with samples
tau_sq = y_t(:,1).*y_t(:,2)./(y_t(:,1)+y_t(:,2));
cutoff_tmp = zeros(1,size(Data,1));
for ind_d = 1:size(Data,1)
    ro = y_t(:,1).*Data(ind_d)./(y_t(:,1)+y_t(:,2));
    %Outside Integration for each experiment
    prob_tmp = zeros(1,N);
    for ind_mu = 1:N
        h_tmp = (1./sqrt(2*pi*tau_sq))...
            .*exp((-1./(2*tau_sq)).*((mu(ind_mu)*ones(size(y_t
               ,1),1)-ro).^2));
        %Inside multiple integration w.r.t importance sampling
           is done
        prob_tmp(ind_mu) = sum(h_tmp.*weights)/sum(weights);
    end
    %prob_tmp = prob_tmp/sum(prob_tmp);%Probability as in
       Equation 10 of Scott & Berger
    cutoff_tmp(ind_d) = sum(abs(mu).*prob_tmp)/sum(prob_tmp);
    %Outside integration finished
end
%Final cutoff as in Equation 15.
cutoff = c*cutoff_tmp./(1+c*cutoff_tmp);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [val] = h_SB(args,x)
V = args(1); sigma2 = args(2); p =args(3);vs2 = V + sigma2;
o = 1 + (1/p - 1)*sqrt(sigma2/vs2)*exp((V*x.^2)/(2*sigma2*vs2));
val = 1./o;
```

```
function [d_index] = PEB(Data,est_p,est_sigma_sq,est_tau_sq)
%PEB takes in the data of size mxn and the estimates for p,
   sigma_sq and tau_sq. It returns the decision made by the
   procedure as d_index. d_index is a vector of dimension mx1
   with values 1 (if null is accepted), 0 (if alternative is
   accepted).
d_index = bayes_oracle(Data,est_p,est_sigma_sq,est_tau_sq);
```

```
function [est_p,est_sigma_sq,est_tau_sq] = em_estimate(Data)
%em_estimate takes in the data and returns the estimates for p,
   sigma_sq and tau_sq.
```

```
[est_p] = fminbnd(@(p) -likelihood_noprior_em(Data,p),0.01,0.99);
[est_sigma_sq,est_tau_sq] = emalgo(Data,est_p,0.5,5,10);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [loglik] = likelihood_noprior_em(Data,p)
[sigma_sq,tau_sq] = emalgo(Data,p,0.5,5,10);
loglik = sum(log(p*prod(normpdf(Data,0,sqrt(sigma_sq+tau_sq)),2)+(1-p)*prod(normpdf(Data,0,sqrt(sigma_sq)),2)));
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [est_sigma_sq,est_tau_sq] = emalgo(Data,p,g_sigma_sq,g_tau_sq,n_iter)
n = size(Data,2);
for ind =1:n_iter
    f_A = normpdf(Data,0,sqrt(g_sigma_sq+g_tau_sq));
    f_0 = normpdf(Data,0,sqrt(g_sigma_sq));
    c1 = [p*prod(f_A,2) (p*prod(f_A,2)+(1-p)*prod(f_0,2))];
    c1 = c1(:,1)./c1(:,2);
    c2 = [(1-p)*prod(f_0,2) (p*prod(f_A,2)+(1-p)*prod(f_0,2))];
    c2 = c2(:,1)./c2(:,2);
    %Estimate of Sigma
    C = sum(c2.*sum(Data.^2,2));
    D = n*sum(c2);
    est_sigma_sq = (C/D);
    %Estimate of Tau
    A = sum(c1.*sum(Data.^2,2));
    B = n*sum(c1);
    est_tau_sq = A/B-est_sigma_sq;
    g_sigma_sq = est_sigma_sq;
    g_tau_sq = est_tau_sq;
end
```

```
function [est_p,est_sigma_sq,est_tau_sq] = mm_estimate(Data,option)
%em_estimate takes in the data and returns the estimates for p, sigma_sq and tau_sq.
[est_p] = fminbnd(@(p) -likelihood_prior_mm(Data,p),0.01,0.99);
[est_sigma_sq,est_tau_sq] = moments(Data,est_p);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [post_loglik] = likelihood_prior_mm(Data,p)
```

```
[sigma_sq,tau_sq] = moments(Data,p);
beta = 5.58;
post_loglik = sum(log(p*prod(normpdf(Data,0,sqrt(sigma_sq+tau_sq
    )),2)+(1-p)*prod(normpdf(Data,0,sqrt(sigma_sq)),2)))-beta*log
    (p)-log(beta+1);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [est_sigma_sq, est_tau_sq] = moments(Data, p)
D = Data(:);
c2 = mean(D.^2);
c4 = mean(D.^4);
a =  3*(1/p-1);
b = -6*(1/p-1)*c2;
c = ((3/p)*c2^2)-c4;
est_sigma_sq = abs((-b-sqrt(b^2-4*a*c))/(2*a));
est_tau_sq = ((1/p)*abs(c2-est_sigma_sq));
```

```
function [fdr_cost,bfdr_cost,br_cost,power_cost,overall_error] =
    cost(index,d_index)
%cost takes in the decisions as d_index and the index vector
    with the real information and returns error rates FDR, BFDR,
    BR, Power and OE
del_0 = 1; del_A = 1;
U = zeros(1,size(index,2));
V = zeros(1,size(index,2));
T = zeros(1,size(index,2));
S = zeros(1,size(index,2));
anti_index = (2*index-1)<0;
anti_d_index = (2*d_index-1)<0;
for ind1 = 1:size(index,2)
    U(ind1) = sum(anti_index(:,ind1).*anti_d_index(:,ind1));
    V(ind1) = sum(anti_index(:,ind1).*d_index(:,ind1));
    T(ind1) = sum(index(:,ind1).*anti_d_index(:,ind1));
    S(ind1) = sum(index(:,ind1).*d_index(:,ind1));
end
t_fdr_cost = zeros(1,size(index,2));
for ind1 = 1:size(index,2)
    if ((V(ind1)+S(ind1))>0)
        t_fdr_cost(ind1) = V(ind1)./(V(ind1)+S(ind1));
    else
```

```
21          t_fdr_cost(ind1) = 0;
22      end
23  end
24  fdr_cost = mean(t_fdr_cost);
25  if(mean(V+S>0)~=0)
26      bfdr_cost = fdr_cost./mean(V+S>0);
27  else
28      bfdr_cost = fdr_cost;
29  end
30  br_cost = (del_0*mean(V)+del_A*mean(T))/size(index,1);
31  t_power_cost = zeros(1,size(index,2));
32  for ind1 = 1:size(index,2)
33      if ((S(ind1)+T(ind1))>0)
34          t_power_cost(ind1) = S(ind1)./(S(ind1)+T(ind1));
35      else
36          t_power_cost(ind1) = 0;
37      end
38  end
39  if(mean(T+S>0)~=0)
40      power_cost = mean(t_power_cost)./mean(T+S>0);
41  else
42      power_cost = mean(t_power_cost);
43  end
44  overall_error = bfdr_cost+(1-power_cost); % P(Type 1 error) + P(
      Type 2 error)
```

Appendix C

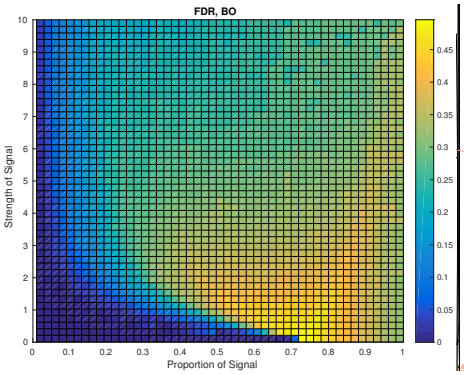
Results

In following figures, the error rates *FDR, BFDR, BR* and *Power* for each multiple testing method are presented. Phase transition is visible in these also.

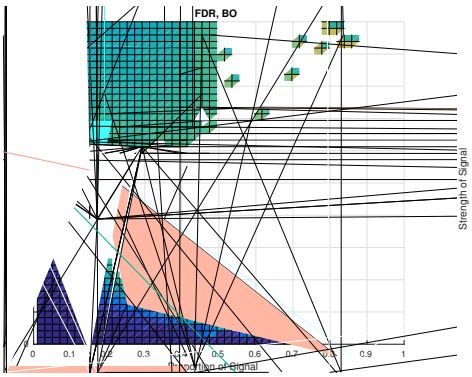
The figures of the Bayes oracle represent the best case scenarios for results and the performance rates of other methods should be compared to those of the oracle.

FDR

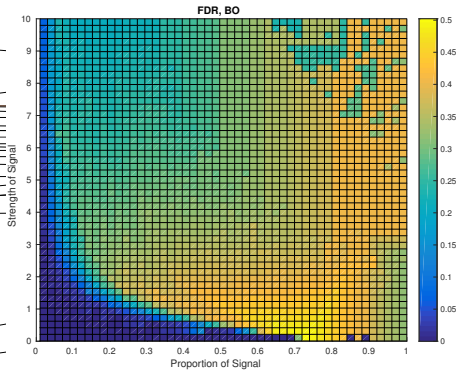
FDR of BO



FDR of BO, m=100

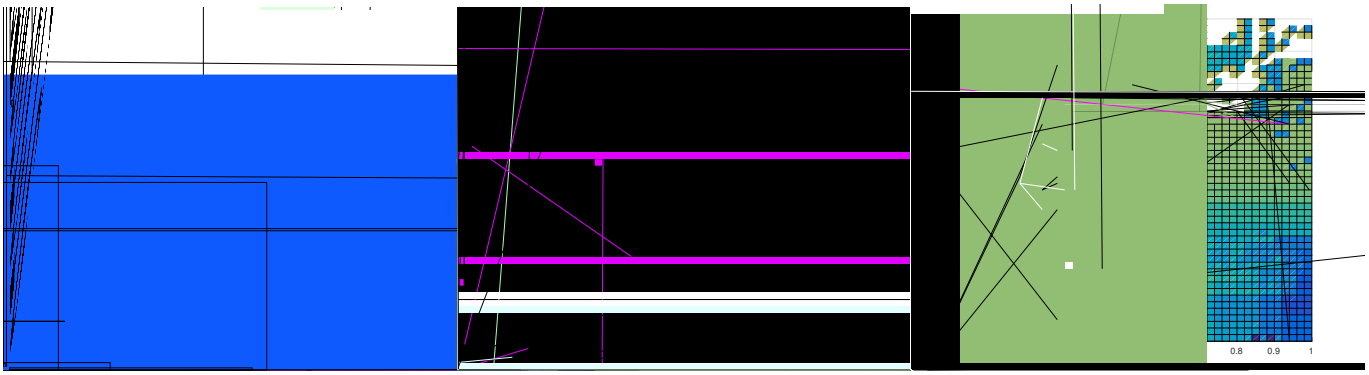


FDR of BO, m=200



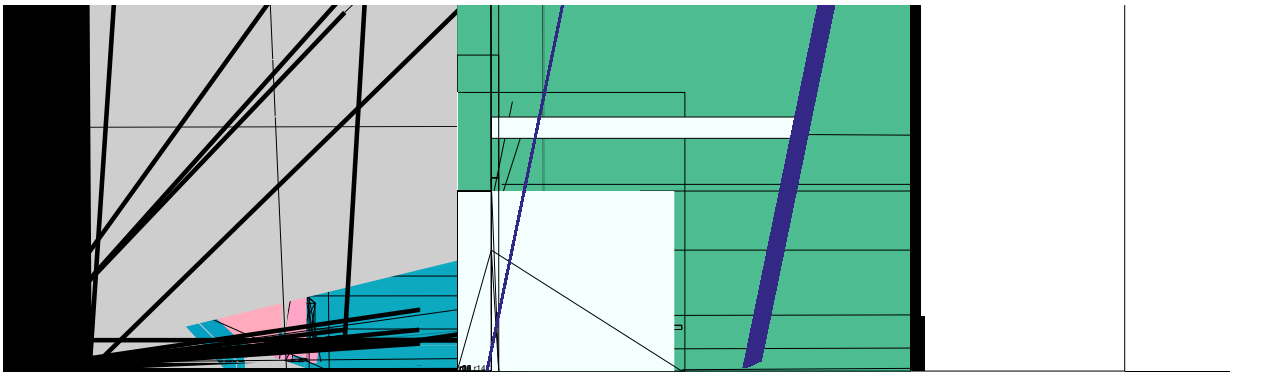
FDR of BO, m=400

FDR of BH



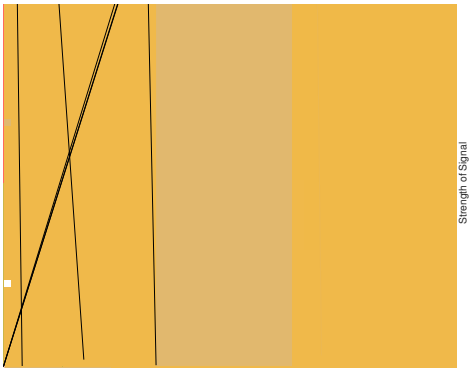
FDR of BH, m=100 FDR of BH, m=200 FDR of BH, m=400

FDR of BH1

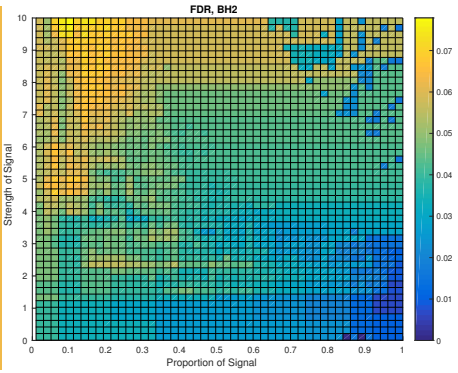


FDR of BH1, m=100 FDR of BH1, m=200 FDR of BH1, m=400

FDR of BH2



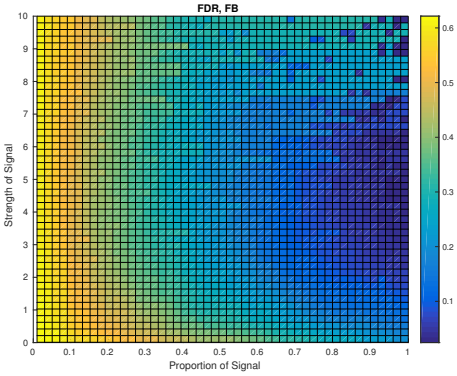
FDR of BH2, m=100



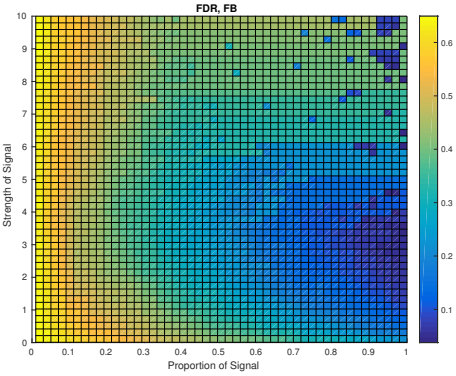
FDR of BH2, m=200

FDR of BH2, m=400

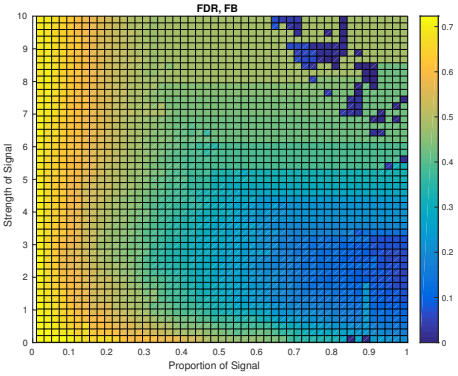
FDR of FB



FDR of FB, m=100

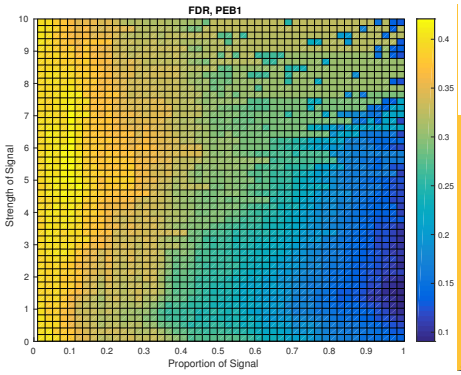


FDR of FB, m=200

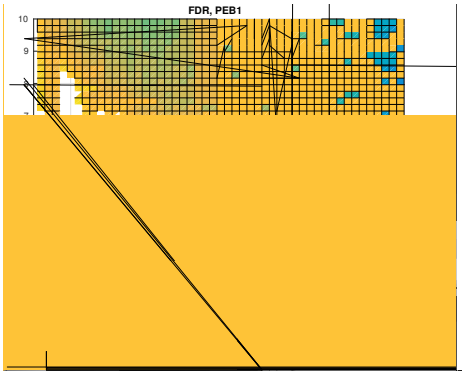


FDR of FB, m=400

FDR of PEB1



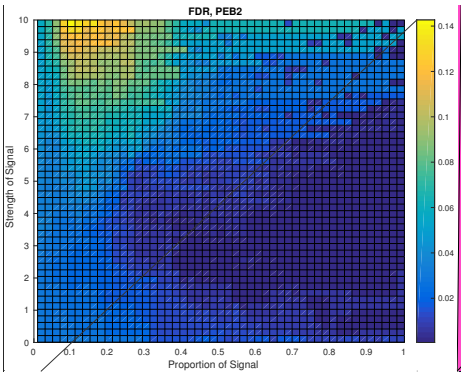
FDR of PEB1, m=100



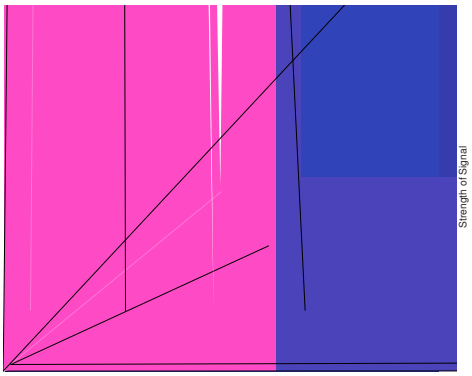
FDR of PEB1, m=200

FDR of PEB1, m=400

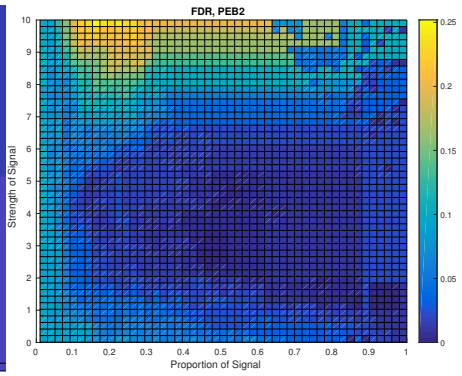
FDR of PEB2



FDR of PEB2, m=100



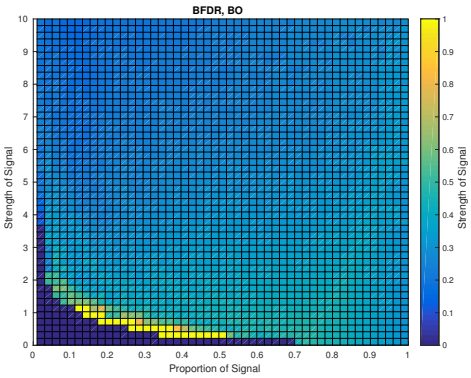
FDR of PEB2, m=200



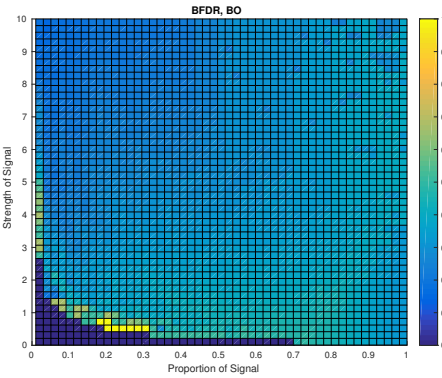
FDR of PEB2, m=400

BFDR

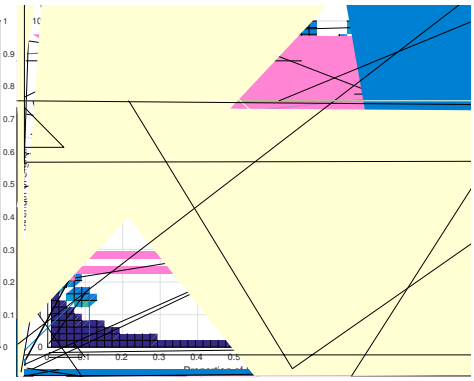
BFDR of BO



BFDR of BO, m=100



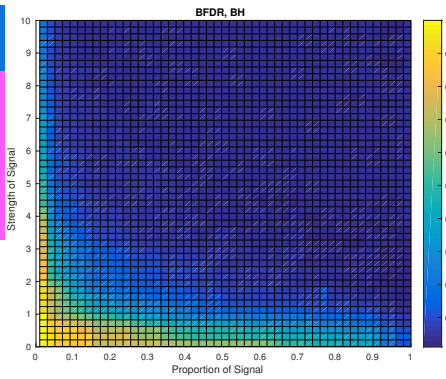
BFDR of BO, m=200



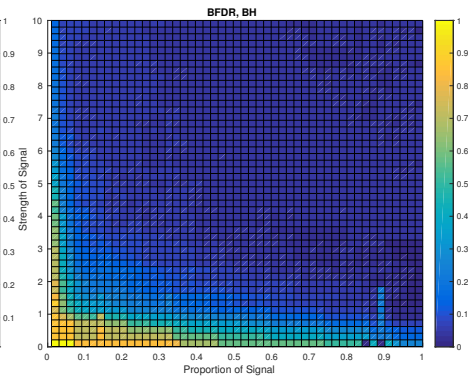
BFDR of BO, m=400

BFDR of BH

BFDR of BH, m=100

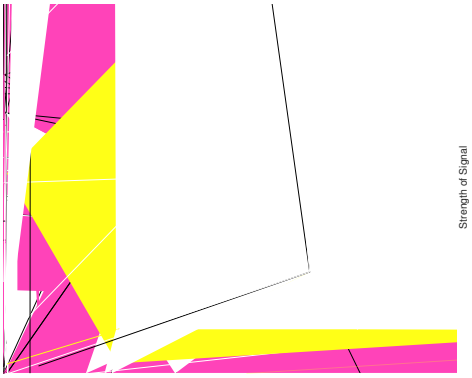


BFDR of BH, m=200

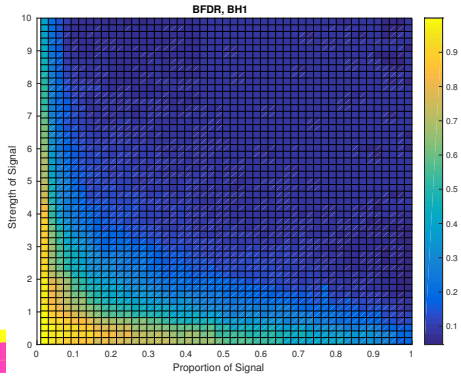


BFDR of BH, m=400

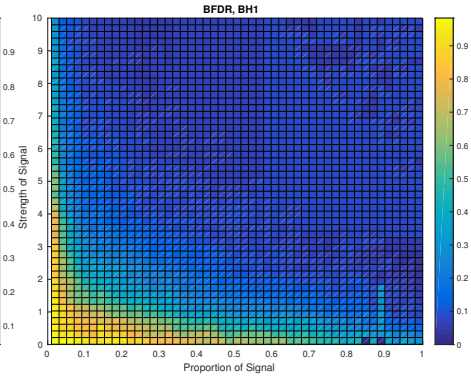
BFDR of BH1



BFDR of BH1, m=100

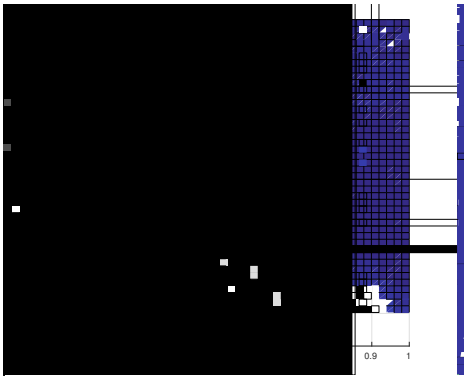


BFDR of BH1, m=200

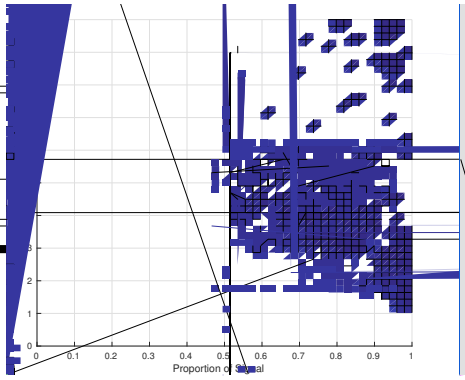


BFDR of BH1, m=400

BFDR of BH2



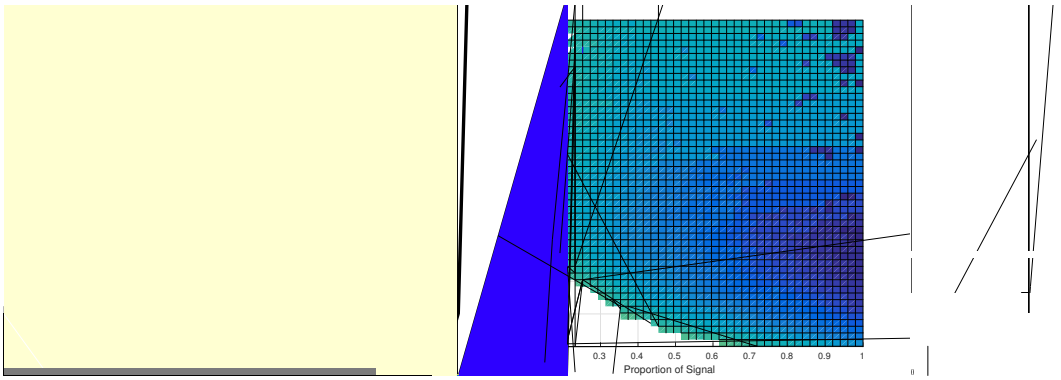
BFDR of BH2, m=100



BFDR of BH2, m=200

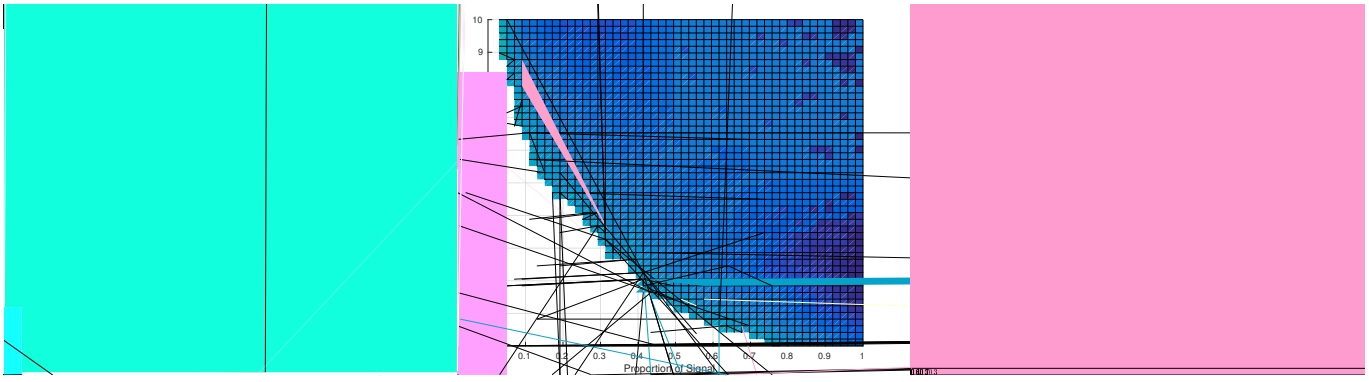
BFDR of BH2, m=400

BFDR of FB



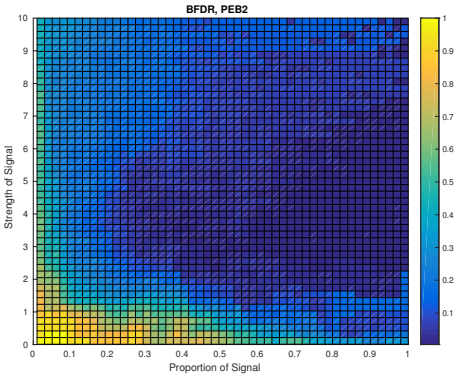
BFDR of FB, m=100 BFDR of FB, m=200 BFDR of FB, m=400

BFDR of PEB1

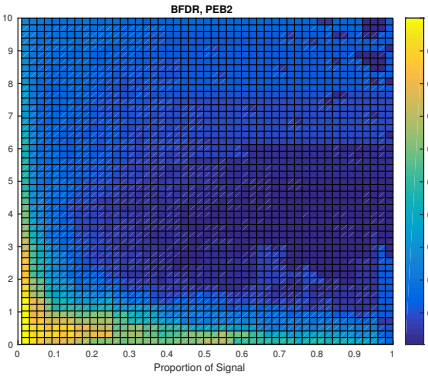


BFDR of PEB1, m=100 BFDR of PEB1, m=200 BFDR of PEB1, m=400

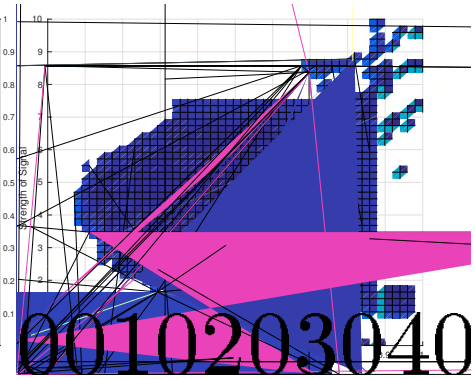
BFDR of PEB2



BFDR of PEB2, m=100

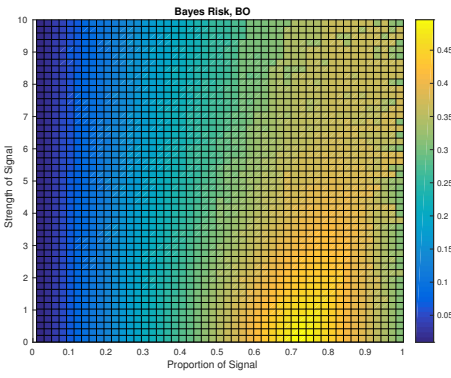


BFDR of PEB2, m=200

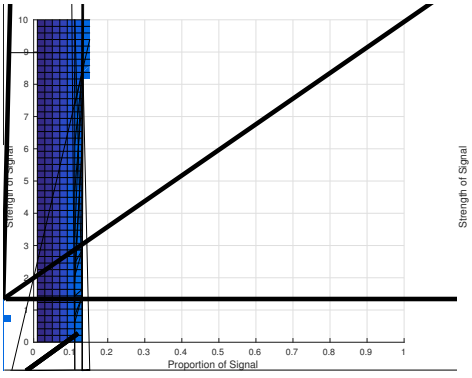


BFDR of PEB2, m=400

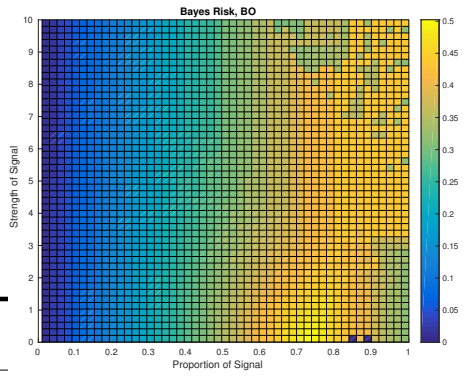
BR



BR of BO, m=100

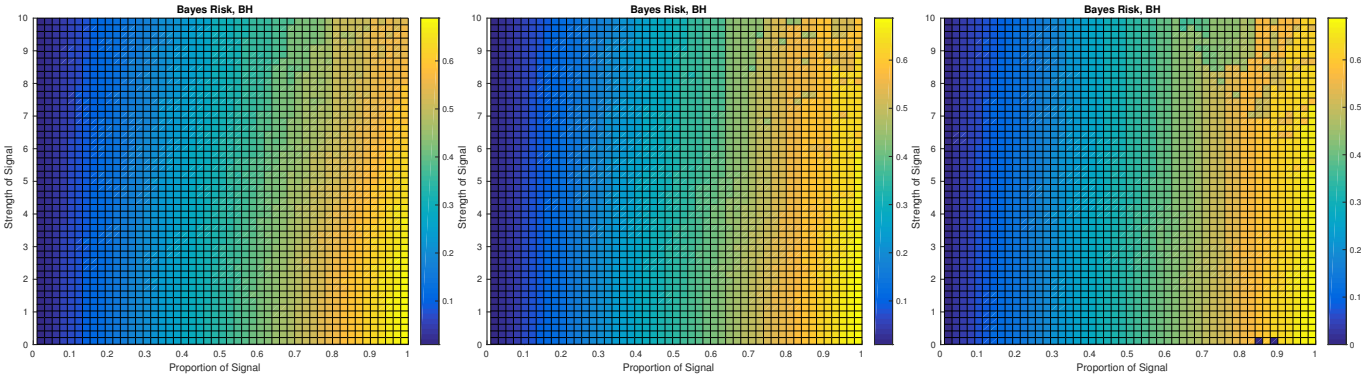


BR of BO, m=200



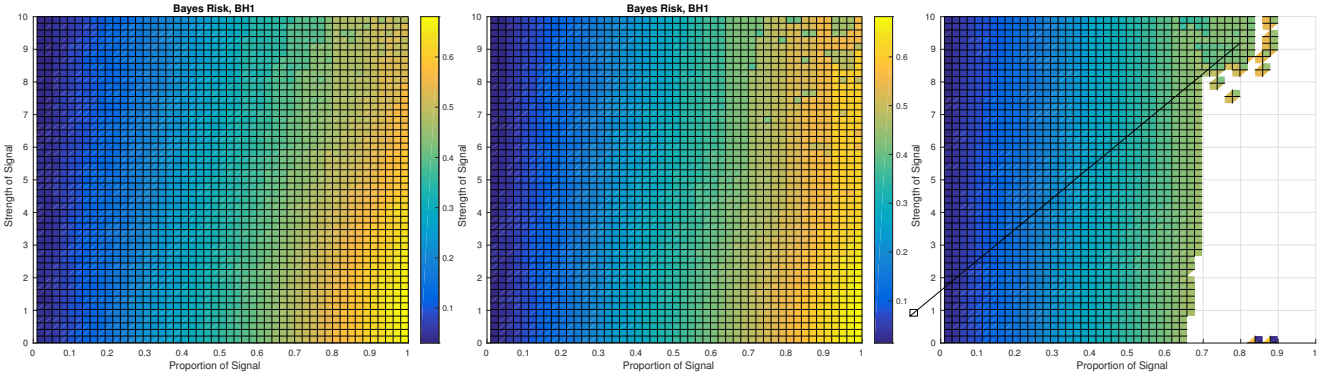
BR of BO, m=400

BR of BH



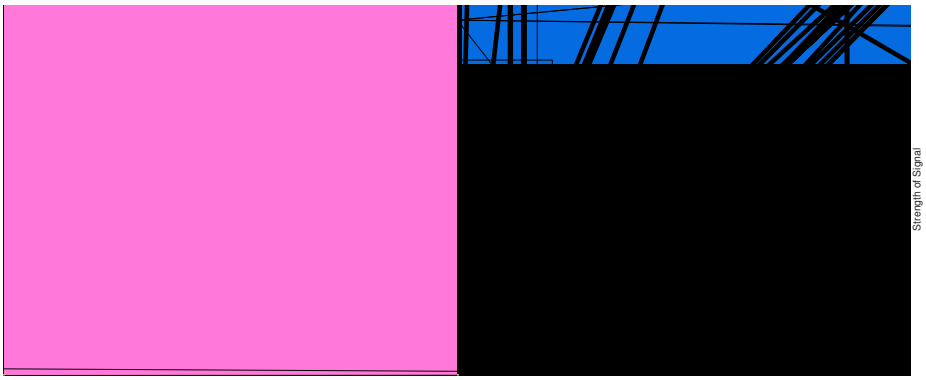
BR of BH, m=100 BR of BH, m=200 BR of BH, m=400

BR of BH1



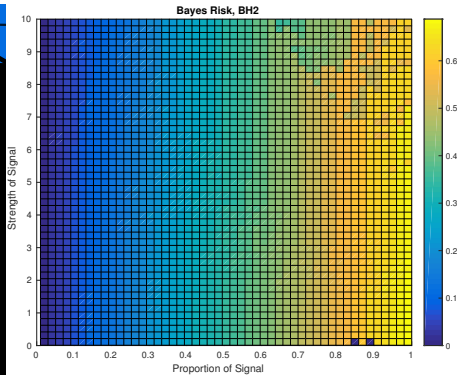
BR of BH1, m=100 BR of BH1, m=200 BR of BH1, m=400

BR of BH2



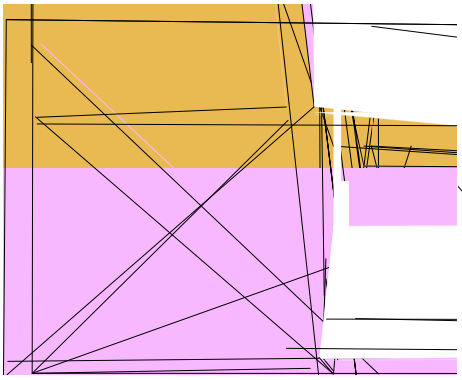
BR of BH2, m=100

BR of BH2, m=200

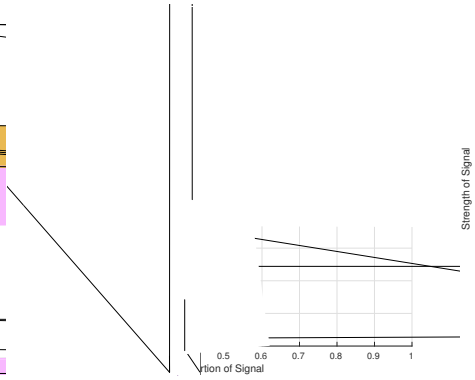


BR of BH2, m=400

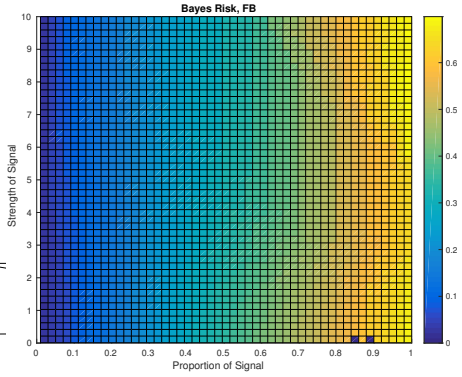
BR of FB



BR of FB, m=100

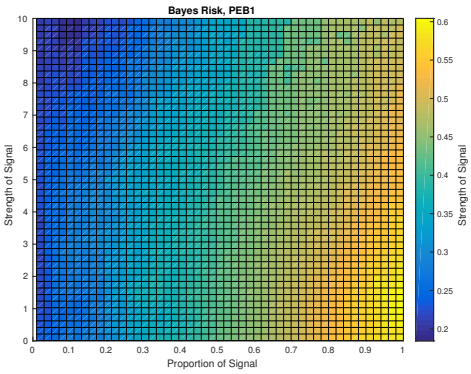


BR of FB, m=200

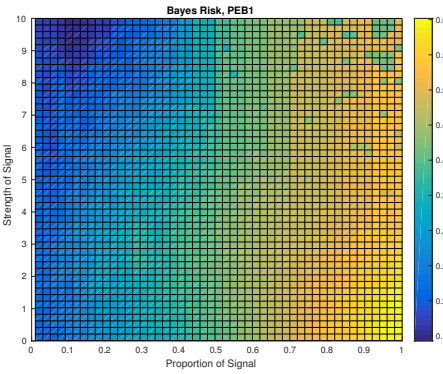


BR of FB, m=400

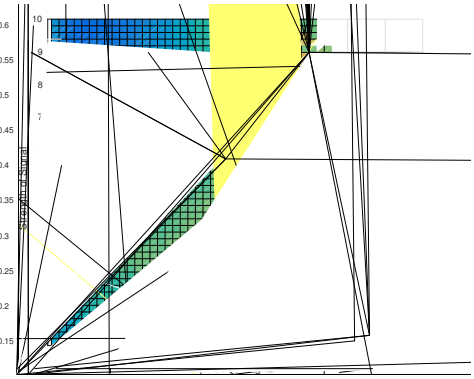
BR of PEB1



BR of PEB1, m=100

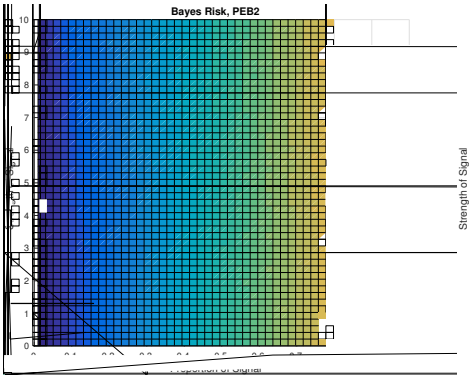


BR of PEB1, m=200

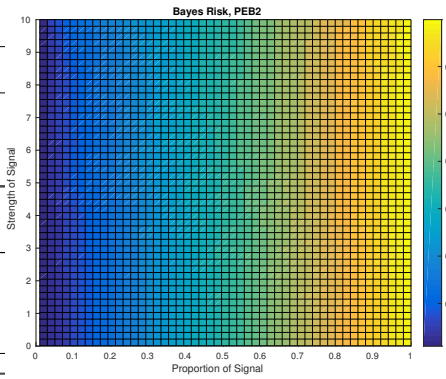


BR of PEB1, m=400

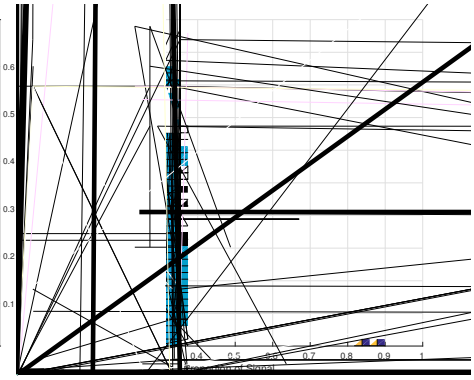
BR of PEB2



BR of PEB2, m=100

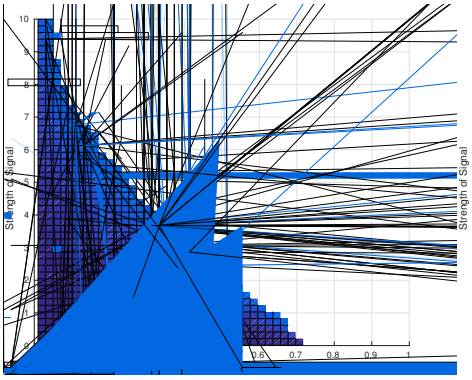


BR of PEB2, m=200

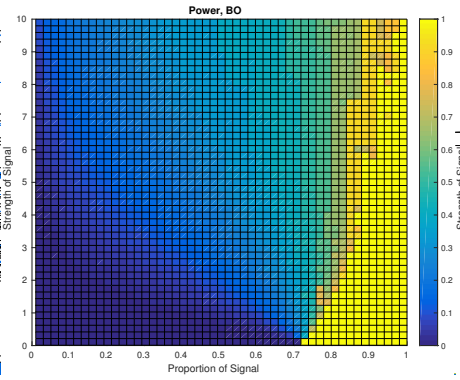


BR of PEB2, m=400

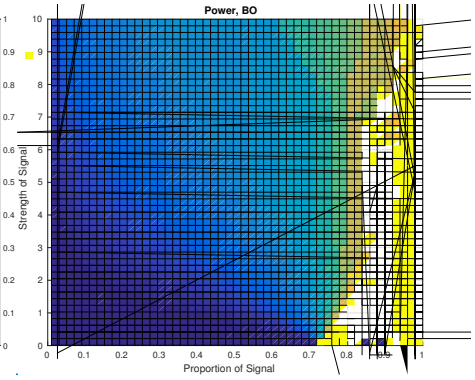
Power



Power of BO, m=100

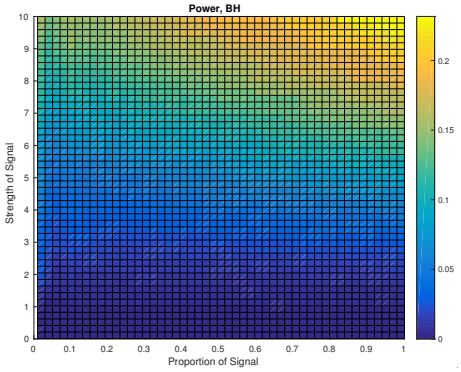


Power of BO, m=200



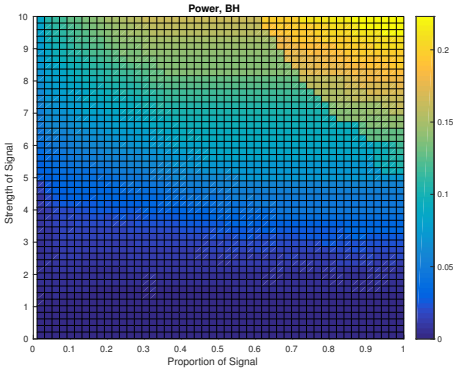
Power of BO, m=400

Power of BH



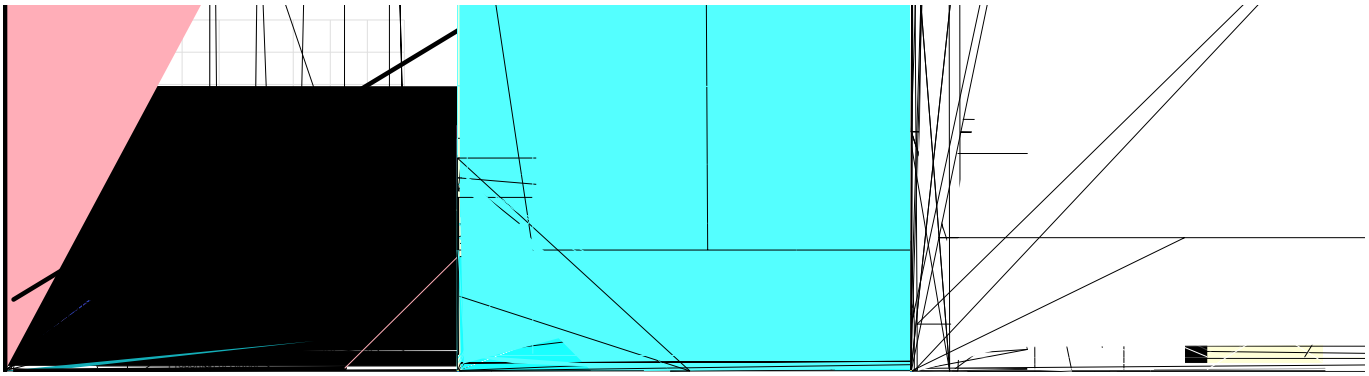
Power of BH, m=100

Power of BH, m=200



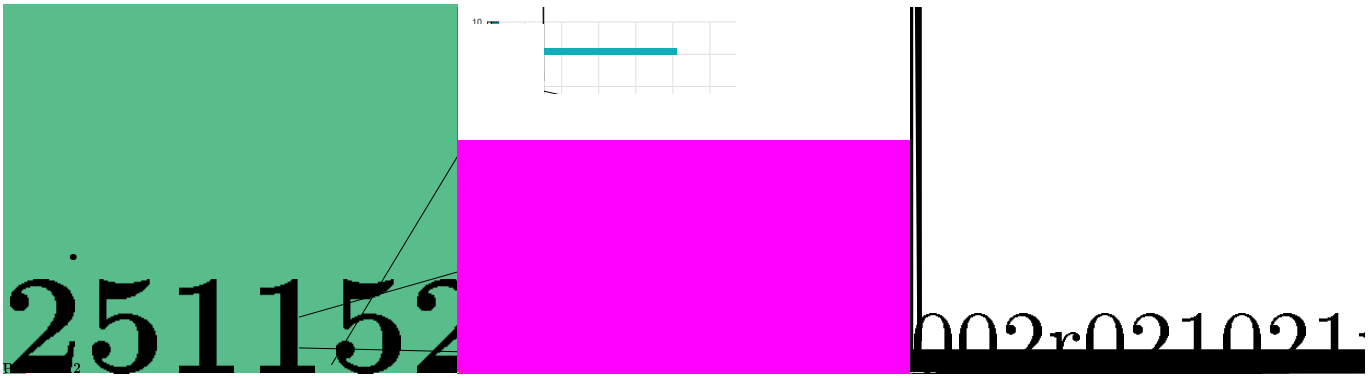
Power of BH, m=400

Power of BH1



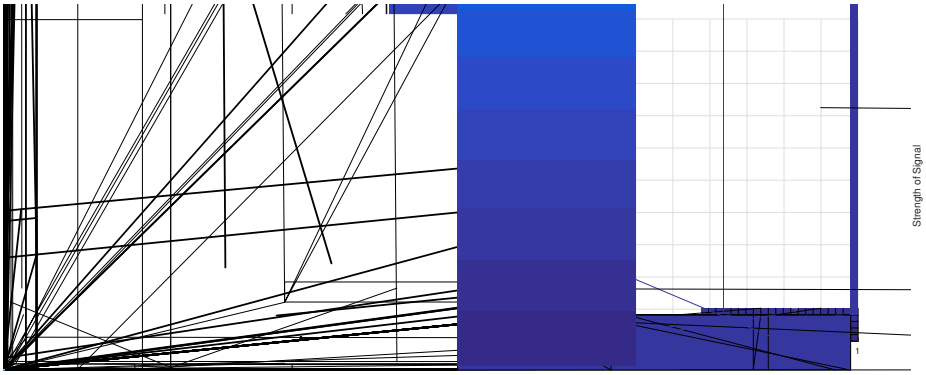
Power of BH1, m=100 Power of BH1, m=200 Power of BH1, m=400

Power of BH2



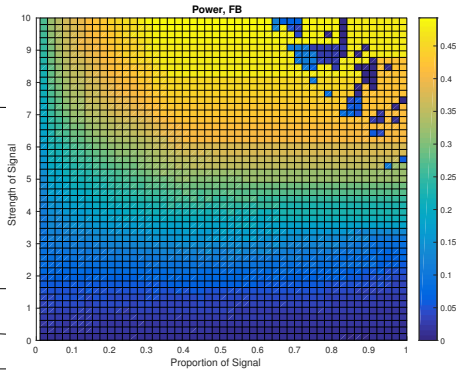
Power of BH2, m=100 Power of BH2, m=200 Power of BH2, m=400

Power of FB



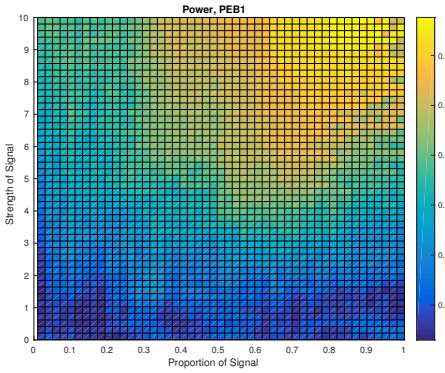
Power of FB, m=100

Power of FB, m=200

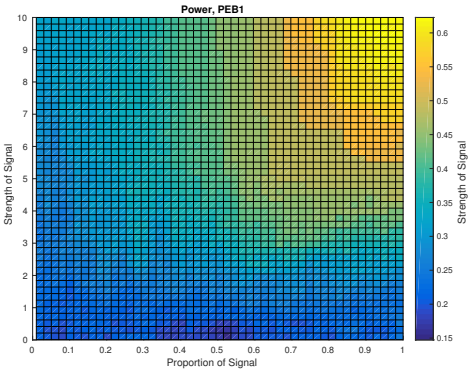


Power of FB, m=400

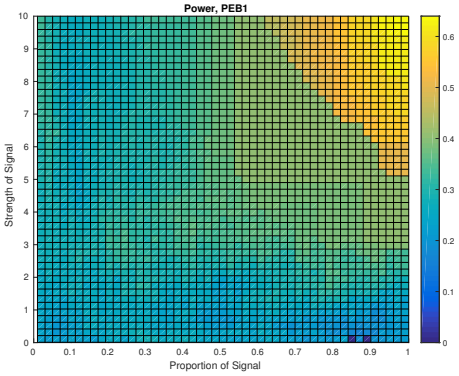
Power of PEB1



Power of PEB1, m=100

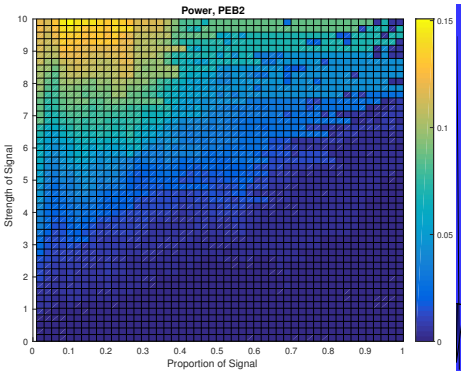


Power of PEB1, m=200

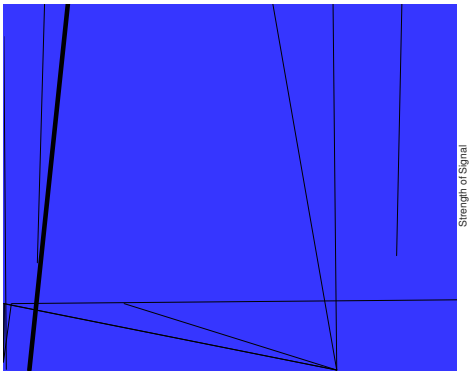


Power of PEB1, m=400

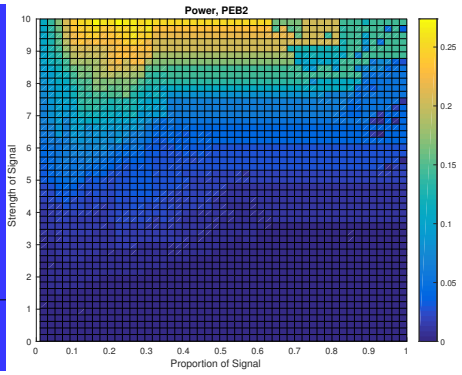
Power of PEB2



Power of PEB2, m=100



Power of PEB2, m=200



Power of PEB2, m=400