8-23-2013

Multiple Testing under Dependence with Approximate Posterior Likelihood and Related Topics

Sairam D. Rayaprolu
University of Connecticut, sairam.rayaprolu@uconn.edu

Multiple Testing Under Dependence with Approximate Posterior Likelihood and Related Topics

Sairam D. Rayaprolu, Ph.D.
University of Connecticut, 2013

ABSTRACT

This dissertation studies large-scale multiple testing which plays an important role in many areas of modern science and technology, such as biomedical imaging and genomic data processing. It has long been recognized that statistical dependence in data poses a significant challenge to large-scale multiple testing. Failure to take the dependence into account can result in severe drop in performance of multiple testing. In particular, the detection power of large-scale multiple tests is known to suffer under dependence when the False Discovery Proportion must be controlled. However, it often happens that the dependence structure is unknown and only a single, albeit very high-dimensional, observation of test statistic is available. This makes large scale multiple testing under dependence considerably harder. This situation can be likened to a signal processing problem with the truth/falsehood of a hypothesis playing the role of an unobservable binary signal and hypothesis-testing becomes analogous to signal detection. To complete the analogy, the signals have an unknown statistical dependence and the test-statistics are the dependent noise-corrupted observations. The typical total number of simultaneous hypotheses in this work can be between a thousand and a million. The target application context is that of large scale 'preliminary sieving', using noisy observations, with the goal of reducing the scale of the problem for further examination. Likewise, the detection of extremely sparse signals lies outside the scope of this work.

Sairam D. Rayaprolu — University of Connecticut, [2013]

Our work addresses this problem for the case of a stationary, ergodic signal vector with low signal-strength, a known noise distribution and a known signal-noise interaction-function. This case has many potential applications as signals embedded in data can often be characterized as a stationary ergodic process with an unknown distribution, while the distribution of the noise that distorts the signals can be accurately inferred beforehand under controlled experiments. Our main contribution in this setting is a new approach for improved recovery of a long sequence of dependent binary signals embedded in noisy observations. The novel aspect of our approach is the approximation and numerical computation of the posterior probabilities of binary signals at individual sites of the process, by drawing strength from observations at nearby sites without assuming the availability of their joint prior distribution. Although we only consider signal vectors registered as a time series, the approach in principle may apply to random fields as well.

A problem closely related to multiple testing under arbitrary dependence is the simulation of random transition matrices. This problem is motivated by the need for 'random' Markov chains in the study on multiple testing. Random transition matrices can also be used to simulate random contingency tables, models of real-world networks, and other high-dimensional data with versatile dependence structures. For example, simulating random stochastic matrices with a specified principal eigenvector or a specified spectral gap facilitates the simulation of markov chains with specified stationary distribution or mixing-time respectively. The exact-simulation problem is known to be hard and consequently simple recipes for the exact simulation of such random matrices, even from a uniform distribution, are unavailable. We use known results to suggest simple heuristics to simulate, from an unknown distribution, stochastic matrices that have a prescribed principal left eigenvector and/or approximately a prescribed spectral gap.

Sairam D. Rayaprolu — University of Connecticut, [2013]

The unifying theme that pervades this dissertation is that of unknown dependence structure in high-dimensional data.

Multiple Testing Under Dependence with Approximate Posterior Likelihood and Related Topics

Sairam D. Rayaprolu

B.Tech., Chemical Engineering, Indian Institute of Technology, Bombay, India, 1996

M.S., Systems Engineering, University of Arizona, AZ, USA, 1998

M.S., Applied Mathematics, University of Arizona, AZ, USA, 2008

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2013

APPROVAL PAGE

Doctor of Philosophy Dissertation

Multiple Testing Under Dependence with Approximate Posterior Likelihood and Related Topics

Presented by

Sairam D. Rayaprolu, B.Tech. Chemical Engineering, M.S. Systems Engineering, M.S. Applied Mathematics

Major Advisor Zhiyi Chi

Associate Advisor Joseph Glaz

Associate Advisor Vladimir Pozdnyakov

University of Connecticut

2013

Acknowledgments

I thank my advisor Professor Zhiyi Chi without whose vision, ideas, insight, patient guidance and inspiration, this dissertation would not have been possible. I thank the members of my dissertation committee Professor Vladimir Pozdnyakov, Professor Joseph Glaz and Professor Jun Yan for their time, input and interest. I thank Professor Dipak Dey for all his help, encouragement and guidance from the beginning through completion of my PhD study in Statistics. I am extremely grateful to Professor Rabindra Bhattacharya for his kindness, wisdom and encouragement. I thank all the faculty members of the Statistics department for their guidance and leadership. I thank my fellow graduate students, too many to name here, who helped me on numerous occasions.

Contents

List of Tables

Chapter 1

Introduction

Simultaneous testing of multiple hypotheses, simply referred to as multiple testing, is a common and important inference problem in many statistical investigations. In large datasets, traditional single inference approaches are well known to fail. Modern multiple testing addresses itself to the general statistical problem of making a small number of false discoveries [incorrect rejections] by controlling a suitable error rate and at the same time maximizing a suitable measure of power [for e.g. number of true discoveries i.e. correct rejections]. The number of hypotheses can range from thousands to millions depending on the application. Informally, the goal is to (i) reject as many false nulls as possible and at the same time (ii) keep the proportion of true nulls in the set of rejections low.

Statistical research on this problem has gained increased relevance because many important application areas involve simultaneous testing of a large number of hypotheses. Such applications include genomics, bioinformatics, signal-processing, brain-imaging, pharmacology, epidemiology, general medicine, psychometrics and marketing. Moreover, multiple testing can be used as an effective tool in statistical procedures such as decision trees, variable selection etc. Farcomeni [17] provides a broad overview of research in modern multiple hypothesis testing and its applications. Multiple testing is often also referred to with terms such as *Multiplicity Adjustment*, *Multiplicity Control*, *Multiple Comparisons*, *Type I Error Control* and *False Discovery Control*. The *Discovery* in the latter term refers to the correct detection of a False Null, which in a scientific context, is usually interesting.

An arbitrary multiple testing problem is represented as a contingency table in 1. The table shows the parameters and the random variables in a multiple testing problem. The Type I

error measure that has received the most attention is the expectation of the unobservable random variable *False Discovery Proportion*[FDP]. It is referred to as the *False Discovery Rate* [FDR]. The FDR concept was introduced by Benjamini and Hochberg [5] to provide a far *less stringent* or *less conservative* error measure than the Family Wise Error Rate (FWER). It is important to note that the nature of the application is important in determining the Type I error measure. If the experimenter wants to control the FWER, i.e. the probability of incorrectly rejecting at least one true null, then controlling FDR is clearly not appropriate. The philosophy of controlling FDR in a multiple test relies on the assumption that the experimenter is willing to control only the *proportion* of rejections that are incorrect and not the very possibility of even a single incorrect rejection. FDR control has been found to be relevant in several applications where simultaneous testing of a large number of hypotheses has to be carried out, for example in DNA microarray experiments as Dudoit et al. [13] explain. The large number of hypotheses means that the goal of not making a single false discovery is almost impossible to achieve if a reasonable number of true discoveries must be made simultaneously.

Hypothesis	*Accept null*	*Reject null*	*Total*
Null True	U	V [false discoveries]	m_0 [unknown parameter]
Alternative True	T [missed discoveries]	S	$m - m_0 = \sum \eta_j$
	$m - R$	$R = \sum d_j$	m [known parameter]

Table 1: **Contingency table for an arbitrary multiple test**
m: total number of hypotheses being tested
m_0: count of true nulls, usually unknown
R: total number of rejections, i.e.*discoveries*
V: count of incorrectly rejected true nulls, i.e. *false discoveries*
U: count of correctly accepted true nulls, i.e. *true non-discoveries*
T: count of incorrectly accepted false nulls, i.e. *false non-discoveries*
S: count of correctly rejected false nulls, i.e. *true discoveries*
d_j: binary decision variable on falsehood of jth null
η_j: binary random variable denoting falsehood of jth null

FDP is the fraction of false rejections in the total number of rejections. It is set to 0 if there are no rejections. The FDP, the FDR and the FWER are formally defined in Chapter 2. Benjamini and Hochberg [5] made the crucial observation that FDR is a type I error measure and that it is the expectation of a random variable that is by definition smaller than the random variable whose expectation is the FWER. Please see

Appendix I for a proof of the inequality. Consequently they proposed that bounding this smaller quantity [FDP] is less conservative that controlling FWER. To be sure, FDR control is not FDP control, it is the control of the average FDP. FDR control is FDP control via its expectation. As Roquain (2011) and others note, controlling the FDR is meaningful only when FDP concentrates well around the FDR. However, the FDR control remains useful, is a simpler criterion and is much more developed at this point. The mathematical expectation here is computed under the unobservable *true configuration* of the hypotheses. However, even if the true configuration is unobservable, Benjamini and Hochberg [5] showed that the FDR of the Benjamini-Hochberg procedure [henceforth referred to as the *BH procedure*] can be bounded above subject to independence or a suitable dependence assumption on the test statistics i.e. the hypotheses. Storey et al. [39] proved the same result using a simpler martingale argument and introduced a new conservative estimator of the FDR.

The BH procedure revitalized research in multiple testing and has inspired other procedures that build on it. The BH procedure guarantees FDR control under the assumption of independence of test statistics. Further, it has been observed that, in practice, the BH procedure is robust to dependence when applied to large scale multiple tests as, for example in Clarke and Hall [12]. However when the observations are strongly correlated, the performance of BH procedure and similar p-value based procedures is known to suffer. Therefore strongly correlated data poses serious challenges to the BH procedure and other similar procedures. The assumption of independence is incorporated into the statistical model so that it can be exploited to prove the FDR controlling property of the BH procedure. This property of the BH procedure was later extended by Benjamini and Yekutieli [6] to *positively dependent* test statistics. The positive dependence is precisely defined in Benjamini and Yekutieli [6]. Several other results and procedures for situations involving some types of dependence have also been put forth. Sarkar [31] and Sarkar [32] also discuss a few such methods for the positive dependence case. Sarkar [32] studies both FDR-control and power of methods under dependence. He notes that strong dependence adversely affects either FDR-control or power. In practice however, the assumption of

independence can rarely be verified. In fact it can be argued that most data on which multiple test procedures give their decisions have some unknown dependence structure. Sun and Cai [40] conclude that:

- The violation of the independence assumption can result in either loss of power [*overly conservative*] or loss of FDR control [*overly liberal*].
- The *validity* i.e. FDR control of multiple test procedures has been overemphasized at the expense of *efficiency* i.e. power.
- A statistical model of the dependence structure can be exploited to improve a multiple test procedure's *optimality* i.e. its efficiency as compared to the maximum attainable efficiency by any procedure working with the same data.

Their conclusions above serve as the background for our work. We use the signal-processing framework that was introduced in Chi [11]. Following Chi [11], we model the falsehood/truth of a null is the presence/absence of a signal. In other words the test statistic is assumed to be a noise corrupted binary variable. The binary variables are unobservable, stationary-ergodic and dependent. The signal-noise interaction function and the [continuous] noise density can be arbitrary but are assumed known. This provides a very general framework for statistical modeling of dependent observations. The dependence structure i.e. the prior joint distribution among the binary random variables remains unknown. This is circumvented in two steps. First, the availability of a single controlled (high dimensional) observation of a signal/null sequence is assumed. This provides empirical prior conditional moments that are used as a proxy for a prior. Second, although the posterior cannot be exactly computed, it can be approximated when signal strength is low, precisely when efficiency of procedures suffers the most. In Chapter 3, we present this two stage mathematical model and the Taylor series approximation of the posterior probabilities. In Chapter 4, we present the results of our numerical simulations and computations for the observed FDP and detection count when using the approximated posteriors on a Bayesian procedure with proven power-optimal properties under a

constraint on FDR. The latter posteriors borrow strength from their nearby sites. These are then compared to the corresponding FDPs and detection counts a) exact posteriors computed without taking dependence into consideration or b)p-values in conjunction with a step-up method.

The definition of power in multiple testing is not unique. The choice of the definition must involve the judgement of decision maker. One natural measure of the power of a multiple testing procedure is the expected number of true discoveries. Another measure of power involves the FNR or the false non-discovery rate. It is a notion dual to FDR and is defined as the expected proportion of signals among the accepted nulls. This Type II error measure penalizes the false non-discoveries only as a proportion of all acceptances. These quantities are formally defined in chapter 2. We use the expected number of true discoveries as our measure of power.

Chapter 2

Background and Related Work

2.1 FDR Control

The FDP is the number of false rejections divided by the total number of rejections. As mentioned in the introduction, FDR control is not FDP control, it is the control of the average FDP. FDR control is FDP control via its expectation. As Roquain (2011) and others note, controlling the FDR is meaningful only when FDP concentrates well around the FDR. However, the FDR control remains useful, is a simpler criterion and is much more developed at this point. Moreover both literature and simulation indicate that the random variable FDP, is well behaved in many large scale multiple tests if α bounded below by a number which depends on the context. The FDP, the FDR and the FWER are formally defined below.

$$
\begin{aligned}
R &= \text{total number of rejections, i.e. } \textit{discoveries} \\
V &= \text{count of incorrectly rejected true nulls, i.e. } \textit{false discoveries} \\
\text{FDP} &= \frac{V}{R \vee 1} \qquad \text{FDR} = E\left(\text{FDP}\right) \\
\text{FWER} &= P\left(V > 0\right) \qquad \text{FDR} \leq \text{FWER}
\end{aligned}
$$

The expectation above is computed under the unobservable *true configuration* of the hypotheses. However, even if the true configuration is unobservable, Benjamini and Hochberg [5] showed that the FDR of the Benjamini-Hochberg procedure can be bounded above subject to independence or a suitable dependence assumption on the test statistics. Storey et al. [39] proved the same result using a simpler martingale argument and

introduced a new conservative estimator of the FDR.

$$H_i : \theta_i = \theta_{0i} \qquad 1 \leq i \leq m \qquad\qquad X_i \sim F_{\theta_i} \qquad X_i \text{ continuous}$$
$$p_i = P_{\theta_{0i}}\left(X_i > T_i\right)$$

The classical frequentist multiple testing framework is shown above. Under the null hypothesis H_i, the p-value p_i is uniformly distributed on $[0, 1]$ because of the probability integral transformation in the definition of the p-value. The BH procedure uses the decision rule below to reject or accept a hypothesis. It assumes that the test statistics corresponding to the true-nulls are independent.

$$\text{reject} \quad \left\{ H_{j:m} \,\middle|\, j \leq \max\{k \,|\, p_{k:m} \leq \frac{k\alpha}{m}\} \right\}$$
$$\implies \qquad \text{FDR} = \frac{m_0\alpha}{m} \leq \alpha \qquad\qquad \text{(Benjamini and Hochberg [5])}$$

where $p_{j:m}$ is the j^{th} ordered p-value and $H_{j:m}$ the corresponding null hypothesis. m_0 is the number o true-nulls. The BH procedure revitalized research in multiple testing and has inspired other procedures that build on it . One of the chief limitations of the BH procedure is its assumption of independence of test statistics. This was later extended by Benjamini and Yekutieli [6] to positively dependent test statistics. Several other results and procedures for situations involving some types of dependence have also been put forth. Sarkar [31] and Sarkar [32] also discuss a few such methods for the positive dependence case. Sarkar [32] studies both FDR-control and power of methods under dependence. He notes that strong dependence makes either FDR-control or power worse.

2.1.1 Other type-1 error rates

The FDX, the k-FWER and the k-FDR, like the FDR, are Type-I error rates that are less restrictive than the FWER. They are defined below.

$$\begin{aligned} \text{FDX}\,(a) &= P\,(\text{FDP} > a) \\ k\text{-FWER} &= P\,(V \geq k) \\ k\text{-FDP} &= \mathbf{1}\,(V \geq k)\,\frac{V}{R \vee 1} \qquad\qquad k\text{-FDR} = E\,(k\text{-FDP}) \end{aligned}$$

Lehmann and Romano [28] argue that the k-FWER may be appropriate, especially when m is large and one can tolerate fewer than k false rejections. This rate is less stringent than FWER ($k = 1$) and still controls the total number of false rejections. Lehmann and Romano [28] also point out that bounding the FDR does not prohibit the FDP from varying and bounding the FDX guarantees that the FDP is bounded with a prescribed probability. The k-FDR was introduced by Sarkar [33] as an analogous generalization of the FDR. Sarkar and Guo [34] develop procedures for k-FDR control for the case of independent p-values. The latter article argues that k-FDR control is less conservative and more powerful than k-FWER control.

2.2 The Validity Efficiency Tradeoff: Power and FDR-control

The power of a multiple testing procedure can be characterized as the expected true discovery count. Another measure of power is the FNR or the false non-discovery rate. It is a notion dual to FDR and is defined as the expected proportion of true alternatives among the hypotheses accepted as nulls. Both the measures are formally defined below.

We use the expected true discovery count as our measure of power.

$$\begin{aligned}
\text{Power} &= E\,(R-V) \qquad [\textit{expected true discovery count}] \\
\text{FNP} &= \frac{T}{(m-R)\vee 1} \\
\text{FNR} &= E\,(\text{FNP}) \qquad [\textit{expected proportion of false nulls in hypotheses accepted as nulls}]
\end{aligned}$$

Applications often involve hypotheses under unknown dependence and as mentioned above, known methods do not necessarily perform well under unknown or strong dependence. At the same time, existing procedures based on marginal-p-values do not exploit what may be known about the dependence structure among the hypotheses. This results in sub-optimal power as Sun and Cai [40] demonstrate. Sun and Cai [40] note that the validity of FDR control has been overemphasized but the issue of efficiency has largely been ignored.

The loss of power may be more acute when conservative procedures that remain valid under dependence are used or when a known dependence structure is ignored. Due to these reasons, much research is needed for developing optimal multiple-testing procedures under general dependence structures. Understanding how the dependence structure affects the performance and optimality of multiple testing procedures is an important aspect of the problem.

One approach to multiple testing, when appropriate, is the Bayesian approach. For example, [16] and Muller et al. [30] develop Bayesian procedures. One of the chief merits of the Bayesian approach is that it can be used without making any specific assumption on the dependence-structure. Sarkar et al. [35] provide a general decision theoretic framework for Bayesian multiple testing with no dependence assumptions. However, Bayesian procedures rely on posterior probabilities and it is not possible to compute the posteriors without knowledge of the joint distribution of the large number test-statistics. Chi [11] investigates this problem by imposing an HMM dependence structure among the hypotheses, which are represented by binary random variables.

2.3 Maximizing power under dependence subject to an upper bound on FDR

Existing literature on multiple testing predominantly addresses independent hypotheses because assuming independence makes analysis more tractable. Generally speaking, the existing methods tend to be more powerful under independence. Almost all well known frequentist methods for multiple-testing only allow special types of dependence such as positive dependence. We keep the dependence structure more flexible by modeling the truth or falsehood of hypotheses as dependent binary signals. Instead of p-values we use posterior-probabilities which, as is shown in Sarkar et al. [35] , take dependence between hypotheses into account.

The Bayesian model of hypothesis testing starts with a prior probability distribution on the sequence of dependent hypotheses. The support is the high-dimensional set of binary strings in the discrete product space $\{0,1\}^m$ whose cardinality is 2^m. The typical value of m here is anywhere between thousands to millions. Sarkar et al. [35] formulate the Bayesian approach as a decision problem with the FDP as a loss function. The frequentist FDR is replaced by *posterior-FDR* and power by *posterior power*. In the Bayesian setting, the test statistic can be considered a vector $\boldsymbol{X} = (X_1, ..., X_m) \sim P(\boldsymbol{X}|\boldsymbol{\eta})$ where the parameter $\boldsymbol{\eta} = (\eta_1, ...\eta_m)$ with $\eta_j = \mathbb{1}\left(H_j \text{ false}\right)$. The binary random variable η_j represents the falsehood of hypothesis H_j. Let $\boldsymbol{d}\left(\boldsymbol{X}\right) = (d_1(\boldsymbol{X}), d_2(\boldsymbol{X}), \ldots d_n(\boldsymbol{X}))$ be the decision rule for the observation $\boldsymbol{X}$, i.e. $d_j(\boldsymbol{X}) = \mathbb{1}(H_j \text{ rejected})$.

The power-optimal procedure that maximizes the posterior mean of the number of correct rejections subject to an upper bound on the posterior FDR can be expressed as a constrained optimization problem. Since we have not made any assumptions about the dependence structure the optimization problem below is applicable to an arbitrary multiple testing problem. In theory, each decision,i.e. each acceptance/rejection, can depend on all the observations in the test vector. The mathematical formulation of the constrained optimization problem is shown below. The derivation of the solution follows

Chi [11].

Maximize the posterior expected number of correct rejections

$$\max \qquad E\left[\, S \,\middle|\, \boldsymbol{X} \,\right]$$

subject to the constraint that the posterior FDR does not exceed α

$$\text{s.t.} \qquad E\left[\, \frac{V}{R} \,\middle|\, \boldsymbol{X} \,\right] \leq \alpha$$

The above optimization problem can be expressed explicitly in terms of the individual 0-1 decisions for the hypotheses, namely the components of $\boldsymbol{d}$, as shown below.

$$\max \qquad E\left[\, \sum_{j=1}^{m} \eta_j \, d_j\left(\boldsymbol{X}\right) \,\middle|\, \boldsymbol{X} \,\right]$$

$$\text{s.t.} \qquad E\left[\, \frac{\sum_{j=1}^{m} \left(1-\eta_j\right) \, d_j\left(\boldsymbol{X}\right)}{\sum_{j=1}^{m} d_j\left(\boldsymbol{X}\right)} \,\middle|\, \boldsymbol{X} \,\right] \qquad \leq \alpha$$

Using the linearity of expectation and the fact that $\boldsymbol{d}$ is a function only of $\boldsymbol{X}$,we again rewrite the problem in terms of the posterior probabilities of the alternative hypotheses and the binary decision variables d_j.

$$\max \qquad \sum_{j=1}^{m} P\left(\eta_j = 1 \,|\, \boldsymbol{X}\right) d_j\left(\boldsymbol{X}\right) \tag{2.1}$$

$$\tag{2.2}$$

$$\text{s.t.} \qquad 1 \; - \; \sum_{j=1}^{m} P\left(\eta_j = 1 | \boldsymbol{X}\right) \frac{d_j\left(\boldsymbol{X}\right)}{\sum_{j=1}^{m} d_j\left(\boldsymbol{X}\right)} \qquad \leq \alpha \tag{2.3}$$

This last problem has a convenient monotonicity property. Let $q_k = P(\eta_k = 1|\boldsymbol{X})$. Sort $q_1, q_2, \ldots q_m$ in increasing order $q_{1:m} \leq q_{2:m} \leq \ldots \leq q_{m:m}$ and rearrange the corresponding nulls as $H_{1:m}, H_{2:m}, \ldots H_{m:m}$. Then, the optimization problem is solved by the steps: (i) start with the largest order statistic $q_{m:m}$ and reject $H_{m:m}$, (ii) proceed in decreasing order of $q_{m:m}, q_{(m-1):m}, \ldots$ and continue rejecting the corresponding nulls, iii) stop when

the constraint (2.3) is violated. (iv) all remaining nulls are accepted. The algorithm terminates when the average posterior probability of the alternatives becomes smaller than $1-\alpha$. In other words, as Sarkar et al. [35] observe, the *greedy algorithm* shown below gives the optimal decision procedure for this problem.

$$\text{reject} \quad \left\{ H_{m-r+1:m} \,\middle|\, r \leq \max\left\{k \mid \frac{\sum_{j=1}^{k} q_{m-j+1:m}}{k} \geq 1-\alpha \right\} \right\} \tag{2.4}$$

Where $H_{k:m}$ corresponds to the order statistic $q_{k:m}$. Sarkar et al. [35] note that this last rule can equivalently be stated as, *reject every family of null hypotheses with average posterior probability, of the nulls, less than* α. The importance of this result lies in the fact that its conclusion remains true for an *arbitrary dependence structure* on the hypotheses. Guindani et al. [21] use a similar Bayesian procedure that maximizes the posterior expected number of true positives with a prescribed upper bound on the posterior expected number of false positives instead of the posterior FDR.

A simple observation about the procedure (2.4) is made below.

Theorem. *The power-optimal policy* (2.4) *will not declare any signals, i.e. not reject any hypotheses, if the maximum posterior probability of a signal is smaller than* $1-\alpha$. *Or equivalently, the power-optimal policy derived above will not declare any signals if the the minimum posterior probability of a null is larger than* α.

Proof. This is a simple consequence of the fact that the average of each subset of a finite set of numbers is bounded above (resp. below) by the maximum (resp. minimum) of the set. Consequently, the rejection criterion is not attained at the very first iteration of the procedure i.e. $R = 0$, $FDP = 0$ but all measures of power are 0 as no signals are detected. □

2.4 Dependence Structure and Multiple Testing

An immediate and important observation about the procedure (2.4) is that its only input is the vector of posterior probabilities $E(\boldsymbol{\eta}|\boldsymbol{X})$. In particular, the dependence structure of $\boldsymbol{X}$ can influence the multiple-test decision vector, i.e. the acceptance/rejection vector, only via the posterior probabilities given the observation vector. In other words, as far as this multiple test procedure is concerned, the vector of posterior probabilities contains all the information available on the the dependence structure. This divides the multiple testing problem into two separate steps, the computation of the posterior probabilities followed by the determination of the decision rule using the multiple test procedure. These observations are summarized below.

1. The dependence structure and the test statistic $\boldsymbol{X}$ are required only for the computation of the posterior probabilities.

2. All dependence structures of $\boldsymbol{X}$ resulting in the same posterior probability vector will result in the same decision-vectors.

3. This approach to multiple testing is applicable to any network topology and dependence structure of $\boldsymbol{X}$ that allows the computation or approximation of the posterior probabilities.

In particular, the network topology need not be a chain although the focus of the analytical and numerical work of the subsequent chapters is linearly ordered observations. Other scenarios include observations arranged as, for example, a stationary random field on a regular graph. Under reasonable conditions, regularity of the graph allows the empirical estimation of conditional moments necessary for approximation of posteriors. Motivated by fMRI data, multiple testing has been used in the context of spatial signals in Benjamini and Heller [4]. Siegmund et al. [36] uses scan statistics in conjunction with a modified FDR to address locally dependent signals and points to applications in fMRI data as well as to genomics. Glaz et al. [19] gives a False Discovery approach to scan potential clusters in a random field.

Chapter 3

Posterior Approximation In a Signal-Noise Setting

As pointed out in the last chapter, the key input to the Bayesian multiple test procedure is the vector of posterior probabilities, $E(\boldsymbol{\eta}|\boldsymbol{X})$, of signals given the observations. By Bayes theorem, the posterior probability of a signal, the prior probability of a signal and the joint conditional density of observations given a signal satisfy:

$$E(\eta_t|\boldsymbol{X}) = P(\eta_t = 1|\boldsymbol{X}) \propto \rho(\boldsymbol{X}|\eta_t = 1)P(\eta_t = 1) \tag{3.1}$$

$$P(\eta_t = 0|\boldsymbol{X}) \propto \rho(\boldsymbol{X}|\eta_t = 0)P(\eta_t = 0) \tag{3.2}$$

$$\frac{P(\eta_t = 1|\boldsymbol{X})}{P(\eta_t = 0|\boldsymbol{X})} = \frac{\rho(\boldsymbol{X}|\eta_t = 1)P(\eta_t = 1)}{\rho(\boldsymbol{X}|\eta_t = 0)P(\eta_t = 0)} \tag{3.3}$$

In a typical multiple test under unknown dependence, the joint prior of $\boldsymbol{\eta}$ is unknown. The mechanism by which the observations are generated is in general unknown. This makes it difficult to compute the ratio of conditional densities in the right hand side of the (3.3). However, when the multiple testing problem can be recast as a signal detection problem where each observation is the outcome of a known signal-noise interaction operator, the situation reduces to Bayesian signal processing. The importance of each observation also depends on signal strength which is modeled using a small constant ϵ. The need for efficiency is greatest when the signal strength is low.

In this chapter, posterior probabilities are approximated without assuming the availability of a joint prior distribution. This is accomplished in two stages.

- Use a single, albeit high dimensional, controlled observation of $\boldsymbol{\eta}$ to compute and store first and second order prior conditional moments of $\boldsymbol{\eta}$ i.e. conditional probabilities and conditional covariances of $\boldsymbol{\eta}$.

- Approximate posterior probabilities using the moments computed in the previous step. The Taylor series approximation in the signal strength parameter is appropriate because the signal strength is low.

Each posterior probability obtained above incorporates information from nearby sites. These posteriors are used in the power-optimal multiple testing procedure provided in Sarkar et al. [35].

3.1 Approximating Posterior Probabilities in the Signal Processing Framework

3.1.1 The Statistical Signal Processing Model

We adopt a statistical signal processing model where the unobservable signal/null sequence is one realization of discrete time sequence, $\boldsymbol{\eta}$, of Bernoulli random variables. The realization $\eta_t = 0$ corresponds to η_t being a null and the realization $\eta_t = 1$ corresponds to η_t being a signal. It must be noted that these Bernoulli random variables are not independent unless otherwise specified. $\boldsymbol{\eta}$ remains hidden from observation. Each η_t is scaled by a positive parameter ϵ termed *signal strength*. Lower signal strength results in a smaller contribution of the signal, when present, to the observation. This deterministic constant is fixed for the multiple test and is assumed to be known. The scaled signal $\epsilon\eta_t$ and a continuously distributed, stochastic, i.i.d. noise Z_t. *interact* i.e. are inputs to a known deterministic function $f(\epsilon\eta_t, Z_t)$ of two variables. The value of this function is the

observed test statistic X_t. For example, $X_t = \epsilon\eta_t + Z_t$ or $X_t = \exp(\epsilon\eta_t)Z_t$. In general,

$$X_t = f(\epsilon\eta_t, Z_t), \ \ \forall t \tag{3.4}$$

Exact computation of the posterior-probabilities requires detailed knowledge about the joint and conditional priors of the signals which are usually unavailable. We approximate the posterior-log likelihood ratios assuming the availability of only the first and second conditional moments with only first order conditioning. Each approximate log likelihood ratio is expressed as a Taylor series expansion in the signal-strength parameter ϵ. This method is aimed at situations where the signal-strength is relatively small, the approximation accuracy depending on the noise logdensity $h(.)$, the signal-noise interaction $f(.,.)$ and prior conditional moments.

While the objective is to approximate posterior probabilities, the posterior log-likelihood ratio turns out to be more convenient to work with. The relationship between the posterior likelihood ratio and the posterior probability is shown below. It can be seen that they have a one-to one relationship shown below. The logarithm of the posterior likelihood ratio is the logit transform of the posterior probability. This again is a one to one transformation.

$$\frac{P\left(\eta_j = 1|\boldsymbol{X}\right)}{P\left(\eta_j = 0|\boldsymbol{X}\right)} = (\frac{1}{P\left(\eta_j = 0|\boldsymbol{X}\right)} - 1) = \frac{P\left(\eta_j = 1|\boldsymbol{X}\right)}{1 - P\left(\eta_j = 1|\boldsymbol{X}\right)}$$

$$\lambda(j, \boldsymbol{X}) = \ln\left[\frac{P\left(\eta_j = 1|\boldsymbol{X}\right)}{P\left(\eta_j = 0|\boldsymbol{X}\right)}\right] \implies P\left(\eta_j = 1|\boldsymbol{X}\right) = \frac{exp(\lambda(j, \boldsymbol{X}))}{1 + exp(\lambda(j, \boldsymbol{X}))}$$

In order to understand the effect of stochastic dependence among the hypotheses on multiple testing, Chi [11] represents the hypotheses with an unobservable dependent chain of Bernoulli random variables. The two values 0 and 1 respectively represent the truth or the falsehood of the corresponding hypothesis, which, in a Bayesian framework acts as the parameter of interest. A single positive parameter ϵ scales all the Bernoulli random variables. The magnitude of the scaling parameter controls the contribution of

each null/signal to the corresponding component of the observation vector. The scaled signal/null is transformed by unobservable i.i.d. noise random variables to produce the observable test statistics.

The approximation problem statement

Given an unobservable Bernoulli sequence $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_m)$, unobservable i.i.d noise $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_m)$ each component of which has a continuous known distribution on $\mathbb{R}$ with logdensity $h(.)$, a known signal strength parameter $\epsilon > 0$, a known signal-noise interaction function[17] $f(\epsilon\eta_t, Z_t)$, a test statistic $\boldsymbol{X} = (X_1, X_2, \ldots X_m)$ with $X_t = f(\epsilon\eta_t, Z_t)$ and prior marginal and conditional moments up to second order i.e. $E(\boldsymbol{\eta}|\eta_t = i)$ and $Cov(\boldsymbol{\eta}|\eta_t = i)$ for $i = 0, 1$, approximate the posterior log-likelihood ratio of each hypothesis η_k up to the second order in signal strength ϵ.

Please see the notation and definitions listed at the end of this subsection.

3.1.2 Taylor series expansion of the posterior log-likelihood-ratio in signal-strength ϵ

We approximate the posterior-log-likelihood ratio using Taylor series expansion of signal-strength parameter around 0. The need for a valid and powerful multiple testing procedure under dependence is greater when the signal strength is low. We expect a power series approximation of the posterior to perform better than statistics which do not borrow strength from nearby sites, for small ϵ. Our numerical results, explained in detail in chapter 4, provided computational evidence that this approximation is superior to a marginal-posteriors and p-values in many cases when strong dependence exists. The notation used in the posterior likelihood approximation is shown below.

[17] this is a binary operator acting on the scaled signal and the noise. For example this interaction function can be addition, multiplication etc.

Notation and definitions (vectors and matrices in bold)

m the total number of hypotheses and the size of the test statistic vector $\boldsymbol{X}$

$\boldsymbol{X}$ the test statistic vector in $\mathbb{R}^m$

X_0 the observed test statistic at the current, i.e. reference/origin, site.

$X_k = f(\epsilon\eta_k, Z_k)$ k^{th} component of the test statistic vector

$f(\epsilon\eta_k, Z_k)$ known function modeling the interaction between signal and noise

$\epsilon > 0$ *signal strength* parameter

$\boldsymbol{Z}$ unobservable noise vector in $\mathbb{R}^m$ with i.i.d. components

$\boldsymbol{\eta}$ unobservable binary vector of dependent hypotheses, η_k a Bernoulli r.v.

$h(.)$ marginal log-density of each component of the noise

$\rho(X)$ joint density of the observation vector $\boldsymbol{X}$

$\boldsymbol{\sigma}$ variable denoting an arbitrary binary string of signals/nulls of same size as $\boldsymbol{X}$

Operator notation

$[\boldsymbol{D}^{(1)}(\boldsymbol{X}, h)]_t = h'(X_t)$ derivative of log density evaluated at X_t

$[\boldsymbol{D}^{(2)}(\boldsymbol{X}, h)]_t = h''(X_t)$ second derivative of log density evaluated at X_t

$[\boldsymbol{G}^{i,(j)}(\boldsymbol{X}, h)]_t = (X_t)^i\, h^{(j)}(X_t)$ product of the j^{th} derivative of log-density at X_t and $(X_t)^i$, $i, j \in \{1, 2\}$

$[\boldsymbol{\Delta E}_0]_t = P(\eta_t = 1|\eta_0 = 1) - P(\eta_t = 1|\eta_0 = 0)$

$[\boldsymbol{\Delta Cov}_0]_{t,s} = Cov(\eta_t, \eta_s|\eta_0 = 1) - Cov(\eta_t, \eta_s|\eta_0 = 0)$

3.2 Additive Noise

This subsection addresses the important special case when the binary signal is scaled by ϵ is added to noise with log-density $h(.)$. The signal processing model and the notation introduced in the previous subsection will be used throughout the discussion. By assumption,

$$X_t = \epsilon\eta_t + Z_t, \quad -n \leq t \leq n \tag{3.5}$$

The posterior probabilities of a signal (resp. null) with index t conditioned on the entire observation vector $\boldsymbol{X}$ are respectively expressed as products of likelihood and prior in the two expressions below.

$$P(\eta_t = 1|\boldsymbol{X}) \propto \rho(\boldsymbol{X}|\eta_t = 1)P(\eta_t = 1)$$
$$P(\eta_t = 0|\boldsymbol{X}) \propto \rho(\boldsymbol{X}|\eta_t = 0)P(\eta_t = 0)$$

Without loss of generality, assume $t = 0$. After taking natural logarithm and subtracting, the *logit transform* of the conditional posterior probability of a signal can be expressed as:

$$\ln\left[\frac{P(\eta_0 = 1|\boldsymbol{X})}{P(\eta_0 = 0|\boldsymbol{X})}\right] = \ln\left[\frac{P(\eta_0 = 1)}{P(\eta_0 = 0)}\right] + \ln\left[\frac{\rho(\boldsymbol{X}|\eta_0 = 1)}{\rho(\boldsymbol{X}|\eta_0 = 0)}\right] \tag{3.6}$$

$$\rho(\boldsymbol{X}|\eta_0 = 1) = \sum_{\boldsymbol{\sigma}} \rho(X, \sigma|\eta_0 = 1) \tag{3.7}$$

$$\rho(\boldsymbol{X}|\eta_0 = 1) = \sum_{\boldsymbol{\sigma}} \rho(\boldsymbol{X}, \boldsymbol{\sigma}|\eta_0 = 1, \boldsymbol{\eta} = \boldsymbol{\sigma})P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1) \tag{3.8}$$

where each $\boldsymbol{\sigma}$ represents one possible true-configuration of the binary hypotheses i.e. a particular sequence of signals/nulls. It follows that

$$\rho(\boldsymbol{X}|\eta_0 = 1) = \sum_{\substack{\boldsymbol{\sigma} \\ \sigma_0=1}} \rho(\boldsymbol{X}, \boldsymbol{\sigma}|\boldsymbol{\eta} = \boldsymbol{\sigma}) P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1) \tag{3.9}$$

Given a particular realization the binary random vector $\boldsymbol{\eta} = \boldsymbol{\sigma}$, each observation X_t is dependent only on the noise random variable Z_t. This follows from the assumption of i.i.d. noise. Using the relationship between the observation and the additive noise given in (3.5), (3.8) can be written as

$$\rho(\boldsymbol{X}|\eta_0 = 1) = \sum_{\boldsymbol{\sigma}:\sigma_0=1} \prod_t \exp(h(X_t - \epsilon\sigma_t)) P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1) \tag{3.10}$$

$$= \sum_{\boldsymbol{\sigma}:\sigma_0=1} \exp(\sum_t h(X_t - \epsilon\sigma_t)) P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1) \tag{3.11}$$

Notice that the index t varies over all hypotheses/sites or equivalently over the vector of observations. For small $\epsilon > 0$, expand $\zeta(\epsilon, \boldsymbol{\sigma}) = \exp \sum_t h(X_t - \epsilon\sigma_t)$, up to second order, in the above expression as a Taylor series in ϵ. First we have,

$$\zeta(0, \boldsymbol{\sigma}) = \exp\left[\sum_t h(X_t)\right], \tag{3.12}$$

$$\zeta'(0, \boldsymbol{\sigma}) = -\zeta(0, \boldsymbol{\sigma})\Big(\sum_t h'(X_t)\sigma_t\Big), \tag{3.13}$$

$$\zeta''(0, \boldsymbol{\sigma}) = \zeta(0, \boldsymbol{\sigma})\left\{\left[\sum_t h'(X_t)\sigma_t\right]^2 + \sum_t h''(X_t)\sigma_t^2\right\} \tag{3.14}$$

Note that $\sigma_t^2 = \sigma_t$. Consequently, we have the second order approximation

$$\zeta(\epsilon, \boldsymbol{\sigma}) = \zeta(0, \boldsymbol{\sigma})\Bigg\{1 - \Big(\sum_t h'(X_t)\sigma_t\Big)\epsilon \tag{3.15}$$

$$+ \Big(\Big[\sum_t h'(X_t)\sigma_t\Big]^2 + \sum_t h''(X_t)\sigma_t\Big)\frac{\epsilon^2}{2} \tag{3.16}$$

$$+ \mathcal{O}(\epsilon^3)\Bigg\} \tag{3.17}$$

Combining this with

$$\rho(\boldsymbol{X}|\eta_0 = 1) = \sum_{\boldsymbol{\sigma}:\sigma_0=1} \zeta(\epsilon, \boldsymbol{\sigma}) P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1),$$

yields

$$\begin{aligned}\rho(\boldsymbol{X}|\eta_0 = 1) = &\sum_{\boldsymbol{\sigma}:\sigma_0=1} \zeta(0, \boldsymbol{\sigma})\Big[1 - \Big(\sum_t h'(X_t)\sigma_t\Big)\epsilon\Big] P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1) \\ &+ \sum_{\boldsymbol{\sigma}:\sigma_0=1} \zeta(0, \boldsymbol{\sigma})\Big[\Big([\sum_t h'(X_t)\sigma_t]^2\Big)\frac{\epsilon^2}{2}\Big] P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1) \\ &+ \sum_{\boldsymbol{\sigma}:\sigma_0=1} \zeta(0, \boldsymbol{\sigma})\Big[\Big(\sum_t h''(X_t)\sigma_t\Big)\frac{\epsilon^2}{2}\Big] P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1) + \mathcal{O}(\epsilon^3)\end{aligned}$$

Interchange the finite sums and use the definition of conditional expectation given $\eta_0 = 1$ over all possible binary sequences $\boldsymbol{\sigma}$ with $\sigma_0 = \eta_0 = 1$. The expression then reduces to

$$\rho(\boldsymbol{X}|\eta_0 = 1) = \zeta(0, \boldsymbol{\sigma})\left(1 + A\epsilon + (B + C)\frac{\epsilon^2}{2} + \mathcal{O}(\epsilon^3)\right) \tag{3.18}$$

where

$$\begin{aligned}A &= -\Big[\sum_t h'(X_t)E(\eta_t|\eta_0 = 1)\Big] \\ B &= \sum_{\boldsymbol{\sigma}:\sigma_0=1} \Big[\sum_t h'(X_t)\sigma_t\Big]^2 P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1) \\ C &= \sum_t h''(X_t)E(\eta_t|\eta_0 = 1)\end{aligned}$$

Expanding $\left[\sum_t h'(X_t)\sigma_t\right]^2$ algebraically, interchanging sums and again using the definition of conditional expectation for the summand in the double summation over t and s gives,

$$B = \sum_t \sum_s \Big(\sum_{\substack{\boldsymbol{\sigma} \\ \sigma_0=0}} [h'(X_t)h'(X_s)\sigma_t\sigma_s P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1)] \Big) \tag{3.19}$$

As a result, we get

$$\begin{aligned} \rho(\boldsymbol{X}|\eta_0 = 1) = \exp(\sum_t h(X_t)) \Bigg(1 &- \Big[\sum_t h'(X_t) E(\eta_t|\eta_0 = 1) \Big] \epsilon \\ &+ \Big[\sum_t \sum_s h'(X_t)h'(X_s) E(\eta_t\eta_s|\eta_0 = 1) \\ &+ \sum_t h''(X_t) E(\eta_t|\eta_0 = 1) \Big] \frac{\epsilon^2}{2} + \mathcal{O}(\epsilon^3) \Bigg) \end{aligned} \tag{3.20}$$

By symmetry, the second order Taylor series approximation given $\eta_0 = 0$ is

$$\begin{aligned} \rho(\boldsymbol{X}|\eta_0 = 0) = \exp(\sum_t h(X_t)) \Bigg(1 &- \Big[\sum_t h'(X_t) E(\eta_t|\eta_0 = 0) \Big] \epsilon \\ &+ \Big[\sum_t \sum_s [h'(X_t)h'(X_s) E(\eta_t\eta_s|\eta_0 = 0)] \\ &+ \sum_t h''(X_t) E(\eta_t|\eta_0 = 0) \Big] \frac{\epsilon^2}{2} + \mathcal{O}(\epsilon^3) \Bigg) \end{aligned} \tag{3.21}$$

Substituting the expressions in (3.20) and (3.21) into the r.h.s. of (3.6), we obtain:

$$\ln\left[\frac{P(\eta_0=1|\boldsymbol{X})}{P(\eta_0=0|\boldsymbol{X})}\right] = \ln\left[\frac{P(\eta_0=1)}{P(\eta_0=0)}\right] + \ln\Bigg\{1-\left[\sum_t h'(X_t)E(\eta_t|\eta_0=1)\right]\epsilon + \left[\sum_t\sum_s [h'(X_t)h'(X_s)E(\eta_t\eta_s|\eta_0=1)] + \sum_t h''(X_t)E(\eta_t|\eta_0=1)\right]\frac{\epsilon^2}{2}+\mathcal{O}(\epsilon^3)\Bigg\} - \ln\Bigg\{1-\left[\sum_t h'(X_t)E(\eta_t|\eta_0=0)\right]\epsilon + \left[\sum_t\sum_s [h'(X_t)h'(X_s)E(\eta_t\eta_s|\eta_0=0)] + \sum_t h''(X_t)E(\eta_t|\eta_0=0)\right]\frac{\epsilon^2}{2}+\mathcal{O}(\epsilon^3)\Bigg\} \tag{3.22}$$

The last two logarithmic terms on the r.h.s. of (3.22) can be simplified by expanding both terms as Taylor series in ϵ around 0 and using the second order approximations. Given a_0, a_1, b_0, b_1,

$$\ln(1+a_1\epsilon+b_1\frac{\epsilon^2}{2}+\mathcal{O}\left(\epsilon^3\right)) - \ln(1+a_0\epsilon+b_0\frac{\epsilon^2}{2}+\mathcal{O}\left(\epsilon^3\right)) = (a_1-a_0)\epsilon + ((b_1-a_1^2)-(b_0-a_0^2))\frac{\epsilon^2}{2}+\mathcal{O}(\epsilon^3) \tag{3.23}$$

The second order approximation (3.22) can be rewritten by using the r.h.s of the second order approximation (3.23) and letting

$$a_i = -\Big[\sum_t h'(X_t)E(\eta_t|\eta_0=i)\Big]$$
$$b_i = \Big(\Big[\sum_t\sum_s [h'(X_t)h'(X_s)E(\eta_t\eta_s|\eta_0=i)]\Big] + \sum_t h''(X_t)E(\eta_t|\eta_0=i)\Big)$$

for $i = 0, 1$. Then,

$$a_1 - a_0 = -\Big[\sum_t h'(X_t)(E(\eta_t|\eta_0 = 1) - E(\eta_t|\eta_0 = 0))\Big] \tag{3.24}$$

and for $i = 0, 1$,

$$\begin{aligned} a_i^2 &= \Big(\Big[\sum_t \sum_s [h'(X_t)h'(X_s)E(\eta_t|\eta_0 = i)E(\eta_s|\eta_0 = i)]\Big]\Big) \\ b_i - a_i^2 &= \Big(\Big[\sum_t \sum_s [h'(X_t)h'(X_s)[E(\eta_t\eta_s|\eta_0 = i) - E(\eta_t|\eta_0 = i)E(\eta_s|\eta_0 = 1)]]\Big] \\ &\quad + \sum_t h''(X_t)[E(\eta_t|\eta_0 = i)]\Big) \\ &= \sum_t \sum_s [h'(X_t)h'(X_s)Cov(\eta_t, \eta_s|\eta_0 = i)] + \sum_t h''(X_t)[E(\eta_t|\eta_0 = i)] \end{aligned} \tag{3.25}$$

For convenience of notation, define the vector difference,

$$\boldsymbol{\Delta E}_t = \boldsymbol{E}\left(\boldsymbol{\eta}|\eta_t = 1\right) - \boldsymbol{E}\left(\boldsymbol{\eta}|\eta_t = 0\right) \tag{3.26}$$

Similarly, define the matrix difference,

$$\boldsymbol{\Delta Cov}_t = \boldsymbol{Cov}\left(\boldsymbol{\eta}|\eta_t = 1\right) - \boldsymbol{Cov}\left(\boldsymbol{\eta}|\eta_t = 0\right) \tag{3.27}$$

The vector of conditional probabilities in (3.26) and the matrix of conditional covariances in (3.27) are both determined by the dependence structure of the signal/null random vector $\boldsymbol{\eta}$. However, these conditional moments can be approximated from a single long controlled observation because:

- *stationarity* and *ergodicity* of $\boldsymbol{\eta}$ allow the approximation of conditional moments by the law of large numbers. Stationarity guarantees that the conditional distributions depend only on the state and remain invariant with respect to the index.

- absence of *long range dependence* allows restriction of the moment evaluation to an appropriately chosen *dependence window/radius*.

- The above two approximation procedures are, in principle, applicable not just to linearly ordered signals/nulls but also to other stationary, ergodic, discrete, random fields. The extension to random fields is possible if the random field has a known periodic/lattice structure without long range dependence or if the dependence in the random field is determined by parameters that can be estimated from an observation.

Next define the vector functions of the observation vector $\boldsymbol{X}$ as shown below:

$$\boldsymbol{D}^{(1)}(\boldsymbol{X}, h) = \Big(\ldots, h'(X_{-t}), \ldots, h'(X_{-1}), h'(X_0), h'(X_1), \ldots, h'(X_t), \ldots\Big)^T$$
$$\boldsymbol{D}^{(2)}(\boldsymbol{X}, h) = \Big(\ldots, h''(X_{-t}), \ldots, h''(X_{-1}), h''(X_0), h''(X_1), \ldots, h''(X_t)\ldots\Big)^T$$

Now using $\boldsymbol{D}^{(1)}(\boldsymbol{X}, h), \boldsymbol{D}^{(2)}(\boldsymbol{X}, h)$ and the notation introduced in (3.26) and (3.27), the equations (3.24) and (3.25) can be expressed in the compact form,

$$a_1 - a_0 = -[\boldsymbol{\Delta E}_0]^T \boldsymbol{D}^{(1)}(\boldsymbol{X}, h) \tag{3.28}$$

$$\begin{aligned} b_i - a_i^2 = {} & \boldsymbol{D}^{(1)}(\boldsymbol{X}, h)^T [\boldsymbol{Cov}(\boldsymbol{\eta}|\eta_0 = i)]\boldsymbol{D}^{(1)}(\boldsymbol{X}, h) \\ & + [\boldsymbol{E}(\boldsymbol{\eta}|\eta_0 = 1)]^T \boldsymbol{D}^{(2)}(\boldsymbol{X}, h) \end{aligned} \tag{3.29}$$

for $i = 0, 1$.

Substituting (3.28) and (3.29) into approximation (3.23) and then using (3.23) to expand the approximation (3.22) gives the final second order approximation,

$$\begin{aligned}\ln\left[\frac{P(\eta_0=1|\boldsymbol{X})}{P(\eta_0=0|\boldsymbol{X})}\right] &= \ln\left[\frac{P(\eta_0=1)}{P(\eta_0=0)}\right] \\ &\quad - \Big([\boldsymbol{\Delta E}_0]^T \boldsymbol{D}^{(1)}(\boldsymbol{X},h)\Big)\epsilon \\ &\quad + \Big([\boldsymbol{\Delta E}_0]^T \boldsymbol{D}^{(2)}(\boldsymbol{X},h)\Big)\frac{\epsilon^2}{2} \\ &\quad + \Big(\boldsymbol{D}^{(1)}(\boldsymbol{X},h)^T[\boldsymbol{\Delta Cov}_0]\boldsymbol{D}^{(1)}(\boldsymbol{X},h)\Big)\frac{\epsilon^2}{2} + \mathcal{O}(\epsilon^3) \qquad (3.30)\end{aligned}$$

the first order coefficient expansion being,

$$[\boldsymbol{\Delta E}_0]^T \boldsymbol{D}^{(1)}(\boldsymbol{X},h) \;=\; \sum_t \Big(P(\eta_t=1|\eta_0=1) - P(\eta_t=1|\eta_0=0)\Big)\,h'(X_t),$$

the first term of the second order coefficient expanded as,

$$[\boldsymbol{\Delta E}_0]^T \boldsymbol{D}^{(2)}(\boldsymbol{X},h) \;=\; \sum_t \Big(P(\eta_t=1|\eta_0=1) - P(\eta_t=1|\eta_0=0)\Big)\,h''(X_t),$$

and finally the second term of the second order coefficient expanded as,

$$\begin{aligned}&\boldsymbol{D}^{(1)}(\boldsymbol{X},h)^T[\boldsymbol{\Delta Cov}_0]\boldsymbol{D}^{(1)}(\boldsymbol{X},h) \\ =\;& \sum_{t,s}\Big(P(\eta_t=1,\eta_s=1|\eta_0=1) - P(\eta_t=1|\eta_0=1)P(\eta_s=1|\eta_0=1)\Big)h'(X_t)h'(X_s) \\ -\;& \sum_{t,s}\Big(P(\eta_t=1,\eta_s=1|\eta_0=0) - P(\eta_t=1|\eta_0=0)P(\eta_s=1|\eta_0=0)\Big)h'(X_t)h'(X_s)\end{aligned}$$

3.3 Multiplicative Noise

Multiplicative noise is an important class of signal-noise interaction. Just as was in the case of additive noise, the signal is scaled by the signal strength parameter. The scaled binary signal is exponentiated and then rescaled by noise. The sign in the exponent is chosen to be negative here but a positive sign can also be dealt with in a similar manner.

Under the model

$$X_t = Z_t \exp(-\epsilon\eta_t), \qquad \forall t \tag{3.31}$$

$$Z_t \sim \exp(h(z_t)), \qquad \text{i.i.d} \tag{3.32}$$

The fundamental expression for the joint conditional density of the observations given a signal at a reference index is again shown below.

$$\rho(\boldsymbol{X}|\eta_t = 1) = \sum_{\boldsymbol{\sigma}:\sigma_t=1} \rho(\boldsymbol{X}, \boldsymbol{\sigma}|\boldsymbol{\eta} = \boldsymbol{\sigma}) P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_t = 1) \tag{3.33}$$

For multiplicative noise in (3.31), the expression in (3.33) can be rewritten in terms of the noise and signal strength as shown in (3.35). The observations are independent given the signal/null sequence as the noise is i.i.d. Using (3.31) and (3.32) and the corresponding Jacobian, the conditional density of X_t given η_t is $\exp(\epsilon\sigma_t)\exp(h(X_t \exp(\epsilon\sigma_t)))$. Then,

$$\rho(\boldsymbol{X}|\boldsymbol{\eta} = \boldsymbol{\sigma}) = \prod_t \exp(\epsilon\sigma_t)\exp(h(X_t \exp(\epsilon\sigma_t))) \tag{3.34}$$

and as a result,

$$\rho(\boldsymbol{X}|\eta_0 = 1) = \sum_{\boldsymbol{\sigma}:\sigma_0=1} \prod_t \exp(\epsilon\sigma_t)\exp(h(X_t \exp(\epsilon\sigma_t)))P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1) \tag{3.35}$$

$$\rho(\boldsymbol{X}|\eta_0 = 1) = \sum_{\boldsymbol{\sigma}:\sigma_0=1} \exp(\sum_t [\epsilon\sigma_t + h(X_t \exp(\epsilon\sigma_t))])P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1) \tag{3.36}$$

Just as was done in (3.15) the function $\gamma(\epsilon, \boldsymbol{\sigma}) = \exp\left(\sum_t \left[\epsilon\sigma_t + h(X_t \exp(\epsilon\sigma_t))\right]\right)$ can be expanded in a Taylor series as a function of ϵ near $\epsilon = 0$.

$$
\begin{aligned}
\gamma(\epsilon, \boldsymbol{\sigma}) &= \exp(\sum_t \left[\epsilon\sigma_t + h(X_t \exp(\epsilon\sigma_t))\right]) \\
\implies \gamma(0, \boldsymbol{\sigma}) &= \exp(\sum_t h(X_t)) \\
\gamma'(\epsilon, \boldsymbol{\sigma}) &= \Big[\sum_t \left(\sigma_t + X_t\sigma_t h'(X_t \exp\left(\epsilon\sigma_t\right)) \exp(\epsilon\sigma_t)\right)\Big]\gamma(\epsilon, \boldsymbol{\sigma}) \\
\implies \gamma'(0, \boldsymbol{\sigma}) &= \Big(\sum_t \left[\sigma_t + h'(X_t)X_t\sigma_t\right]\Big)\gamma(0, \boldsymbol{\sigma}) \\
\gamma''(\epsilon, \boldsymbol{\sigma}) &= \Big(\Big[\sum_t \left(\sigma_t + X_t\sigma_t h'(X_t \exp\left(\epsilon\sigma_t\right)) \exp(\epsilon\sigma_t)\right)\Big]^2\Big)\gamma(\epsilon, \boldsymbol{\sigma}) \\
&\quad + \Big(\sum_t h''(X_t \exp(\epsilon\sigma_t))X_t^2{\sigma_t}^2 \exp(2\epsilon\sigma_t)\Big)\gamma(\epsilon, \boldsymbol{\sigma}) \\
&\quad + \Big(\sum_t h'(X_t \exp(\epsilon\sigma_t))X_t{\sigma_t}^2 \exp(\epsilon\sigma_t)\Big)\gamma(\epsilon, \boldsymbol{\sigma}) \\
\implies \gamma''(0, \boldsymbol{\sigma}) &= \gamma(0, \boldsymbol{\sigma}) \left(\left[\sum_t \sigma_t + h'(X_t)X_t\sigma_t\right]^2 + \sum_t \left[h''(X_t)X_t^2\sigma_t^2 + h'(X_t)X_t\sigma_t^2\right]\right)
\end{aligned}
$$

Define

$$
\begin{aligned}
g(x) &= 1 + h'(x)x \\
u(x) &= h'(x)x + h''(x)x^2
\end{aligned}
$$

Then,

$$
\begin{aligned}
\gamma'(0, \boldsymbol{\sigma}) &= \Big[\sum_t g(X_t)\sigma_t\Big]\gamma(0, \boldsymbol{\sigma}) \\
\gamma''(0, \boldsymbol{\sigma}) &= \left\{\Big[\sum_t g(X_t)\sigma_t\Big]^2 + \sum_t u(X_t)\sigma_t^2\right\}\gamma(0, \boldsymbol{\sigma})
\end{aligned}
$$

Note that $\sigma_t^2 = \sigma_t$. We then have the second order approximation for $\gamma(\epsilon, \boldsymbol{\sigma})$ below:

$$
\gamma(\epsilon, \boldsymbol{\sigma}) = \gamma(0, \boldsymbol{\sigma})\left[1 + \phi(\boldsymbol{\sigma})\epsilon + \Big((\phi(\boldsymbol{\sigma}))^2 + \psi(\boldsymbol{\sigma})\Big)\frac{\epsilon^2}{2} + \mathcal{O}(\epsilon^3)\right] \tag{3.37}
$$

where

$$\phi(\boldsymbol{\sigma}) = \sum_t [\sigma_t + h'(X_t)X_t\sigma_t] = \sum_t g(X_t)\sigma_t \tag{3.38}$$

$$\psi(\boldsymbol{\sigma}) = \sum_t [h''(X_t)X_t^2\sigma_t^2 + h'(X_t)X_t\sigma_t^2] = \sum_t u(X_t)\sigma_t \tag{3.39}$$

Therefore, using (3.37) along with the exact expression for $\rho(\boldsymbol{X}|\eta_0 = 1)$ in (3.36) gives the second order approximation below:

$$\rho(\boldsymbol{X}|\eta_0 = 1) = \sum_{\boldsymbol{\sigma}:\sigma_0=1} \left\{ \gamma(0,\boldsymbol{\sigma}) \Big[1 + \phi(\boldsymbol{\sigma})\epsilon + \Big((\phi(\boldsymbol{\sigma}))^2 + \psi(\boldsymbol{\sigma}) \Big) \frac{\epsilon^2}{2} \Big] P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1) + \mathcal{O}(\epsilon^3) \right\}$$

Just as was done in the additive case, interchanging the finite sums and using the definition of conditional expectation given $\eta_0 = 1$ over all possible binary sequences $\boldsymbol{\sigma}$ such that $\sigma_0 = 1$ simplifies all terms except the squared term. The expression then reduces to one analogous to (3.18):

$$\begin{aligned} \rho(\boldsymbol{X}|\eta_0 = 1) = \exp(\sum_t h(X_t)) \Bigg(1 + \Big[\sum_t E(\eta_t|\eta_0 = 1) + \sum_t h'(X_t)X_t E(\eta_t|\eta_0 = 1) \Big] \epsilon \\ + \Big[\sum_{\boldsymbol{\sigma}:\sigma_0=1} [\sum_t (1 + h'(X_t)X_t)\sigma_t]^2 P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1) \\ + \sum_t [h''(X_t)X_t^2 E(\eta_t|\eta_0 = 1) + h'(X_t)X_t E(\eta_t|\eta_0 = 1)] \Big] \frac{\epsilon^2}{2} \\ + \mathcal{O}(\epsilon^3) \Bigg) \end{aligned} \tag{3.40}$$

Like in (3.19), after expanding the double sum and using the definition of conditional expectation,

$$\begin{aligned} &\sum_{\boldsymbol{\sigma}:\sigma_0=1} \Big[\sum_t [(1 + h'(X_t)X_t)\sigma_t] \Big]^2 P(\boldsymbol{\eta} = \boldsymbol{\sigma}|\eta_0 = 1) \\ &\quad = \sum_t \sum_s (1 + h'(X_t)X_t)(1 + h'(X_s)X_s) E(\eta_t\eta_s|\eta_0 = 1) \end{aligned} \tag{3.41}$$

Then,

$$\rho(\boldsymbol{X}|\eta_0=1)=\exp(\sum_t h(X_t))\Bigg(1+\Big[\sum_t (1+h'(X_t)X_t)E(\eta_t|\eta_0=1)\Big]\epsilon \\ +\Big[\sum_t\sum_s (1+h'(X_t)X_t)(1+h'(X_s)X_s)E(\eta_t\eta_s|\eta_0=1) \\ +\sum_t h''(X_t)X_t^2E(\eta_t|\eta_0=1)+\sum_t h'(X_t)X_tE(\eta_t|\eta_0=1)\Big]\frac{\epsilon^2}{2} \\ +\mathcal{O}(\epsilon^3)\Bigg) \tag{3.42}$$

By symmetry the corresponding expression given $\eta_0=0$ is below.

$$\rho(\boldsymbol{X}|\eta_0=0)=\exp(\sum_t h(X_t))\Bigg(1+\Big[\sum_t (1+h'(X_t)X_t)E(\eta_t|\eta_0=0)\Big]\epsilon \\ +\Big[\sum_t\sum_s (1+h'(X_t)X_t)(1+h'(X_s)X_s)E(\eta_t\eta_s|\eta_0=0) \\ +\sum_t h''(X_t)X_t^2E(\eta_t|\eta_0=0)+\sum_t h'(X_t)X_tE(\eta_t|\eta_0=0)\Big]\frac{\epsilon^2}{2} \\ +\mathcal{O}(\epsilon^3)\Bigg) \tag{3.43}$$

Take logarithms and subtract (3.43) from (3.42). Then substitute the difference into the r.h.s. of (3.6) rewritten below as (3.44) for convenience.

$$\ln\left[\frac{P(\eta_0=1|\boldsymbol{X})}{P(\eta_0=0|\boldsymbol{X})}\right]=\ln\left[\frac{P(\eta_0=1)}{P(\eta_0=0)}\right]+\ln\left[\frac{\rho(\boldsymbol{X}|\eta_0=1)}{\rho(\boldsymbol{X}|\eta_0=0)}\right] \tag{3.44}$$

The substitution yields the second order approximation below for the posterior log likelihood ratio:

$$\begin{aligned}
\ln[\frac{P(\eta_0=1|\boldsymbol{X})}{P(\eta_0=0|\boldsymbol{X})}] &= \ln\left[\frac{P(\eta_0=1)}{P(\eta_0=0)}\right] \\
&\quad + \ln\Bigg\{1+\left[\sum_t g(X_t)E(\eta_t|\eta_0=1)\right]\epsilon \\
&\qquad + \Bigg[\sum_t\sum_s [g(X_t)g(X_s)E(\eta_t\eta_s|\eta_0=1)] \\
&\qquad\quad + \sum_t u(X_t)E(\eta_t|\eta_0=1)\Bigg]\frac{\epsilon^2}{2}+\mathcal{O}(\epsilon^3)\Bigg\} \\
&\quad - \ln\Bigg\{1+\left[\sum_t (g(X_t))E(\eta_t|\eta_0=0)\right]\epsilon \\
&\qquad + \Bigg[\sum_t\sum_s [g(X_t)g(X_s)E(\eta_t\eta_s|\eta_0=0)] \\
&\qquad\quad + \sum_t u(X_t)E(\eta_t|\eta_0=0)\Bigg]\frac{\epsilon^2}{2}+\mathcal{O}(\epsilon^3)\Bigg\} \qquad (3.45)
\end{aligned}$$

The last two logarithmic terms on the r.h.s. of (3.45) can be simplified by expanding both terms as Taylor series in ϵ around 0 and using the second order approximations. Given a_0, a_1, b_0, b_1, the approximation (3.23) is stated again here

$$\begin{aligned}
&\ln(1+a_1\epsilon+b_1\frac{\epsilon^2}{2}+\mathcal{O}\left(\epsilon^3\right)) - \ln(1+a_0\epsilon+b_0\frac{\epsilon^2}{2}+\mathcal{O}\left(\epsilon^3\right)) \\
&\qquad = (a_1-a_0)\epsilon + ((b_1-a_1^2)-(b_0-a_0^2))\frac{\epsilon^2}{2}+\mathcal{O}(\epsilon^3) \qquad (3.46)
\end{aligned}$$

Using (3.45), the corresponding constants in (3.46) for $i = 0, 1$,

$$
\begin{aligned}
a_i &= \Big[\sum_t g(X_t)E(\eta_t|\eta_0 = i)\Big] \\
b_i &= \Big(\Big[\sum_t \sum_s [g(X_t)g(X_s)E(\eta_t\eta_s|\eta_0 = i)]\Big] + \sum_t u(X_t)E(\eta_t|\eta_0 = i) \\
a_1 - a_0 &= \Big[\sum_t g(X_t)(E(\eta_t|\eta_0 = 1) - E(\eta_t|\eta_0 = 0))\Big] \qquad (3.47) \\
a_i^2 &= \Big(\Big[\sum_t \sum_s [g(X_t)g(X_s)E(\eta_t|\eta_0 = i)E(\eta_s|\eta_0 = i)]\Big]\Big) \\
b_i - a_i^2 &= \Big(\Big[\sum_t \sum_s [g(X_t)g(X_s)[E(\eta_t\eta_s|\eta_0 = i) - E(\eta_t|\eta_0 = i)E(\eta_s|\eta_0 = i)]]\Big] \\
&\qquad + \sum_t u(X_t)[E(\eta_t|\eta_0 = i)]\Big)
\end{aligned}
$$

Using the definition of conditional covariance between η_t and η_s, $Cov(\eta_t, \eta_s|\eta_0 = i) = E(\eta_t\eta_s|\eta_0 = i) - E(\eta_t|\eta_0 = i)E(\eta_s|\eta_0 = i)$, the last equation can be rewritten for $i = 0, 1$ as

$$
\begin{aligned}
b_i - a_i^2 &= \sum_t \sum_s [(1 + h'(X_t)X_t)(1 + h'(X_s)X_s)Cov(\eta_t, \eta_s|\eta_0 = i)] \\
&\qquad + \sum_t h''(X_t)X_t^2[E(\eta_t|\eta_0 = i)] + \sum_t h'(X_t)X_t[E(\eta_t|\eta_0 = i)] \qquad (3.48)
\end{aligned}
$$

Recall the notation introduced in (3.26) and (3.27) rewritten below. These quantities do not involve the observations and can be estimated using a controlled observation.

$$
\begin{aligned}
\boldsymbol{\Delta E}_t &= \boldsymbol{E}\left(\boldsymbol{\eta}|\eta_t = 1\right) - \boldsymbol{E}\left(\boldsymbol{\eta}|\eta_t = 0\right) \\
\boldsymbol{\Delta Cov}_t &= \boldsymbol{Cov}\left(\boldsymbol{\eta}|\eta_t = 1\right) - \boldsymbol{Cov}\left(\boldsymbol{\eta}|\eta_t = 0\right)
\end{aligned}
$$

Next define the vector functions of the observation vector $\boldsymbol{X}$ as shown below:

$$
\begin{aligned}
\boldsymbol{G}^{1,(1)}\left(\boldsymbol{X}, h\right) &= \Big(\ldots, X_{-t}\, h'(X_{-t}), \ldots, X_{-1}\, h'(X_{-1}), X_0\, h'(X_0), X_1\, h'(X_1), \ldots, X_t\, h'(X_t), \ldots\Big)^T \\
\boldsymbol{D}^{2,(2)}\left(\boldsymbol{X}, h\right) &= \Big(\ldots, X_{-t}^2\, h''(X_{-t}), \ldots, X_{-1}^2\, h''(X_{-1}), X_0^2\, h''(X_0), X_1^2\, h''(X_1), \ldots, X_t^2\, h''(X_t) \ldots\Big)^T \\
\mathbf{1} &= (\ldots, 1, 1, 1 \ldots, 1, \ldots)^T
\end{aligned}
$$

Using the notation just introduced i.e. $\boldsymbol{G}^{1,(1)}(\boldsymbol{X},h)$, $\boldsymbol{G}^{2,(2)}(\boldsymbol{X},h)$ and the notation introduced in (3.26) and (3.27), the equations (3.47) and (3.48) can be expressed in the compact form

$$a_1 - a_0 = [\boldsymbol{\Delta E}_0]^T[\mathbf{1} + \boldsymbol{G}^{1,(1)}(\boldsymbol{X},h)] \tag{3.49}$$

$$\begin{aligned} b_i - a_i^2 = {} & [\mathbf{1} + \boldsymbol{G}^{1,(1)}(\boldsymbol{X},h)]^T[\boldsymbol{Cov}(\boldsymbol{\eta}|\eta_0 = i)][\mathbf{1} + \boldsymbol{G}^{(1)}(\boldsymbol{X},h)] \\ & + [\boldsymbol{E}(\boldsymbol{\eta}|\eta_0 = i)]^T\boldsymbol{G}^{2,(2)}(\boldsymbol{X},h) \\ & + [\boldsymbol{E}(\boldsymbol{\eta}|\eta_0 = i)]^T\boldsymbol{G}^{1,(1)}(\boldsymbol{X},h) \end{aligned} \tag{3.50}$$

$$\begin{aligned} b_0 - a_0^2 = {} & [\mathbf{1} + \boldsymbol{G}^{1,(1)}(\boldsymbol{X},h)]^T[\boldsymbol{Cov}(\boldsymbol{\eta}|\eta_0 = 0)][\mathbf{1} + \boldsymbol{G}^{1,(1)}(\boldsymbol{X},h)] \\ & + [\boldsymbol{E}(\boldsymbol{\eta}|\eta_0 = 0)]^T\boldsymbol{G}^{2,(2)}(\boldsymbol{X},h) \\ & + [\boldsymbol{E}(\boldsymbol{\eta}|\eta_0 = 0)]^T\boldsymbol{G}^{1,(1)}(\boldsymbol{X},h) \end{aligned} \tag{3.51}$$

Replacing the sums by their equivalents in the vector/matrix notation, the r.h.s. of (3.46), i.e. the second order Taylor series approximation of the log-likelihood ratio of the

posterior probabilities for the case of multiplicative noise can be written as

$$X_t = Z_t \exp\left(-\epsilon \eta_t\right), \quad \forall t$$

$$\begin{aligned}
\ln\left[\frac{P(\eta_0 = 1|\boldsymbol{X})}{P(\eta_0 = 0|\boldsymbol{X})}\right] &= \ln\left[\frac{P(\eta_0 = 1)}{P(\eta_0 = 0)}\right] \\
&+ \left([\boldsymbol{\Delta E}_0]^T[\mathbf{1} + \boldsymbol{G}^{1,(1)}(\boldsymbol{X}, h)]\right)\epsilon \\
&+ \left([\boldsymbol{\Delta E}_0]^T \boldsymbol{G}^{1,(1)}(\boldsymbol{X}, h)\right)\frac{\epsilon^2}{2} \\
&+ \left([\boldsymbol{\Delta E}_0]^T \boldsymbol{G}^{2,(2)}(\boldsymbol{X}, h)\right)\frac{\epsilon^2}{2} \\
&+ \left([\mathbf{1} + \boldsymbol{G}^{1,(1)}(\boldsymbol{X}, h)]^T[\boldsymbol{\Delta Cov}_0][\mathbf{1} + \boldsymbol{G}^{1,(1)}(\boldsymbol{X}, h)]\right)\frac{\epsilon^2}{2} \\
&+ \mathcal{O}(\epsilon^3)
\end{aligned} \tag{3.52}$$

Chapter 4

Numerical Simulation, Computation and Comparison of Error Rates

4.1 The Computational Setting and Objectives

The general setting for computational experiments, comparisons and simulation is described here. The starting point for the large-scale multiple test is a large unobservable binary signal/null vector $\boldsymbol{\eta}$ that needs sifting/sieving. *Large-scale* in this context can usually mean anywhere between from thousands to millions of hypotheses. Such large numbers are quite typical in micro-array testing and in biomedical imaging. In this chapter, statistical dependence and its effects within the signal/null vector $\boldsymbol{\eta}$ (and consequently the test vector $\boldsymbol{X}$) are quantified, measured and compared for insight into the relative merits/demerits of multiple-testing methodologies. The goal is to, when possible, computationally identify the drawbacks and/or the advantages of a particular procedure in a particular situation. The performance of a multiple test can be summarized by the false discovery proportion (FDP) and the number of true discoveries (NTD). As was mentioned in the preceding chapters, it has long been recognized that statistical dependence and correlation can severely affect the performance of multiple testing procedures. It also has been recognized that some multiple testing methods are remarkably robust in some settings. Benjamini and Yekutieli [6], Chi [11], Sun and Cai [40] and Clarke and Hall [12], just to name a few studies, all address the issue of dependence from different angles.

Dependence can worsen validity (Type I error control) alone, detection power alone, or both. Several aspects of the multiple test together with statistical dependence determine the Type I and Type II errors of the test. Some important aspects are listed below.

- The *multiple-testing procedure* that is applied to the probability vector as described below.

- The *measure of type I error*, for example FWER or FDR, and the *measure of type II error* used and the corresponding control levels, denoted by α and θ, respectively.

- The *probability vector*, computed using the observations. Each probability in the vector is a p-value or a posterior probability of a single hypothesis. These probabilities may or may not take dependence on other signals/nulls into account.

- The *proportion of signals* in the unobservable true configuration of $\boldsymbol{\eta}$ and the *total number of hypotheses/observations*. Signal sparsity usually poses a challenge in signal detection and multiple testing is often used as preliminary strategy to reduce signal sparsity.

- The *dependence structure* of the underlying signal/null vector and the induced dependence structure of the observations.

- The *signal strength* of the observations, which is an explicit or implicit parameter of the random process that generates observations. High signal-strength corresponds to observations with a high contribution of the signal and low signal-strength corresponds to very noisy observations.

- The *range of dependence* represented by a parameter w which is a nonnegative integer that represents the half width of a window of influence to each signal/null to be taken into account. Small values for w mean only short range dependence is taken into account. $w = 0$ represents hypotheses are treated as independent.

The overall goal of the numerical experiments is to gain insight into the effects of dependence while controlling for the other factors listed and understanding their role.

4.2 Simulation of Dependent, Noisy Observations and Posterior Computation

The steps involved in the simulation and multiple testing are as follows.

1. Define a Markov chain over a state space E with $r = 5$ states and a desired stationary probability vector $\boldsymbol{\pi}$.

2. A stationary, ergodic Markov chain $\boldsymbol{\mathcal{M}} = (\mathcal{M}_t)$ of length $m = 100,000$ is simulated using a randomly generated transition matrix $\boldsymbol{P}$ that has the desired stationary probability distribution,i.e. $\boldsymbol{\pi}^T\boldsymbol{P} = \boldsymbol{\pi}^T$.

3. A binary Hidden-Markov chain $\boldsymbol{\eta} = (\eta_t)$ is defined over the state space $\{0, 1\}$ by a function $\boldsymbol{\tau}: E \to \{0, 1\}$ that maps a subset $F \subset E$ to $\{0\}$ and F^c to $\{1\}$ such that $\eta_t = \tau(\mathcal{M}_t)$. It must be noted that this chain does not possess the Markov property that its parent chain enjoys. The function τ is used only for the simulation of the hypotheses chain. Hence, each realization of $\boldsymbol{\eta}$ represents a large scale multiple hypothesis chain with an unknown, nontrivial and non-Markov dependence structure.

4. The last procedure can also be used to generate a random field of nulls and signals with unknown and nontrivial dependence structure by using a parent stationary, ergodic Markov random field on a regular graph over a finite set of states. As noted in Chapter 2 and as evidenced by the derivation of posteriors the hypotheses need not be linearly ordered in order to compute posteriors.

5. An i.i.d. vector $\boldsymbol{Z} = (Z_t)$ of continuously distributed noise components each with the same log density $h(.)$ is generated. $\boldsymbol{\eta}$ and $\boldsymbol{Z}$ interact componentwise to produce the observation $\boldsymbol{X}$. For this simulation the noise is chosen to be standard normal and the interaction is chosen to be additive as shown in (3.5) and signal-strength ϵ is given various values between 0.25 and 1.25.

$$X_t = \epsilon\eta_t + Z_t, \quad , \qquad \forall t$$
$$Z_t \sim \exp(h(z_t)), \qquad \text{i.i.d}$$

6. The observation vector $\boldsymbol{X}$ is used to compute a posterior probability vector. The posteriors may or may not borrow strength from nearby observations. In the absence of long range dependence, correlation of X_t and X_s approaches zero as the distance between t and s increases. A half window length w is chosen to represent the range that is considered sufficiently strong dependence. The expression for the posteriors given in (3.30) is used. The coefficients required in the equations are computed using a controlled known $\boldsymbol{\eta}$ from the same distribution.

7. The frequentist p-value and the *local-FDR* as defined in Efron [15] do not borrow strength from nearby observations. The p-value performances are used as benchmarks to compare the performance of the posteriors that take dependence into account.

8. Finally, the power-optimal multiple testing procedure under dependence described in (2.4) is used for the approximated posteriors and the local FDR-based posteriors for the given Type I error rate α. For p-values, the Benjamini Hochberg procedure is used for the required Type I error rate.

9. For the purposes of accurate comparison a modified *step-up procedure* that attains a target FDP or a target number of true discoveries is used for the p-value vector. The cutoff for these p-value based procedure is chosen so that they attain the best possible FDP when number of true discoveries is controlled or vice-versa.

4.3 Multiple Testing Algorithms

An arbitrary multiple test can be thought of as a function that maps the vector of probabilities $\boldsymbol{p} = (p_t)$ to a unique decision vector $\boldsymbol{d} = (d_t)$ of acceptances/rejections [output]. The Benjamini Hochberg procedure proposed in Benjamini and Hochberg [5] is shown below in its algorithmic form. This procedure is a step-up procedure

Algorithm 1 Benjamini-Hochberg Procedure (Storey et al. [39])

sort the p-values $p_1, p_2, \ldots, p_m$ in ascending order
denote sorted p-values by $q_1 \leq q_2 \leq \ldots \leq q_m$
find $\hat{k} = max\{1 \leq k \leq m : q_k \leq \frac{k\alpha}{m}\}$
if $\hat{k}$ exists, reject hypotheses corresponding to $q_1 \leq q_2 \leq \ldots \leq q_{\hat{k}}$
if $\hat{k}$ does not exist, do not reject any hypothesis

The power-optimal Bayesian algorithm that we use for multiple-testing uses the posteriors we approximated in Chapter 3.

Algorithm 2 Bayesian Power-Optimal Procedure (Sarkar et al. [35])

sort $p(\eta_1 = 1|\boldsymbol{X}), p(\eta_2 = 1|\boldsymbol{X}, \ldots, p(\eta_m = 1|\boldsymbol{X})$ in ascending order
denote sorted posterior probabilities by $q_1 \leq q_2 \leq \ldots \leq q_m$
if $q_m < 1 - \alpha$, do not reject any hypotheses
otherwise find $\hat{k} = max\{1 \leq k \leq m : \frac{\sum_{j=1}^{k} q_{m-j+1}}{k} \geq 1 - \alpha\}$
reject hypotheses corresponding to $q_{\hat{k}} \leq q_{\hat{k}+1} \leq \ldots \leq q_m$

In order to accurately compare the BH algorithm with the Bayesian power-optimal procedure, the stopping criterion for the BH procedure is modified so that it 1)stops as soon as it matches the number of true discoveries made by the Bayesian procedure or 2) stops as soon its FDP exceeds the FDP attained by the Bayesian procedure.

4.4 Numerical Results

Table 2: Notation for Numerical Computation. All simulations and computations were implemented in MATLAB.

α	FDR control level
$\mathcal{FDP}_{\mathcal{POST}}$	Observed *false discovery proportion* using *approximate posterior probabilities*. Takes dependence into account and borrows strength from nearby observations. Uses the power optimal Bayesian procedure described in section 1.
$\mathcal{S}_{\mathcal{POST}}$	Observed *number of true discoveries*(NTD) using *approximate posterior probabilities*. Takes dependence into account and borrows strength from nearby observations. Uses the power optimal Bayesian procedure described in section 1.
$\boldsymbol{w}$	*Half window length* for borrowing strength from nearby observations. Usually varies from an integer from 1 to 5
$\mathcal{FDP}_{\mathcal{LOC}}$	Observed *false discovery proportion* using *Efron's Local FDR*. Each posterior probability is based on one observation and ignores dependence among hypotheses. Uses the power optimal Bayes procedure described in section 1.
$\mathcal{S}_{\mathcal{LOC}}$	Observed *number of true discoveries*(NTD) using *Efron's Local FDR*. Each posterior probability is based on one observation and ignores dependence among hypotheses. Uses the power optimal Bayesian procedure described in section 1.
$\mathcal{FDP}_{\mathcal{PVAL}}$	Observed *false discovery proportion* using *p-values* Uses one observation and ignores dependence among hypotheses. A step-up procedure acting on p-values, like the Benjamini-Hochberg procedure, is used.
$\mathcal{S}_{\mathcal{PVAL}}$	Observed *number of true discoveries*(NTD) using *p-values* Uses one observation and ignores dependence among hypotheses. A step-up procedure acting on p-values, like the Benjamini-Hochberg procedure, is used.
$\boldsymbol{P}$	The randomly simulated transition matrix generated to create a hidden-markov binary chain. This matrix helps generate a chain with a "random" and unknown dependence structure.
$\mathcal{SLEM}$	*second largest eigen-modulus* of the transition matrix $\mathcal{P}$ underlying the (hidden) Markov chain. Larger values correspond to stronger dependence among the observations
$\boldsymbol{\pi}$	*Stationary probability distribution vector* of the transition matrix P. The first two states correspond to the null and the last three states correspond to a signal.

4.4.1 Comparison of FDP and True Discoveries: Strong Dependence

Table:3 displays the false discovery proportion and number of true discoveries of (a) the power-optimal Bayesian procedure after borrowing strength from nearby sites and (b) p-value based step-up procedure that is forced to attain either the NTD or the FDP attained by (a). The intervals shown, in the column for a random variable y, are $\mu(y) \pm \frac{\sigma(y)}{\sqrt{1000}}$. The $\mathcal{SLEM}$ of the parent Markov chain is 0.8262 and it corresponds to strong dependence in the hypotheses. It can be observed that when $\epsilon = 1$,$\alpha \geq 0.2$ or when $\epsilon = 0.5$,$\alpha \geq 0.3$ the posteriors outperform p-values in both NTD and FDP.

Table 3: Comparison of the False Discovery Proportion and the number of true discoveries of posteriors using a Bayesian power optimal procedure using either approximate multiple-site-based posterior likelihood from conditioning nearby observations to p-value based multiple testing using a step-up procedure.

$X_t = \epsilon\eta_t + Z_t$(additive noise), chain Length = 100,000, $w = 3$ iterations = 1000							
$\mathcal{SLEM}$	$Signal\%$	ϵ	α	$\mathcal{FDP}_{\mathcal{POST}}$	$\mathcal{FDP}_{\mathcal{PVAL}}$	$\mathcal{S}_{\mathcal{POST}}$	$\mathcal{S}_{\mathcal{PVAL}}$
0.8262	10.4%	1	0.3	(0.5000, 0.5005)	(0.6751, 0.6754)	(3730.4, 3736.3)	(1110.4, 1118.5)
		1	0.2	(0.4213,0.4218)	(0.6268,0.6272)	(2750.3,2756.1)	(566.4,572.9)
		1	0.1	(0.3194,0.3202)	(0.5548,0.5554)	(1678.9,1683.9)	(198.8,203.4)
0.8262	10.4%	0.5	0.1[10]	(0.2716,0.3002)	(0.2090,0.2283)	(0.73,0.78)	(2806.5,3103.3)
		0.5	0.2[10]	(0.2719,0.2954)	(0.3304,0.3492)	(2.6,2.8)	(1624.8,1869.5)
		0.5	0.3	(0.3660,0.3746)	(0.5803,0.5871)	(19.2,19.8)	(31.6,73.3)
		0.5	0.4	(0.4529,0.4566)	(0.6559,0.6581)	(82.0,83.2)	(7.1,7.7)
		0.5	0.5	(0.5402,0.5422)	(0.7037,0.7048)	(266.8,269.0)	(15.34,16.4)

$$P = \begin{pmatrix} 0.9085 & 0.0606 & 0.0156 & 0.0115 & 0.0037 \\ 0.0760 & 0.8607 & 0.0473 & 0.0057 & 0.0103 \\ 0.1400 & 0.1976 & 0.6376 & 0.0104 & 0.0145 \\ 0.5543 & 0.1503 & 0.0970 & 0.1895 & 0.0089 \\ 0.1734 & 0.3763 & 0.1285 & 0.0141 & 0.3077 \end{pmatrix} \quad \boldsymbol{\pi} = \begin{pmatrix} 0.5219 & 0.3781 & 0.0784 & 0.0113 & 0.0102 \end{pmatrix}^T$$

$\tau(i) = 0, i \in \{1, 2\} \equiv$ Null and $\tau(i) = 1, i \in \{3, 4, 5\} \equiv$ Signal, $P(\{3, 4, 5\}) = 0.104$

[10]This α is low and results in a very high standard deviation for the observed FDP and the corresponding number of true discoveries for p-values. The primary reason is that the observed FDP is often 1 and the corresponding R for the p-value vector is m, i.e. all hypotheses are rejected resulting in perfect discovery.

4.4.2 Comparison of FDP and True Discoveries: Moderate Dependence

Table:4 displays the false discovery proportion and number of true discoveries of (a) the power-optimal Bayesian procedure after borrowing strength from nearby sites and (b)p-value based step-up procedure that is forced to attain either the NTD or the FDP attained by (a). The intervals shown, in the column for a random variable y, are $\mu(Y) \pm \frac{\sigma(Y)}{\sqrt{1000}}$. The $\mathcal{SLEM}$ of the parent Markov chain is 0.634 and it corresponds to moderate dependence in the hypotheses. The relatively high signal proportion 23.87% improves the performance of both procedures. It can be observed that the posteriors outperform p-values in both NTD and FDP in all cases except when for $\epsilon = 0.5$,$\alpha = 0.1$. In the latter case the p-value based algorithm has become unstable with very high FDP.

Table 4: Comparison of observed FDP and the NTD of the Multiple Testing with Posteriors using a Bayesian power-optimal procedure that borrows strength from nearby observations to multiple testing with p-values using a step-up procedure. Implemented on MATLAB

$X_t = \epsilon\eta_t + Z_t$(additive noise), chain Length = 100,000, $w = 3$ iterations = 1000							
$SLEM$	$Signal\%$	ϵ	α	$\mathcal{FDP}_{\mathcal{POST}}$	$\mathcal{FDP}_{\mathcal{PVAL}}$	$\mathcal{S}_{\mathcal{POST}}$	$\mathcal{S}_{\mathcal{PVAL}}$
0.634	23.87%	0.5	0.3	(0.3183, 0.3211)	(0.3862, 0.3889)	(88.23, 89.23)	(44.8, 48.47)
		0.5	0.4	(0.4143,0.4153)	(0.4762,0.4771)	(698.05,700.8)	(211.35,218.27)
		0.5	0.1[23]	(0.1991,0.2249)	(0.7751,0.8009)	(0.1044,0.1190)	(4754.8,5372.2)
		0.5	0.5	(0.5400,0.5402)	(0.5851,0.5852)	(16397,16406)	(13926,13938)
0.634	23.87%	1	0.2	(0.2696,0.2700)	(0.3316,0.3320)	(4384.0,4391.0)	(2534.8,2548.5)
		1	0.3	(0.3631,0.3634)	(0.4264,0.4267)	(8184.3,8192.9)	(5514.0,5528.2)
		1	0.4	(0.4524,0.4526)	(0.5093,0.5095)	(12272,12281)	(9391.6,9405.4)
0.634	23.87%	0.75	0.3	(0.3369,0.3374)	(0.4054,0.4059)	(2479.8,2485.5)	(1060.1,1071.6)
		0.75	0.2	(0.2409,0.2419)	(0.3084,0.3094)	(689.8,692.7)	(227.4,234.4)
		0.75	0.1	(0.1390,0.1418)	(0.1875,0.1906)	(57.6,58.5)	(46.09,49.86)

$$P = \begin{pmatrix} 0.2232 & 0.2360 & 0.1169 & 0.1568 & 0.2672 \\ 0.0167 & 0.8874 & 0.0034 & 0.0338 & 0.0586 \\ 0.1380 & 0.4382 & 0.2392 & 0.0681 & 0.1167 \\ 0.2454 & 0.1314 & 0.1290 & 0.3795 & 0.1146 \\ 0.3954 & 0.1145 & 0.3045 & 0.0501 & 0.1355 \end{pmatrix} \quad \boldsymbol{\pi} = \begin{pmatrix} 0.0988 & 0.6612 & 0.0690 & 0.0762 & 0.0948 \end{pmatrix}^T$$

$\tau(i) = 0, i \in \{1, 2\} \equiv$ Null and $\tau(i) = 1, i \in \{3, 4, 5\} \equiv$ Signal, $P(\{3, 4, 5\}) = 0.2387$

[23]This α is low and results in a very high standard deviation for the observed FDP and the corresponding number of true discoveries for p-values. The primary reason is that the observed FDP is often

4.4.3 FDP and True Discoveries: Moderate Dependence and Independence

Table:5 displays the false discovery proportion and number of true discoveries of (a) the power-optimal Bayesian procedure after borrowing strength from nearby sites and (b)p-value based step-up procedure that is forced to attain either the NTD or the FDP attained by (a). The intervals shown, in the column for a random variable y, are $\mu(Y) \pm \frac{\sigma(Y)}{\sqrt{1000}}$. The $\mathcal{SLEM}$ of the parent Markov chain is either 0.5157 or 0. The former value, 0.5157, gives hypotheses that are weakly-to-moderately dependent whereas $\mathcal{SLEM} = 0$ corresponds to independent hypotheses.

Table 5: Comparison of observed FDP and the number of true discoveries of the Multiple Testing with Posteriors using a Bayesian power-optimal procedure that borrows strength from nearby observations to multiple testing with p-values using a step-up procedure. The $\mathcal{SLEM}$ of the parent Markov chain is 0.5157 which corresponds to weak-to-moderate dependence. It can be observed that the posteriors slightly outperform the p-values when there is weak-to-moderate dependence. The p-values very slightly outperform posteriors for independent hypotheses.

$X_t = \epsilon\eta_t + Z_t$(additive noise), chain Length = 100,000, $w = 3$ for P_1 iterations = 1000							
$SLEM$	$Signal\%$	ϵ	α	$\mathcal{FDP}_{\mathcal{POST}}$	$\mathcal{FDP}_{\mathcal{PVAL}}$	$\mathcal{S}_{\mathcal{POST}}$	$\mathcal{S}_{\mathcal{PVAL}}$
0.5157 $(P_1, \boldsymbol{\pi}_1)$	9.62%	1	0.3	(0.3262, 0.3281)	(0.3429, 0.3447)	(175.94, 177.12)	(156.28, 159.56)
		1	0.2^9	(0.2214,0.2252)	(0.2360,0.2398)	(41.22,41.79)	(54.36,57.71)
		1	0.4	(0.4249,0.4260)	(0.4407,0.4418)	(488.16,490.06)	(423.58,428.09)
		1	0.5	(0.5214,0.5221)	(0.5354,0.5361)	(1091.9,1094.5)	(981.1,985.86)
0 $(P_2, \boldsymbol{\pi}_2)$ i.i.d	19.85%	1	0.3	(0.2988,0.2994)	(0.2986,0.2992)	(1741,1744.2)	(1759.7,1763.1)
		1	0.4	(0.3988,0.3992)	(0.3987,0.3991)	(3984.6,3988.9)	(3995.4,3999.7)
		1	0.2	(0.1991,0.2001)	(0.1986,0.1996)	(494.88,496.51)	(530.56,534.26)
		1	0.5	(0.4990,0.4992)	(0.4989,0.4992)	(7309.5,7314.6)	(7314.9,7320)

$$P_1 = \begin{pmatrix} 0.7229 & 0.1948 & 0.0585 & 0.0225 & 0.0014 \\ 0.2913 & 0.6267 & 0.0027 & 0.0742 & 0.0052 \\ 0.7798 & 0.0245 & 0.0986 & 0.0938 & 0.0033 \\ 0.1255 & 0.6204 & 0.0254 & 0.2227 & 0.0061 \\ 0.1821 & 0.6763 & 0.0479 & 0.0859 & 0.0079 \end{pmatrix} \quad \boldsymbol{\pi}_1 = \begin{pmatrix} 0.5274 & 0.3764 & 0.0371 & 0.0560 & 0.0032 \end{pmatrix}^T$$

1 and the corresponding R for the p-value vector is m, i.e. all hypotheses are rejected resulting in perfect discovery.

$\tau(i) = 0, i \in \{1,2\} \equiv$ Null and $\tau(i) = 1, i \in \{3,4,5\} \equiv$ Signal, $P(\{3,4,5\}) = 0.0962$

$$P_2 = \begin{pmatrix} 0.6476 & 0.1546 & 0.1036 & 0.0737 & 0.0204 \\ 0.6476 & 0.1546 & 0.1036 & 0.0737 & 0.0204 \\ 0.6476 & 0.1546 & 0.1036 & 0.0737 & 0.0204 \\ 0.6476 & 0.1546 & 0.1036 & 0.0737 & 0.0204 \\ 0.6476 & 0.1546 & 0.1036 & 0.0737 & 0.0204 \end{pmatrix} \quad \boldsymbol{\pi}_2 = \begin{pmatrix} 0.6476 & 0.1546 & 0.1036 & 0.0737 & 0.0204 \end{pmatrix}^T$$

$\tau(i) = 0, i \in \{1,2\} \equiv$ Null and $\tau(i) = 1, i \in \{3,4,5\} \equiv$ Signal, $P(\{3,4,5\}) = 0.1985$

[9]This α is low and results in a very high standard deviation for the observed FDP and the corresponding number of true discoveries for p-values. The primary reason is that the observed FDP is often 1 and the corresponding R for the p-value vector is m, i.e. all hypotheses are rejected resulting in perfect discovery.

Chapter 5

Random Sampling of Stochastic Matrices

5.1 Motivation

In chapter 4, we tested our posterior approximations on simulated large-scale hypothesis chains with differing strengths of nontrivial dependence. The dependence structure had to be nontrivial and controllable so that it could be altered for testing effects of dependence on the performance of posterior approximations and multiple testing procedures. At the same time dependence structure in the data had to be as 'structureless' as possible to avoid confounding artifacts. This was achieved by simulating homogeneous Markov chains with various dependence strengths and then hiding them to create binary hidden-Markov Bernoulli sequences. The dependence structure of a homogenous Markov chain, $MC \equiv (\mu_t)$ over the states in $\{a_1, a_2 \dots a_d\}$ is determined by its transition matrix. The transition matrix $\boldsymbol{P}$ of a homogeneous Markov chain is the matrix of one-step transition probabilities where $P_{ij} = p_{a_k,a_j} = P(\mu_{t+1} = a_j | \mu_t = a_k)$. The dependence structure of a hidden Markov chain is a function of its parent Markov chain. Therefore hypotheses with reproducible, nontrivial dependence structures without confounding artifacts is closely tied to the problem of randomly sampling transition matrices. First, a few relevant classes of matrices are defined.

Definition 5.1.1. *The $d-1$ dimensional* simplex *in $\mathbb{R}^d$, denoted here by Δ_d, is the intersection of the unit $\|.\|_1$ sphere in $\mathbb{R}^d$ and the nonnegative orthant in $\mathbb{R}^d_+ = \{\boldsymbol{u} \mid u_k \geq 0, k \in \{1, 2, \ldots, , d\}\}$. i.e. $\Delta_d = \{\boldsymbol{v} \in [0,1]^d \mid \sum_{k=1}^d v_k = 1\}$.*

Definition 5.1.2. *A nonnegative matrix is a matrix $\boldsymbol{M}$ each entry of which is nonnegative. The set of all $d \times d$ matrices nonnegative matrices is denoted by $\mathcal{N}_d$. The nonnegativity of a matrix $\boldsymbol{M}$ is denoted $\boldsymbol{M} \geq 0$.*

Definition 5.1.3. *A row-stochastic matrix is a nonnegative matrix each of whose rows sums to 1. Each row $\boldsymbol{v}$ of a $d \times d$ row-stochastic matrix $\boldsymbol{S}$ belongs to the simplex in $\mathbb{R}^d_+$ defined by $\Delta_d = \{\boldsymbol{v} \in \mathbb{R}^d_+ | \mathbf{1}^T \boldsymbol{v} = 1\}$ or equivalently $\boldsymbol{S}\,\mathbf{1} = \mathbf{1}$, $\boldsymbol{S} > 0$. The set of all $d \times d$ row-stochastic matrices is denoted by $\mathcal{M}_d \subset \mathcal{N}_d$.*

Definition 5.1.4. *A doubly stochastic matrix is a row-stochastic matrix each of whose columns sums to 1. Each row and each column of a $d \times d$ doubly stochastic matrix $\boldsymbol{D}$ belongs to the simplex in $\mathbb{R}^d_+$ defined as $\Delta_d = \{\boldsymbol{v} \in \mathbb{R}^d_+ | \mathbf{1}^T \boldsymbol{v} = 1\}$, or equivalently $\boldsymbol{D}\,\mathbf{1} = \mathbf{1}$, $\boldsymbol{D}^T\,\mathbf{1} = \mathbf{1}$, $\boldsymbol{D} > 0$. The set of all $d \times d$ doubly-stochastic matrices is denoted by $\mathcal{B}_d \subset \mathcal{M}_d \subset \mathcal{N}_d$.*

As mentioned above, row-stochastic matrices used to generate the dependent binary chains in chapter-4 were simulated randomly in MATLAB in order to represent a random but unknown dependence structure. The transition matrices were then used to generate HMM binary chains with dependence structures that

1. can stand as proxies for unknown dependence structures in high-dimensions
2. are reproducible, non-Markovian and nontrivial
3. controllable so that they can be varied to test effects of dependence on large-scale multiple testing
4. as 'structureless' as possible in order to avoid confounding artifacts

While the transition matrix completely determines the dependence structure of a stationary, homogeneous Markov chain, two parameters of the transition matrix are particularly

important in determining the Markov chain's long-term behavior and the properties of the large-scale multiple tests generated by hiding the Markov chain. They are

1. **Multiple test dependence strength and second largest (in magnitude) eigenvalue ($\mathcal{SLEM}$):** The 'strength of dependence' within a stationary, homogeneous Markov chain is determined by the second largest (in magnitude) eigenvalue of its transition matrix. It is a well known result in the theory of Markov chains that the stationary, homogeneous, ergodic Markov chain mixes at rate $\mathcal{O}(\rho^k)$ where ρ where is the Second-Largest-Eigen-Modulus, $\mathcal{SLEM}$ for short and k is the number of steps. Consequently, larger the $\mathcal{SLEM}$ of the parent Markov chain, longer the dependence range in hypotheses generated by hiding the Markov chain. Hiding a Markov chain with a small $\mathcal{SLEM}$ gives rise to hypotheses chains with weak dependence. If $\mathcal{SLEM} = 0$, the simulated hypotheses will be realizations of an independent Bernoulli chains. The spectral gap of a transition matrix is $1 - \mathcal{SLEM}$. See Levin et al. [29] for example.

2. **Multiple test signal proportion and principal eigenvector (Perron vector):** Markov chains with randomly generated stationary distributions were generated by randomly varying the principal *left* eigenvectors of their transition matrices. The stationary distribution was varied based on the desired proportion of signals in the hypotheses. So the signal proportion is the sum of the stationary probabilities of the states that correspond to signals. In other words, the long run proportion of the time spent in a known subset of the state space is the proportion of signals. Consequently, controllable stationary distribution vectors can be hidden to produce controllable signal proportions in multiple tests.

In chapter 4, performance of approximated posteriors in multiple testing was tested by using the dependence structures that were sampled. However, the $\mathcal{SLEM}$ and/or the left-principal eigenvector were varied by constrained random sampling of Markov chains i.e. without ad-hoc fixing of the entries of transition matrix of the Markov chain to

suitable numbers. The transition matrices were sampled from the space of irreducible nonnegative (row) stochastic matrices.

Random row-stochastic matrices can also be used to simulate random walks on graphs and to simulate random contingency tables. They can also be used to statistically test for 'atypical' structure in a given contingency table or a given network. For example, Barvinok [3] investigates the properties and behavior of random contingency tables of non-negative integers with given row-sum and column-sum vectors. Random stochastic

The goal of this chapter is to address the following related random sampling problems.

1. **Problem-I**: Randomly sample a transition matrix with a prescribed stationary distribution.

2. **Problem-II**: Randomly sample a transition matrix with a prescribed spectral gap or with a spectral gap in a prescribed interval.

3. **Problem-III**: Randomly sample a transition matrix with both a prescribed stationary distribution and with an SLEM in a prescribed interval.

5.2 Basic Results on Square Nonnegative and Stochastic Matrices

In this chapter, a 'nonnegative matrix', unless otherwise specified, refers to a *nonnegative square matrix*. Also, a probability vector or a stationary distribution $\boldsymbol{\pi}$ will always be positive i.e. without zero probability states. Many results, that we will need, on stochastic matrices, follow directly from the Perron-Frobenius theory for irreducible nonnegative matrices. Bapat and Raghavan [2] provides an introduction to nonnegative matrices, Perron-Frobenius theory for irreducible nonnegative matrices and some applications. Chapter 8 of Horn and Johnson [26] also introduces nonnegative matrices and the Perron-Frobenius theory for irreducible non-negative matrices.

An irreducible square matrix is a matrix that *cannot* be reduced to a block form of four matrices, shown below, by a similarity transformation with a permutation matrix.

$$\boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P} = \begin{bmatrix} \boldsymbol{B}_{r\times r} & \boldsymbol{C}_{r\times(d-r)} \\ \boldsymbol{0}_{(d-r)\times r} & \boldsymbol{D}_{(n-r)\times(d-r)} \end{bmatrix}$$

The fundamental result, often known as the Perron-Frobenius[51] theorem for irreducible nonnegative matrices, is stated below without proof.

Theorem 5.2.1. *Let $\boldsymbol{A} \in \boldsymbol{\mathcal{N}_d}$ be an irreducible matrix in, $\boldsymbol{\mathcal{N}_d}$, the space of $d \times d$ nonnegative matrices. Let $\rho(\boldsymbol{A})$ denote the spectral radius of $\boldsymbol{A}$, i.e. absolute value of the largest eigenvalue of $\boldsymbol{A}$in magnitude. Then,*

(a) $\rho(\boldsymbol{A}) > 0$

(b) $\rho(\boldsymbol{A})$ is an eigenvalue of $\boldsymbol{A}$

(c) There is a positive vector $v > 0$ such that $\boldsymbol{A}v = \rho(\boldsymbol{A})v$ and

(d) $\rho(\boldsymbol{A})$ is an algebraically (and hence geometrically) simple eigenvalue of of $\boldsymbol{A}$.

Corollary 5.2.2. *Let $\boldsymbol{S} \in \mathcal{M}_d$ be an irreducible matrix where $\mathcal{M}_d$ is the space of $d \times d$ nonnegative, row-stochastic matrices. Let $\rho(\boldsymbol{S})$ denote the spectral radius of $\boldsymbol{S}$, i.e. absolute value of the largest eigenvalue in magnitude. Then,*

(a) $\rho(\boldsymbol{S}) = 1$

(b) $\lambda = 1$ is an eigenvalue of $\boldsymbol{S}$ with algebraic and geometric multiplicities 1.

(c) $\boldsymbol{1}_d$ is the eigenvector of $\boldsymbol{S}$ corresponding to the eigenvalue 1, i.e. $\boldsymbol{S}\boldsymbol{1}_d = \boldsymbol{1}_d$.

(d) There is a unique positive (probability) vector $\boldsymbol{\pi} \in int(\Delta_d)$ such that $\boldsymbol{\pi}^T\boldsymbol{S} = \boldsymbol{\pi}^T$, where $int(\Delta_d)$ is the interior of the $(d-1)$ simplex, in the cone $\mathbb{R}^d_+$, consisting of all positive probability vectors in $\mathbb{R}^d_+$.

[51] Birkhoff [7] showed that the Perron Frobenius theorem is a corollary to the contraction mapping theorem under *Hilbert's projective metric* on the cone $\mathbb{R}^d_+$. Hilbert's projective metric is defined by $distance(\boldsymbol{x},\boldsymbol{y}) = log\left(\frac{max_i(\frac{x_i}{y_i})}{min_i(\frac{x_i}{y_i})}\right)$. Kohlberg and Pratt [27] and Bapat and Raghavan [2] provide an exposition of Hilbert's projective metric and its geometric interpretation. Birkhoff's observation was that, viewed using another metric, an arbitrary nonnegative matrix acts like a contraction, on the cone $\mathbb{R}^d_+$, provides a different perspective on the action of nonnegative matrices on positive vectors.

(e) 1 *is an algebraically (and hence geometrically) simple eigenvalue of* $\boldsymbol{S}$ *with left eigenvector* $\boldsymbol{\pi}$ *and (right) eigenvector* $\mathbf{1}_d$.
(f) All other left eigenvectors of $\boldsymbol{S}$ *are orthogonal to* $\mathbf{1}_d$ *and all other right eigenvectors of S are orthogonal to* $\boldsymbol{\pi}$.

Proof. (a),(b) and (c) direct consequences of the theorem 5.2.1 and the definition of a stochastic matrix, $\boldsymbol{S}\mathbf{1}_d = \mathbf{1}_d$. The fact that $\boldsymbol{S}^T$ is also a nonnegative matrix, applying theorem 5.2.1 to it and dividing the eigenvector v by its sum prove (d) and hence (e). The fact that Left eigenvectors and right eigenvectors, corresponding to distinct eigenvalues, of an arbitrary matrix are always orthogonal and Part (d) of theorem 5.2.1 together prove (f). □

Corollary 5.2.3. *Let* $\boldsymbol{D} \in \mathcal{M}_d$ *be a doubly stochastic matrix, i.e. each row and each column sums to 1. Then,*
(a) $\mathbf{1}_d$ *is both a left and a right eigenvector of* $\boldsymbol{S}$ *corresponding to the eigenvalue 1.*
(b) The uniform distribution over the d states is the stationary distribution for a Markov chain with a transition matrix $\boldsymbol{S}$.

5.3 Convex Polytopes of Matrices

Let $\mathcal{M}_d$ be the set of $d \times d$ stochastic matrices. If each matrix is treated as point $\mathcal{M}_d$, it is a convex, compact polytope with $d(d-1)$ degrees of freedom in $\mathbb{R}^{d^2}$. It has zero Lebesgue measure in $\mathbb{R}^{d^2}$. However, being compact, the trace of the Lebesgue measure on it is well defined, in this case it is identical to the Hausdoff measure. Further, the polytope is compact and the measure can be normalized to define a probability distribution on it. Chafaï [10] provides more details. This corresponds to the uniform distribution on $\mathcal{M}_d$. The sample space of all stochastic matrices, equipped with the uniform distribution, is referred to as the *Dirichlet-Markov Ensemble*. Chafaï [10] surveys this ensemble. Henk et al. [24] provides a concise introduction to convex polytopes and their properties along with several examples.

5.3.1 Exact Sampling from the Dirichlet-Markov Ensemble

Sampling a row-stochastic matrix (sometimes referred to as Markov matrix) uniformly from all such matrices is a basic sampling problem in simulation of random Markov chains. It is also repeatedly carried out in many other simulation problems.

Each row of a $d \times d$ row-stochastic matrix S is a point in the simplex $\Delta_d == \{\boldsymbol{v} \in [0,1]^d \mid \sum_{k=1}^{d} v_k = 1\}$. The matrix S can be treated as point $\mathbb{R}^{d^2}$. The set $\mathcal{M}_d$ of all such matrices satisfies the properties below.

1. $\mathcal{M}_d$ defined as: $\{S|S\mathbf{1}_d = \mathbf{1}_d\}$ and $S \geq 0_{d\times d}$. $\mathcal{M}_d$ is bounded by linear constraints (an intersection of half-spaces)

2. If $S \in \mathcal{M}_d$ and $T \in \mathcal{M}_d$, $0 < \kappa < 1 \implies \kappa S + (1-\kappa)T \in \mathcal{M}_d$, i.e. $\mathcal{M}_d$ is convex.

3. $\mathcal{M}_d$ is bounded, hence using 1 and 2, it is a convex polytope of dimension $d(d-1)$ in $\mathbb{R}^{d^2}$.

4. The product of row-stochastic matrices is again row-stochastic. I_d is a stochastic matrix. Therefore, $\mathcal{M}_d$ is a also multiplicative semigroup. So is $\boldsymbol{\mathcal{N}_d}$, the set of nonnegative matrices.

As the detailed survey article Chafaï [10] points out, the polytope of row-stochastic matrices $\mathcal{M}_d$ can be equipped with the uniform distribution by the following procedure:

Exact sampling procedure for a uniformly distributed row-stochastic matrix (Chafaï [10])

1. Simulate d iid exponentially distributed random variables, $a_1, a_2, \ldots a_d$, each with mean 1.

2. The vector $(\frac{a_1}{\sum_{j=1}^{d} a_j}, \frac{a_2}{\sum_{j=1}^{d} a_j}, \ldots \frac{a_d}{\sum_{j=1}^{d} a_j})$ is distributed as the uniform-Dirichlet distribution on the unit simplex in $\mathbb{R}^d$, with mean $(\frac{1}{d}, \frac{1}{d}, \ldots \frac{1}{d})$.

3. Simulate d iid Dirichlet vectors by repeating the 1 and 2.

4. The vectors simulated in step 3 form the independent rows of a row-stochastic matrix that has the uniform distribution on $\mathcal{M}_d$.

5. The columns of the sampled matrix are exchangeable but they not independent.

$\mathcal{M}_d$ equipped with the uniform distribution on it is referred to as the Dirichlet-Markov ensemble. The rows of a matrix sampled from this ensemble are independent and hence exchangeable. Its columns are also exchangeable. While not pursued in this study, a probability measure on $\mathcal{M}_d$ can be seen as a probability measure on a semigroup. Högnäs and Mukherjea [25] details the properties of measures on $\mathcal{M}_d$ and related semigroups from the semigroup viewpoint.

5.3.2 The Transportation Polytope

The convex polytope of nonnegative $m \times n$ matrices $\boldsymbol{M}$ with *row-sum-vector* $\boldsymbol{M}\mathbf{1}_n = \boldsymbol{r} > \mathbf{0}$ and *column-sum-vector* $\boldsymbol{M}^T\mathbf{1}_m = \boldsymbol{c} > \mathbf{0}$, where $r_1 + r_2 + \ldots + r_m = c_1 + c_2 + \ldots + c_n$, is called the (two-way) transportation polytope and will be denoted here by $\mathcal{T}_{m,n}(\boldsymbol{r}, \boldsymbol{c})$ (See Ziegler [41]). Its dimension is $(m-1)(n-1)$. The integer counterpart of this polytope, when $\boldsymbol{r}$ and $\boldsymbol{c}$ are vectors of positive integers, is the set of all contingency tables with *margins* $\boldsymbol{r}$ and $\boldsymbol{c}$. Barvinok [3] investigates the properties of a table sampled uniformly from the latter space.

The Symmetric Transportation Polytope

The symmetric transportation polytope is a special case of the transportation polytope where $m = n$ and $\boldsymbol{r} = \boldsymbol{c}$, so that $\mathcal{T}_{m,n}(\boldsymbol{r}, \boldsymbol{c}) = \mathcal{T}_{m,m}(\boldsymbol{r}, \boldsymbol{r})$.

The Birkhoff Polytope

The Birkhoff polytope is a very special case of the transportation polytope where $\boldsymbol{r} = \boldsymbol{c} = \mathbf{1}$. It is the set of all *doubly stochastic* matrices. The Markov chain with a doubly stochastic transition matrix necessarily has the uniform stationary distribution over its states $(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m})$. Denoting the Birkhoff polytope by $\mathcal{B}_m$ and using the notation introduced in this chapter so far:

$$\mathcal{B}_d = \mathcal{T}_{d,d}(\mathbf{1}_d, \mathbf{1}_d) \tag{5.1}$$

The classical result below is due to Birkhoff and von Neumann independently. See Ziegler [41] for example.

Theorem 5.3.1. *The extreme points of the Birkhoff polytope are the $d!$ permutation matrices. The Birkhoff polytope is the convex hull of permutation matrices.*

The next theorem is known as Caratheódory's theorem.

Theorem 5.3.2. *Each point in a convex polytope can be expressed as a convex combination of at most $d+1$ extreme points, where d is the dimension of the polytope.*

5.4 The Convex Polytope of Stochastic Matrices with Prescribed Principal Left Eigenvector

The Dirichlet-Markov ensemble is easy to sample from in spite of its high-dimensionality of $d(d-1)$ because the probability measure is a product measure of independent probability measures on the rows. A natural and important sampling problem is that of sampling a stochastic matrix with a given stationary distribution $\boldsymbol{\pi}$. In order to do this, we first

define the sample space as the set of matrices $\boldsymbol{P} \in \mathcal{Q}_d(\boldsymbol{\pi}) \subset \mathcal{M}_d$ such that:

$$\boldsymbol{\pi}^T \boldsymbol{P} = \boldsymbol{\pi}^T \tag{5.2}$$

$$\boldsymbol{P}\, \mathbf{1}_d = \mathbf{1}_d \tag{5.3}$$

5.4.1 Properties of the Polytope $\mathcal{Q}_d$

Some relevant observations about the polytope $\mathcal{Q}_d$ are:

1. $\mathcal{Q}_d(\boldsymbol{\pi})$ is $(d-1)^2$ dimensional object in $\mathbb{R}^{d^2}$. This can be seen by observing that fixing a $(d-1)\times(d-1)$ square sub matrix fixes the remaining entries in the matrix.

2. $\mathcal{Q}_d(\boldsymbol{\pi})$ is a convex, sub-polytope of $\mathcal{M}_d$, the polytope of all row-stochastic matrices. It has $d-1$ fewer dimensions than $\mathcal{M}_d$.

3. For $\boldsymbol{\pi} = (\frac{1}{d}, \frac{1}{d}, \ldots, \frac{1}{d})$, $\mathcal{Q}_d(\boldsymbol{\pi}) = \mathcal{B}_d$, the set of $d \times d$ *doubly stochastic matrices*, i.e. the Birkhoff polytope.

4. Unlike the dirichlet-Markov ensemble, the rows of a matrix sampled from $\mathcal{Q}_d(\boldsymbol{\pi})$ are not independent.

5. For each fixed $\boldsymbol{\pi}$, $\mathcal{Q}_d(\boldsymbol{\pi})$ is a subsemigroup of the multiplicative semigroup $\mathcal{M}_d$.

6. (Hartfiel [23]) Let $\boldsymbol{\Pi} = diag(\boldsymbol{\pi})$, then the system $\boldsymbol{\pi}^T\boldsymbol{P} = \boldsymbol{\pi}^T$ can be written as $\boldsymbol{P}^T\boldsymbol{\Pi}\mathbf{1}_d = \boldsymbol{\pi}$. Also, multiplying (5.3) by the invertible matrix $\boldsymbol{\Pi}$ and letting $\boldsymbol{L} = \boldsymbol{\Pi}\boldsymbol{P}$, (5.2) and (5.3) give a system of equations in $\boldsymbol{L}$ shown below

$$\boldsymbol{L}^T\, \mathbf{1}_d = \boldsymbol{\pi} \tag{5.4}$$

$$\boldsymbol{L}\, \mathbf{1}_d = \boldsymbol{\pi} \tag{5.5}$$

A square matrix satisfying a system of the type (5.4) and (5.5) is called *line-sum-symmetric* because each row sum is equal to the corresponding column sum (Eaves et al.

[14]). Doubly stochastic matrices are trivially line-sum-symmetric. The theorem below is an immediate consequence of property 6.

Theorem 5.4.1. *(Hartfiel [23]) The convex polytope of stochastic matrices, $\mathcal{Q}_d$, each matrix in which has the same stationary probability vector $\boldsymbol{\pi}$, is (linearly) isomorphic to the convex polytope of matrices with row-sum-vector $\boldsymbol{\pi}$ and column-sum-vector $\boldsymbol{\pi}$.*

5.5 Sampling Approaches to Problem-I

Unlike in the case $\mathcal{M}_d$, there is no known direct method to sample uniformly from the polytopes $\mathcal{B}_d$, $\mathcal{T}_{d,d}(\boldsymbol{r}, \boldsymbol{c})$, $\mathcal{Q}_d(\frac{1}{d}\,\boldsymbol{\pi})$. The alternative is to use random walk based methods. These methods have been known to work until the dimension reaches a certain threshold. It must be kept in mind that the dimensionality of this problem for $d \times d$ matrices is $\mathcal{O}(d^2)$. For example, for a Markov chain with 100 states, the sampling problem is 10,000 dimensional.

5.5.1 Possible Exact Sampling Approaches

To the best of the author's knowledge, at the time of this writing, there are no known methods for exact sampling from the Birkhoff polytope or the symmetric transportation polytope.

Equivalence to Exact Sampling of the Symmetric Transportation Polytope

Transformation, when feasible, is often useful in exact sampling. The idea is to use an appropriate bijective transformation of the polytope so that it can be sampled by sampling the image set for which exact sampling is feasible. A natural transformation of the polytope $\mathcal{Q}_d(\boldsymbol{\pi})$ is to a symmetric transportation polytope $\mathcal{T}_{d,d}(\boldsymbol{\pi}, \boldsymbol{\pi})$. Recall the definition and properties of $\mathcal{Q}_d(\boldsymbol{\pi})$, established in Section 5.4. The matrix $\boldsymbol{L} = \boldsymbol{\Pi P}$ satisfies (5.4) and (5.5) rewritten below.

$$\boldsymbol{L}^T\ \mathbf{1}_d = \boldsymbol{\pi} \tag{5.6}$$

$$\boldsymbol{L}\,\mathbf{1}_d = \boldsymbol{\pi} \tag{5.7}$$

This system of equations, by definition, shown in 5.3.2, is the symmetric transportation polytope $\mathcal{T}_{d,d}(\boldsymbol{\pi}, \boldsymbol{\pi})$. So, if a symmetric transportation polytope can be exactly sampled uniformly, the polytope $\mathcal{Q}_d(\boldsymbol{\pi})$ being the image of the bijective linear, can also be sampled uniformly because the *Jacobian* is constant for a linear transformation. The simplest case of this situation is when $\boldsymbol{\pi} = \frac{1}{d}\mathbf{1}_d$ when the polytope corresponds to the Birkhoff polytope. At the time of this writing, exact sampling of the Birkhoff polytope remains unsolved.

Diagonal Scaling Algorithms

In [37], Sinkhorn proposed an iterative scaling procedure that maps each positive square matrix to a unique doubly stochastic matrix. The scheme simply divides each row by its sum and then each column by its sum iteratively until convergence. This result was later extended by Brualdi et al. [9] to a set containing all irreducible nonnegative matrices. Later Hartfiel [22] and Eaves et al. [14]. further generalized the result. The algorithm below is described in Franklin and Lorenz [18] in a slightly different form (also see Hartfiel [22] and Eaves et al. [14]) and builds on Sinkhorn's basic idea. This algorithm partitions the set of irreducible nonnegative matrices that converge to the same matrix in $\mathcal{Q}_d((\pi))$. However, this equivalence class is not characterized in closed form and is determined by a nonlinear operator whose Jacobian cannot be computed exactly. Therefore the distribution of the matrix sampled using this algorithm is not known.

Algorithm 3 Iterative scaling of an irreducible nonnegative matrix to obtain its corresponding irreducible stochastic matrix with left Perron vector $\boldsymbol{\pi}$. This algorithm was used in the numerical simulations in chapter 4 to generate stochastic matrices with left Perron vector $\boldsymbol{\pi}$.

$\boldsymbol{\pi} \in interior(\Delta_d)$

$\boldsymbol{A} > 0$, sampled uniformly from the set of all stochastic matrices.

$\boldsymbol{A}_1^R \leftarrow \boldsymbol{A}$

$\boldsymbol{A}_1^L \leftarrow \boldsymbol{A}$

$k \leftarrow 1$

while $\|\boldsymbol{A}_k^R \mathbf{1}_n - \mathbf{1}_d\| > \epsilon$ or $\|\boldsymbol{\pi}^T \boldsymbol{A}_k^L - \boldsymbol{\pi}^T\| > \epsilon$ **do**

$\boldsymbol{D}_k^L \leftarrow [diag(\boldsymbol{A}_k^R \mathbf{1}_n)]^{-1}$

$\boldsymbol{A}_k^L \leftarrow \boldsymbol{D}_k^L \boldsymbol{A}_k^R$

$\boldsymbol{D}_k^R \leftarrow [diag(\boldsymbol{\pi} \boldsymbol{A}_k^L)]^{-1} diag(\boldsymbol{\pi})$

$\boldsymbol{A}_{k+1}^R \leftarrow \boldsymbol{A}_k^L \boldsymbol{D}_k^R$

$k \leftarrow k + 1$

end while

5.5.2 Asymptotically Exact Sampling Approaches

Hit-and-Run, Gibbs/Metropolis and Random-walk Approaches

Among the fastest methods, in practice, to sample uniformly from a convex polytope is the *Hit and Run sampler* (Andersen and Diaconis [1]). This algorithm can also been used for sampling from non-convex bounded regions with target distributions other than the uniform distribution. Smith [38] introduced the algorithm in the sampling context. Andersen and Diaconis [1] generalizes the hit-and-run algorithm and unifies several Monte Carlo algorithms under its unmbrella. The MATLAB function cprnd$(N, A, b, options)$ can use either the hit-and-run sampler or the Gibbs sampler to uniformly sample from a convex polytope. The polytope is specified as a system of linear equalities and inequalities. The difficulty in using this is the large dimension of $(d-1)^2$. The polytope expressed as $\boldsymbol{Aq} <= \boldsymbol{b}$ requires a vector $\boldsymbol{q}$ of size d^2. For a Markov chain with with 100 states the dimension of the problem is 99,801.

5.6 Sampling Approaches to Problem II

The *convergence rate*, *strength of dependence*, *'range' of dependence* and *mixing time* of a homogeneous Markov chain all refer to how fast the distribution of the states approaches the stationary distribution of the chain. This parameter is controlled by the second largest (in magnitude) eigenvalue of the transition matrix. This is a standard result in linear algebra and Markov chain theory. See, for example, Levin et al. [29]. This quantity will be referred to here as $SLEM(\boldsymbol{P})$, where $\boldsymbol{P}$ is the transition matrix. The k-step transition probabilities are of the order $\mathcal{O}(SLEM(\boldsymbol{P})^k)$. The ability to prescribe, control or influence this number in a simulated Markov chain transition matrix gives the corresponding flexibility to prescribe the mixing rate and the range of dependence in the chain. The following theorem is fundamental to the simulation problem.

5.6.1 Eigenvalue Concentration

For a uniformly sampled stochastic matrix, as d grows, all except one eigenvalue approach 0. The result below was shown in Goldberg and Neumann [20] and was later improved upon by the Circular Law theorem in Bordenave et al. [8].

Theorem 5.6.1. *(Goldberg and Neumann [20], Bordenave et al. [8]) As $d \to \infty$, the $SLEM(\boldsymbol{P}_d)$ of a matrix $\boldsymbol{P}_d$ sampled from the $d \times d$ Dirichlet-Markov ensemble converges in probability to 0. Hence as $d \to \infty$, the spectral-gap of $\boldsymbol{P}_d$, $1 - SLEM(\boldsymbol{P}_d)$ converges in probability to 1.*

In other words, if $\boldsymbol{P}_d$ is uniformly sampled from the $d \times d$ stochastic matrices, as $d \to \infty$, for each fixed c, $0 < c < 1$, $P(SLEM(\boldsymbol{P}_d) > c) \to 0$.

Some numerical simulations of the value of SLEM for uniformly sampled stochastic matrices are shown in the Table: 6.

Table 6: Second Largest Eigen-Modulus of uniformly sampled row-stochastic (Dirichlet-Markov ensemble) matrices of various dimensions. Computations performed using MATLAB.

d	$\frac{1}{\sqrt{d}}$	$\mu(SLEM) \pm \frac{\sigma(SLEM)}{\sqrt{1000}}$
5	0.4472	(0.3703,0.3769)
10	0.3162	(0.3021,0.3058)
25	0.2	(0.2069,0.2083)
50	0.1414	(0.1475,0.1481)
100	0.1	(0.1042,0.1044)
500	0.0447	(0.0460,0.0460)

5.6.2 Behavior of SLEM and Consequences for Sampling

As can be seen from Table 6, the eigenvalues are sharply concentrated near $\frac{1}{\sqrt{d}}$ for even relatively small values of d. This makes it particularly challenging to control the $SLEM$ of a sampled matrix if larger eigenvalues are desired. Eigenvalue concentration has been observed and studied in many different ensembles of random matrices. Sampling matrices with eigenvalues larger than a prescribed number c involves sampling from an asymptotically measure 0 set. Naive rejection sampling does not work because the measure of the set is very small.

The SLEM is a continuous function of the matrix entries and but not a differentiable function of the entries. We plotted the value of the the second largest eigen-modulus along chords, drawn by connecting randomly chosen pairs of matrices, through the polytope $\mathcal{Q}_d(\boldsymbol{\pi})$. The plots numerically indicate that the value of $SLEM$ maybe very small deeper in the interior of the polytope and sometimes grows larger closer to the boundary. It also appears it displays monotonically increasing behavior as one gets closer to the boundary from the interior. A well designed rejection sampling approach, that exploits the behavior of the SLEM in the polytope, may be effective.

Appendix A

Definitions, Proofs and Notation

A.1 Proofs

Theorem. *FDR is by definition bounded above by FWER.*

Proof. This proof was originally stated in Benjamini and Hochberg [5]. Please refer to Table: 1 for notation.

$$
\begin{aligned}
V &\leq R \\
\mathrm{FDP} &= \frac{V}{R \vee 1} \leq 1 \\
\mathrm{FDR} &= E(\mathrm{FDP}) \\
\mathrm{FWER} &= P(V > 0) = E(\mathbb{1}_{V>0}) \\
\frac{V}{R \vee 1} &= \frac{V}{R \vee 1}\mathbb{1}_{\{V>0\}} \leq \mathbb{1}_{\{V>0\}} \\
\implies E(\mathrm{FDP}) &\leq E\left(\mathbb{1}_{\{V>0\}}\right) \\
\implies \mathrm{FDR} &\leq \mathrm{FWER} \qquad \square
\end{aligned}
$$

$\square$

A.2 Definitions

A.2.1 Measures of Type II Error and Power of a Multiple Test

The power of a multiple testing procedure can be characterized in multiple ways.A measures of power and a measure of Type II error are formally defined below. We use the expected number of true discoveries as our measure of power.

$$Power = E(R - V) \quad [\textit{expected number of true discoveries}]$$

$$FNP = \frac{T}{(m - R) \vee 1}$$

$$FNR = E(FNP) \quad [\textit{expected proportion ofsignals in hypotheses accepted as nulls}]$$

Bibliography

[1] Andersen, H. C. and Diaconis, P. (2007). Hit and run as a unifying device. *J. Soc. Fr. Stat. & Rev. Stat. Appl.*, 148(4):5–28.

[2] Bapat, R. and Raghavan, T. (1997). *Nonnegative Matrices and Applications*. Encyclopedia of Mathematics and its Applications. Cambridge University Press.

[3] Barvinok, A. (2010). What does a random contingency table look like? *Combin. Probab. Comput.*, 19(4):517–539.

[4] Benjamini, Y. and Heller, R. (2007). False discovery rates for spatial signals. *J. Amer. Statist. Assoc.*, 102(480):1272–1281.

[5] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300.

[6] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29 No.4:1165–1188.

[7] Birkhoff, G. (1957). Extensions of Jentzsch's theorem. *Trans. Amer. Math. Soc.*, 85:219–227.

[8] Bordenave, C., Caputo, P., and Chafaï, D. (2012). Circular law theorem for random Markov matrices. *Probab. Theory Related Fields*, 152(3-4):751–779.

[9] Brualdi, R. A., Parter, S. V., and Schneider, H. (1966). The diagonal equivalence of a nonnegative matrix to a stochastic matrix. *J. Math. Anal. Appl.*, 16:31–50.

[10] Chafaï, D. (2010). The Dirichlet Markov ensemble. *J. Multivariate Anal.*, 101(3):555–567.

[11] Chi, Z. (2011). Effects of statistical dependence on multiple testing under a hidden markov model. *Annals of Statistics*, 39, No 1:439–473.

[12] Clarke, S. and Hall, P. (2009). Robustness of multiple testing procedures against dependence. *Ann. Statist.*, 37(1):332–358.

[13] Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18 No. 1:71–103.

[14] Eaves, B. C., Hoffman, A. J., Rothblum, U. G., and Schneider, H. (1985). Line-sum-symmetric scalings of square nonnegative matrices. *Math. Programming Stud.*, (25):124–141. Mathematical programming, II.

[15] Efron, B. (2010). *Large-scale inference*, volume 1 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press, Cambridge. Empirical Bayes methods for estimation, testing, and prediction.

[16] Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2009). A bayesian discovery procedure. *J. R. Statist. Soc. B*, 71 Part 5:905–925.

[17] Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 17:347–388.

[18] Franklin, J. and Lorenz, J. (1989). On the scaling of multidimensional matrices. *Linear Algebra Appl.*, 114/115:717–735.

[19] Glaz, J., Pozdnyakov, V., and Wallenstein, S. (2009). *Scan Statistics: Methods and Applications*. Statistics for Industry and Technology. Birkhäuser. Perone-Pacifico, M. and Verdinelli, I., False Discovery Control for Scan Clustering pgs 271–288.

[20] Goldberg, G. and Neumann, M. (2003). Distribution of subdominant eigenvalues of matrices with random rows. *SIAM J. Matrix Anal. Appl.*, 24(3):747–761 (electronic).

[21] Guindani, M., Muller, P., and Zhang, S. (2009). A bayesian discovery procedure. *J. R. Statist. Soc. B*, 71 Part 5:905–925.

[22] Hartfiel, D. J. (1971). Concerning diagonal similarity of irreducible matrices. *Proc. Amer. Math. Soc.*, 30:419–425.

[23] Hartfiel, D. J. (1974). A study of convex sets of stochastic matrices induced by probability vectors. *Pacific J. Math.*, 52:405–418.

[24] Henk, M., Richter-Gebert, J., and Ziegler, G. M. (1997). Basic properties of convex polytopes. In *Handbook of discrete and computational geometry*, CRC Press Ser. Discrete Math. Appl., pages 243–270. CRC, Boca Raton, FL.

[25] Högnäs, G. and Mukherjea, A. (2010). *Probability Measures on Semigroups: Convolution Products, Random Walks and Random Matrices*. Probability and Its Applications. Springer, 2 edition.

[26] Horn, R. and Johnson, C. (1997). *Matrix Analysis*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2 edition.

[27] Kohlberg, E. and Pratt, J. W. (1982). The contraction mapping approach to the Perron-Frobenius theory: why Hilbert's metric? *Math. Oper. Res.*, 7(2):198–210.

[28] Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics*, 33 No. 3:1138–1154.

[29] Levin, D. A., Peres, Y., and Wilmer, E. L. (2008). *Markov Chains and Mixing Times*, volume 1. American Mathematical Society. Modern approach to the theory of Markov chains.

[30] Muller, P., Parmigiani, G., and Rice, K. (2006). FDR and bayesian multiple comparisons rules. *Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics*, Benidorm (Alicante, Spain), June 1st6th.

[31] Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *The Annals of Statistics*, 30 No.1:239–257.

[32] Sarkar, S. K. (2006). False discovery and false non-discovery rates in single-step multiple testing procedures. *The Annals of Statistics*, 34 No.1:394–415.

[33] Sarkar, S. K. (2007). Stepup procedures controlling generalized fwer and generalized fdr. *The Annals of Statistics*, 35 No. 6:2405–2420.

[34] Sarkar, S. K. and Guo, W. (2009). On a generalized false discovery rate. *The Annals of Statistics*, 37 No. 3:1545–1565.

[35] Sarkar, S. K., Zhou, T., and Ghosh, D. (2008). A general decision theoretic formulation of procedures controlling fdr and fnr from a bayesian perspective. *Statistica Sinica*, 18:925–945.

[36] Siegmund, D. O., Zhang, N. R., and Yakir, B. (2011). False discovery rate for scanning statistics. *Biometrika*, 98(4):979–985.

[37] Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879.

[38] Smith, R. L. (1984). Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Oper. Res.*, 32(6):1296–1308.

[39] Storey, Taylor, and Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc. B*, 66 Part 1:187–205.

[40] Sun, W. and Cai, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Statist. Soc. B*, 71 Part 2:393–424.

[41] Ziegler, G. M. (1994). *Lectures on Polytopes*. Graduate Texts in Mathematics. Springer, 1 edition.