False Discovery Variance Reduction in Large Scale Simultaneous Hypothesis Tests

Sairam Rayaprolu1 · Zhiyi Chi1

Received: 30 April 2018 / Revised: 29 November 2019 / Accepted: 1 December 2019 / Published online: 21 February 2020 © Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Statistical dependence between hypotheses poses a significant challenge to the stability of large scale multiple hypotheses testing. Ignoring it often results in an unacceptably large spread in the false positive proportion even though the average value is acceptable (Fan et al., J Amer Statist Assoc 107(499): 1019-1035, 2012; Owen J R Stat Soc Ser B 67(3): 411–426, 2005; Qiu et al., Stat Appl Genet Mol Biol 4: 32, 2005 and Schwartzman and Lin Biometrika 98(1): 199–214, 2011). However, the statistical dependence structure of data is often unknown. Using a generic signal-processing model, Bayesian multiple testing, and simulations, we demonstrate that the variance of the false positive proportion can be substantially reduced even under unknown short range dependence. We do this by modeling the data generating process as a stationary ergodic binary signal process embedded in noisy observations. We derive conditional probabilities needed for the Bayesian multiple testing by incorporating nearby observations into a second order Taylor series approximation. Simulations under general conditions are carried out to assess the validity and the variance reduction of the approach. Along the way, we address the problem of sampling a random Markov matrix with specified stationary distribution and lower bounds on the top absolute eigenvalues, which is of interest in its own right.

Keywords Multiple hypothesis testing · FDR · HMM

Mathematics Subject Classification (2010) Primary 62H15 · 62M02 · 62M07

1 Introduction

The problem of multiplicity in simultaneous hypothesis testing has been well recognized in experimental research and in statistics for more than half a century. However,
contemporary technologies have multiplied the scale of the problem by several orders of magnitude. Whereas classical problems were concerned with simultaneous inference of several or tens of hypotheses, present-day data intensive technologies such as DNA microarray in genomics and functional magnetic resonance imaging (fMRI) in brain imaging routinely produce tens or even hundreds of thousands of simultaneous hypotheses.

To get an idea on the severity of the problem of multiplicity, consider, for example, fMRI used to detect brain areas that activate in response to a specific task. A single fMRI can contain hundreds of thousands of multivariate observations, each being a time series of measurements at a spatial location, or “voxel”, in the brain (Lindquist and Mejia 2015). A typical analysis on such a dataset uses a “massively univariate approach”. For each voxel, a separate hypothesis test is performed on the time series of its measurements with the null hypothesis being that there is no activation in the voxel. Even with no brain activation in any voxel, if there are, say 100,000 voxels, a voxel-wise significance level of 0.05 would result in roughly 5000 significant $p$-values, in other words, 5000 false positives. Therefore it is necessary to choose a more stringent, sufficiently small $p$-value threshold. However, a $p$-value threshold that is too stringent will result in unacceptably many false negatives often resulting in failure to detect any voxels with significant brain activation. Determining a $p$-value threshold that strikes a balance between the two extremes is at the heart of the large scale multiple testing problem. See Farcomeni (2008) and Roquain (2011) for a broad overview of research in multiple testing.

False positives are usually referred to as false discoveries in multiple testing literature. The outcomes of multiple testing are summarized in the table below. Here, $V$ is the number of Type I errors (false discoveries), $T$ the number of false non-discoveries (false negatives), $S$ the number of true discoveries (NTD) or true positives and $U$ the number of true non-discoveries (true negatives). $R$ is the total number of discoveries (positives) and $A$ the total number of non-discoveries (negatives). Finally $m_0$ is the number of true nulls and $m$ the total number of hypotheses.

| Hypothesis | Accept | Reject | Total |
|------------|--------|--------|-------|
| True null  | $U$    | $V$    | $m_0$ |
| False null | $T$    | $S$    | $m_1$ |
|            | $A$    | $R$    | $m$  |

Procedures that reliably control Type I errors have been a focus in large scale multiple testing. To start with, a numerical criterion on the testing result has to be specified. Since its introduction by Benjamini and Hochberg (Benjamini and Hochberg 1995), the false discovery rate (FDR) has been widely used. Consequently, a great deal of research has been focused in particular on procedures that tend to have a low FDR irrespective of the distribution of data or the application.

By definition, the FDR is the expectation of $V / \max\{R, 1\}$. The latter, known as the false discovery proportion (FDP), is a random variable and summarizes the Type I errors of the multiple testing on the data at hand. When its variance is high there will be a high probability of it being substantially larger than the FDR, making the testing result less reliable even if on average it is at an acceptable level. Since it is usually not feasible to improve the reliability by repeating a large scale multiple testing procedure numerous times, it is important to achieve low variability of the FDP. In a similar spirit, a performance indicator of practical importance is stability, defined as the the standard deviation of the total number of discoveries (Gordon et al. 2007; Li 2015).
Assuming independence of test statistics of the hypotheses, the FDR can be controlled by the well-known Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg 1995; Storey et al. 2004). However, in practice the data in large scale multiple testing almost always has certain statistical dependence. Under various assumptions of dependence, the BH procedure is still valid, i.e., able the bound the FDR at a desired level (Wu 2008; Benjamini and Yekutieli 2001; Sarkar 2002; 2006; Clarke and Hall 2009). On the other hand, the FDR is mostly, though not entirely, concerned with Type I errors. There are other important metrics to evaluate a multiple testing procedure. Besides the aforementioned metrics on variability, a key metric is power, defined as $E(S)$ or $E(S/m_1)$. When several aspects are taken into account, in the presence of substantial dependence between the nulls, the overall performance of the BH procedure deteriorates, sometimes severely (Efron 2007; Fan et al. 2012; Owen 2005; Qiu et al. 2005; Roquain and Villers 2011).

To tackle statistical dependence in multiple testing, some studies have taken a probabilistic approach by modeling the data as stochastic processes with nontrivial dependence structure. They use specific models such as hidden Markov models (Sun and Cai 2009; Wu 2008), Markov random fields (Nguyen et al. 2014; Liu et al. 2016), or Gaussian random fields (Cheng and Schwartzman 2017; Schwartzman et al. 2011). The main strength of these models is their tractability. However, the statistical dependence structure of a dataset is often unknown or only partially known, which is in fact a major obstacle to multiple testing. For example, recent research has shown that that the principal cause of the invalid cluster inferences in fMRI is spatial autocorrelation functions that do not follow the assumed Gaussian shape (Eklund et al. 2016).

In this paper, we consider the FDR control with acceptable variability in the FDP, when the knowledge about the dependence structure of the nulls is very limited. We propose a Bayesian approach that is based on a generic signal-processing model and demonstrate through simulations that the variance of the FDP can be substantially reduced even under unknown short range dependence. While the nulls are modeled by a hidden process, no assumptions are made on the dependence structure of the process. Thus, the approach is nonparametric regarding the dependence between the nulls. Under the Bayesian setting, it is known that multiple testing based on the conditional probabilities of the nulls is optimal under certain criteria ((Chi 2011; Sarkar et al. 2008); cf. Section 2). However, the conditional probabilities are not available due to the unknown dependence structure of the nulls. The idea of our approach is to treat the conditional probabilities as functions of the signal strength and approximates them by a second order Taylor series in the weak-signal regime. In addition to providing much technical convenience, consideration of the weak-signal regime is of practical relevance due to the challenge it poses to multiple testing for very noisy data (Lindquist and Mejia 2015). The resulting approximation has two novel features. First, it naturally incorporates clusters of neighboring observations. Second, it does this without knowledge of the joint distribution of neighboring nulls. Instead, it utilizes their empirical moments. Expansion in terms of moments or cumulants is a defining feature in well known approximations such as normal approximation and Edgeworth approximation. It is interesting that our approximation has a similar feature, even though it results from a Taylor expansion in a parameter not directly related to the hidden process of the nulls.

After a brief literature review in Section 2, in Section 3, we introduce the signal-processing model used in the Bayesian approach and then derive the second order Taylor approximations to the conditional probabilities. In Section 4, we report numerical studies to assess multiple testing based on approximated conditional likelihoods in the presence of dependence. The results show that the multiple testing can control the FDR while having a much smaller variance in the FDP comparing to the BH procedure. To conduct the numerical
study, we have to address the problem of sampling a random Markov matrix with specified stationary distribution and lower bounds on the top absolute eigenvalues. This sampling issue is of interest in its own right. Finally, Section 5 makes some concluding remarks.

2 Related Works and Basic Setup

There is now a large literature to address statistical dependence in multiple testing (Fan and Han 2017; Fan et al. 2012; Schwartzman and Lin 2011; Efron 2010; Sun and Cai 2009; Clarke and Hall 2009; Leek and Storey 2008; Efron 2007; Sarkar 2006; Owen 2005; Qiu et al. 2005; Sarkar 2002; Benjamini and Yekutieli 2001; Storey et al. 2004; Wu 2008; Genovese and Wasserman 2004). Sun and Cai (2009) proposed to exploit the dependence structure of hypotheses from a decision-theoretical point of view and considered a Hidden Markov Model (HMM). Let \( H_1, \ldots, H_m \) be a set of hypotheses and \( \eta_j = 0 \) if \( H_j \) is true, and 1 otherwise. Assume that \( \eta = (\eta_1, \ldots, \eta_m) \) form a Markov chain and \( X = (X_1, \ldots, X_m) \) is a set of observations that are independent conditional on \( \eta \). Then they showed that a thresholding procedure for the conditional likelihoods \( P(H_j \text{ is false} \mid X) = P(\eta_j = 1 \mid X) \) minimizes \( E(\lambda V + T) \) for some \( \lambda > 0 \). From a Bayesian perspective, the Markov process is a prior on \( \eta \) and \( P(\eta_j = 1 \mid X) \) the posterior likelihood for \( H_j \) being true. In an earlier work Sarkar et al. (2008), it was shown that for any prior on \( \eta \) and any type of data \( X \), the following BH-type procedure, which will be referred to as the Bayes BH procedure, controls the FDR at a fixed target level \( \alpha \in (0, 1) \).

Bayes BH Procedure

1. Sort \( P(\eta_j = 1 \mid X) \) in ascending order as \( q_1 \leq q_2 \leq \ldots \leq q_m \).
2. If \( q_m < 1 - \alpha \), then accept all \( H_j \), otherwise reject all \( H_j \) with \( P(\eta_j = 1 \mid X) \geq q_R \), where \( R = \max\{1 \leq k \leq m : k^{-1} \sum_{j=1}^{k} q_{m-j+1} \geq 1 - \alpha \} \).

It was shown in Chi (2011) that the Bayes BH procedure is power optimal \textit{a posteriori}, that is, it has the largest \( E(S \mid X) \) among all procedures with \( E(\text{FDP} \mid X) \leq \alpha \). For comparison and later use, the BH procedure is based on \( p \)-values of hypotheses and can be described as follows Simes (1986), (Benjamini and Hochberg 1995; Storey et al. 2004).

BH Procedure

1. Sort the \( p \)-values in ascending order as \( p_1 \leq p_2 \leq \ldots \leq p_m \).
2. If \( p_1 > \alpha / m \), then accept all \( H_j \), otherwise reject all \( H_j \) with \( p_j \leq p_R \), where \( R = \max\{1 \leq k \leq m : p_k \leq k\alpha / m \} \).

The above results highlight the importance of \( P(\eta_j = 1 \mid X) \) in multiple testing. Once they are computed, the Bayes BH procedure can be immediately applied. However, to precisely evaluate the conditional likelihoods either by closed-form calculation or simulation, it is necessary to elaborate on the joint dependence structure of \( \eta \) and \( X \), which is typically impossible. In addition, from a signal processing point of view, multiple testing becomes useful typically when signals are relatively weak. We therefore wish to approximate \( P(\eta_j = 1 \mid X) \) using partial information on the dependence structure that is easy to infer, with reasonable accuracy especially when the signal is relatively weak.
In Chi (2011), under the HMM model, $P(\eta_j | X)$ was approximated by a second order Taylor series in signal strength. The Taylor series only uses the joint distribution of triples of nulls. It turns out that under suitable conditions, these distributions can be non-parametrically estimated provided that a sample of uncontaminated signals is available. The Markovian structure is not necessary for the Taylor series approximation. This observation is the starting point of the development in the following sections.

3 Approximate Conditional Likelihoods for Multiple Testing

As mentioned at the end of last section, our Bayesian approach is to apply the Bayes BH procedure to the conditional likelihoods of false nulls given the data. In this section, we first describe a generic signal processing framework for the Bayesian multiple testing. Then we derive a second order Taylor series to approximate the conditional likelihoods. Finally, we specialize to the case where the data consists of observations that can be described as outcomes of localized interactions between attenuated signals and noise. We assume that the parametric form of the signal-noise interaction is known. However, throughout the derivation, no parametric form of the joint distribution of the signals is assumed.

3.1 Signal Processing Point of View of Multiple Testing

Let $(H_t)_{t \in T}$ be a set of hypotheses indexed by a finite set $T$, which can be as simple as $\{1, \ldots, m\}$, or endowed with certain topology to incorporate, say spatial, relationships between the hypotheses. The “signal” is a set of Bernoulli random variables indexed by $T$, henceforth denoted by $\eta = (\eta_t, \in T)$. For each $t \in T$, let $\eta_t = 0$ if $H_t$ is true and 1 otherwise. The values of $\eta_t$, however, are unobservable, and the task is to determine which $H_t$ are false, or equivalently, which $\eta_t$ are equal to 1. In contrast to Sun and Cai (2009) and Chi (2011), no dependence structure of $\eta$ is assumed.

On the other hand, let $X$ denote a vector of observations. In the most general setting, $X$ can be of any type. We assume that conditional on $\eta$, $X$ has a probability density of the form

$$
\rho(x | \eta) = e^{q(x, \epsilon \eta)},
$$

where the function $q$ is assumed to be known. The parameter $\epsilon > 0$ in Eq. 1 will be referred to as “signal strength”. Observe that as $\epsilon = 0$, $X$ is independent of $\eta$, and hence contains no information about the hypotheses. Under the setting, for each $t \in T$, the conditional likelihood of $H_t$ being true given the data $X$ is $P(\eta_t = 1 | X)$. Thus, making decisions on the nulls is equivalent to recovering $\eta$ from the data $X$.

Notation We will treat each $(x_t)_{t \in T}$ as a column vector. Thus, for $u = (u_t)_{t \in T}$ and $v = (v_t)_{t \in T}$, $u^T v = \sum_{t \in T} u_t v_t$. If $p$ is a function on $\mathbb{R}$, then $p(x)$ is a shorthand for $(p(x_t))_{t \in T}$. $(M_{st})_{s,t \in T}$ denotes a matrix with rows indexed by $s$ and columns indexed by $t$. Finally, $\text{diag}(x)$ denotes the diagonal matrix $(M_{st})_{s,t \in T}$ with $M_{tt} = x_t$ and $M_{st} = 0$ for $s \neq t$.

3.2 Taylor Series Approximation of Conditional Likelihoods

We next derive an approximation to $P(\eta_t = 1 | X)$ using a second order Taylor series in $\epsilon$. The method, however, can attain higher order Taylor series as well. The approximation is valid when the signal strength $\epsilon$ is low, precisely when the power of multiple testing
suffers the most. While our objective is to approximate $P(\eta_t = 1 \mid X)$, the conditional log-likelihood ratios

$$r_t(X) = \ln \frac{P(\eta_t = 1 \mid X)}{P(\eta_t = 0 \mid X)}, \quad t \in T$$

(2)

are more convenient to work with. Each $r_t(X)$ is simply the logit transform of $P(\eta_t = 1 \mid X)$. Conversely, the latter is the logistic transform of the former, i.e.,

$$P(\eta_t = 1 \mid X) = \frac{1}{1 + \exp\{-r_t(X)\}}.$$  

(3)

Henceforth, quantities derived from taking expectation conditional on $\eta_t = i$ will be indexed by $i$ and $t$, e.g.,

$$E_{it}(\cdot) = E(\cdot \mid \eta_t = i), \quad \rho_{it}(\cdot) = \rho(\cdot \mid \eta_t = i), \quad \text{Cov}_{it}(\cdot) = \text{Cov}(\cdot \mid \eta_t = i).$$

By Bayes rule, for $t \in T$ and $i = 0, 1$, $P(\eta_t = i \mid X) \propto P(\eta_t = i)\rho_{it}(X)$. Then

$$r_t(X) = \ln \frac{P(\eta_t = 1)}{P(\eta_t = 0)} + \ln \frac{\rho_{1t}(X)}{\rho_{0t}(X)}$$

(4)

and it is essential to evaluate the second term on the r.h.s. of Eq. 4. By conditioning,

$$\rho_{it}(X) = \sum_{\sigma: \sigma_t = i} P_{it}(\eta = \sigma)\rho(X \mid \eta = \sigma).$$

(5)

**Lemma 3.1** Let $\eta$ be a binary process on $T$. Given $g \in C^2(\mathbb{R}^T, \mathbb{R})$, denote its gradient and Hessian by $g'$ and $g''$, respectively. Then as $\epsilon \to 0$,

$$\ln E[e^{g(\epsilon \eta)}] = g(0) + [g'(0)']E(\eta)\epsilon + \frac{1}{2}[E(\eta^1g''(0)\eta) + g'(0)'\text{Cov}(\eta)g'(0)]\epsilon^2 + o(\epsilon^2).$$  

(6)

Moreover, if $\pi = (\pi_t)_{t \in T}$ and $J = (J_{st})_{s,t \in T}$, with $\pi_t = P(\eta_t = 1)$ and $J_{st} = P(\eta_s = \eta_t = 1)$, then

$$E(\eta) = \pi, \quad E(\eta^1g''(0)\eta) = \text{Tr}(g''(0)J), \quad \text{Cov}(\eta) = J - \pi\pi^\dagger.$$  

(7)

**Proof** Equation 6 is well known and can be checked by standard calculation. Since $\eta_t$ is binary, $E(\eta_t) = P(\eta_t = 1) = \pi_t$, yielding the first equation in Eq. 7. Likewise, $E(\eta^1) = J$. Then the second equation in Eq. 7 follows from

$$E(\eta^1g''(0)\eta) = E(\text{Tr}(\eta^1g''(0)\eta)) = E(\text{Tr}(g''(0)\eta\eta^1)) = \text{Tr}(E(g''(0)\eta\eta^1)) = \text{Tr}(g''(0)J).$$

The last equation in Eq. 7 is also easy to verify.

The desired second order Taylor series approximation to $r_t(X)$ is as follows. Once an approximate value is obtained, an approximate value of $P(\eta_t = 1 \mid X)$ can be obtained by Eq. 3.
Theorem 3.2 Let \( q \) be as in Eq. 1. Suppose that for each \( x \), \( q(x, \theta) \) as a function of \( \theta \in \mathbb{R}^T \) belongs to \( C^2 \). Let \( \gamma(x) \) and \( H(x) \) be the gradient and Hessian of \( q(x, \theta) \) at \( \theta = 0 \), respectively. Then

\[
    r_t(X) = \ln \frac{p(\eta_t = 1)}{p(\eta_t = 0)} + \gamma(X)^\dagger [E_{1t}(\eta) - E_{0t}(\eta)]\epsilon + \frac{1}{2} [E_{1t}(\eta^t H(X) \eta) - E_{0t}(\eta^t H(X) \eta)]\epsilon^2 + \frac{1}{2} \gamma(X)^\dagger [\text{Cov}_{1t}(\eta) - \text{Cov}_{0t}(\eta)]\gamma(X)\epsilon^2 + o_{t,X}(\epsilon^2)
\]

as \( \epsilon \to 0 \), where \( o_{t,X}(\cdot) \) means that the implicit constant depends on \( t \) and \( X \).

Proof From Eq. 5 and the assumption, \( \rho_{it}(X) = E_{it}[\rho(X | \eta)] = E_{it}[\exp[q(X, \eta)]] \). Note that \( X \) is fixed. Applying Lemma 3.1 to \( g(\cdot) := q(X, \cdot) \), for \( i = 0, 1 \),

\[
    \ln \rho_{it}(X) = q(X, 0) + [\gamma(X)^\dagger E_{it}(\eta)]\epsilon + \frac{1}{2} [E_{it}(\eta^t H(X) \eta) + \gamma(X)^\dagger \text{Cov}_{it}(\eta)\gamma(X)]\epsilon^2 + o_{i,X}(\epsilon^2).
\]

This together with Eq. 4 yields the claimed expansion. \( \square \)

### 3.3 Localized Dependence of Data on Hypotheses

We now consider the case where the dependence of \( X \) on \( \eta \) is localized. Specifically, suppose \( X = (X_t)_{t \in T} \) and the function \( q \) in Eq. 1 can be written as

\[
    q(x, \theta) = \sum_t q_t(x_t, \theta_t).
\]

Under Eq. 9, \( X_t \) are independent conditionally on \( \eta \). Then from Theorem 3.2, we have the following.

Corollary 3.3 Suppose that for each \( x \) and \( t \in T \), \( q_t(x, \theta) \) as a function of \( \theta \in \mathbb{R} \) is twice continuously differentiable. Let \( \gamma_t(x) \) and \( k_t(x) \) be the first and second derivatives of \( q_t(x, \theta) \) at \( \theta = 0 \), respectively. Let \( \gamma(x) = (\gamma(x_t))_{t \in T} \) and \( k(x) = (k_t(x_t))_{t \in T} \). Then as \( \epsilon \to 0 \),

\[
    r_t(X) = \ln \frac{p(\eta_t = 1)}{p(\eta_t = 0)} + \gamma(X)^\dagger [E_{1t}(\eta) - E_{0t}(\eta)]\epsilon + \frac{1}{2} k(X)^\dagger [E_{1t}(\eta) - E_{0t}(\eta)]\epsilon^2 + \frac{1}{2} \gamma(X)^\dagger [\text{Cov}_{1t}(\eta) - \text{Cov}_{0t}(\eta)]\gamma(X)\epsilon^2 + o_{t,X}(\epsilon^2).
\]

Proof Comparing Eqs. 8 and 10, it suffices to show

\[
    E_{1t}(\eta^t H(X) \eta) - E_{0t}(\eta^t H(X) \eta) = k(X)^\dagger [E_{1t}(\eta) - E_{0t}(\eta)].
\]

From Eq. 9, it is seen that \( H(x) = \text{diag}(k(x)) \). Then for \( i = 0, 1 \),

\[
    E_{it}(\eta^t H(X) \eta) = E_{it}(\text{Tr}(\eta^t H(X) \eta)) = E_{it}(\text{Tr}(\text{diag}(k(x)) \eta \eta^t)) = \sum_s k_s(x_s) p_{it}(\eta_s = 1) = k(x)^\dagger E_{it}(\eta),
\]
and hence the desired identity.

Note that $E_{it} (\eta)$ is the vector of $P(\eta_s = 1 \mid \eta_t = i), s \in T$. Therefore, it is determined by the first and second moments of $\eta$. Likewise, $\text{Cov}_{it} (\eta)$ is determined by the first three moments of $\eta$. On the other hand, $y_t (x)$ and $k_t (x)$ are determined by how $X_t$ depends on $\eta_t$, regardless of the statistical dependence at other $s \in T$. Under many settings of signal processing, these quantities can be estimated directly. One can regard $\eta$ as a message sent through a noisy environment, and $X$ the noise-corrupted message being received. Before the communication formally starts, one can first estimate the moments of the distribution of $\eta$.

For example, if $\eta$ is a long contiguous segment of a stationary and ergodic binary process, then the moments may be estimated using a sample of $\eta$ by the law of large numbers. If in addition, there is no significant long range dependence in the distribution of $\eta$, then the evaluation of the moments, say $E(\eta^0 \eta_t)$, can be restricted to $t$ within an appropriately chosen radius from 0, which significantly reduces the number of moments to be estimated. On the other hand, one can use pre-selected test signals to estimate the dependence of $X_t$ on $\eta_t$. If for all $t$, the dependence is the same, i.e., the function $q_t$ is the same for all $t$, then an estimation may be obtained from a single observation of the pair $(\eta, X)$, again by the law of large numbers.

Sometimes it is more convenient to express $X_t$ as a result of interaction between $\eta_t$ and $Z_t$, where $Z_t$ are i.i.d. and regarded as noise. Specifically, suppose

$$X_t = \varphi_t (\epsilon \eta_t, Z_t), \text{ with } Z_t \in T \text{ independent of } \eta$$

and $Z_t$ i.i.d. with density $e^{b(z)}$.

Then the distribution of $X_t$ conditional on $\eta_t$ can be derived from the distribution of $Z_t$ and the functional form of $\varphi_t$. We next consider two important examples, where the signal noise interaction is additive and multiplicative, respectively. The approximation for the additive interaction will be used in the next section.

Example 3.4 (Additive noise) Let $\varphi_t (u, z) = u + z$ for all $t$. We then find that

$$\rho(X \mid \eta) = \prod_t \exp(h(X_t - \epsilon \eta_t)) = e^{q(X, \epsilon \eta)},$$

with $q(x, u) = \sum_t h(x_t - u_t)$. Then $\gamma(x) = -h'(x)$ and $H(x) = \text{diag}(h''(x))$, so by Theorem 3.2,

$$r_t (X) = \ln \frac{P(\eta_t = 1)}{P(\eta_t = 0)} - h'(X)[E_{1t}(\eta) - E_{0t}(\eta)]\epsilon$$

$$+ \frac{1}{2} h''(X)[\text{Cov}_{1t}(\eta) - \text{Cov}_{0t}(\eta)]h'(X)\epsilon^2$$

$$+ \frac{1}{2} h''(X)[E_{1t}(\eta) - E_{0t}(\eta)]\epsilon^2 + o_{t, X}(\epsilon^2), \text{ as } \epsilon \to 0. \quad (12)$$

Example 3.5 (Multiplicative noise) Let $\varphi_t (u, z) = ze^{-u}$ for all $t$. Then,

$$\rho(X \mid \eta) = \prod_t \exp(\epsilon \eta_t) \exp(h(X_t \exp(\epsilon \eta_t))) = e^{q(X, \epsilon \eta)},$$
with $q(x, u) = \sum_t [u_t + h(x_t e^{u_t})]$. Define $g(x) = 1 + xh'(x)$ and $\omega(x) = xh'(x) + x^2h''(x)$. It is easy to get $\gamma(x) = g(x)$ and $H(x) = \text{diag}(\omega(x))$. Then, as in the additive case,

$$r_t(X) = \ln \frac{P(\eta_t = 1)}{P(\eta_t = 0)} + g(X)[E_{1t}(\eta) - E_{0t}(\eta)]\epsilon + \frac{1}{2} g(X)\epsilon^2 + o_t, x(\epsilon^2), \text{ as } \epsilon \to 0.$$

### 4 Numerical Experiments

In this section, we report simulation studies to assess multiple testing based on approximated conditional likelihoods in the presence of dependence. We compare the Bayes BH procedure and the BH procedure. Recall that the former is based on the approximated conditional likelihoods of hypotheses given the data while the latter is based on marginal $p$-values of hypotheses.

#### 4.1 Sampling of Chain of Null Hypotheses

In the simulations, $\eta = (\eta_1, \ldots, \eta_m)$ is a long binary hidden Markov chain. This setting is very flexible since hidden Markov models are dense among essentially all finite-state stationary processes (Künsch et al. 1995). The chain is sampled as follows.

1. Specify a finite state space $E$, a nonempty strict subset $F$ of $E$, and a function $\tau : E \to \{0, 1\}$ that maps $F$ to 0 and $E \setminus F$ to 1.
2. Randomly sample a probability vector $\pi = (\pi_s)_{s \in E}$ and a transition matrix $P$ on the states of $E$ such that $\pi^t P = \pi^t$ (13)
3. Simulate a stationary Markov chain $M = (M_1, \ldots, M_m)$ on $E$ with stationary distribution $\pi$ and transition matrix $P$. Set $\eta_t = \tau(M_t)$ for each $t$. We will abbreviate the transform as $\eta = \tau(M)$

and refer to $M$ as the “parent” Markov chain. In all the simulations of the study, $E = \{1, 2, 3, 4, 5\}$, $F = \{1, 2\}$, and $m = 10^5$. The function $\tau$ is used only for the simulation of $\eta$ and is not available for multiple testing. Thus, each realization of $\eta$ is regarded as a large set of hypotheses with an unknown dependence structure. To avoid artifacts that may confound the comparison of multiple testing procedures, both $\pi$ and $P$ are randomly sampled. Denoting by $\psi$ the proportion of false nulls, we need $\pi$ to be such that the proportion of 1’s in $\eta$ is $\psi$. This is achieved by setting $\pi_s = (1 - \psi)\xi_s / \sum_{s \in F}\xi_s$ for $s \in F$ and $\pi_s = \psi \xi_s / \sum_{s \notin F}\xi_s$ for $s \notin F$, with $\xi_s$ i.i.d. uniformly distributed on (0,1).

Given $\pi$, the key component of the simulation is the transition matrix $P$, which is to be sampled from positive matrices, i.e., matrices with all entries being positive. Consequently, $M$ is irreducible and aperiodic. In addition to the constraint (13), since we need to control the strength of dependence within $\eta$, it is necessary that the mixing rate of $M$ is controllable.
As is well known, 1 is an eigenvalue of $P$ with multiplicity one, and all the other eigenvalues of $P$ have absolute values, i.e. eigenmoduli, strictly less than 1. Sort the eigenmoduli in decreasing order as $1 = \varrho_1(P) > \varrho_2(P) \geq \ldots \geq \varrho_d(P) \geq 0$. It is also well-known that the second largest eigenmodulus is an important parameter in determining the mixing rate of a Markov Chain (Rosenthal 1995). However the convergence of a Markov chain to its long term behavior is not always determined by a single eigenmodulus. The precise relationship of a Markov Chain’s mixing time to the spectrum of its transition matrix is surprisingly complicated (Diaconis 1995; Levin et al. 2009). Thus in addition to second, here we also consider the third largest eigenmodulus. Since the transition matrices that we simulate in the numerical experiments are all 5-by-5, these two parameters of a simulated matrix provide significant information about its spectrum. We next describe a heuristic procedure to sample $P$ that has a prescribed stationary distribution $\pi$.

### 4.2 Sampling of Random Transition Matrices

The ensemble of $d \times d$ uniformly distributed transition matrices, known as the Dirichlet Markov ensemble, can be represented by the uniform distribution on $\Delta_d$, the $d$-fold Cartesian product of $\Delta_1 = \{ (x_1, \ldots, x_d) \in [0, 1]^d : x_1 + \ldots + x_d = 1 \}$. The support of this ensemble is a convex $d(d-1)$-dimensional polytope. Uniform sampling of $P$ from the ensemble is easy because the rows of $P$ are i.i.d. $\sim (\xi_1, \ldots, \xi_d)/(\xi_1 + \ldots + \xi_d)$ with $\xi_i$ i.i.d. exponentially distributed; see Chafaï (2010) for more detail. However, in our study, we need to sample from the $(d-1)^2$ dimensional polytope of transition matrices that have a prescribed stationary distribution $\pi$. To the best of our knowledge, exact uniform sampling from this polytope is unknown for general $\pi$ (Cappellini et al. 2009). An alternative to exact sampling is to use asymptotically exact methods. For example, Diaconis et al. (2012) constructs a Markov chain whose stationary distribution is the uniform distribution on a convex polytope. In Chatterje et al. (2010), a Gibbs sampler is provided for the uniform distribution on the set of doubly stochastic matrices, i.e., transition matrices whose transposes are also transition matrices. However, doubly stochastic matrices are restrictive for our study, as their stationary distribution is always the uniform one on $1, \ldots, d$. To sample from the polytope of transition matrices with a specified stationary distribution, we use a heuristic described below. It should be pointed out that matrices sampled this way are not uniformly distributed on the polytope associated with $\pi$. Also, the complexity of the sampling increases fast as $d$ increases, as the dimension of the polytope grows at the order of $d^2$. We mention that the properties of a “typical” matrix from the uniform distribution on a closely related polytope of matrices was studied in Barvinok (2010).

Sinkhorn’s algorithm (Sinkhorn 1964; 1967) is an iterative scaling procedure that maps each positive matrix, not necessarily square, to a unique positive matrix with prescribed row and column sums. Its scheme is simply to scale the rows of a matrix to obtain the prescribed row sums, then scale the columns of the resulting matrix to obtain the prescribed column sums, and so on until convergence. The algorithm has been extended in many works (cf. Brualdi et al. 1966; Hartfiel 1971; Eaves et al. 1985; Marshall and Olkin 1968; Knight 2008; Rothblum and Schneider 1989), and its geometric rate of convergence and validity for irreducible nonnegative matrices have been established (Franklin and Lorenz 1989). We will use the following version of the algorithm (Sinkhorn 1967), which, given a $d$-dimensional probability vector $\pi$ with all entries being positive, samples a positive matrix $A$ whose row sum and column sum vectors are both equal to $\pi$ up to a prescribed precision level. Consequently, the final output $P$ is a transition matrix with stationary distribution equal to $\pi$ (Hartfiel 1974).
Sinkhorn’s Algorithm

Input: $\pi \in \text{Interior}(\Delta_d)$
Sample $A$ uniformly from the set of $d \times d$ transition matrices. ($A$ is positive w.p. 1.)
While $\|A1_d - \pi\| \geq \epsilon$ or $\|1_t A - \pi^t\| \geq \epsilon$

\[
D^L \leftarrow [\text{diag}(A1_d)]^{-1}\text{diag}(\pi)
\]
\[
A \leftarrow D^L A
\]
\[
D^R \leftarrow [\text{diag}(1_t A)]^{-1}\text{diag}(\pi)
\]
\[
A \leftarrow A D^R
\]

Return $P = \text{diag}(\pi)^{-1} A$

Sinkhorn’s algorithm partitions the set of irreducible nonnegative matrices into equivalent classes, each consisting of matrices that converge to the same transition matrix with stationary distribution $\pi$. However, since there is no closed form characterization of the equivalence, the sampling distribution of the algorithm in general is unknown. For the special case of $\pi = 1_d/d$, the sampling distribution is in fact not uniform but “locally flat” at the center of the polytope of doubly stochastic matrices (Cappellini et al. 2009).

Despite its intractable sampling distribution, Sinkhorn’s algorithm has several advantages. First, it allows prescribing the stationary distribution of $P$ exactly. Second, it is computationally tractable due to its scalability to high dimensions, geometric convergence, and simplicity. Third, it allows generation of irreducible transition matrices containing one or more 0’s by starting with a nonnegative matrix with 0’s in exactly the same entries (Brualdi et al. 1966). Consequently, it is applicable to simulation of Markov chains on strongly connected finite directed graphs. It is known that under the uniform distribution on $\Delta_d$, all the eigenvalues of a $d \times d$ transition matrix, except 1, converge in probability to 0 as $d \to \infty$ (Goldberg and Neumann 2003; Bordenave et al. 2012). Our numerical experiments indicate that this is also the case under the sampling distribution of Sinkhorn’s algorithm.

To specify the size of the leading eigenmoduli of $P$, choose $\lambda \in [0, 1)$ as a lower bound on $\varrho_2(P)$ and $\mu \in [0, 1)$, $\mu < \lambda$ as a lower bound on $\varrho_3(P)$. For $\lambda = 0$, we simply set $P = 1\pi^t$, while for $\lambda \geq 0$ and $\mu \geq 0$ we repeatedly sample $P$ by Sinkhorn’s algorithm until $\varrho_2(P) \geq \lambda$ and $\varrho_3(P) \geq \mu$.

4.3 Comparing Bayes BH and BH Procedures

After $\pi$ and $P$ are sampled and fixed, the simulation proceeds in two steps. First, one realization of $\eta$ is sampled and used to estimate $E(\eta_0)$, $E(\eta_i | \eta_0)$, and $E(\eta_i \eta_j | \eta_0)$. We denote this realization by $\theta$. Then, an independent realization of $\eta$, still denoted by $\eta$, is sampled and used to generate data $X = \epsilon \eta + Z$, where the entries of $Z$ are i.i.d. $\sim N(0, 1)$ and $\epsilon > 0$ is the signal strength. In the simulation, $\eta$ can be viewed as a signal that needs to be sent through a noisy environment. To do this, we first estimate the moments and conditional moments of the signal up to order two using a sample $\theta$. Meanwhile, we estimate the distribution of the noise $Z$, for example, by sending a long sequence of 0’s. In principle, we can have very good knowledge about the noise environment by repeatedly testing it. Then, based on the estimates, multiple testing is used to recover new signals sent through the environment.

Given $X$, $P(\eta_t = 1 | X)$ are approximately evaluated using the result in Example 3.4. To reduce computation, only $E(\eta_s | \eta_t)$ and $E(\eta_s \eta_u | \eta_t)$ with $|s - t| \leq w$ and $|u - t| \leq w$ are evaluated, where the “half window length” $w \geq 0$ determines the number of moments that
need to be incorporated in the approximation. All the other moments are treated as 0. In the absence of long range dependence, \( w \) can be small. The approximated \( P(\eta_t = 1 \mid X) \) are then used by the Bayes BH procedure. On the other hand, the marginal one-sided \( p \)-values based on individual \( X_t \) are used by the BH procedure.

Let \( \alpha \) be the nominal FDR control level for both the Bayes BH and the BH procedures. The Bayes BH procedure can estimate the population proportion of false nulls using the observed sample \( \theta \). However, the BH procedure does not rely on \( \theta \). To level the ground of comparison, the nominal FDR control level for the BH procedure is increased to

\[
\tilde{\alpha} = \alpha/(1 - \tilde{\psi}) \quad \text{with} \quad \tilde{\psi} = (\eta_1 + \ldots + \eta_m)/m,
\]

i.e., \( \tilde{\psi} \) is the actual proportion of false nulls underlying the data \( X = \epsilon \eta + Z \). It is well known that the BH procedure under the augmented FDR control level is more powerful while still controlling the FDR at the nominal level (Benjamini and Hochberg 1995). In contrast, the Bayes BH procedure has no access to statistics of \( \eta \) or \( Z \). It uses \( \hat{\psi} = m^{-1} \sum \theta_t \) as the estimate of the proportion of false nulls, and derives all the estimates of moments and conditional moments from the observed sample \( \theta \).

For a single instance of multiple testing, its performance can be measured by the FDP and the number of true discoveries (NTD). Both measures are random variables as they are functions of the data as well as the procedure being used. Their population means and variances can be used to assess the overall performance of a multiple testing procedure, in particular, its validity, power, and stability. To estimate these parameters, the Bayes BH and the BH procedures are applied to multiple replications of data, all of which are generated with the same \( \eta \) but with independent realizations of noise. To be specific, given \( \eta \), the following steps are taken.

1. Sample \( X = \epsilon \eta + Z \). At nominal FDR control level \( \alpha \), apply the Bayes BH procedure to the approximated \( P(\eta_t = 1 \mid X) \). On the other hand, at the augmented nominal FDR control level \( \tilde{\alpha} \), apply the BH procedure to the marginal \( p \)-values of \( X_t \) under the null \( \eta_t = 0 \).
2. Repeat the above step \( n = 1000 \) times. Compare the FDP and NTD of the Bayes BH procedure and those of the BH procedure, respectively, in terms of sample mean, sample standard error (SE), and empirical density.

### 4.4 Results

First, the main parameters used in the simulations are as follows; see Sections 4.1–4.3 for detail. The proportion of false nulls \( \psi = P(\eta_t = 1) \) takes values in \{0.05, 0.1\}, the signal strength \( \epsilon \) in \{0.75, 1, 1.25\}, and the nominal FDR control level \( \alpha \) in \{0.1, 0.2, 0.3, 0.4, 0.5\}. For the purpose of these numerical experiments, an eigenmodulus of about 0.5 to 0.6 is considered to be moderate, an eigenmodulus of larger than 0.9 is considered to be large and an eigenmodulus smaller than 0.2 is considered to be small. Given half window size \( w \), only \( E(\eta_s \mid \eta_t), E(\eta_s \eta_u \mid \eta_t) \) with \(|s - t| \leq w\) and \(|u - t| \leq w\) are used to approximate \( P(\eta_t = 1 \mid X) \), where \( \eta = \tau(M) \) and \( X = \epsilon \eta + Z \). If \( \varrho_2(P) > 0 \), then \( w = 3 \). On the other hand, if \( \varrho_2(P) = 0 \), i.e., \( \eta_t \) are i.i.d. \( \pi \), then \( w = 0 \). The length of \( \eta \) is \( m = 10^5 \). Sample mean, sample SE, and empirical density of the outcomes of multiple testing are based on \( n = 1000 \) replications of \( X \) with \( \eta \) being fixed.
### 4.4.1 Strong Dependence

Table 1 displays the results of a set of simulations in which $\varrho_2(P)$ and $\varrho_3(P)$ are large and small respectively. The parent Markov chain $M$ has $\varrho_2(P) = 0.9074$ and $\varrho_3 = 0.0645$,

$$
P = \begin{pmatrix}
0.9286 & 0.0122 & 0.0276 & 0.0063 & 0.0253 \\
0.0150 & 0.8548 & 0.0651 & 0.0310 & 0.0340 \\
0.0250 & 0.9629 & 0.0097 & 0.0008 & 0.0016 \\
0.1653 & 0.7046 & 0.0955 & 0.0267 & 0.0079 \\
0.2716 & 0.6988 & 0.0257 & 0.0007 & 0.0033 \\
\end{pmatrix},
$$

(14)

$\pi = (0.3006, 0.5994, 0.0505, 0.0211, 0.0283)^t$, and $\psi = 0.1$. Both the Bayes BH and the BH procedures control the FDR around the nominal level $\alpha$, with the former having somewhat smaller variance of the FDP. On the other hand, while both procedures have similar average NTD, the Bayes BH procedure has substantially less variation in the NTD. Thus, the approximate conditional probabilities provide better stability than the marginal $p$-values. In spite of the large $\varrho_2(P)$, the half window length $w = 3$ seems to work well.

| $\epsilon$ | $\alpha$ | $\hat{\mu}_{\text{FDP}}$ | $\text{SE}(\hat{\mu}_{\text{FDP}})$ | $\hat{\mu}_{\text{NTD}}$ | $\text{SE}(\hat{\mu}_{\text{NTD}})$ |
|---|---|---|---|---|---|
| 0.75 | 0.4 | 0.41246 | 0.40276 | 0.069156 | 0.12349 |
| 0.75 | 0.5 | 0.50544 | 0.49928 | 0.030706 | 0.051165 |
| 1 | 0.1 | 0.0469 | 0.0999 | 0.1682 | 0.1633 |
| 1 | 0.15 | 0.1478 | 0.1465 | 0.1321 | 0.1324 |
| 1 | 0.2 | 0.2095 | 0.2043 | 0.0672 | 0.0995 |
| 1 | 0.2 | 0.2034 | 0.2010 | 0.0719 | 0.0968 |
| 1 | 0.3 | 0.3085 | 0.3027 | 0.0325 | 0.0422 |
| 1.5 | 0.1 | 0.10743 | 0.0994 | 0.016124 | 0.017342 |
| 1.5 | 0.2 | 0.20871 | 0.19912 | 0.010182 | 0.012625 |

| $\epsilon$ | $\alpha$ | $\hat{\mu}_{\text{NTD}}$ | $\text{SE}(\hat{\mu}_{\text{NTD}})$ |
|---|---|---|---|
| 0.75 | 0.4 | 29.259 | 34.002 | 6.6888 | 23.973 |
| 0.75 | 0.5 | 134.2 | 131.19 | 1.4192 | 57.623 |
| 1 | 0.1 | 1.0530 | 4.9680 | 1.4192 | 5.1406 |
| 1 | 0.15 | 8.0090 | 12.5270 | 3.8180 | 10.5292 |
| 1 | 0.2 | 28.7990 | 30.8470 | 6.6985 | 18.7630 |
| 1 | 0.2 | 24.7470 | 30.1690 | 6.1466 | 18.8499 |
| 1 | 0.3 | 137.0970 | 132.3660 | 13.4961 | 43.9223 |
| 1.5 | 0.1 | 358.46 | 322.19 | 3.8180 | 10.5292 |
| 1.5 | 0.2 | 1096.5 | 1028.1 | 43.9223 | 18.8499 |

Here $\varrho_2(P)$ is large and $\varrho_3(P)$ is small. The transition matrix of the parent Markov chain is Eq. 14 with $\varrho_2(P) = 0.9074$, $\varrho_3(P) = 0.0645$ and $\psi = 10\%$.
Table 2  Comparison of the Bayes BH and BH procedures in the presence of strong dependence.

|   | BBH | BH | BBH | BH |
|---|-----|----|-----|----|
| $\epsilon$ | $\alpha$ | $\hat{\mu}_{\text{FDP}}$ | $\text{SE}(\hat{\mu}_{\text{FDP}})$ | $\hat{\mu}_{\text{NTD}}$ | $\text{SE}(\hat{\mu}_{\text{NTD}})$ |
| 0.75 | 0.3 | 0.29201 | 0.3071 | 0.20283 | 0.20708 |
| 0.75 | 0.4 | 0.3953 | 0.39839 | 0.066775 | 0.13137 |
| 1 | 0.2 | 0.21889 | 0.20237 | 0.057124 | 0.096743 |
| 1 | 0.2 | 0.21572 | 0.19602 | 0.056001 | 0.094159 |
| 1 | 0.25 | 0.25369 | 0.2529 | 0.040186 | 0.062352 |

$\varrho_2(P)$ and $\varrho_3(P)$ are again large and small respectively. The transition matrix of the parent Markov chain is $\varrho_2 = 0.9566$, $\varrho_3 = 0.1155$ and $\psi = 10\%$.

Table 2 displays the results of another set of simulations in which $\varrho_2(P)$ is even larger, but still shows similar phenomenon observed in Table 1. In the simulations, $\varrho_2(P) = 0.9566$,

$$P = \begin{pmatrix}
0.8993 & 0.0017522 & 0.0098301 & 0.031722 & 0.057401 \\
0.015641 & 0.96471 & 0.0086498 & 0.0077613 & 0.0032366 \\
0.66064 & 0.11572 & 0.075542 & 0.067761 & 0.080337 \\
0.71019 & 0.090088 & 0.06356 & 0.13104 & 0.0051248 \\
0.57536 & 0.025292 & 0.27334 & 0.10032 & 0.025692
\end{pmatrix}$$

with $\pi = (0.67187, 0.22813, 0.02414, 0.033354, 0.042506)^t$, and $\psi = 0.1$.

In addition to summary statistics such as sample mean and sample SE, density curves also show better stability of the Bayes BH procedure. Figure 1 displays the kernel smoothing density curves of the FDP and the NTD from several additional sets of simulations with strong dependence (transition matrices not shown). Consistent with the above results, the distribution of the NTD of the Bayes BH procedure is substantially more concentrated than that of the BH procedure.

4.4.2 Additional Effects of Dependence Structure on Multiple Testing

Table 3 displays the results of a set of simulations in which the dependence of the hypotheses is strong with $\varrho_2(P) = 0.9591$. However, the third largest eigenmodulus, i.e. $\varrho_3(P)$, is also sampled to be relatively large at 0.5232. This makes the sampling of $P$ more difficult. In contrast, in the simulations of Section 4.4.1, there is no control on $\varrho_3(P)$ and as a result, $\varrho_3(P)$ is small. In most cases, the value is less than 0.3. For example, for the transition
matrices in Eq. 14 and Eq. 15, the value of $\varphi_3(P)$ is 0.0645 and 0.1155, respectively. The transition matrix of the parent Markov chain for Table 3 is

$$P = \begin{pmatrix}
0.0014 & 0.9653 & 0.0124 & 0.0080 & 0.0129 \\
0.9506 & 0.0011 & 0.0056 & 0.0155 & 0.0272 \\
0.3236 & 0.1203 & 0.5312 & 0.0238 & 0.0011 \\
0.8080 & 0.1589 & 0.0190 & 0.0025 & 0.0115 \\
0.1898 & 0.7959 & 0.0053 & 0.0024 & 0.0066
\end{pmatrix}$$

(16)
Table 3  Comparison of the Bayes BH and BH procedures in the presence of strong dependence with large $\varrho_2$ and a relatively large $\varrho_3$.

| $\epsilon$ | $\alpha$ | $\hat{\mu}_{FDP}$ | SE($\hat{\mu}_{FDP}$) | $\hat{\mu}_{NTD}$ | SE($\hat{\mu}_{NTD}$) |
|---|---|---|---|---|---|
| 1 | 0.2 | 0.29271 | 0.19 | 0.10223 | 0.23451 |
| 1 | 0.3 | 0.3676 | 0.30591 | 0.06527 | 0.19306 |
| 1 | 0.4 | 0.47105 | 0.40827 | 0.039127 | 0.11613 |
| 1.25 | 0.2 | 0.35611 | 0.1958 | 0.032559 | 0.086935 |
| 1.25 | 0.356 | 0.42145 | 0.35488 | 0.02103 | 0.041823 |

The transition matrix of the parent Markov chain is Eq. 16 with $\varrho_2 = 0.9591$, $\varrho_3 > 0.5$, and $\psi = 5\%$.

with $\pi = (0.47319, 0.47681, 0.018907, 0.01173, 0.019363)^t$ and $\psi = 0.05$. Note that the proportion of false nulls is only half as large as the one in Section 4.4.1, making error control more difficult. Table 3 shows that in this case, even though the Bayes BH procedure has a harder time to control the FDR, it has substantially less variation in the FDP than the BH procedure does. Further, the power of the Bayes BH procedure is clearly superior to the BH procedure. For example, at a realized FDP $\approx 0.35$, the former makes about 204 true discoveries whereas the latter makes about 149. The Bayes BH procedure also has significantly lower SE-to-mean ratio of the NTD than the BH procedure. The phenomenon was repeatedly observed when $\varrho_3(P)$ was sampled to be large. Figure 2 provides examples of the phenomenon. In contrast to Fig. 1, the FDP density curves of the two procedures exhibit pronounced differences, and their NTD density curves are far apart instead of overlapping with each other.

As noted in the introduction, it is known that the BH procedure is often able to control the FDR under dependence, and certain properties of dependence are sufficient for this to happen (Wu 2008; Benjamini and Yekutieli 2001; Sarkar 2002; 2006; Clarke and Hall 2009). Much less is known about the relationship between properties of dependence and other aspects of performance of multiple testing, such as stability and power. One advantage of random sampling from a large space of dependence structures is that it enables experimental investigation of the relationship. The properties of statistical dependence as characterized by $\varrho_2(P)$ and $\varrho_3(P)$ have been the focus here. The above results indicate that $\varrho_2(P)$ alone is not sufficient to characterize the relative performances of the Bayes BH and the BH procedures. This numerical observation is consistent with what is now known about the mixing times of Markov chains (Diaconis 1995; Levin et al. 2009). The results in Table 3 and Fig. 2 indicate that $\varrho_3(P)$ may sometimes play a role in shaping the differences.
In our simulations, $P$ is a $5 \times 5$ matrix. Thus $\varrho_2(P)$ and $\varrho_3(P)$ together provide a significant part of information on the spectrum of the transition matrix. This points to a complex relationship between the dependence structure of the underlying hypotheses and the relative performances of the Bayes BH and the BH procedures.
Table 4 Comparison of the Bayes BH and BH procedures in the presence of moderate dependence.

| $\epsilon$ | $\alpha$ | $\hat{\mu}_{\text{FDP}}$ | SE($\hat{\mu}_{\text{FDP}}$) | $\hat{\mu}_{\text{NTD}}$ | SE($\hat{\mu}_{\text{NTD}}$) |
|---|---|---|---|---|---|
| 0.75 | 0.3 | 0.29201 | 0.3071 | 0.20283 | 0.20708 |
| 0.75 | 0.4 | 0.22492 | 0.3014 | 0.2719 | 0.2124 |
| 0.75 | 0.5 | 0.4966 | 0.50118 | 0.03134 | 0.045887 |
| 1 | 0.2 | 0.192 | 0.2012 | 0.075112 | 0.084956 |
| 1 | 0.3 | 0.294 | 0.30074 | 0.035074 | 0.040486 |

The transition matrix of the parent Markov chain is Eq. 17 with $\varrho_2 = 0.6019$ and $\psi = 10\%$

4.4.3 Moderate Dependence

Table 4 displays results for the case of moderate dependence. In this set of simulations, $\varrho_2(P) = 0.6019$,

$$P = \begin{pmatrix} 0.1999 & 0.6548 & 0.0537 & 0.0561 & 0.0354 \\ 0.8932 & 0.0765 & 0.0137 & 0.0088 & 0.0076 \\ 0.4431 & 0.4365 & 0.0721 & 0.0284 & 0.0199 \\ 0.1159 & 0.8271 & 0.0371 & 0.0129 & 0.0070 \\ 0.6080 & 0.0597 & 0.0505 & 0.0710 & 0.2108 \end{pmatrix}$$

with $\pi = (0.4969, 0.4031, 0.0376, 0.0349, 0.0275)^t$ and $\psi = 0.1$. From the table, the NTD of the Bayes BH procedure is clearly more stable than that of the BH procedure. This can also be seen from the densities of the NTD shown in Fig. 3. To a lesser extent, the FDP

![Fig. 3](image-url) Densities of the FDP and NTD of the Bayes BH and BH procedures in the presence of moderate dependence of hypotheses.
Table 5 Comparison of the Bayes BH and BH procedures when there is no dependence in data

| $\epsilon$ | $\alpha$ | $\hat{\mu}_{FDP}$ | SE($\hat{\mu}_{FDP}$) | $\hat{\mu}_{NTD}$ | SE($\hat{\mu}_{NTD}$) |
|------------|----------|---------------------|------------------------|---------------------|------------------------|
| 1          | 0.3      | 0.29028             | 0.29932                | 0.19437             | 0.18847                |
| 1          | 0.4      | 0.39102             | 0.39308                | 0.081393            | 0.123                  |
| 1          | 0.5      | 0.50282             | 0.50157                | 0.039488            | 0.065857               |
| 1.25       | 0.2      | 0.19442             | 0.19951                | 0.072535            | 0.083292               |
| $\epsilon$ | $\alpha$ | $\hat{\mu}_{FDP}$ | SE($\hat{\mu}_{FDP}$) | $\hat{\mu}_{NTD}$ | SE($\hat{\mu}_{NTD}$) |
| 1          | 0.3      | 5.322               | 11.16                  | 2.9969              | 9.6522                 |
| 1          | 0.4      | 22.881              | 32.613                 | 5.9599              | 19.899                 |
| 1          | 0.5      | 79.951              | 87.779                 | 10.701              | 36.661                 |
| 1.25       | 0.2      | 24.838              | 31.121                 | 5.9329              | 15.893                 |

of the Bayes BH procedure is also more stable than that of the BH procedure except when $\epsilon = 0.75$, $\alpha = 0.3$ or 0.4. In the latter situation, the variance of the FDP of the Bayes BH procedure is still competitive with that of the BH procedure.

Fig. 4 Densities of the FDP and NTD of the Bayes BH and BH procedures when hypotheses are independent
4.4.4 Independence

Finally, as a test on the validity of the Bayes BH procedure, Table 5 displays results when $\varrho_2(P) = 0$ and, as a result, the entries of $\eta$ are i.i.d. In the simulations, $\psi = 0.05$ and $P = 1_5 \pi^t$, where the stationary probability vector $\pi$ is randomly sampled as in Section 4.1. As can be seen, under independence, both the Bayes BH procedure and the BH procedure are valid (Fig. 4). In terms of power, although the NTD of the Bayes BH procedure has a smaller average than that of the BH procedure, its stability is clearly superior. Similar phenomenon can be seen from the density plot in Fig. 3.

5 Conclusion

We presented a flexible signal processing model of large scale multiple testing and used probability approximation to conduct multiple testing. This allowed us to empirically demonstrate that even when very little is known about the dependence structure of the hypotheses, substantial reduction in variance can be achieved by using approximated conditional probabilities in multiple testing. Using simulations designed to generate very general dependence structures, we provided evidence for the effectiveness of the local conditioning approach to FDP variance reduction and NTD variance reduction in large scale multiple testing. Conditional likelihood has been fruitfully exploited in multiple testing (Efron et al. 2001; Efron 2004; 2007). Indeed, the notion of Local False Discovery Rate introduced in Efron et al. (2001) is a conditional likelihood that does not directly borrow strength from nearby observations. In contrast, our approach explicitly incorporates nearby observations into the conditional likelihood. Our analysis indicates that the reduction in the standard deviations of the FDP and of the number of true discoveries can be substantial when conditional likelihood is used. In particular, for very noisy data, even when there is some increase in the FDR as compared to using $p$-values, the reduction in the standard deviation of FDP may justify a Bayesian approach.

The effects of dependence on large scale multiple tests has been a topic of research at least since the beginning of this century (Benjamini and Yekutieli 2001). Various special dependence structures between the hypotheses have been studied and continue to be studied to ascertain the effectiveness of both $p$-value based and Empirical Bayes methods (Benjamini and Yekutieli 2001; Wu 2008; Stephens 2017) in multiple testing. In addition, with some notable exceptions (Owen 2005; Fan et al. 2012; Qiu et al. 2005; Schwartzman and Lin 2011; Gordon et al. 2007; Li 2015), the standard deviations of the FDP and of the number of true discoveries have usually taken a second place to their means. However, given that it is often impractical to repeat a multiple test numerous times, not controlling the standard deviation of the number of true discoveries and of the FDP can lead to highly biased estimates of the FDP and the number of true discoveries. The need for simulating general dependence structures is crucial to testing the performance of our approach. This led us to the random HMM dependence structures that can be generated by using random Markov matrices using Sinkhorn’s algorithm.

Our approach strikes a compromise between generality and tractability by using the signal processing framework that models data as a function of signal, signal strength, i.i.d. noise and signal noise interaction. The flexibility can be useful in applications. For instance, additive noise is not the only signal-noise interaction that is seen in fMRI practice (Marx et al. 2013), so a different signal-noise interaction function may improve the analysis. Assuming a fixed signal noise interaction and a fixed noise distribution is not unrealistic in
many applications. The challenge is that they are usually unknown. Two important issues that we have not addressed here are modeling real data with this approach and obtaining a “noiseless” sample of the data for measuring moments. These two issues are intimately connected to the application in question. Obtaining a noiseless sample of the signals is not very difficult in a signal processing context. However, in DNA microarray analysis and in fMRI it remains a challenge to not just find a noiseless sample but also to arrive at the signal strength and the signal noise interaction. A possible solution might be to incorporate a parametric model to the underlying binary process that identifies false nulls.

Sometimes, the modeling framework developed here can be applied to real data even if the application does not render itself to a direct and explicit signal noise interaction formulation. For instance, if a real data application allows the estimation of conditional densities of observations given the null and the alternate, there may be no need to hypothesize an explicit noise distribution. For example, the epidemic model in Sun and Cai (2009) estimates null and alternate densities using a parametric model. A “noiseless” sample from the known history of the epidemic can be used for measuring conditional moments. Such data provide an avenue for extending the application of our modeling framework beyond simulated data. Our next goal is the development of such ideas to successfully apply this method to suitable multiple testing applications.

References

Barvinok A (2010) What does a random contingency table look like? Combin Probab Comput 19(4):517–539
Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300
Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Statist 29(4):1165–1188
Bordenave C, Caputo P, Chafaï D (2012) Circular law theorem for random Markov matrices. Probab Theory Related Fields 152(3–4):751–779
Brualdi RA, Parter SV, Schneider H (1966) The diagonal equivalence of a nonnegative matrix to a stochastic matrix. J Math Anal Appl 16:31–50
Cappellini V, Sommers H-J, Bruzda W, Życzkowski K (2009) Random bistochastic matrices. J Phys A 42(36):365209–23
Chafaï D (2010) The Dirichlet Markov ensemble. J Multivariate Anal 101(3):555–567
Chatterje S, Diaconis P, Sly A (2010) Properties of uniform doubly stochastic matrices. ArXiv e-prints
Cheng D, Schwartzman A (2017) Multiple testing of local maxima for detection of peaks in random fields. Ann Statist 45(2):529–556
Chi Z (2011) Effects of statistical dependence on multiple testing under a hidden Markov model. Ann Statist 39(1):439–473
Clarke S, Hall P (2009) Robustness of multiple testing procedures against dependence. Ann Statist 37(1):332–358
Diaconis P (1995) The cutoff phenomenon in finite Markov chains. Proc Natl Acad Sci USA 93:1659–1664
Diaconis P, Lebeau G, Michel L (2012) Gibbs/Metropolis algorithms on a convex polytope. Math Z 272(1–2):109–129
Eaves BC, Hoffman AJ, Rothblum UG, Schneider H (1985) Line-sum-symmetric scalings of square nonnegative matrices. Math Programming Stud (25): 124–141. Mathematical programming, II
Efron B (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Amer Statist Assoc 99(465):96–104
Efron B (2007) Correlation and large-scale simultaneous significance testing. J Amer Statist Assoc 102(477):93–103
Efron B (2010) Correlated z-values and the accuracy of large-scale statistical estimates. J Amer Statist Assoc 105(491):1042–1055
Efron B, Tibshirani R, Storey JD, Tusher VG (2001) Empirical Bayes analysis of a microarray experiment. J Amer Statist Assoc 96(456):1151–1160
Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. Proc Natl Acad Sci 113(28):7900–7905
Fan J, Han X (2017) Estimation of the false discovery proportion with unknown dependence. J R Stat Soc Ser B 79(4):1143–1164
Fan J, Han X, Gu W (2012) Estimating false discovery proportion under arbitrary covariance dependence. J Amer Statist Assoc 107(499):1019–1035
Farcomeni A (2008) A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Stat Methods Med Res 17(4):347–388
Franklin J, Lorenz J (1989) On the scaling of multidimensional matrices. Linear Algebra Appl 114/115:717–735
Genovese C, Wasserman L (2004) A stochastic process approach to false discovery control. Ann Statist 32(3):1035–1061
Goldberg G, Neumann M (2003) Distribution of subdominant eigenvalues of matrices with random rows. SIAM J Matrix Anal Appl 24(3):747–761. (electronic)
Gordon A, Glazko G, Qiu X, Yakovlev A (2007) Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. Ann Appl Statist 1(1):179–190
Hartfiel DJ (1971) Concerning diagonal similarity of irreducible matrices. Proc Amer Math Soc 30:419–425
Hartfiel DJ (1974) A study of convex sets of stochastic matrices induced by probability vectors. Pacific J Math 52:405–418
Knight PA (2008) The Sinkhorn-Knopp algorithm: convergence and applications. SIAM J Matrix Anal Appl 30(1):261–275
Künsch H, Geman S, Kehagias A (1995) Hidden Markov random fields. Ann Appl Probab 5(3):577–602
Leek JT, Storey JD (2008) A general framework for multiple testing dependence. Proc Natl Acad Sci USA 105(48):18718–18723
Levin DA, Peres Y, Wilmer EL (2009) Markov chains and mixing times. American Mathematical Society, Providence, RI. With a chapter by James G. Propp and David B. Wilson
Li D (2015) Power and stability comparisons of multiple testing procedures with false discovery rate control. J Stat Comput Simulation 85(14):2808–2822
Lindquist MA, Mejia A (2015) Zen and the art of multiple comparisons. Psychosom Med 77(2):114–125
Liu J, Zhang C, Page D (2016) Multiple testing under dependence via graphical models. Ann Appl Stat 10(3):1699–1724
Marshall AW, Olkin I (1968) Scaling of matrices to achieve specified row and column sums. Numer Math 12:83–90
Marx M, Pauly KB, Chang C (2013) A novel approach for global noise reduction in resting-state fMRI: APPLCOR. Neuroimage 64:19–31
Nguyen HD, McLachlan GJ, Cherbuin N, Janke AL (2014) False discovery rate control in magnetic resonance imaging studies via Markov random fields. IEEE Trans Med Imaging 33(8):1735–1748
Owen AB (2005) Variance of the number of false discoveries. J R Stat Soc Ser B 67(3):411–426
Qiu X, Klebanov L, Yakovlev A (2005) Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. Stat Appl Genet Mol Biol, De Gruyter 4(1):1–32. https://ideas.repec.org/a/bpj/sagmbi/v4y2005i1n34.html
Roquain E (2011) Type I error rate control for testing many hypotheses: a survey with proofs. Journal de la Societe Francaise de Statistique 152(2):3–38
Roquain E, Villers F (2011) Exact calculations for false discovery proportion with application to least favorable configurations. Ann Statist 39(1):584–612
Rosenthal JS (1995) Convergence rates for Markov chains. SIAM Rev 37(3):387–405
Rothblum UG, Schneider H (1989) Scalings of matrices which have prespecified row sums and column sums via optimization. Linear Algebra Appl 114/115:737–764
Sarkar SK (2002) Some results on false discovery rate in stepwise multiple testing procedures. Ann Statist 30(1):239–257
Sarkar SK (2006) False discovery and false non-discovery rates in single-step multiple testing procedures. Ann Statist 34(1):394–415
Sarkar SK, Zhou T, Ghosh D (2008) A general decision theoretic formulation of procedures controlling FDR and FNR from a Bayesian perspective. Statistica Sinica 18:925–945
Schwartzman A, Gavrilo Y, Adler RJ (2011) Multiple testing of local maxima for detection of peaks in 1D. Ann Statist 39(6):3290–3319
Schwartzman A, Lin X (2011) The effect of correlation in false discovery rate estimation. Biometrika 98(1):199–214
Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73(3):751–754

Springer
Sinkhorn R (1964) A relationship between arbitrary positive matrices and doubly stochastic matrices. Ann Math Statist 35:876–879
Sinkhorn R (1967) Diagonal equivalence to matrices with prescribed row and column sums. Amer Math Monthly 74:402–405
Stephens M (2017) False discovery rates: a new deal. Biostatistics 18(2):275–294
Storey JD, Taylor JE, Siegmund DO (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J R Stat Soc Ser B 66(1):187–205
Sun W, Cai TT (2009) Large-scale multiple testing under dependence. J R Stat Soc Ser B 71(2):393–424
Wu W-B (2008) On false discovery control under dependence. Ann Statist 36(1):364–380

Publisher’s Note  Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.