Multiple Testing under Dependence via Semiparametric Graphical Models

Jie Liu JIELIU@CS.WISC.EDU
Department of Computer Sciences, University of Wisconsin-Madison

Chunming Zhang CMZHANG@STAT.WISC.EDU
Department of Statistics, University of Wisconsin-Madison

Elizabeth Burnside EBURNSIDE@UWHEALTH.ORG
Department of Radiology, University of Wisconsin-Madison

David Page PAGE@BIOSTAT.WISC.EDU
Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison

Abstract

It has been shown that graphical models can be used to leverage the dependence in large-scale multiple testing problems with significantly improved performance (Sun & Cai, 2009; Liu et al., 2012). These graphical models are fully parametric and require that we know the parameterization of f_1 — the density function of the test statistic under the alternative hypothesis. However in practice, f_1 is often heterogeneous, and cannot be estimated with a simple parametric distribution. We propose a novel semiparametric approach for multiple testing under dependence, which estimates f_1 adaptively. This semiparametric approach exactly generalizes the local FDR procedure (Efron et al., 2001) and connects with the BH procedure (Benjamini & Hochberg, 1995). A variety of simulations show that our semiparametric approach outperforms classical procedures which assume independence and the parametric approaches which capture dependence.

1. Introduction

High-throughput computational biology studies, such as gene expression analysis and genome-wide association studies, often involve large-scale multiple testing problems which exhibit dependence in the sense that whether the null hypothesis of one test is true or not depends on the ground truth of other tests. Recently, new multiple testing procedures have been proposed with such dependence explicitly captured by graphical models such as hidden Markov models (Sun & Cai, 2009) and Markov-random-field-coupled mixture models (Liu et al., 2012). These graphical models are fully parametric, and they assume that we know not only the parameterization form of f_0, but also the parameterization form of f_1. [1] Eventually, a fully parametric graphical model is learned, and the multiple testing problem becomes an inference problem on the graphical model. This parametric approach is effective in some simple situations, but the assumptions for f_1 often make it impractical, as discussed next.

A long tradition in hypothesis testing is to derive test statistics and calculate P-values all under the null hypothesis $\mathcal{H}_0$. The distribution of the test statistic under $\mathcal{H}_1$ sometimes can be difficult to derive. Take for instance a two-proportion z-test, which tests whether two Bernoulli variables have the same parameter (i.e. $P(\text{head})$ in coin-flippings); the two-proportion z-test is widely used in case-control studies (e.g. comparing the minor allele frequencies in cases and controls). Under $\mathcal{H}_0$ (the two proportions are the same), the test statistic X asymptotically follows a standard normal $\mathcal{N}(0,1)$. Under $\mathcal{H}_1$ (the two proportions are different), X asymptotically follows a standardized non-centered normal $\mathcal{N}(\mu,1)$ ($\mu \neq 0$) where μ depends on the odds-ratio of this genetic marker. When there are multiple genetic markers to be tested, f_0 remains $\mathcal{N}(0,1)$, but f_1 becomes a mixture of Gaussians because these associated markers can have different odds-ratios and therefore

[1] f_0 and f_1 are the probability density functions of the test statistic under the null hypothesis $\mathcal{H}_0$ and the alternative hypothesis $\mathcal{H}_1$, respectively. In the HMM model (Sun & Cai, 2009) and the MRF-coupled mixture model (Liu et al., 2012), f_0 and f_1 are the emitting probabilities for state 0 and state 1 respectively.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

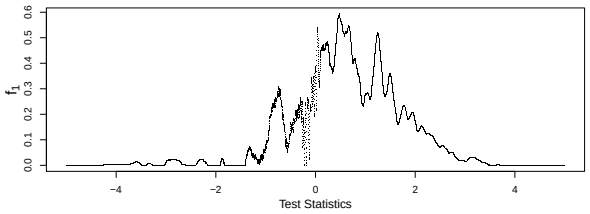


Figure 1. Estimated f_1 in a real-world genome-wide association study on breast cancer.

different μ values (i.e. different effect sizes). In this situation, f_1 is no longer a simple parametric distribution. In a real-world genome-wide association study on breast cancer, we plot the estimated f_1 in Figure 1; obviously it is inappropriate to estimate f_1 with a simple parametric distribution. Note that this is not a problem for classical multiple testing procedures such as the BH procedure, whose calculations of P-values are done under $\mathcal{H}_0$, but this is a serious problem for the graphical-model-based procedures which require f_1 to be estimated parametrically. Therefore, the key question is *whether we can still make use of the graphical models to leverage the dependence among the hypotheses without making assumptions about* f_1.

In this paper, we propose a semiparametric graphical model to leverage the dependence among the hypotheses. In our model, f_1 is estimated nonparametrically and the remaining parts are estimated parametrically. More algorithmic details are introduced in Section 3 after we summarize the terminology in Section 2. Section 4 shows that the two widely-used multiple testing procedures, the BH procedure (Benjamini & Hochberg, 1995) and the local FDR procedure (Efron et al., 2001), estimate their parameters in the same semiparametric way to avoid assumptions about f_1. This unification demonstrates that the most appropriate way of using graphical models to capture the dependence is the semiparametric model in our paper rather than the fully parametric models (Sun & Cai, 2009; Liu et al., 2012). Simulations in Section 5 show that our semiparametric approach controls false discovery rate and reduces false non-discovery rate, compared with the baseline procedures. We apply the procedure to a real-world genome-wide association study on breast cancer in Section 6 and identify a number of genetic variants.

2. Preliminaries

FDR, FNR, Validity and Efficiency: When we test m hypotheses simultaneously, various outcomes can be described by Table 1 based on their ground truth and whether the hypotheses are rejected. *False discovery rate* (FDR), $E(N_{10}/R|R{>}0)P(R{>}0)$, is the expected proportion of incorrectly rejected null hypotheses (Benjamini & Hochberg, 1995). *False non-discovery rate* (FNR),

Table 1. Classification of tested hypotheses

	RETAINED	REJECTED	TOTAL
$\mathcal{H}_0$ IS TRUE	N_{00}	N_{10}	m_0
$\mathcal{H}_0$ IS FALSE	N_{01}	N_{11}	m_1
TOTAL	S	R	m

$E(N_{01}/S|S{>}0)P(S{>}0)$, is the expected proportion of false non-rejections in those tests whose null hypotheses are not rejected (Genovese & Wasserman, 2002). An FDR procedure is *valid* if it controls FDR at a nominal level α. One valid procedure is more *efficient* than another if it has a smaller FNR. In multiple testing problems, we would like to control FDR at the nominal level and reduce FNR as much as possible.

Dependence in Multiple Testing: Classical multiple testing procedures usually assume independence among the hypotheses. The effects of dependence on multiple testing have been investigated with a focus on the validity issue, namely how to control FDR at the nominal level when dependence exists (Benjamini & Yekutieli, 2001; Finner & Roters, 2002; Reiner et al., 2003; Owen, 2005; Sarkar, 2006; Efron, 2007; Farcomeni, 2007; Romano et al., 2008; Strimmer, 2008; Wu, 2008; Blanchard & Roquain, 2009). Despite FDR-control challanges, dependence also brings opportunities for decreasing FNR. This efficiency issue has been investigated (Yekutieli & Benjamini, 1999; Genovese et al., 2006; Benjamini & Heller, 2007; Zhang et al., 2011), indicating FNR could be decreased by leveraging the dependence among hypotheses. Several approaches have been proposed, such as dependence kernels (Leek & Storey, 2008), factor models (Friguet et al., 2009) and principal factor approximation (Fan et al., 2012). Sun & Cai (2009) use a hidden Markov model to explicitly leverage chain dependence structures (Sun & Cai, 2009). Liu et al. (2012) extend such graphical-model-based approaches to general dependence structures via a Markov-random-field-coupled mixture model. Capturing the dependence in multiple testing in such an explicit manner is innovative, but it relies on the strong assumption that we know the parameterization of f_1, which is unrealistic in all but the simplest situations. Improper assumption of f_1 may make the testing procedure too liberal, e.g. Figure 4 of Sun & Cai (2009), or conservative, e.g. Figure 3 of Liu et al. (2012). In this paper, we build on the approach of Liu et al. (2012) and take the major step of relaxing this assumption by estimating f_1 adaptively.

3. Methods

3.1. Graphical models for Multiple Testing

Let $\mathbf{x} = (x_1, ..., x_m)$ be a vector of test statistics from hypotheses $(\mathcal{H}_1, ..., \mathcal{H}_m)$ with their ground truth denoted by a latent Bernoulli vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_m) \in \{0, 1\}^m$,

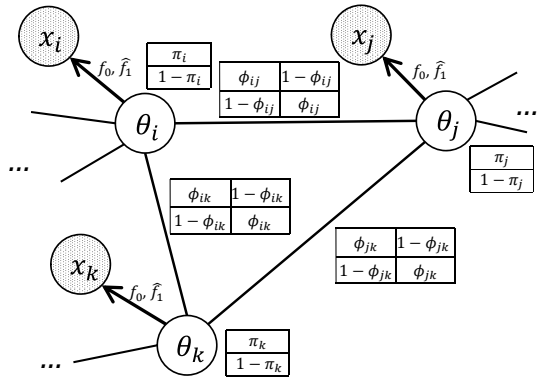


Figure 2. The semiparametric graphical model for hypotheses $\mathcal{H}_i$, $\mathcal{H}_j$ and $\mathcal{H}_k$ with observed test statistics (x_i, x_j, x_k) and latent ground truth (θ_i, θ_j, θ_k).

with $\theta_i = 0$ denoting that the hypothesis $\mathcal{H}_i$ is null and $\theta_i = 1$ denoting that the hypothesis $\mathcal{H}_i$ is non-null. In the work of Liu et al. (2012), the dependence among these hypotheses is represented as a binary Markov random field (MRF) on $\boldsymbol{\theta}$. The structure of the MRF is assumed to be known, and described by an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with the node set $\mathcal{V}$ and the edge set $\mathcal{E}$. The dependence between $\mathcal{H}_i$ and $\mathcal{H}_j$ is denoted by an edge connecting node_i and node_j. The strength of dependence is captured by a potential function (parametrized by ϕ_{ij}, $0<\phi_{ij}<1$) on this edge. The degree of prior belief that $\mathcal{H}_i$ is null is captured by the node potential function (parametrized by π_i, $0<\pi_i<1$). Suppose that the probability density function of the test statistic $x_i|\theta_i=0$ is f_0, and the density of $x_i|\theta_i=1$ is f_1. Then $(\mathbf{x}, \boldsymbol{\theta}; \boldsymbol{\pi}, \boldsymbol{\phi}, f_0, f_1)$ forms an MRF-coupled mixture model where $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ are node potential functions and edge potential functions in the MRF. In the MRF-coupled mixture model, $\mathbf{x}$ is observed, and $\boldsymbol{\theta}$ is hidden. We also need to estimate $\boldsymbol{\pi}$, $\boldsymbol{\phi}$ and f_1. [2]

For the reasons discussed in Section 1, it is often difficult to estimate f_1 with a simple parametric distribution. In order to avoid the f_1 assumption, we estimate f_1 adaptively via an indirect, nonparametric way, as introduced in Section 3.2. Then we estimate $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ via a contrastive divergence style algorithm, as introduced in Section 3.3. Therefore the graphical model is learned semiparametrically — f_1 is learned nonparametrically and the MRF part is learned by estimating parameters $\boldsymbol{\phi}$ and $\boldsymbol{\pi}$. Finally, we perform marginal inference of $\boldsymbol{\theta}|\mathbf{x}$ with the learned model and reject hypotheses with a step-up procedure to control FDR, as introduced in Section 3.4. Figure 2 [3] shows the semiparametric MRF-coupled mixture model for the three dependent hypotheses $\mathcal{H}_i$, $\mathcal{H}_j$ and $\mathcal{H}_k$.

[2] f_0 is usually known to us in hypothesis testing.

[3] We slightly modify Figure 1 of Liu et al. (2012).

3.2. Nonparametric Estimation of f_1

We cannot directly estimate f_1 from observed $\mathbf{x}$ because the ground truth $\boldsymbol{\theta}$ is hidden. However, we can estimate f from observed $\mathbf{x}$ nonparametrically via kernel density estimation. Therefore, we can estimate f_1 indirectly using the rule of total probability

$$f(x) = p_0 f_0(x) + (1 - p_0) f_1(x), \tag{1}$$

where p_0 is the proportion of null hypotheses. Since we know f_0 in advance (e.g. $\mathcal{N}(0, 1)$), we only need to estimate f and p_0 so as to estimate f_1.

Estimating p_0: We can estimate p_0 with the method in the work of Storey (2002), namely

$$\hat{p}_0(\lambda) = \frac{W(\lambda)}{(1 - \lambda)m}, \tag{2}$$

where $\lambda \in [0, 1)$ is a tuning parameter, and $W(\lambda)$ is the total number of hypotheses whose P-values are above λ. The motivation of this estimation is that the P-values of null hypotheses are uniformly distributed on the interval $(0, 1)$. If we assume all the hypotheses with P-values greater than λ are from null hypotheses, then $W(\lambda)/(1 - \lambda)$ is the total number of null hypotheses. Therefore the right hand side of (2) is an estimate of p_0. Obviously, $\hat{p}_0(\lambda)$ overestimates p_0 because there might be nonnull hypotheses whose P-values are greater than λ, especially when λ is small. Therefore, a bias-variance trade-off presents in the choice of λ — a larger λ value yields less bias but brings in more variance. Storey et al. (2004) showed that the BH procedure coupled with $\hat{p}_0(\lambda)$ maintains strong control of FDR under mild conditions. In simulations, we test different λ values, and the results show that the performance of our multiple testing procedure is insensitive to different choices of λ.

Estimating f: Since we can observe all the test statistics $\mathbf{x}$, we can estimate f directly via kernel density estimation (Rosenblatt, 1956). One may choose any kernel function and bandwidth parameter as long as they provide a reasonable estimate. A Gaussian kernel would be a natural choice. Nevertheless in our experiments, we use the Epanechnikov kernel because its computation burden is low, and it is optimal in a minimum variance sense (Epanechnikov, 1969). Finally we can get $\hat{f}$, the nonparametric estimate of f.

Estimating f_1: With the estimated $\hat{p}_0$ and $\hat{f}$, we estimate f_1 as

$$\hat{f}_1(x) = \frac{\hat{f}(x) - \hat{p}_0 f_0(x)}{1 - \hat{p}_0}. \tag{3}$$

3.3. Parametric Estimation of ϕ and π

The pairwise potential functions $\boldsymbol{\phi}$ and the node potential functions $\boldsymbol{\pi}$ parametrize the Markov random field part of the model. In the simulations, we tie all the pairwise potential functions together, i.e. $\boldsymbol{\phi}=\{\phi\}$. In the real-world application in Section 6, we assume there are three types of edges (high correlation edges, medium correlation edges and low correlation edges), and there are three parameters, $\boldsymbol{\phi}=\{\phi_h, \phi_m, \phi_l\}$, corresponding to the three levels of correlation. We also tie all the node potentials in both the simulations and the real-world application, i.e. $\boldsymbol{\pi}=\{\pi\}$.

Parameter learning for MRFs is generally difficult due to the partition function. So far, the state-of-the-art parameter learning algorithms are based on contrastive divergence (Hinton, 2002), such as the persistent contrastive divergence (PCD) algorithm (Tieleman, 2008). Contrastive divergence algorithms are iterative algorithms which gradually update parameters by generating particles based on current estimates of parameters and then comparing the moments from the particles with the moments from the data. Contrastive divergence is related to pseudo-likelihood (Besag, 1975) and ratio matching (Hyvärinen, 2007a;b). However, contrastive divergence algorithms cannot be directly applied to our model because $\boldsymbol{\theta}$ is hidden. Therefore, we modify the PCD algorithm as follows. Suppose we already generate particles for $\boldsymbol{\theta}$ in the normal PCD algorithm. We further generate the particles for $\mathbf{x}$ using f_0 and $\hat{f}_1$ conditional on the generated particles for $\boldsymbol{\theta}$. Then we update the parameters by comparing the moments from particles for $\mathbf{x}$ and the moments from the observed $\mathbf{x}$. One systematic review of learning parameters in hidden Markov random fields is in the prior work of Liu et al. (2014).

3.4. Inference of θ and FDR Control

After we estimate f_1, $\boldsymbol{\phi}$ and $\boldsymbol{\pi}$, the MRF-coupled mixture model is fully specified, and the next importance step is to calculate the posterior probability that $\mathcal{H}_i$ is null given all the observed statistics $\mathbf{x}$, namely $P(\theta_i{=}0|\mathbf{x})$ for $i = 1, ..., m$. This quantity is termed the *local index of significance* (LIS) (Sun & Cai, 2009), which reduces to *local false discovery rate* $P(\theta_i{=}0|x_i)$ when the hypotheses are independent. In our simulations and the real-world application, we use a Markov chain Monte Carlo (MCMC) algorithm to perform posterior inference for $P(\theta_i{=}0|\mathbf{x})$.

After we calculate the posterior marginal probabilities of $\boldsymbol{\theta}$ (i.e. LIS), we use a step-up procedure (Sun & Cai, 2009) to decide which of the hypotheses should be rejected so as to control FDR at the nominal level α. We first sort LIS from the smallest value to the largest value. Suppose $\text{LIS}_{(1)}$, $\text{LIS}_{(2)}$, ..., and $\text{LIS}_{(m)}$ are the ordered LIS, and the corresponding hypotheses are $\mathcal{H}_{(1)}$, $\mathcal{H}_{(2)}$,..., and $\mathcal{H}_{(m)}$. Let

$$k = \max\left\{ i : \frac{1}{i}\sum\nolimits_{j=1}^{i} \text{LIS}_{(j)} \leq \alpha \right\}. \tag{4}$$

Then we reject $\mathcal{H}_{(i)}$ for $i = 1, ..., k$.

4. Connections with Classical Multiple Testing Procedures

We show that both the local FDR procedure (Efron et al., 2001) and the BH procedure (Benjamini & Hochberg, 2000; Genovese & Wasserman, 2004) can be regarded as semiparametric graphical models which do not consider dependence among the hypotheses. The local FDR procedure uses Bayes Theorem to calculate the posterior probability that $\mathcal{H}_i$ is null given its observed test statistic x_i, namely

$$P(\mathcal{H}_i \text{ is null}|X_i{=}x_i) = \frac{p_0 f_0(x_i)}{p_0 f_0(x_i) + p_1 f_1(x_i)}. \tag{5}$$

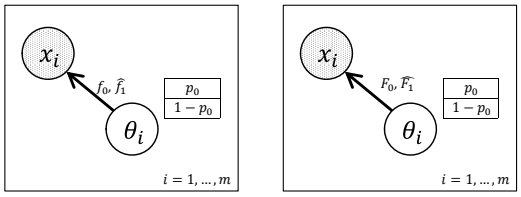
This posterior probability is termed the *local false discovery rate* (Efron & Tibshirani, 2002). Note that our LIS reduces to local false discovery rate under the assumption of independence. Efron & Tibshirani (2002) recommend using empirical Bayes inference (Robbins, 1956) to calculate local false discovery rate as

$$P(\mathcal{H}_i \text{ is null}|X_i{=}x_i) = \frac{\hat{p}_0 f_0(x_i)}{\hat{f}(x_i)}, \tag{6}$$

where $\hat{f}$ is the empirical density of the test statistic and $\hat{p}_0$ is an estimate of p_0. If we use θ_i to denote the ground truth of $\mathcal{H}_i$, its local false discovery rate is $P(\theta_i = 0|X_i{=}x_i)$. Therefore, we can use the graphical model in Figure 3(a) to denote it. Obviously, this model is exactly our semiparametric model in Figure 2, except that there are no pairwise potentials capturing the dependence because the local FDR procedure assumes independence among the hypotheses. The model for the local FDR procedure is also semiparametric because f_1 is nonparametrically estimated. Also note that the parameter π in our model reduces to the prior parameter p_0 in this simplified model.

The following shows that the BH procedure is also a semiparametric model, but the observed statistic is modeled by a cumulative distribution function (CDF). Let $P_{(1)}{<}...{<}P_{(m)}$ be the ordered P-values from the m tests and $P_{(0)}{=}0$. The BH procedure rejects any hypothesis whose P-value satisfies $P \leq P^*$ with

$$P^* = max\{P_{(i)}|P_{(i)} \leq \frac{i}{m}\frac{\alpha}{p_0}\}, \tag{7}$$



(a) The local FDR procedure (b) The BH procedure

Figure 3. The plate presentation of the semiparametric graphical models for local FDR and the BH procedure.

which controls FDR at the level α (Benjamini & Hochberg, 1995; Storey, 2002; Genovese & Wasserman, 2002). The inequality in (7) can be rewritten as

$$\frac{p_0 P_{(i)}}{i/m} \leq \alpha. \tag{8}$$

Because a P-value is the CDF of f_0 at the value of its test statistic x, and i/m is the empirical CDF of f at the test statistic of $\mathcal{H}_{(i)}$, (8) is further rewritten as

$$\frac{p_0 F_0(x)}{\hat{F}(x)} \leq \alpha, \tag{9}$$

where F_0 and F are the CDFs of f_0 and f respectively, and $\hat{F}$ is an empirical version of F. Note that the left hand side of (9) is also an empirical Bayes inference, similar to (6). Therefore, both the BH procedure and the local FDR procedure can be interpreted as empirical Bayes inference, and the difference is that the BH procedure uses the CDFs whereas the local FDR procedure uses the density functions. Thus, we can present the BH procedure as the graphical model in Figure 3(b). This model is also semiparametric because F_1 is nonparametrically estimated. Therefore, both the local FDR procedure and the BH procedure are semiparametric graphical models which do not consider dependence among the hypotheses.

5. Simulations

We explore the empirical performance of our multiple testing procedure and *three* baseline procedures, including the local FDR procedure (Efron et al., 2001), the BH procedure (Benjamini & Hochberg, 2000; Genovese & Wasserman, 2004) and the procedure based on a parametric graphical model (Liu et al., 2012). Because we have the ground truth parameters, we have two versions of our multiple testing approach, namely an oracle procedure and a data-driven procedure. The oracle procedure knows the true parameters in the graphical model (including $\boldsymbol{\phi}$, $\boldsymbol{\pi}$ and f_1), whereas the data-driven procedure does not and has to estimate the graphical model in the semiparametric way introduced in Sections 3.2 and 3.3. Both the BH procedure and the local FDR procedure need an estimate of p_0; we use the same estimating method in Section 3.2 for a fair comparison. The local FDR procedure also needs an estimate of f, and we estimate it in the same way as in our data-driven procedure.

We choose the setup to be consistent with previous work (Sun & Cai, 2009; Liu et al., 2012) when possible. We consider *two dependence structures*, namely a chain structure and a grid structure. For the chain structure, we choose the number of hypotheses m=10,000. For the grid structure, we choose a 100×100 grid, which also yields 10,000 hypotheses. We test *two levels of dependence strength*, i.e. ϕ=0.8 and ϕ=0.6. We set π to be 0.4. We first simulate the ground truth of the hypotheses $\boldsymbol{\theta}$ from $P(\boldsymbol{\theta}; \boldsymbol{\phi}, \boldsymbol{\pi})$ and then simulate the test statistics $\mathbf{x}$ from $P(\mathbf{x}|\boldsymbol{\theta}; f_0, f_1)$. We assume that the observed x_i under the null hypothesis (namely $\theta_i = 0$) is from a standard normal $\mathcal{N}(0,1)$. We test *two different models* for x_i under the alternative hypothesis (namely $\theta_i = 1$) as follows.

Model 1: $x_i|\theta_i$=1 comes from a mixture of Gaussians

$$\frac{1}{3}\mathcal{N}(1,1) + \frac{1}{3}\mathcal{N}(\mu,1) + \frac{1}{3}\mathcal{N}(5,1). \tag{10}$$

In total, we test nine values for μ, namely 1.4, 1.8, 2.2, 2.6, 3.0, 3.4, 3.8, 4.2 and 4.6. Different μ values yield different f_1 with different shapes.

Model 2: $x_i|\theta_i$=1 comes from a Gaussian $\mathcal{N}(\mu,1)$ and μ has a prior of $Gamma(2.0, \beta)$ where β is the scale parameter. We test six different values for β, namely 1.0, 1.2, 1.4, 1.6, 1.8 and 2.0. This model is designed to mimic the common situation in GWAS that common genetic variants have small effect sizes and rare genetic variants have large effect sizes (Manolio et al., 2009).

We compare *three measures* from these procedures. First, we check whether these procedures are valid, namely whether the FDR yielded from these procedures is controlled at the nominal level α. The nominal FDR level α is 0.10, which is consistent with the multiple testing literature (Efron, 2010). Second, we compare the FNR yielded from these procedures. The third measure is the average number of true positives (ATP) of these procedures. Valid procedures with a lower FNR and a higher ATP are considered to be more efficient (or powerful). In the simulations, each experiment is replicated 500 times and the average results are reported.

Performance under chain structure: The performance of the five procedures under the chain dependence structure is shown in Figures 4 and 5, which correspond to Model 1 and Model 2, respectively. It is observed that all five procedures are valid. The parametric procedure (Liu et al., 2012) is conservative, which agrees with the observations in Figure 3(1d) of Liu et al. (2012). Our semiparametric

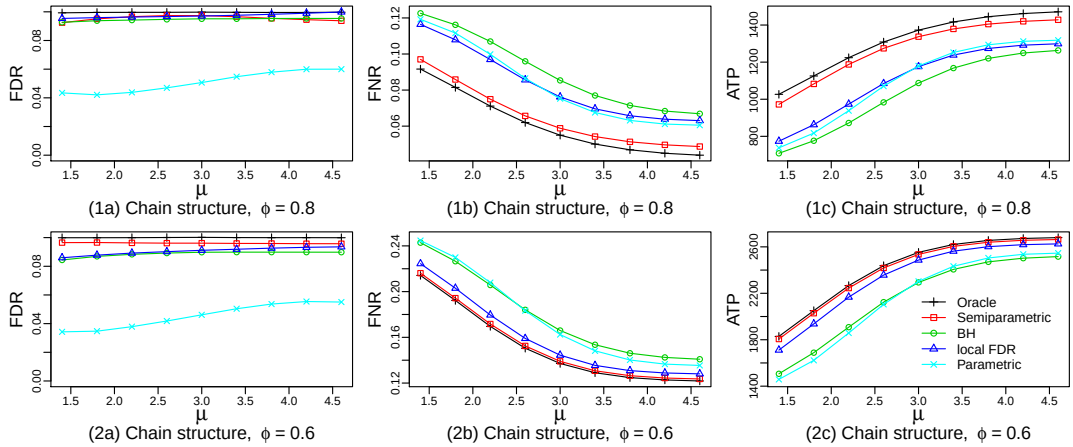


Figure 4. Performance of the procedures under **Model 1** when (1) $\phi = 0.8$ and (2) $\phi = 0.6$ in terms of (a) FDR (b) FNR and (c) ATP when the dependence structure is chain.

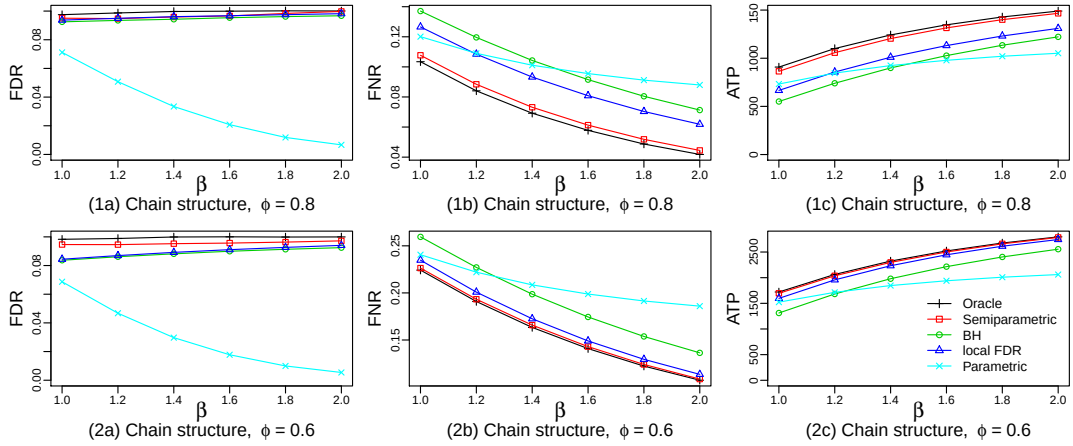


Figure 5. Performance of the procedures under **Model 2** when (1) $\phi = 0.8$ and (2) $\phi = 0.6$ in terms of (a) FDR (b) FNR and (c) ATP when the dependence structure is chain.

data-driven procedure, the BH procedure and the local FDR procedure are slightly conservative. The oracle procedure slightly outperforms the semiparametric data-driven procedure based the plots for FNR and ATP. These two completely dominate the three baselines, which indicates the benefit of leveraging dependence among the hypotheses via the semiparametric graphical model. We also observe that the advantage of the oracle procedure and our semiparametric data-driven procedure over the local FDR procedure is larger when $\phi = 0.8$ than when $\phi = 0.6$. The reason is that as ϕ decreases from 0.8 to 0.6, the dependence strength among the hypotheses decreases, and we benefit less from leveraging the dependence. When $\phi = 0.5$, the edge potentials in our graphical model are no longer informative, and the node potentials become the priors in the local FDR procedure, and our procedure exactly reduces to the local FDR procedure.

Performance under grid structure: The performance of the five procedures under the grid dependence structure is shown in Figures 6 and 7, which correspond to Model 1 and Model 2, respectively. All five procedures are valid. The parametric procedure is considerably conservative, which agrees with the observations in Figure 3(3d) of Liu et al. (2012). Again, our semiparametric data-driven procedure significantly outperforms the three baselines in all the configurations, demonstrating the benefit of leveraging dependence among the hypotheses via the semiparametric graphical model. The difference between our semiparametric data-driven procedure and the baselines is even larger compared with simulations under the chain structure. The reason is that in the grid structure, each hypothesis has more neighbors than in the chain structure, and we can benefit more from leveraging the dependence among the hypotheses.

Robustness of λ: In the previous simulations, λ is fixed at 0.8. We test another two values for λ, namely 0.2 and 0.5, and repeat previous simulations. The performance of our semiparametric procedure under the chain dependence

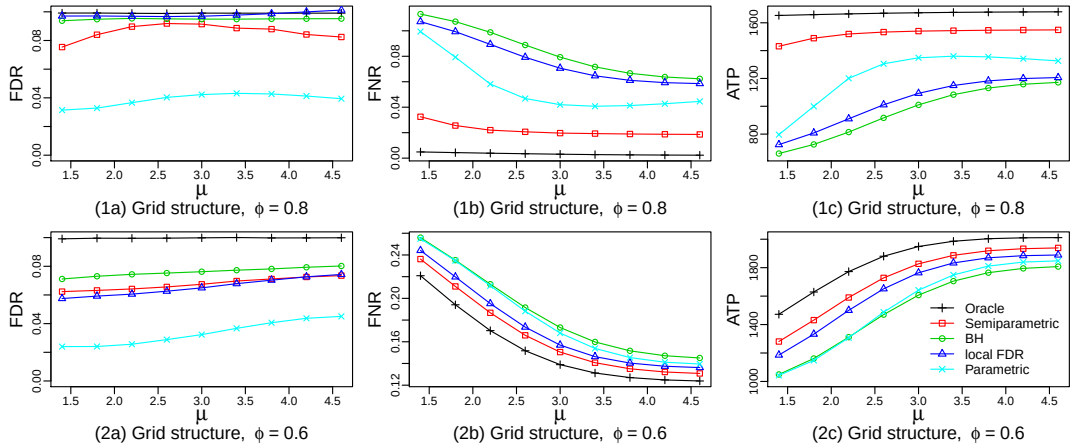


Figure 6. Performance of the procedures under **Model 1** when (1) $\phi = 0.8$ and (2) $\phi = 0.6$ in terms of (a) FDR (b) FNR and (c) ATP when the dependence structure is grid.

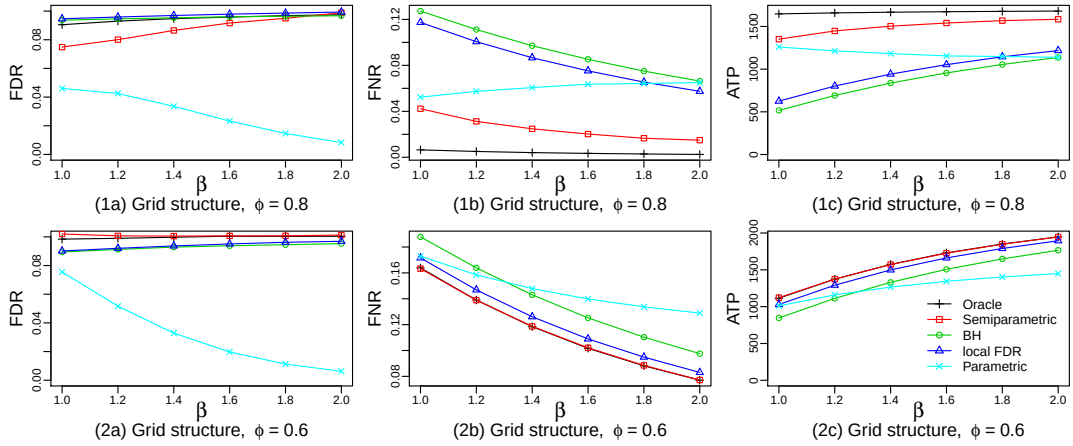


Figure 7. Performance of the procedures under **Model 2** when (1) $\phi = 0.8$ and (2) $\phi = 0.6$ in terms of (a) FDR (b) FNR and (c) ATP when the dependence structure is grid.

structure and Model 1 with $\phi = 0.8$ is provided in Figure 8. Quite surprisingly, our data-driven semiparametric procedure is valid for the three values of λ and is slightly conservative for most of the configurations. However, the FNR and ATP of our data-driven procedure for the three different values of λ are almost the same. Therefore although our approach needs to pick a λ parameter, its performance is robust for different choices of λ. The robustness of λ was also observed in the work of Storey (2002). The sensitivity analysis of λ in other configurations yield similar observations, and is given in Appendix 1 (in the supplementary materials).

Efficiency of Ranking: Although ranking the hypotheses by the probability that $\mathcal{H}_0$ is false is a secondary goal in multiple testing, readers may wonder how well our semiparametric procedure performs in terms of ranking the hypotheses. For the oracle procedure, the parametric procedure (Liu et al., 2012) and our semiparametric procedure, we rank the hypotheses by the posterior probability that $\mathcal{H}_0$ is false, namely $1 - \text{LIS}$. For BH, we use $1-P$-value. For local FDR procedure, we use $1 - lfdr$. Here we plot the ROC curves and PR curves yielded by the five procedures in Figure 9 for $\mu = 1.4$ and $\phi = 0.8$ in the chain structure under model 1. We observe that the oracle procedure produces the most efficient ranking, followed by the semiparametric procedure and the parametric procedure. The rankings yielded by local FDR and BH procedure are less effi-

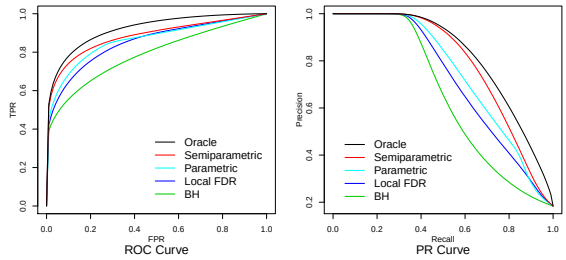


Figure 9. ROC/PR curves from these procedures.

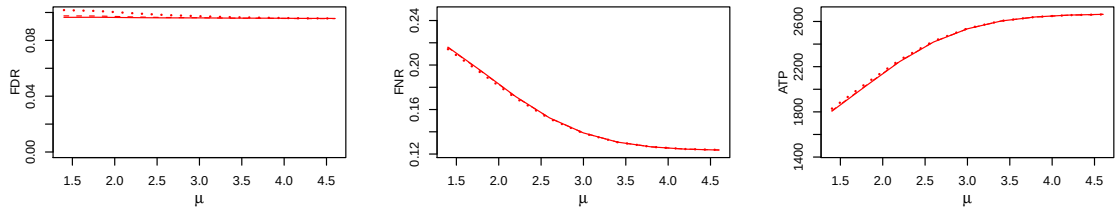


Figure 8. Performance of our procedure when $\lambda = 0.2$ (dotted lines), 0.5 (dashed lines) and 0.8 (solid lines).

cient. The ROC curves and PR curves of these procedures under other configurations show similar behavior, and are provided in Appendix 2 (in the supplementary materials).

Run Time: In the chain-structure simulations, it took our data-driven procedure about 10 hours to finish the 500 replications sequentially (for one μ value in (10)) on one 3GHz CPU. In the grid-structure simulations, it took our procedure around 30 hours to finish the 500 replications sequentially (for one μ value in (10)) on one 3GHz CPU.

6. Application

We apply our procedure to a real-world GWAS on breast cancer (Hunter et al., 2007) which involves 528,173 SNPs for 1,145 cases and 1,142 controls. In total, we test 528,173 hypotheses, and they are dependent because SNPs nearby tend to be highly correlated. We query the squared correlation coefficients (r^2 values) among the SNPs from HapMap (International HapMap Consortium, 2003), and build the dependence structure as follows. Each SNP becomes a node in the graph. For each SNP, we connect it with the SNP having the highest r^2 value with it. We further categorize the edges into a high correlation edge set $\mathcal{E}_h$ (r^2 above 0.8), a medium correlation edge set $\mathcal{E}_m$ (r^2 between 0.5 and 0.8) and a low correlation edge set $\mathcal{E}_l$ (r^2 between 0.25 and 0.5). We have three parameters (ϕ_h, ϕ_m, and ϕ_l) for the three sets of edges.

When we apply our procedure on the dataset, the individual test is a two-proportion z-test. We set $\lambda = 0.8$, and the value of p_0 is estimated to be 0.978, which means that about 2.2% of the SNPs are associated to breast cancer. The estimated f_1 in this study is plotted in Figure 1. The whole experiment takes around 30 hours on a single processor. Our procedure reports 20 SNPs with LIS value below 0.01. There are five clusters covering 18 of them. All 18 SNPs have very small P-values from the two-proportion z-test and locate near one another in the same cluster. The first cluster on Chr2, the cluster on Chr4, the cluster on Chr9 and the cluster on Chr10 are identified in the studies of Hunter et al. (2007) and Satrom et al. (2009). The second cluster on Chr2 is associated to a telomere and telomeres are known to be related to breast cancer (Svenson et al., 2008). We further use a second cohort to validate the 18 SNPs, and 16 of them show a moderate level of association on the second cohort. More details are provided in Appendix 3 (in the supplementary materials). We also would like to mention that there is some work on estimating less conservative significance thresholds for controlling family-wise error rate in GWAS (Salyakina et al., 2005; Han & Eskin, 2010).

7. Conclusion

We propose a novel semiparametric graphical model to leverage the dependence in multiple testing problems. Although our semiparametric approach seems incremental over previous fully parametric approach (Sun & Cai, 2009; Liu et al., 2012) from the viewpoint of graphical models, such a modification is nontrivial to the multiple testing area, for both a methodological reason and an application reason. From the methodological standpoint, our semiparametric approach naturally generalizes the local FDR procedure and connects with the BH procedure — we show that both the BH procedure and the local FDR procedure estimate their parameters in the same semiparametric way to avoid assumptions about f_1. The methodological unification demonstrates that such a modification is necessary for multiple testing. From the application aspect, our semiparametric approach no longer requires the investigators to know the parameterization of f_1, which is generally unknown in practical problems. Improper parameterization assumptions for f_1 can make the fully parametric approach either too liberal which makes the procedure invalid, or too conservative which makes the procedure lose power, as illustrated by both our simulations and previous work (Sun & Cai, 2009; Liu et al., 2012). Our semiparametric approach better controls FDR and is more powerful. For these reasons, we suggest that investigators choose the semiparametric approach for their large-scale multiple testing problems if (i) they speculate that there exists dependence among the hypotheses, and (ii) there is no suitable parametric distribution for f_1.

Acknowledgements

The authors gratefully acknowledge the support of NIH grants R01GM097618, R01LM011028, R01CA165229, R01LM010921, P30CA014520, UL1TR000427, NSF grants DMS-1106586 and DMS-1308872, and Wisconsin Alumni Research Foundation.

References

Benjamini, Yoav and Heller, Ruth. False discovery rates for spatial signals. *JASA*, 102:1272–1281, 2007.

Benjamini, Yoav and Hochberg, Yosef. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *JRSS-B*, 57(1):289–300, 1995.

Benjamini, Yoav and Hochberg, Yosef. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J EDUC BEHAV STAT*, 25(1):60–83, 2000.

Benjamini, Yoav and Yekutieli, Daniel. The control of the false discovery rate in multiple testing under dependency. *ANN STAT*, 29:1165–1188, 2001.

Besag, Julian. Statistical analysis of non-lattice data. *JRSS-D*, 24 (3):179–195, 1975.

Blanchard, Gilles and Roquain, Étienne. Adaptive false discovery rate control under independence and dependence. *J MACH LEARN RES*, 10:2837–2871, 2009.

Efron, Bradley. Correlation and large-scale simultaneous significance testing. *JASA*, 102(477):93–103, 2007.

Efron, Bradley. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2010.

Efron, Bradley and Tibshirani, Robert. Empirical Bayes methods and false discovery rates for microarrays. *GENET EPIDEMIO*, 23(1):70–86, 2002.

Efron, Bradley, Tibshirani, Robert, Storey, John D., and Tusher, Virginia. Empirical Bayes analysis of a microarray experiment. *JASA*, 96:1151–1160, 2001.

Epanechnikov, V. A. Non-parametric estimation of a multivariate probability density. *THEOR PROBAB APPL*, 14(1):153–158, 1969.

Fan, Jianqing, Han, Xu, and Gu, Weijie. Control of the false discovery rate under arbitrary covariance dependence. *JASA*, 107(499):1019–1045, 2012.

Farcomeni, Alessio. Some results on the control of the false discovery rate under dependence. *SCAND J STAT*, 34(2):275–297, 2007.

Finner, H. and Roters, M. Multiple hypotheses testing and expected number of type I errors. *ANN STAT*, 30:220–238, 2002.

Friguet, Chloé, Kloareg, Maela, and Causeur, David. A factor model approach to multiple testing under dependence. *JASA*, 104(488):1406–1415, 2009.

Genovese, Christopher and Wasserman, Larry. Operating characteristics and extensions of the false discovery rate procedure. *JRSS-B*, 64:499–517, 2002.

Genovese, Christopher and Wasserman, Larry. A stochastic process approach to false discovery control. *ANN STAT*, 32:1035–1061, 2004.

Genovese, Christopher, Roeder, Kathryn, and Wasserman, Larry. False discovery control with p-value weighting. *BIOMETRIKA*, 93:509–524, 2006.

Han, Buhm and Eskin, Eleazar. Multiple testing in genetic epidemiology. *Encyclopedia of Life Sciences*, 2010.

Hinton, Geoffrey. Training products of experts by minimizing contrastive divergence. *NEURAL COMPUT*, 14:1771–1800, 2002.

Hunter, David J., Kraft, Peter, ... , and Chanock, Stephen J. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *NAT GENET*, 39(7):870–874, 2007.

Hyvärinen, Aapo. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE T NEURAL NETWOR*, 18(5):1529–1531, 2007a.

Hyvärinen, Aapo. Some extensions of score matching. *COMPUT STAT DATA AN*, 51(5):2499–2512, 2007b.

International HapMap Consortium. The international HapMap project. *NATURE*, 426:789–796, 2003.

Leek, Jeffrey T. and Storey, John D. A general framework for multiple testing dependence. *P NATL ACAD SCI USA*, 105 (48):18718–18723, 2008.

Liu, Jie, Zhang, Chunming, McCarty, Catherine, Peissig, Peggy, Burnside, Elizabeth, and Page, David. Graphical-model based multiple testing under dependence, with applications to genome-wide association studies. In *UAI*, 2012.

Liu, Jie, Zhang, Chunming, Burnside, Elizabeth, and Page, David. Learning heterogeneous hidden Markov random fields. In *AISTATS*, 2014.

Manolio, Teri A, Collins, Francis S, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

Owen, Art B. Variance of the number of false discoveries. *JRSS-B*, 67:411–426, 2005.

Reiner, Anat, Yekutieli, Daniel, and Benjamini, Yoav. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, 2003.

Robbins, Herbert. An empirical Bayes approach to statistics. In *The 3rd Berkeley Symposium I*, pp. 157–163, 1956.

Romano, Joseph, Shaikh, Azeem, and Wolf, Michael. Control of the false discovery rate under dependence using the bootstrap and subsampling. *TEST*, 17:417–442, 2008.

Rosenblatt, Murray. Remarks on some nonparametric estimates of a density function. *ANN MATH STAT*, 27(3):832–837, 1956.

Salyakina, Daria, Seaman, Shaun R, Browning, Brian L, Dudbridge, Frank, and Müller-Myhsok, Bertram. Evaluation of nyholts procedure for multiple testing correction. *Human heredity*, 60(1):19–25, 2005.

Sarkar, Sanat K. False discovery and false nondiscovery rates in single-step multiple testing procedures. *ANN STAT*, 34(1): 394–415, 2006.

Satrom, Pal, Biesinger, Jacob, ..., and Larson, Garrett P. A risk variant in an mir-125b binding site in bmpr1b is associated with breast cancer pathogenesis. *CANCER RES*, 69(18):7459–7465, 2009.

Storey, John D. A direct approach to false discovery rates. *JRSS-B*, 64:479–498, 2002.

Storey, John D, Taylor, Jonathan E, and Siegmund, David. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *JRSS-B*, 66(1):187–205, 2004.

Strimmer, Korbinian. A unified approach to false discovery rate estimation. *BMC bioinformatics*, 9(1):303, 2008.

Sun, Wenguang and Cai, T. Tony. Large-scale multiple testing under dependence. *JRSS-B*, 71:393–424, 2009.

Svenson, Ulrika, Nordfjall, Katarina, ..., and Roos, Goran. Breast cancer survival is associated with telomere length in peripheral blood cells. *CANCER RES*, 68(10):3618–3623, 2008.

Tieleman, Tijmen. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *ICML*, pp. 1064–1071, 2008.

Wu, Wei Biao. On false discovery control under dependence. *ANN STAT*, 36(1):364–380, 2008.

Yekutieli, Daniel and Benjamini, Yoav. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J STAT PLAN INFER*, 82:171–196, 1999.

Zhang, Chunming, Fan, Jianqing, and Yu, Tao. Multiple testing via FDR_L for large-scale imaging data. *ANN STAT*, 39(1): 613–642, 2011.