The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1699–1724
DOI: 10.1214/16-AOAS956

MULTIPLE TESTING UNDER DEPENDENCE VIA GRAPHICAL MODELS

BY JIE LIU[1], CHUNMING ZHANG[2] AND DAVID PAGE[1]

University of Wisconsin-Madison

Large-scale multiple testing tasks often exhibit dependence. Leveraging the dependence between individual tests is still one challenging and important problem in statistics. With recent advances in graphical models, it is feasible to use them to capture the dependence among multiple hypotheses. We propose a multiple testing procedure which is based on a Markov-random-field-coupled mixture model. The underlying true states of hypotheses are represented by a latent binary Markov random field, and the observed test statistics appear as the coupled mixture variables. The model can be learned by a novel EM algorithm. The next step is to infer the posterior probability that each hypothesis is null (termed *local index of significance*), and the false discovery rate can be controlled accordingly. We also provide a semiparametric variation of the graphical model which is useful in the situation where f_1 (the density function of the test statistic under the alternative hypothesis) is heterogeneous among multiple hypotheses. This semiparametric approach exactly generalizes the local FDR procedure [*J. Amer. Statist. Assoc.* **96** (2001) 1151–1160] and connects with the BH procedure [*J. Roy. Statist. Soc. Ser. B* **57** (1995) 289–300]. Simulations show that the numerical performance of multiple testing can be improved substantially by using our procedure. We apply the procedure to a real-world genome-wide association study on breast cancer, and we identify several SNPs with strong association evidence.

1. Introduction. Observations from large-scale multiple testing problems often exhibit dependence in the sense that whether the null hypothesis of one test is true or not (termed the *underlying true state*) depends on the underlying true states of other tests. For instance, in genome-wide association studies, researchers collect hundreds of thousands of highly correlated genetic markers (single-nucleotide polymorphisms, or SNPs) with the purpose of identifying the subset of markers associated with a heritable disease or trait. In functional magnetic resonance imaging studies of the brain, thousands of spatially correlated voxels are collected while subjects are performing certain tasks, with the purpose of detecting the relevant voxels. The most popular family of large-scale multiple testing procedures

Received September 2014; revised May 2016.

[1]Supported in part by NIH Grants R01GM097618 and R01LM011028.

[2]Supported in part by NSF Grants DMS-11-06586, DMS-12-08872, DMS-15-21761, and the Wisconsin Alumni Research Foundation.

Key words and phrases. Multiple testing under dependence, graphical models, Markov random field, local index of significance, genome-wide association study.

is the false discovery rate analysis, such as the p-value thresholding procedures [Benjamini and Hochberg (1995, 2000), Genovese and Wasserman (2004)], the local false discovery rate procedure [Efron et al. (2001)] and the positive false discovery rate procedure [Storey (2002, 2003)]. However, all these classical multiple testing procedures ignore the correlation structure among the individual factors, and the question is *whether we can reduce the false nondiscovery rate by leveraging the dependence*, *while still controlling the false discovery rate in multiple testing*.

Graphical models provide an elegant way of representing dependence. With recent advances in graphical models, especially more efficient algorithms for inference and parameter estimation, it is feasible to use these models to leverage the dependence between individual tests in multiple testing problems. More specifically, we can use graphical models to explicitly model the underlying true states of the hypotheses as random variables to encode the dependence, and then model the observed test statistics independently given their underlying true states. For example, one influential paper [Sun and Cai (2009)] uses a hidden Markov model to represent the dependence structure, and has shown its optimality under certain conditions and its strong empirical performance. In the model, the underlying true states of the hypotheses form a Markov chain, and the observed test statistics are assumed to be independent given the underlying true states. It is the first graphical model that explicitly specifies the dependence between the hypotheses in multiple testing problems, and the procedure methodologically differs from other works on multiple testing under dependence [Benjamini and Yekutieli (2001), Blanchard and Roquain (2009), Efron (2007), Farcomeni (2007), Finner and Roters (2002), Owen (2005), Romano, Shaikh and Wolf (2008), Sarkar (2006)] which only explicitly model either test statistics or p-values.

Nevertheless, the procedure of Sun and Cai (2009) can only deal with a sequential dependence structure, and the dependence parameters are homogeneous. In this paper, we propose a multiple testing procedure based on a Markov-random-field-coupled mixture model which allows arbitrary dependence structures. In our model, the underlying true states of the hypotheses form a Markov random field, and the observed test statistics are assumed to be independent given the underlying true states. This extension requires more sophisticated algorithms for parameter estimation and inference. For parameter estimation, we design a novel EM algorithm with MCMC in the E-step and a contrastive divergence style algorithm [Tieleman (2008)] in the M-step. We show that there is a lower bound of the log likelihood which nondecreases over the EM iterations except for some MCMC error introduced in the E-step. We use the MCMC algorithm to infer the posterior probability that each hypothesis is null (termed *local index of significance* or LIS). Finally, the false discovery rate can be controlled by thresholding the LIS.

Another extension to the work of Sun and Cai (2009) is that we design a semiparametric variation of the graphical model which nonparametrically estimates the

f_1 function—the density function of the test statistic under the alternative hypothesis. This is particularly important in some practical problems where f_1 is heterogeneous among multiple hypotheses, and thus cannot be estimated with a simple parametric distribution. The remainder of the graphical model is still estimated parametrically. The inference of the posterior probability and the false discovery rate control in this semiparametric variation remain the same as the parametric procedure. More importantly, this semiparametric approach exactly generalizes the local FDR procedure [Efron et al. (2001)] and connects with the BH procedure [Benjamini and Hochberg (1995)].

The rest of the paper is organized as follows. Section 2 introduces terminology and previous multiple testing procedures. Sections 3 and 4 introduce the graphical model-based multiple testing procedures, including the details about the parametric and semiparametric estimation of the graphical model, the inference of the posterior probability, the control of the false discovery rate and the connection with previous procedures. Section 5 evaluates our procedure on a variety of simulations, and the empirical results show that the numerical performance can be improved substantially by using our procedure. In Section 6, we apply the semiparametric procedure to a real-world genome-wide association study (GWAS) on breast cancer, and we identify several SNPs with strong association evidence. In Section 7, we provide the details of the EM algorithm, and show that there is a lower bound of the log likelihood which nondecreases over the EM iterations. We finally conclude in Section 8.

2. Terminology and previous procedures. Suppose that we carry out m tests whose results can be categorized as in Table 1. *False discovery rate* (FDR), defined as $E(N_{10}/R|R>0)P(R>0)$, depicts the expected proportion of incorrectly rejected null hypotheses [Benjamini and Hochberg (1995)]. *False nondiscovery rate* (FNR), defined as $E(N_{01}/S|S>0)P(S>0)$, depicts the expected proportion of false nonrejections in those tests whose null hypotheses are not rejected [Genovese and Wasserman (2002)]. An FDR procedure is *valid* if it controls FDR at a prespecified level, and *optimal* if it has the smallest FNR among all valid FDR procedures [Sun and Cai (2009)].

The effects of correlation on multiple testing have been discussed, under different assumptions, with a focus on the validity issue [Benjamini and Yekutieli (2001), Blanchard and Roquain (2009), Efron (2007), Farcomeni (2007),

TABLE 1
Classification of tested hypotheses

	Not rejected	**Rejected**	**Total**
Null	N_{00}	N_{10}	m_0
Non-null	N_{01}	N_{11}	m_1
Total	S	R	m

Finner and Roters (2002), Owen (2005), Romano, Shaikh and Wolf (2008), Sarkar (2006), Wu (2008)]. The efficiency issue has also been investigated [Benjamini and Heller (2007), Genovese, Roeder and Wasserman (2006), Yekutieli and Benjamini (1999), Zhang, Fan and Yu (2011)], indicating FNR could be decreased by considering dependence in multiple testing. Several approaches have been proposed, such as dependence kernels [Leek and Storey (2008)], factor models [Friguet, Kloareg and Causeur (2009)] and principal factor approximation [Fan, Han and Gu (2012)]. Sun and Cai (2009) explicitly use a hidden Markov model (HMM) to represent the dependence structure over the underlying true states of the hypotheses and analyze the optimality under the compound decision framework [Sun and Cai (2007)]. However, their procedure and its extensions, SLIS [Wei et al. (2009)], PLIS [Wei et al. (2009)] and RSPLIS [Xiao, Zhu and Guo (2013)], can only deal with sequential dependence. In this paper, we replace the HMM with a Markov-random-field-coupled mixture model, which allows richer and more flexible dependence structures.

3. The parametric procedure. Let $\mathbf{x} = (x_1, \ldots, x_m)$ be a vector of test statistics from a set of m hypotheses $(\mathcal{H}_1, \ldots, \mathcal{H}_m)$. The underlying true states of these hypotheses are denoted by a latent Bernoulli vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m) \in \{0, 1\}^m$, 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. Conditionally on $\boldsymbol{\theta}$, x_i's are independent. The dependence among these hypotheses is represented as a binary Markov random field (MRF) on $\boldsymbol{\theta}$. The structure of the MRF can be 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$_j$ in $\mathcal{E}$, and the strength of dependence is parameterized by the potential function $\Psi_l^{\mathcal{E}}$ (parametrized by ϕ_l, $0 < \phi_l < 1$) on this edge (indexed by l). The degree of prior belief that $\mathcal{H}_i$ is null is captured by the node potential function $\Psi_i^{\mathcal{V}}$ (parametrized by π_i, $0 < \pi_i < 1$). The probability of $\boldsymbol{\theta}$ from the MRF with parameters $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ is

$$P(\boldsymbol{\theta}; \boldsymbol{\pi}, \boldsymbol{\phi}) = \frac{1}{Z(\boldsymbol{\pi}, \boldsymbol{\phi})} \prod_{i=1}^{m} \Psi_i^{\mathcal{V}}(\boldsymbol{\theta}; \pi_i) \prod_{l=1}^{|\mathcal{E}|} \Psi_l^{\mathcal{E}}(\boldsymbol{\theta}; \phi_l), \tag{3.1}$$

where $Z(\boldsymbol{\pi}, \boldsymbol{\phi})$ is the normalizing constant. Suppose that the probability density function of the test statistic x_i given $\theta_i = 0$ is f_0, and the density of x_i given $\theta_i = 1$ is f_1. Then $\mathbf{x}$ is an MRF-coupled mixture. Figure 1 shows the MRF-coupled mixture model for three dependent hypotheses $\mathcal{H}_i$, $\mathcal{H}_j$ and $\mathcal{H}_k$.

For now, let us assume for simplicity that the mixture model is parameterized by a parameter set $\vartheta = (\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})$, where $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ parameterize the binary MRF, and $\boldsymbol{\psi}$ parameterizes f_0 and f_1. For example, if f_0 is standard normal $\mathcal{N}(0, 1)$ and f_1 is noncentered normal $\mathcal{N}(\mu, 1)$, then $\boldsymbol{\psi}$ only contains parameter μ. This multiple testing procedure is termed *the parametric procedure*. In Section 4, we introduce *the semiparametric procedure* which is designed for the situations where

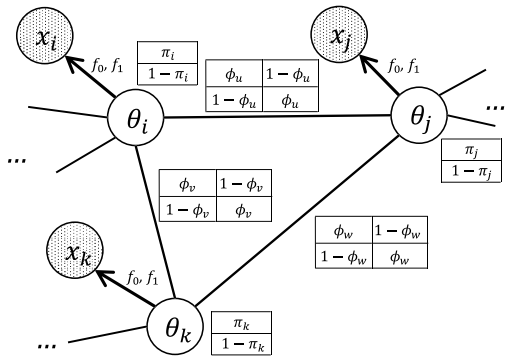


FIG. 1. *The MRF-coupled mixture model for three dependent hypotheses $\mathcal{H}_i$, $\mathcal{H}_j$ and $\mathcal{H}_k$ with* observed *test statistics (x_i, x_j and x_k) and underlying true states (θ_i, θ_j and θ_k). The MRF is parameterized by π_i, π_j and π_k, ϕ_u, ϕ_v and ϕ_w, and the coupled mixtures are parameterized by f_0 and f_1.*

f_1 is heterogeneous among multiple hypotheses and needs to be estimated nonparametrically.

In our MRF-coupled mixture model, $\mathbf{x}$ is observable, and $\boldsymbol{\theta}$ is hidden. For a given parameter set $\vartheta = (\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})$, the joint probability density over $\mathbf{x}$ and $\boldsymbol{\theta}$ is

$$P(\mathbf{x}, \boldsymbol{\theta}|\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}) = P(\boldsymbol{\theta}; \boldsymbol{\pi}, \boldsymbol{\phi}) \prod_{i=1}^{m} P(x_i|\theta_i; \boldsymbol{\psi}). \tag{3.2}$$

We define the marginal probability that $\mathcal{H}_i$ is null given all observed statistics $\mathbf{x}$ under the parameters in ϑ, $P_\vartheta(\theta_i = 0|\mathbf{x})$, to be the *local index of significance* (LIS) for $\mathcal{H}_i$ [Sun and Cai (2009)]. Sun and Cai (2009) insightfully discussed the properties of LIS and its relationship with the p-value and local FDR [Efron et al. (2001)].

There are three steps in using this graphical model to capture the dependence in multiple hypotheses. First, we have to estimate the parameters $\vartheta = (\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})$. Second, we have to compute the posterior marginal distribution of the hidden variables θ_i given the test statistics $\mathbf{x}$, namely, LIS for each hypothesis. Last, we have to link the LIS values with FDR and control FDR. The three steps are introduced in Sections 3.1, 3.2 and 3.3, respectively. In Section 3.4, the optimality of the procedure is discussed under the compound decision theoretic framework [Robbins (1951), Sun and Cai (2007)].

3.1. *Parameters and parameter estimation.* In our model, the dependence among these hypotheses is represented by a Markov random field on the latent vector $\boldsymbol{\theta}$ parameterized by $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$, and the observed test statistics $\mathbf{x}$ are represented by the coupled mixture parameterized by $\boldsymbol{\psi}$. Estimating $(\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})$ is difficult for two reasons. First, parameter estimation is difficult by nature in undirected graphical models due to the global normalization constant $Z(\boldsymbol{\pi}, \boldsymbol{\phi})$ in Formula (3.1)

[Wainwright, Jaakkola and Willsky (2003a), Welling and Sutton (2005)]. Second, $\boldsymbol{\theta}$ is latent and we only have one observed training sample $\mathbf{x}$.

For a given vector $\mathbf{x}$, the log likelihood of the parameters $\vartheta = (\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})$ is

$$\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}) = \log P(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}) = \log \sum_{\boldsymbol{\theta} \in \{0,1\}^m} P(\mathbf{x}, \boldsymbol{\theta}; \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}). \tag{3.3}$$

Since we only have one instantiation $(\mathbf{x}, \boldsymbol{\theta})$, we usually have to assume that π_i's are the same for $i = 1, \ldots, m$ and that ϕ_l's are the same for all edges in the edge set $\mathcal{E}$, for effective parameter estimation. This *homogeneity* assumption is similar to the assumption in the work of Sun and Cai (2009) that the transition parameter and the emission parameter stay the same for i $(i = 1, \ldots, m)$ in their HMM model. To alleviate this assumption in GWAS, three improved HMM based procedures, SLIS [Wei et al. (2009)], PLIS [Wei et al. (2009)] and RSPLIS [Xiao, Zhu and Guo (2013)], are designed to estimate different parameters for different chromosomes. In our real-world GWAS application in Section 6, we have different parameters for SNP pairs with different levels of correlation.

We use an EM algorithm to solve this problem of the hidden vector $\boldsymbol{\theta}$. In the E-step, we run our MCMC algorithm in Section 3.2 to infer the latent $\boldsymbol{\theta}$ based on the currently estimated parameters $\vartheta = (\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})$. In the M-step, we run a gradient ascent algorithm, similar to the persistent contrastive divergence (PCD) algorithm [Tieleman (2008)], to estimate $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ from the currently inferred $\boldsymbol{\theta}$. We also perform maximum likelihood estimation of $\boldsymbol{\psi}$ from currently inferred $\boldsymbol{\theta}$ and observed $\mathbf{x}$ in the M-step. We run the EM algorithm until both $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ converge. Although this EM algorithm involves intensive computation in both E-step and M-step, it converges very quickly in our experiments. Similar to other EM algorithms, our algorithm only converges to a local maximum of the likelihood $\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})$, but the lower bound nondecreases over the EM iterations (except for some MCMC error introduced in the E-step). The details of the EM algorithm and the explanation are provided in Section 7.

3.2. *Posterior inference.* After we estimate the parameters, we are interested in calculating $P_\vartheta(\theta_i = 0|\mathbf{x})$ for a given parameter set ϑ. One popular family of inference algorithms is the sum-product family [Kschischang, Frey and Loeliger (2001)], which is also known as belief propagation [Yedidia, Freeman and Weiss (2000)]. For loop-free graphs, belief propagation algorithms provide exact inference results with a computational cost linear in the number of variables. In our MRF-coupled mixture model, the structure of the latent MRF is described by a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$. When $\mathcal{G}$ is chain structured, the instantiation of belief propagation is the forward-backward algorithm [Baum et al. (1970)]. When $\mathcal{G}$ is tree structured, the instantiation of belief propagation is the upward-downward algorithm [Crouse, Nowak and Baraniuk (1998)]. For graphical models with cycles, loopy belief propagation [Murphy, Weiss and Jordan (1999), Weiss (2000)] and the tree-reweighted

algorithm [Wainwright, Jaakkola and Willsky (2003b)] can be used for approximate inference. Other inference algorithms for graphical models include junction trees [Lauritzen and Spiegelhalter (1988)], sampling methods [Gelfand and Smith (1990)] and variational methods [Jordan et al. (1999)]. Recent papers [Schraudolph (2010), Schraudolph and Kamenetsky (2009)] discuss exact inference algorithms on binary Markov random fields which allow loops. In our simulations, we use belief propagation when the graph $\mathcal{G}$ has no loops. When $\mathcal{G}$ has loops (e.g., in the simulations on genetic data and the real-world application), we use a Markov chain Monte Carlo (MCMC) algorithm to perform inference for $P_\vartheta(\theta_i = 0|\mathbf{x})$.

3.3. *FDR control.* After we calculate the posterior marginal probabilities of $\boldsymbol{\theta}$ (or LIS), we have to decide which of these hypotheses should be rejected. Intuitively, we should reject the hypotheses with small LIS values, but we have to associate these marginal probabilities with FDR and be able to control FDR at a prespecified level. We use the step-up procedure in the work of Sun and Cai (2009) to control FDR at the prespecified level α. We first sort LIS from the smallest value to the largest value. Suppose $\mathrm{LIS}_{(1)}, \mathrm{LIS}_{(2)}, \ldots,$ and $\mathrm{LIS}_{(m)}$ are the ordered LIS, and the corresponding hypotheses are $\mathcal{H}_{(1)}, \mathcal{H}_{(2)}, \ldots,$ and $\mathcal{H}_{(m)}$. Let

$$k = \max\left\{ i : \frac{1}{i}\sum_{j=1}^{i} \mathrm{LIS}_{(j)} \le \alpha \right\}. \tag{3.4}$$

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

3.4. *Optimality analysis.* There are two types of optimality for these graphical model-based multiple testing procedures. The first optimality is for the *oracle procedure* which knows the ground truth of the parameters in the graphical model. The optimality of the oracle procedure is in the sense that it minimizes the marginal FNR subject to a constraint on the marginal FDR. By exploring the connection between multiple testing and weighted classification under the compound decision theoretic framework [Robbins (1951), Sun and Cai (2007)], Sun and Cai (2009) proved that their oracle procedure is optimal under a mild monotone ratio condition (MRC). Although the proof is for hidden Markov models (HMM), it can be easily generalized to our MRF-coupled mixture model. Therefore, the optimality of the oracle procedure can be proved under the compound decision framework [Sun and Cai (2007, 2009)], as long as an exact inference algorithm exists or an approximate inference algorithm can be guaranteed to converge to the correct marginal probabilities (see Section 3.2). The second type of optimality is the *asymptotic optimality* of the *data-driven* parametric procedure in the sense that it attains both the FDR and FNR levels of the oracle procedure asymptotically (as the number of tests $m \to \infty$). Such a proof requires that the parameters in the graphical model can be estimated consistently. To the best of our knowledge, there is no such consistence guarantee for the estimators of MRF-coupled mixture models in the literature up to now. Therefore, the asymptotic optimality of the data-driven procedure is still unknown, and will be an important problem in future work.

4. The semiparametric procedure. The graphical model in the parametric procedure is effective to leverage the dependence in multiple testing problems, but it makes a strong assumption that the f_1 function can be estimated parametrically. The work of Sun and Cai (2009) makes the same assumption. However, a long tradition in hypothesis testing is to derive test statistics and calculate p-values all under the null hypothesis $\mathcal{H}_0$. Statisticians avoid making assumptions about f_1 because 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, that is, P(head) in coin-flippings; the two-proportion z-test is widely used in case-control studies, for example, 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 noncentered 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 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 2; 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 procedure in Section 3 which requires f_1 to be estimated parametrically. Therefore, the key question is *how to use the graphical models to leverage the dependence among the hypotheses without making assumptions about* f_1.

In this section, we make one modification to the graphical model—f_1 is learned nonparametrically and the MRF part is learned parametrically by estimating parameters $\boldsymbol{\phi}$ and $\boldsymbol{\pi}$. With the learned model, we use the same approach in Section 3.2 to perform marginal inference of $\boldsymbol{\theta}|\mathbf{x}$, and then use the same step-up procedure in Section 3.3 to control FDR. This multiple testing procedure is named *the semiparametric procedure*. More algorithmic details are introduced in Section 4.1.

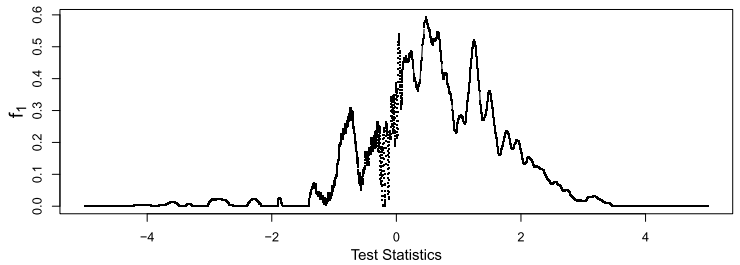


FIG. 2. *Estimated f_1 in a real-world genome-wide association study on breast cancer.*

Section 4.2 further shows that the two widely used multiple testing procedures, the BH procedure [Benjamini and 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 semiparametric procedure exactly reduces to the local FDR procedure [Efron et al. (2001)] when the tests are independent. This unification demonstrates that it is sensible to use this semiparametric model to capture the dependence in multiple testing problems.

4.1. *Nonparametric estimation of* f_1. We cannot directly estimate f_1 from the observed test statistics $\mathbf{x}$ because the underlying true state vector $\boldsymbol{\theta}$ 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

$$f(x) = p_0 f_0(x) + (1 - p_0) f_1(x), \tag{4.1}$$

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.

We can estimate p_0 with the method in Storey (2002), namely,

$$\hat{p}_0(\lambda) = \frac{W(\lambda)}{(1 - \lambda)m}, \tag{4.2}$$

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 the null hypotheses, then $W(\lambda)/(1 - \lambda)$ is the total number of null hypotheses. Therefore, the right-hand side of (4.2) is an estimate of p_0. Obviously, $\hat{p}_0(\lambda)$ overestimates p_0 because there may be non-null 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, Taylor and Siegmund (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 reasonable choices of λ. Note that there are several alternative methods [Kim and Zhang (2014), Liang and Nettleton (2012)] which can be used to improve Storey's estimator of p_0 in Formula (4.2).

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.

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

$$\hat{f}_1(x) = \frac{\hat{f}(x) - \hat{p}_0 f_0(x)}{1 - \hat{p}_0}. \tag{4.3}$$

Several iterative and more sophisticated approaches have been proposed to estimate f_1 in a similar semiparametric fashion, including the weighed kernel estimator [Guedj et al. (2009), Robin et al. (2007)], the randomly weighed kernel estimator [Nguyen and Matias (2014)] and the maximum smoothed likelihood estimator [Nguyen and Matias (2014)]. These approaches may produce a better estimate of f_1 at the cost of additional computation.

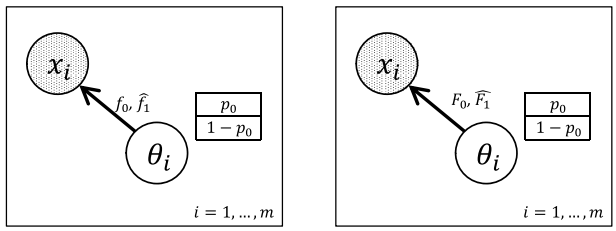
4.2. *Connections with classical multiple testing procedures.* We show that both the local FDR procedure [Efron et al. (2001)] and the BH procedure [Benjamini and Hochberg (2000), Genovese and Wasserman (2004)] can be regarded as semiparametric graphical models which do not consider dependence among 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,

$$P(\mathcal{H}_i \text{ is null} | X_i = x_i) = \frac{p_0 f_0(x_i)}{p_0 f_0(x_i) + p_1 f_1(x_i)}. \tag{4.4}$$

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

$$P(\mathcal{H}_i \text{ is null} | X_i = x_i) = \frac{\hat{p}_0 f_0(x_i)}{\hat{f}(x_i)}, \tag{4.5}$$

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 underlying true state of $\mathcal{H}_i$, then 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



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

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

in Figure 1, 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 $\boldsymbol{\pi}$ 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 $P_{(1)} < \cdots < P_{(m)}$ be the ordered p-values from the m tests and $P_{(0)} = 0$. The BH procedure rejects any hypothesis whose p-value satisfies $P \le P^*$ with

$$P^* = \max\left\{ P_{(i)} \middle| P_{(i)} \le \frac{i}{m}\frac{\alpha}{p_0} \right\}, \tag{4.6}$$

which controls FDR at the level α [Benjamini and Hochberg (1995), Genovese and Wasserman (2002), Storey (2002)]. The inequality in (4.6) can be rewritten as

$$\frac{p_0 P_{(i)}}{i/m} \le \alpha. \tag{4.7}$$

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 $\mathcal{H}_{(i)}$, (4.7) is further rewritten as

$$\frac{p_0 F_0(x)}{\hat{F}(x)} \le \alpha, \tag{4.8}$$

where F_0 and F are the CDFs of f_0 and f, respectively, and $\hat{F}$ is an empirical version of F. 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 approach and two baseline procedures, the local FDR procedure [Efron et al. (2001)] and the BH procedure [Benjamini and Hochberg (2000), Genovese and Wasserman (2004)]. Because we have the ground truth parameters and two different ways of estimating the graphical model, there are three versions of our multiple testing approach, namely, an oracle procedure, a data-driven parametric procedure and a data-driven semiparametric procedure. The oracle procedure knows the true parameters in the graphical model (including $\boldsymbol{\pi}$, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$), whereas the data-driven procedures do not and have to estimate the graphical model in the parametric and semiparametric ways introduced in Sections 3 and 4.

We choose the setup to be consistent with previous work of Sun and Cai (2009) 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*, that is, $\phi = 0.8$ and $\phi = 0.6$. We set π to be 0.4. We first simulate the underlying true states 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 normals

$$\frac{1}{3}\mathcal{N}(1, 1) + \frac{1}{3}\mathcal{N}(\mu, 1) + \frac{1}{3}\mathcal{N}(5, 1). \tag{5.1}$$

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)].

The oracle procedure knows the true parameters in the graphical model, including $\boldsymbol{\pi}$, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$. For the data-driven parametric procedure, f_1 is assumed to be a simple Gaussian. For the data-driven semiparametric procedure, f_1 is estimated in the semiparametric way introduced in Section 4 with the Epanechnikov kernel (bandwidth is 1.0). Both the BH procedure and the local FDR procedure need an estimate of p_0; we use the same estimating method in Formula (4.2) 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 semiparametric procedure.

We compare *three measures* from these procedures. First, we check whether the five procedures are valid, namely, whether the FDR yielded from these procedures is controlled at the prespecified level α. The prespecified FDR level α is 0.10, which is consistent with the multiple testing literature [Efron (2010)]. Second, we compare the FNR yielded by 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 is conservative. Our semiparametric 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 on the plots for FNR and ATP. These two completely dominate the parametric procedure, the BH procedure and the local FDR procedure,

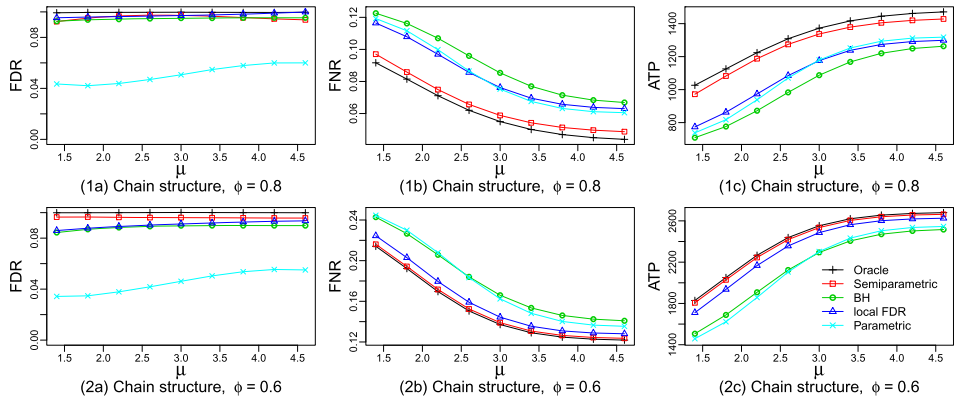


FIG. 4. *Performance of the procedures under Model* 1 *when* (1) $\phi = 0.8$ *and* (2) $\phi = 0.6$ *in terms of* (a) *FDR*, (b) *FNR and* (c) *ATP when the dependence structure is chain*.

indicating 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 para-

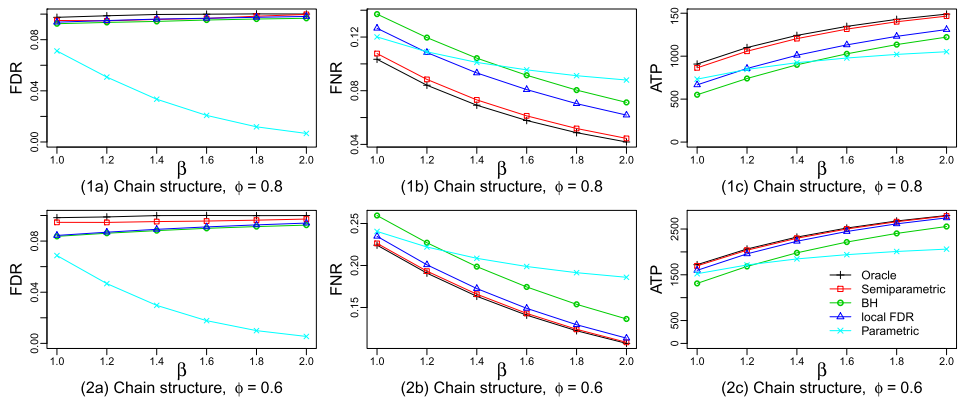


FIG. 5. *Performance of the procedures under Model* 2 *when* (1) $\phi = 0.8$ *and* (2) $\phi = 0.6$ *in terms of* (a) *FDR*, (b) *FNR and* (c) *ATP when the dependence structure is chain*.

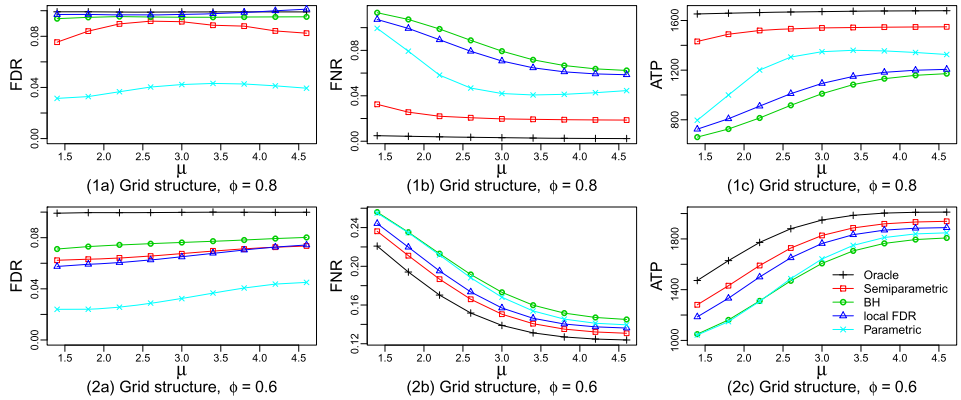


FIG. 6. *Performance of the procedures under Model* 1 *when* (1) $\phi = 0.8$ *and* (2) $\phi = 0.6$ *in terms of* (a) *FDR*, (b) *FNR and* (c) *ATP when the dependence structure is grid.*

metric procedure is considerably conservative. Again, our semiparametric data-driven procedure significantly outperforms the three baselines in all 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 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 structure and Model 1 with $\phi = 0.8$ is provided in Figure 8. Quite surprisingly, our data-driven

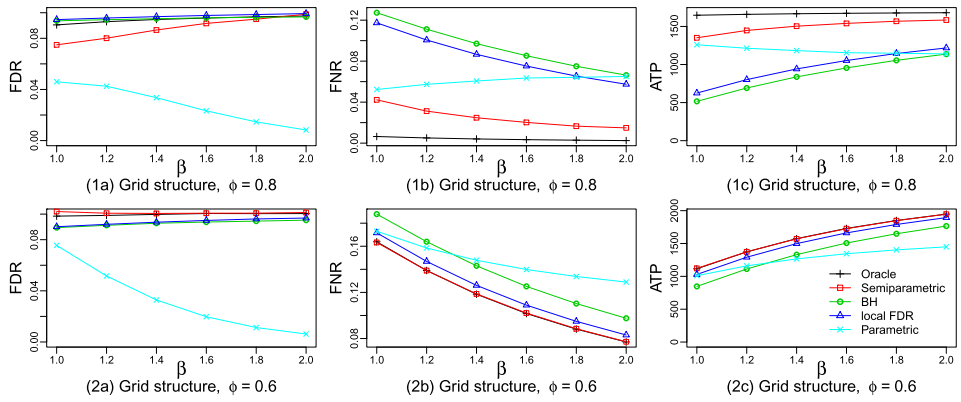


FIG. 7. *Performance of the procedures under Model* 2 *when* (1) $\phi = 0.8$ *and* (2) $\phi = 0.6$ *in terms of* (a) *FDR*, (b) *FNR and* (c) *ATP when the dependence structure is grid.*

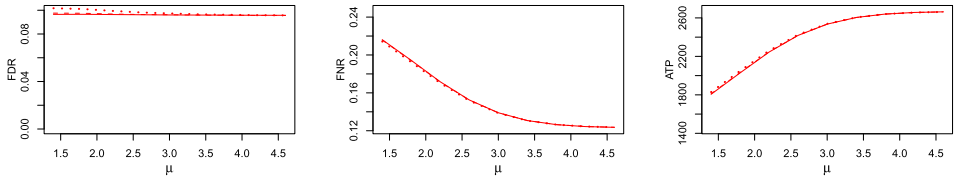


FIG. 8. *Performance of our procedure when* $\lambda = 0.2$ (*dotted lines*), 0.5 (*dashed lines*) *and* 0.8 (*solid lines*).

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, our approach is robust for different choices of λ. The robustness of λ was also observed in Storey (2002). The sensitivity analysis of λ in other configurations yields similar observations.

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 and the semiparametric procedure, we rank the hypotheses by the posterior probability that $\mathcal{H}_0$ is false, namely, $1 - \mathrm{LIS}$. For BH, we use $1 - p$-value. For local FDR procedure, we use $1 - \mathrm{lfdr}$. 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 the local FDR and the BH procedure are less efficient. The ROC curves and PR curves of these procedures under other configurations show similar behavior.

Run time: In the chain-structure simulations, it takes our data-driven procedures about 10∼20 hours to finish the 500 replications sequentially [for one μ value in

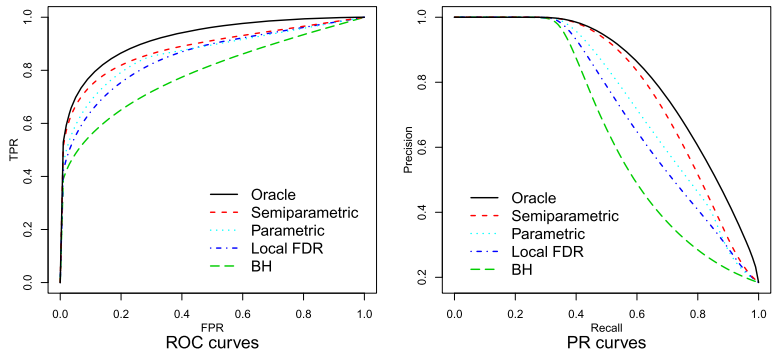


FIG. 9. *ROC and PR curves from these procedures.*

(5.1)] on one 3GHz CPU. In the grid-structure simulations, it takes our data-driven procedures around 30∼50 hours to finish the 500 replications sequentially [for one μ value in (5.1)] on one 3 GHz CPU.

6. Application. We apply our semiparametric procedure to a real-world GWAS on breast cancer [Hunter et al. (2007)] which involves 528,173 SNPs for 1145 cases and 1142 controls. In total, we test 528,173 hypotheses, and they are dependent because SNPs nearby tend to be highly correlated. By using our semiparametric procedure, we assume that each of the hypotheses has one of the two underlying true states, null (the SNP is not associated with breast cancer) or non-null (the SNP is associated with breast cancer). We assume that the dependence among the hypotheses is captured by a Markov random field over the underlying true states of the hypotheses. We build the structure of the MRF from the HapMap database [International HapMap Consortium (2003)], which quantifies the linkage disequilibrium (LD) of the human genome, the phenomenon that alleles that are close together in the genome tend to be inherited together. We assume that two SNPs in LD of greater strength tend to have a larger probability of sharing the same underlying true state. We query the squared correlation coefficients (r^2 values, which measures LD) 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. The probability of the underlying true state vector $\boldsymbol{\theta}$ from the MRF is

$$
\begin{aligned}
P(\boldsymbol{\theta}; \pi, \phi_h, \phi_m, \phi_l) = {} & \frac{1}{Z(\pi, \phi_h, \phi_m, \phi_l)} \prod_{i=1}^{m} \Psi_i^{\mathcal{V}}(\boldsymbol{\theta}; \pi) \\
& \times \prod_{j=1}^{|\mathcal{E}_h|} \Psi_j^{\mathcal{E}_h}(\boldsymbol{\theta}; \phi_h) \prod_{j=1}^{|\mathcal{E}_m|} \Psi_j^{\mathcal{E}_m}(\boldsymbol{\theta}; \phi_m) \prod_{j=1}^{|\mathcal{E}_l|} \Psi_j^{\mathcal{E}_l}(\boldsymbol{\theta}; \phi_l),
\end{aligned} \tag{6.1}
$$

where $Z(\pi, \phi_h, \phi_m, \phi_l)$ is the normalizing constant. $\Psi_i^{\mathcal{V}}$ and $\Psi_j^{\mathcal{E}}$ are the potential function on node i and the potential function on edge j, respectively.

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 2. 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, as listed in Table 2. 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

TABLE 2

Details of the SNP clusters identified by our semiparametric procedure, including the chromosome (Chr) and the physical position (PhyPos) they locate, the LIS value yielded by our semiparametric procedure, the p-value yielded from the individual test and the odds-ratio calculated on the second GWAS dataset

dbSNP ID	Chr	PhyPos	LIS	*p*-value	Odds-ratio
rs2288118	2	86,221,768	0	1.8E−04	1.18
rs1991106	2	86,227,832	0.0048	8.4E−04	1.17
rs1075622	2	86,249,588	0.0040	7.5E−05	1.15
rs2367202	2	86,257,194	0.0025	1.7E−04	1.18
rs1025104	2	86,262,322	0.0025	1.8E−04	1.20
rs4398317	2	13,6817,773	0	5.3E−04	1.17
rs4954580	2	13,6820,035	0.0047	9.4E−04	1.15
rs4440020	2	13,6824,059	0.0039	8.3E−04	1.17
rs4075810	2	13,6836,877	0.0058	8.8E−04	1.15
rs1970801	4	96,427,703	0.0072	1.2E−04	1.02
rs11097457	4	96,433,991	0.0083	1.9E−04	1.02
rs10819865	9	100,730,611	0	3.2E−04	1.06
rs1338733	9	100,737,703	0.0020	1.5E−04	1.08
rs1571581	9	100,738,024	0.0038	1.9E−04	1.07
rs12553370	9	100,756,745	0.0040	7.0E−04	1.07
rs11200014	10	123,324,920	0.0071	2.3E−05	1.20
rs1219648	10	123,336,180	0.0065	2.8E−05	1.15
rs2420946	10	123,341,314	0.0023	2.8E−05	1.15

cluster on Chr9 and the cluster on Chr10 are identified in the works 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 dataset to validate the 18 SNPs. The second GWAS dataset comes from the Marshfield Clinic. The Personalized Medicine Research Project [McCarty et al. (2005)], sponsored by the Marshfield Clinic, is used as the sampling frame to identify 162 breast cancer cases and 162 controls. The project is reviewed and approved by the Marshfield Clinic IRB. Subjects are selected using clinical data from the Marshfield Clinic Cancer Registry and Data Warehouse. Cases are defined as women having a confirmed diagnosis of breast cancer, which is obtained from the institutional cancer registry. Controls are confirmed through the Marshfield Clinic electronic medical records as never having had a breast cancer diagnosis by ICD-9 diagnosis code. Cases include both invasive breast cancer and ductal carcinoma in situ. We use an age matching strategy to construct case and control groups that are similar in age distribution. Specifically, we select a control whose age is within five years of the age of each case. The DNA samples

are genotyped using the Illumina HumanHap660 array as part of the eMERGE (electronic MEdical Records and Genomics) network [McCarty et al. (2011)].

On the second dataset, we calculate the odds ratio of the 18 SNPs, as listed in Table 2. It turns out that 16 of them show a moderate level of association. The five SNPs in the first cluster (on chromosome 2) have an odds ratio around 1.17–1.20. The four SNPs in the second cluster (on chromosome 2) have an odds ratio around 1.15–1.17. The two SNPs in the third cluster (on chromosome 4) have an odds ratio around 1.02. The four SNPs in the fourth cluster (on chromosome 9) have an odds ratio around 1.06–1.08. The three SNPs in the last cluster (on chromosome 10) have an odds ratio around 1.15–1.20.

7. The expectation–maximization algorithm. In this section, we provide details of the EM algorithm we use for the parametric procedure (in Section 3.1), and show that the lower bound of the log likelihood function $\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})$ nondecreases over the EM iterations (except for some MCMC error introduced in the E-step).

We begin with the lower bound of the log likelihood function, and then introduce the EM algorithm. Let $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ be any distribution on $\boldsymbol{\theta} \in \{0,1\}^m$. It is well known that there exists a lower bound of the log likelihood $\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})$ in (3.3), which is provided by an auxiliary function $\mathcal{F}(q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \{\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}\})$ defined as follows:

$$\begin{aligned}\mathcal{F}\big(q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \{\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}\}\big) &= \sum_{\boldsymbol{\theta} \in \{0,1\}^m} q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{P(\boldsymbol{\theta}, \mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})}{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \\ &= \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}) - \mathrm{KL}\big[q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) | P\big(\boldsymbol{\theta} | \mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}\big)\big],\end{aligned} \tag{7.1}$$

where $\mathrm{KL}[q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) | P(\boldsymbol{\theta} | \mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})]$ is the Kullback–Leibler divergence between $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $P(\boldsymbol{\theta} | \mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})$, the posterior distribution of the hidden variables. This Kullback–Leibler divergence is the distance between $\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})$ and $\mathcal{F}(q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \{\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}\})$.

Expectation–maximization: We maximize $\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})$ with an EM algorithm which iteratively maximizes its lower bound $\mathcal{F}(q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \{\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}\})$. We first initialize $\boldsymbol{\pi}^{(0)}$, $\boldsymbol{\phi}^{(0)}$ and $\boldsymbol{\psi}^{(0)}$. In the tth iteration, the updates in the expectation (E) step and the maximization (M) step are

$$q_{\boldsymbol{\theta}}^{(t)} = \operatorname*{argmax}_{q_{\boldsymbol{\theta}}} \mathcal{F}\big(q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \big\{\boldsymbol{\pi}^{(t-1)}, \boldsymbol{\phi}^{(t-1)}, \boldsymbol{\psi}^{(t-1)}\big\}\big) \qquad \text{(E)},$$

$$\boldsymbol{\pi}^{(t)}, \boldsymbol{\phi}^{(t)}, \boldsymbol{\psi}^{(t)} = \operatorname*{argmax}_{\{\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}\}} \mathcal{F}\big(q_{\boldsymbol{\theta}}^{(t)}, \{\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}\}\big) \qquad \text{(M)}.$$

In the E-step, we maximize $\mathcal{F}(q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \{\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\psi}^{(t-1)}\})$ with respect to $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. Because the difference between $\mathcal{F}(q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \{\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}\})$ and $\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})$ is $\mathrm{KL}[q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) | P(\boldsymbol{\theta} | \mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})]$, the maximizer in the E-step $q_{\boldsymbol{\theta}}^{(t)}$ is $P(\boldsymbol{\theta} | \mathbf{x}; \boldsymbol{\pi}^{(t-1)}, \boldsymbol{\phi}^{(t-1)}, \boldsymbol{\psi}^{(t-1)})$, namely, the posterior distribution of $\boldsymbol{\theta} | \mathbf{x}$ under the current estimated parameters $\boldsymbol{\pi}^{(t-1)}$, $\boldsymbol{\phi}^{(t-1)}$ and $\boldsymbol{\psi}^{(t-1)}$. This posterior distribution can be

calculated by Markov chain Monte Carlo for general graphs. For graphs with special structures (such as trees), exact algorithms with computational cost linear in the number of variables (such as sum-product algorithm) may be available. Please refer to Section 3.2 for more details.

In the M-step, we maximize $\mathcal{F}(q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}), \{\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}\})$ with respect to $\{\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}\}$, which can be rewritten as

$$\begin{aligned}
&\operatorname*{argmax}_{\{\boldsymbol{\pi},\boldsymbol{\phi},\boldsymbol{\psi}\}} \mathcal{F}\big(q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}), \{\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}\}\big) \\
&\quad= \operatorname*{argmax}_{\{\boldsymbol{\pi},\boldsymbol{\phi},\boldsymbol{\psi}\}} \sum_{\boldsymbol{\theta}\in\{0,1\}^m} q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \log P(\boldsymbol{\theta}, \mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}) \\
&\quad= \operatorname*{argmax}_{\{\boldsymbol{\pi},\boldsymbol{\phi},\boldsymbol{\psi}\}} \sum_{\boldsymbol{\theta}\in\{0,1\}^m} q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \big\{\log P(\boldsymbol{\theta}; \boldsymbol{\pi}, \boldsymbol{\phi}) + \log P(\mathbf{x}|\boldsymbol{\theta}; \boldsymbol{\psi})\big\}.
\end{aligned}$$

It is obvious that this function can be maximized with respect to $\{\boldsymbol{\pi}, \boldsymbol{\phi}\}$ and $\boldsymbol{\psi}$ separately as

$$\begin{aligned}
\boldsymbol{\pi}^{(t)}, \boldsymbol{\phi}^{(t)} &= \operatorname*{argmax}_{\boldsymbol{\pi},\boldsymbol{\phi}} \sum_{\boldsymbol{\theta}\in\{0,1\}^m} q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \log P(\boldsymbol{\theta}; \boldsymbol{\pi}, \boldsymbol{\phi}), \\
\boldsymbol{\psi}^{(t)} &= \operatorname*{argmax}_{\boldsymbol{\psi}} \sum_{\boldsymbol{\theta}\in\{0,1\}^m} q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \log P(\mathbf{x}|\boldsymbol{\theta}; \boldsymbol{\psi}).
\end{aligned} \tag{7.2}$$

Estimating $\boldsymbol{\psi}$: Estimating $\boldsymbol{\psi}$ in this maximum likelihood manner is straightforward because the maximization can be rewritten as follows:

$$\begin{aligned}
&\operatorname*{argmax}_{\boldsymbol{\psi}} \sum_{\boldsymbol{\theta}\in\{0,1\}^m} q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \log P(\mathbf{x}|\boldsymbol{\theta}; \boldsymbol{\psi}) \\
&\quad= \operatorname*{argmax}_{\boldsymbol{\psi}} \sum_{i=1}^{m} \sum_{\theta_i\in\{0,1\}} q_{\theta_i}^{(t)}(\theta_i) \log P(x_i|\theta_i; \boldsymbol{\psi}),
\end{aligned} \tag{7.3}$$

where $q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) = \prod_{i=1}^m q_{\theta_i}^{(t)}(\theta_i)$. Because we usually know the density function of $x_i|\theta_i = 0$, Formula (7.3) can be simplified as

$$\begin{aligned}
&\operatorname*{argmax}_{\boldsymbol{\psi}} \sum_{\boldsymbol{\theta}\in\{0,1\}^m} q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \log P(\mathbf{x}|\boldsymbol{\theta}; \boldsymbol{\psi}) \\
&\quad= \operatorname*{argmax}_{\boldsymbol{\psi}} \sum_{i=1}^{m} q_{\theta_i}^{(t)}(\theta_i = 1) \log P(x_i|\theta_i = 1; \boldsymbol{\psi}).
\end{aligned} \tag{7.4}$$

For many parametric forms of f_1 (such as a Gaussian density), this estimation step can be solved in a maximum likelihood manner since the likelihood function in Formula (7.4) is log-concave.

Estimating $\{\boldsymbol{\pi}, \boldsymbol{\phi}\}$: Estimating $\{\boldsymbol{\pi}, \boldsymbol{\phi}\}$ in Formula (7.2) is difficult due to the intractable $Z(\boldsymbol{\pi}, \boldsymbol{\phi})$. Some approaches [Celeux, Forbes and Peyrard (2003), Zhang,

Brady and Smith (2001)] use the pseudo-likelihood [Besag (1975)] to estimate $\{\boldsymbol{\pi}, \boldsymbol{\phi}\}$ in the M-step. It can be shown that $\sum_{\boldsymbol{\theta}\in\{0,1\}^m} q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \log P(\boldsymbol{\theta}; \boldsymbol{\pi}, \boldsymbol{\phi})$ is *concave* with respect to $\{\boldsymbol{\pi}, \boldsymbol{\phi}\}$. Therefore, we can use the gradient ascent to find the MLE of $\{\boldsymbol{\pi}, \boldsymbol{\phi}\}$, which is similar to using contrastive divergence [Hinton (2002)] to learn MRFs, except we have to reweigh it to $q_{\boldsymbol{\theta}}^{(t)}$.

We run the EM algorithm until both $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ converge. Therefore, it is easy to tell that the lower bound of the log likelihood function $\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi})$ nondecreases over the EM iterations (except for some MCMC error introduced in the E-step).

8. Discussion. In this paper, we use an MRF-coupled mixture model to leverage the dependence in multiple testing problems, and show the improved numerical performance on a variety of simulations and its applicability in a real-world GWAS problem. We provide two versions of this approach—one parametric procedure which can be used in the situation where f_1 can be estimated parametrically, and one semiparametric procedure which can be used in more general situations. From the methodological standpoint, the 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, the semiparametric approach no longer requires the investigators to know the parameterization of f_1, which is generally unknown in practical problems. Nevertheless, the semiparametric procedure may not perform well when there are only a small number of non-null hypotheses and it is challenging to reliably estimate f_1 (see Appendix). 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, (ii) there is no suitable parametric distribution for f_1, and (iii) it is expected there are enough non-null hypotheses (e.g., $m_1 \geq 1{,}000$) to estimate f_1 reliably. Otherwise, we suggest that investigators choose a proper parametric form for f_1 and use the parametric procedure to leverage the dependence among the hypotheses.

Theoretically, one question of interest is whether this graphical model-based procedure is optimal in the sense that it has the smallest FNR among all valid procedures. The optimality of the oracle procedure can be proved under the compound decision framework [Sun and Cai (2007, 2009)], as long as an exact inference algorithm exists or an approximate inference algorithm can be guaranteed to converge to the correct marginal probabilities. The asymptotic optimality of the data-driven procedures (the FNR yielded by the data-driven procedures approaches the FNR yielded by the oracle procedure as the number of tests $m \to \infty$) requires consistent estimates of the unknown parameters in the graphical models. Parameter estimation in undirected graphical models is more complicated than in directed

graphical models due to the normalization constant. To the best of our knowledge, asymptotic properties of parameter estimation for MRF-coupled mixture models have not been investigated. Therefore, we cannot prove the asymptotic optimality of the data-driven procedure so far, although we can observe its close-to-oracle performance in the basic simulations. Note that our conclusion here agrees with the remark from Sun and Cai (2009) that "the optimality of the LIS procedure may be lost in the estimation step" because "theoretical results (consistency of the estimates) for other dependence structures have not been developed in the literature." We believe that the asymptotic optimality of the data-driven procedure in general dependence structures will be an important problem in future work.

APPENDIX: INVESTIGATION ON SITUATIONS WHEN THERE ARE NOT ENOUGH NON-NULL HYPOTHESES

In this section, we investigate the performance of our parametric and semiparametric procedures when there are only a small number of non-null hypotheses via additional simulations. In the simulations, we use the grid structure with $\phi = 0.6$. In order to simulate different proportions of non-null hypotheses, we use six different values for π, including 0.05, 0.10, 0.15, 0.20, 0.25 and 0.30, which yield 1.1%, 2.5%, 4.1%, 6.2%, 9.0% and 12.7% non-null hypotheses, respectively. In the simulations, we test a grid structure of three sizes, namely, 100×100, 200×200 and 300×300, yielding three m values, namely, 10,000, 40,000 and 90,000. For the null hypotheses, the test statistics are simulated from a standard normal distribution $\mathcal{N}(0, 1.0)$. For the non-null hypotheses, the test statistics are simulated from a normal distribution $\mathcal{N}(3.0, 1.0)$.

After we simulate the underlying true states of the hypotheses and the test statistics, we apply Storey's estimator of p_0 in Formula (4.2) and our estimator of f_1 in Formula (4.3). In Storey's estimator of p_0, we set λ to be 0.5. Then, we compare their performance with three measures, as follows:

1. $|\hat{p}_0 - p_0|$: the difference between Storey's estimator of p_0 and ground truth of p_0,
2. IMSE($\hat{f}_1$): integrated mean squared error of the estimator of f_1 in our semiparametric procedure, and
3. IMSE($\tilde{f}_1$): integrated mean squared error of $\tilde{f}_1$, the oracle nonparametric estimator of f_1 which knows the underlying true states of the hypotheses (i.e., $\tilde{f}_1$ directly estimates f_1 via kernel density estimation from the test statistics of these non-null hypotheses).

The performance of different estimators is provided in Figure 10. We have three observations as follows:

1. Storey's estimator of p_0 performs better when there are a smaller proportion of non-null hypotheses and when there are more hypotheses to test.

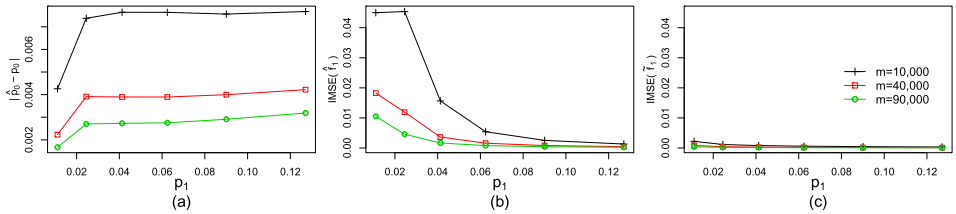


FIG. 10. *The performance of different estimators as the true proportion of non-null hypotheses (denoted by p_1) increases*, (a) *the difference between Storey's estimator of p_0 (denoted by $\hat{p}_0$) and ground truth of p_0*. (b) *IMSE of $\hat{f}_1$, the estimator of f_1 in our semiparametric procedure, and* (c) *IMSE of the oracle nonparametric estimator of f_1 which knows the underlying true states of the hypotheses*.

2. Our estimator of f_1 performs better when there are a larger proportion of non-null hypotheses and when there are more hypotheses to test.
3. If there are more than 1,000 non-null hypotheses (e.g., 10% of 10,000 hypotheses are non-null), our estimator of f_1 yields satisfactory performance (comparable with the performance of the oracle nonparametric estimator $\tilde{f}_1$).

Furthermore, we apply the five procedures, including an oracle procedure, local FDR procedure, BH procedure, our parametric procedure and our semiparametric procedure. The performance (FDR, FNR and ATP) of the five procedures ($m = 10{,}000$) is provided in Figure 11. It is observed that our semiparametric procedure becomes overliberal, as there are fewer non-null hypotheses. However, it is observed that our parametric procedure performs well, controlling FDR at the prespecified level of 0.10 and reducing FNR (outperforming the local FDR procedure and BH procedure). Therefore, the numerical results suggest that the semiparametric procedure may not perform well when there are only a small number of non-null hypotheses and it is challenging to reliably estimate f_1. In this situation, the parametric procedure can be considered if f_1 can be properly parametrized.

AVAILABLE SOFTWARE

The software implementation of the multiple testing procedure is available via http://www.cs.wisc.edu/~jieliu/mtd/software.html.

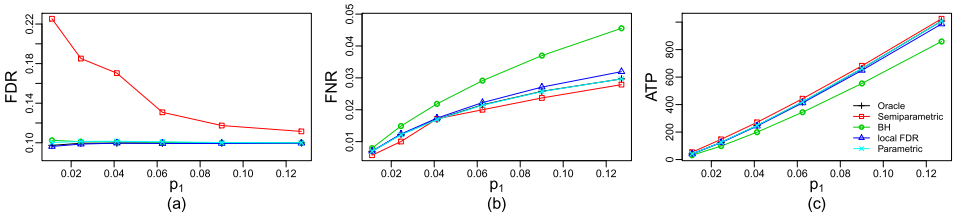


FIG. 11. *Performance of different procedures in the new added simulations as the true proportion of non-null hypotheses (denoted by p_1) increases in terms of* (a) *FDR*, (b) *FNR and* (c) *ATP*.

REFERENCES

BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* **41** 164–171. MR0287613

BENJAMINI, Y. and HELLER, R. (2007). False discovery rates for spatial signals. *J. Amer. Statist. Assoc.* **102** 1272–1281. MR2412549

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BENJAMINI, Y. and HOCHBERG, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* **25** 60–83.

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

BESAG, J. (1975). Statistical analysis of non-lattice data. *J. R. Stat. Soc., Ser. D Stat.* **24** 179–195.

BLANCHARD, G. and ROQUAIN, É. (2009). Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.* **10** 2837–2871. MR2579914

CELEUX, G., FORBES, F. and PEYRARD, N. (2003). EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recogn.* **36** 131–144.

CROUSE, M. S., NOWAK, R. D. and BARANIUK, R. G. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Process.* **46** 886–902. MR1665651

EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302

EFRON, B. (2010). *Large-Scale Inference. Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758

EFRON, B. and TIBSHIRANI, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **23** 70–86.

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571

EPANECHNIKOV, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory Probab. Appl.* **14** 153–158.

FAN, J., HAN, X. and GU, W. (2012). Control of the false discovery rate under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1045.

FARCOMENI, A. (2007). Some results on the control of the false discovery rate under dependence. *Scand. J. Stat.* **34** 275–297. MR2346640

FINNER, H. and ROTERS, M. (2002). Multiple hypotheses testing and expected number of type I errors. *Ann. Statist.* **30** 220–238. MR1892662

FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415. MR2750571

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. MR1141740

GENOVESE, C. R., ROEDER, K. and WASSERMAN, L. (2006). False discovery control with p-value weighting. *Biometrika* **93** 509–524. MR2261439

GENOVESE, C. and WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 499–517. MR1924303

GENOVESE, C. and WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32** 1035–1061. MR2065197

GUEDJ, M., ROBIN, S., CELISSE, A. and NUEL, G. (2009). Kerfdr: A semi-parametric kernel-based approach to local false discovery rate estimation. *BMC Bioinformatics* **10** 84.

HINTON, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14** 1771–1800.

HUNTER, D. J., KRAFT, P., JACOBS, K. B., COX, D. G., YEAGER, M., HANKINSON, S. E., WACHOLDER, S., WANG, Z., WELCH, R., HUTCHINSON, A., WANG, J., YU, K., CHATTERJEE, N., ORR, N., WILLETT, W. C., COLDITZ, G. A., ZIEGLER, R. G., BERG, C. D., BUYS, S. S., MCCARTY, C. A., FEIGELSON, H. S., CALLE, E. E., THUN, M. J., HAYES, R. B., TUCKER, M., GERHARD, D. S., FRAUMENI, J. F., HOOVER, R. N., THOMAS, G. and CHANOCK, S. J. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39** 870–874.

INTERNATIONAL HAPMAP CONSORTIUM (2003). The international HapMap project. *Nature* **426** 789–796.

JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.

KIM, D. and ZHANG, C. (2014). Adaptive linear step-up multiple testing procedure with the bias-reduced estimator. *Statist. Probab. Lett.* **87** 31–39. MR3168932

KSCHISCHANG, F. R., FREY, B. J. and LOELIGER, H.-A. (2001). Factor graphs and the sum-product algorithm. *IEEE Trans. Inform. Theory* **47** 498–519. MR1820474

LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Roy. Statist. Soc. Ser. B* **50** 157–224. MR0964177

LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105** 18718–18723.

LIANG, K. and NETTLETON, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 163–182. MR2885844

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A. et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.

MCCARTY, C. A., WILKE, R. A., GIAMPIETRO, P. F., WESBROOK, S. D. and CALDWELL, M. D. (2005). Marshfield clinic personalized medicine research project (PMRP): Design, methods and recruitment for a large population-based biobank. *Personalized Medicine* **2** 49–79.

MCCARTY, C. A., CHISHOLM, R. L., CHUTE, C. G., KULLO, I. J., JARVIK, G. P., LARSON, E. B., LI, R., MASYS, D. R., RITCHIE, M. D., RODEN, D. M. et al. (2011). The eMERGE network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Medical Genomics* **4** 13.

MURPHY, K. P., WEISS, Y. and JORDAN, M. I. (1999). Loopy belief propagation for approximate inference: An empirical study. In *UAI* 467–475.

NGUYEN, V. H. and MATIAS, C. (2014). Nonparametric estimation of the density of the alternative hypothesis in a multiple testing setup. Application to local false discovery rate estimation. *ESAIM Probab. Stat.* **18** 584–612. MR3334005

OWEN, A. B. (2005). Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 411–426. MR2155346

ROBBINS, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1950 131–148. Univ. California Press, Berkeley. MR0044803

ROBBINS, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. i* 157–163. Univ. California Press, Berkeley. MR0084919

ROBIN, S., BAR-HEN, A., DAUDIN, J.-J. and PIERRE, L. (2007). A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Comput. Statist. Data Anal.* **51** 5483–5493. MR2407654

ROMANO, J. P., SHAIKH, A. M. and WOLF, M. (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling. *TEST* **17** 417–442. MR2470085

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.* **27** 832–837. MR0079873

SARKAR, S. K. (2006). False discovery and false nondiscovery rates in single-step multiple testing procedures. *Ann. Statist.* **34** 394–415. MR2275247

SATROM, P., BIESINGER, J., LI, S. M., SMITH, D., THOMAS, L. F., MAJZOUB, K., RIVAS, G. E., ALLUIN, J., ROSSI, J. J., KRONTIRIS, T. G., WEITZEL, J., DALY, M. B., BENSON, A. B., KIRKWOOD, J. M., ODWYER, P. J., SUTPHEN, R., STEWART, J. A., JOHNSON, D. and LARSON, G. P. (2009). A risk variant in an miR-125b binding site in BMPR1B is associated with breast cancer pathogenesis. *Cancer Res.* **69** 7459–7465.

SCHRAUDOLPH, N. N. (2010). Polynomial-time exact inference in NP-hard binary MRFs via reweighted perfect matching. In *AISTATS*.

SCHRAUDOLPH, N. N. and KAMENETSKY, D. (2009). Efficient exact inference in planar Ising models. In *NIPS*.

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498. MR1924302

STOREY, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2013–2035. MR2036398

STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 187–205. MR2035766

SUN, W. and CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657

SUN, W. and CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 393–424. MR2649603

SVENSON, U., NORDFJÄLL, K., STEGMAYR, B., MANJER, J., NILSSON, P., TAVELIN, B., HENRIKSSON, R., LENNER, P. and ROOS, G. (2008). Breast cancer survival is associated with telomere length in peripheral blood cells. *Cancer Res.* **68** 3618–3623.

TIELEMAN, T. (2008). Training restricted Boltzmann machines using approximations to the likelihood gradient. In *ICML* 1064–1071.

WAINWRIGHT, M. J., JAAKKOLA, T. S. and WILLSKY, A. S. (2003a). Tree-reweighted belief propagation algorithms and approximate ML estimation via pseudo-moment matching. In *AISTATS*.

WAINWRIGHT, M. J., JAAKKOLA, T. S. and WILLSKY, A. S. (2003b). Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Trans. Inform. Theory* **49** 1120–1146. MR1984817

WEI, Z., SUN, W., WANG, K. and HAKONARSON, H. (2009). Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics* **25** 2802–2808.

WEISS, Y. (2000). Correctness of local probability propagation in graphical models with loops. *Neural Comput.* **12** 1–41.

WELLING, M. and SUTTON, C. (2005). Learning in Markov random fields with contrastive free energies. In *AISTATS*.

WU, W. B. (2008). On false discovery control under dependence. *Ann. Statist.* **36** 364–380. MR2387975

XIAO, J., ZHU, W. and GUO, J. (2013). Large-scale multiple testing in genome-wide association studies via region-specific hidden Markov models. *BMC Bioinformatics* **14** 282.

YEDIDIA, J. S., FREEMAN, W. T. and WEISS, Y. (2000). Generalized belief propagation. In *NIPS* 689–695. MIT Press, Cambridge.

YEKUTIELI, D. and BENJAMINI, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Statist. Plann. Inference* **82** 171–196. MR1736442

ZHANG, Y., BRADY, M. and SMITH, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imag.* **20** 45–57.

ZHANG, C., FAN, J. and YU, T. (2011). Multiple testing via FDR_L for large-scale imaging data. *Ann. Statist.* **39** 613–642. MR2797858

J. LIU
UNIVERSITY OF WISCONSIN-MADISON
1300 UNIVERSITY AVENUE, RM 6725
MADISON, WISCONSIN 53706
USA
E-MAIL: jieliu@cs.wisc.edu

CHUNMING ZHANG
UNIVERSITY OF WISCONSIN-MADISON
1300 UNIVERSITY AVENUE, RM 1155
MADISON, WISCONSIN 53706
USA
E-MAIL: cmzhang@stat.wisc.edu

D. PAGE
UNIVERSITY OF WISCONSIN-MADISON
330 N. ORCHARD ST. RM 3174
MADISON, WISCONSIN 53715
USA
E-MAIL: page@biostat.wisc.edu

The code together with instructions and the data files can be downloaded here

- The codes are provided AS IS, without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose and noninfringement. Do Not Distribute.
- For questions and special need, please contact Jie Liu by email.