Annals of the Institute of Statistical Mathematics (2021) 73:891–920
https://doi.org/10.1007/s10463-020-00770-3

Asymptotic theory of dependent Bayesian multiple testing procedures under possible model misspecification

Noirrit Kiran Chandra[1] · **Sourabh Bhattacharya**[2]

Received: 13 May 2020 / Revised: 4 September 2020 / Accepted: 29 September 2020 /
Published online: 13 November 2020

Abstract

We study asymptotic properties of Bayesian multiple testing procedures and provide sufficient conditions for strong consistency under general dependence structure. We also consider a novel Bayesian multiple testing procedure and associated error measures that coherently accounts for the dependence structure present in the model. We advocate posterior versions of FDR and FNR as appropriate error rates and show that their asymptotic convergence rates are directly associated with the Kullback–Leibler divergence from the true model. The theories hold regardless of the class of postulated models being misspecified. We illustrate our results in a variable selection problem with autoregressive response variables and compare our procedure with some existing methods through simulation studies. Superior performance of the new procedure compared to the others indicates that proper exploitation of the dependence structure by multiple testing methods is indeed important. Moreover, we obtain encouraging results in a maize dataset, where we select influential marker variables.

Keywords Bayesian multiple testing · Variable selection · False discovery rate · Kullback–Leibler · Misspecified model · Posterior convergence

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-020-00770-3) contains supplementary material, which is available to authorized users.

✉ Noirrit Kiran Chandra
noirritchandra@gmail.com

[1] Department of Statistics and Data Science, University of Texas at Austin, 2317 Speedway D9800, Austin, TX 78712-1823, USA

[2] Interdisciplinary Statistical Research Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata, WB 700108, India

1 Introduction

In recent times, there has been a tremendous growth in the area of multiple hypothesis testing as simultaneous inference on several parameters is often necessary. Benjamini and Hochberg (1995) introduced a powerful approach to handle this problem in their landmark paper. However, in most real-life situations the test statistics are generally dependent. Benjamini and Yekutieli (2001) showed that the Benjamini–Hochberg procedure is valid under positive dependence. Berry and Hochberg (1999) have given a Bayesian perspective on multiple testing where the tests are related through a dependent prior. Scott and Berger (2010) discussed how empirical Bayes and fully Bayes methods adjust multiplicity.

There are many works in the statistical literature on optimality and asymptotic behaviour of multiple testing methods in dependent cases. Sun and Cai (2007) have proposed an optimal adaptive procedure where the data are generated from a two-component mixture model. Finner and Roters (2002), Efron (2007) discussed the effects of dependence of error rates, among others. Finner et al. (2009) proposed new step-up and step-down procedures which asymptotically maximize power while controlling the *false discovery rate* (FDR). Xie et al. (2011) have proposed an asymptotic optimal decision rule for short-range dependent data with dependent test statistics.

In this article, we study asymptotic properties of loss function-based Bayesian multiple testing procedures under general dependence setup. We show that under mild conditions such procedures are consistent in the sense that the decision rules converge to the truth with increasing sample size, even under dependence. We also show that the derived results hold even when the class of postulated models do not contain the true data-generating process, that is, when the class of proposed models is misspecified.

Finner et al. (2007) discussed the effect of dependent test statistics on the FDR. Schwartzman and Lin (2011) and Fan et al. (2012) discussed estimation of FDR under correlation. In the frequentist multiple testing domain, the common practice is to control FDR or the *false non-discovery rate* (FNR). Therefore, in that domain, asymptotic study of FDR or FNR in dependent cases has been done under different setups. However, in the Bayesian literature, asymptotic study of the aforementioned error rates is not regular, although in practice, it is necessary to control those error rates. In this article, we conduct asymptotic analyses on these error rates under general dependent setup. We show that these error rates are directly associated with the Kullback–Leibler (KL) divergence from the true model in terms of their asymptotic convergence rates.

In the frequentist multiple testing setup, the decision rule for a hypothesis generally depends only on the corresponding test statistics. Bayesian loss function-based multiple testing methods are generally based on marginal posterior probabilities of a null hypothesis being true or false. Most of the existing methods are marginal in the sense that the decision rule for a hypothesis does not depend on decisions of other hypotheses. Indeed, an important issue that seems to have received relatively less attention is that by proper utilization of the dependence

structure among different hypotheses, the efficiency of multiple testing procedures can be significantly improved. Sun and Cai (2009) have showed that incorporating the dependence structure of the parameters in the testing procedure increases efficiency.

The aforementioned discussion points toward taking decisions regarding the hypotheses jointly. In this regard, Chandra and Bhattacharya (2019) developed a novel Bayesian multiple testing method which coherently takes the dependence structure among the hypotheses into consideration. In their method, the decisions are obtained jointly, as functions of appropriate joint posterior probabilities, and hence, the method is referred to as a non-marginal Bayesian procedure. The procedure is based on new notions of error and non-error terms associated with breaking up the total number of hypotheses. They have shown that by virtue of the joint decision-making principle, the non-marginal procedure has the desirable compound decision theoretic properties and for large samples minimizes the KL divergence from the true data-generating process, under general dependence models. Further, with extensive simulation studies they demonstrate significant gain in power over the existing marginal multiple testing methods, both classical and Bayesian. Application of this method to a deregulated micro-RNA discovery problem yielded insightful results which could not be obtained otherwise (Chandra et al. 2019). In the following section, we briefly describe the multiple testing procedure.

1.1 A new non-marginal Bayesian multiple testing procedure

Let $\boldsymbol{X}_n = \{X_1, \ldots, X_n\}$ denote the available data set. Suppose the data are modelled by the family of distributions $P_{\boldsymbol{X}_n|\boldsymbol{\xi}}$ (which may also be nonparametric). For $M > 1$, let us denote by $\boldsymbol{\Xi} = \Theta_1 \times \cdots \times \Theta_M$ the relevant parameter space associated with $\boldsymbol{\xi} = (\theta_1, \ldots, \theta_M)$, where we allow M to be infinity as well. Let $P_{\boldsymbol{\xi}|\boldsymbol{X}_n}(\cdot)$ and $E_{\boldsymbol{\xi}|\boldsymbol{X}_n}(\cdot)$ denote the posterior distribution and expectation, respectively, of $\boldsymbol{\xi}$ given $\boldsymbol{X}_n$, and let $P_{\boldsymbol{X}_n}(\cdot)$ and $E_{\boldsymbol{X}_n}(\cdot)$ denote the marginal distribution and expectation of $\boldsymbol{X}_n$, respectively. Let us consider the problem of testing m hypotheses simultaneously corresponding to the actual parameters of interest, where $1 < m \leq M$. In this work, however, we assume m to be finite.

Without loss of generality, let us consider testing the parameters associated with Θ_i; $i = 1, \ldots, m$, formalized as:

$$H_{0i} : \theta_i \in \Theta_{0i} \text{ versus } H_{1i} : \theta_i \in \Theta_{1i},$$

where $\Theta_{0i} \bigcap \Theta_{1i} = \emptyset$ and $\Theta_{0i} \bigcup \Theta_{1i} = \Theta_i$, for $i = 1, \cdots, m$.

Let

$$d_i = \begin{cases} 1 & \text{if the } i\text{-th hypothesis is rejected;} \\ 0 & \text{otherwise;} \end{cases}$$

$$r_i = \begin{cases} 1 & \text{if } H_{1i} \text{ is true;} \\ 0 & \text{if } H_{0i} \text{ is true.} \end{cases}$$

In many real-life situations, dependent prior structure is envisaged on the parameter space based on available domain knowledge. For example in spatial statistics, Gaussian process prior is often considered. In fMRI data, Gaussian Markov random field prior is a common prior. In such cases, the additional information on the parameters is incorporated in the model through the prior distribution. Various applications in recent times in fields as diverse as spatial statistics and environment (Risser et al. 2019), time series (Scott 2009), neurosciences (Brown et al. 2014), biological sciences (Jensen et al. 2009), to name only a few, consider Bayesian models with dependent prior structures. The basic idea behind the new multiple testing methodology is to incorporate such information, when available, in the testing procedure to obtain improved decision rule. This principle is in accordance with the traditional Bayesian philosophy that when prior information is available, inference can be enhanced.

Let G_i be the set of hypotheses (including hypothesis i) where the parameters are dependent on θ_i. In the new procedure, the decision of each hypothesis is penalized by incorrect decisions regarding other dependent parameters resulting in a compound criterion where all the decisions in G_i deterministically depend upon each other. Define the following quantity

$$z_i = \begin{cases} 1 \text{ if } H_{d_j j} \text{ is true for all } j \in G_i \setminus \{i\}; \\ 0 \text{ otherwise.} \end{cases} \tag{1}$$

If for any $i \in \{1, \ldots, m\}$, $G_i = \{i\}$, a singleton, then we define $z_i = 1$. The notion of true positives (TP) is modified as follows

$$TP = \sum_{i=1}^{m} d_i r_i z_i. \tag{2}$$

The posterior expectation of TP is maximized subject to controlling the posterior expectation of the error term

$$E = \sum_{i=1}^{m} d_i (1 - r_i z_i). \tag{3}$$

It follows that the decision configuration can be obtained by minimizing the function

$$\begin{aligned} \imath(\boldsymbol{d}) &= -\sum_{i=1}^{m} d_i E_{\xi|\boldsymbol{X}_n}(r_i z_i) + \lambda_n \sum_{i=1}^{m} d_i E_{\xi|\boldsymbol{X}_n}(1 - r_i z_i) \\ &= -(1 + \lambda_n) \sum_{i=1}^{m} d_i \left(w_{in}(\boldsymbol{d}) - \frac{\lambda_n}{1 + \lambda_n} \right), \end{aligned}$$

with respect to all possible decision configurations of the form $\boldsymbol{d} = \{d_1, \ldots, d_m\}$, where $\lambda_n > 0$ and

$$w_{in}(\boldsymbol{d}) = E_{\xi|\boldsymbol{X}_n}(r_i z_i) = P_{\xi|\boldsymbol{X}_n}\left(H_{1i} \cap \left\{ \cap_{j \neq i, j \in G_i} H_{d_j j} \right\} \right),$$

is the posterior probability of the decision configuration $\{d_1, \ldots, d_{i-1}, 1, d_{i+1}, \ldots, d_m\}$ being correct. Letting $\beta_n = \lambda_n/(1+\lambda_n)$, one can equivalently maximize

$$f_{\beta_n}(\boldsymbol{d}) = \sum_{i=1}^{m} d_i \big(w_{in}(\boldsymbol{d}) - \beta_n\big) \tag{4}$$

with respect to $\boldsymbol{d}$ and obtain the optimal decision configuration.

Definition 1 Let $\mathbb{D}$ be the set of all m-dimensional binary vectors denoting all possible decision configurations. Define

$$\widehat{\boldsymbol{d}} = \underset{\boldsymbol{d}\in\mathbb{D}}{\operatorname{argmax}} f_{\beta}(\boldsymbol{d})$$

where $0 < \beta < 1$. Then, $\widehat{\boldsymbol{d}}$ is the *optimal decision configuration* obtained as the solution of the non-marginal multiple testing method.

Note that in the definitions of both TP and E, we penalize d_i by incorrect decisions in the same group. Thus, we design a compound criterion where decisions regarding dependent parameters deterministically depend upon each other adjudging other dependent parameters.

It is to be noted that there exist several cluster-based methods in the literature of multiple hypotheses testing. The works of Benjamini and Heller (2007), Sun et al. (2015) are important to mention in this respect, among others. However, the G_is in (1) are not to be confused with the notion of clusters in the aforementioned works. In their approaches, a particular cluster of parameters is regarded as a signal or not. Essentially, the decisions regarding the parameters in their clusters are same. However, that is not the case for our non-marginal method. The motivation behind our grouping is to borrow strength through the dependence structure across dependent parameters. This is a common practice in various applications (Zhang et al. 2011; Liu et al. 2016).

1.2 Choice of $G_1, \ldots, G_m$

Note that the non-marginal method depends on the choice of G_is. However, in implementation of the method, forming groups based on all dependent parameters might be disadvantageous in high-dimensional cases. Keeping very weakly dependent parameters in G_i would increase the complexity of the method without providing much extra information about the dependence structure. It would incur over-penalization levying high posterior probability of $z_i = 0$. This might turn the method to be overly conservative. Therefore, we recommend to restrict the group sizes proportional to the correlation among the parameters. Chandra and Bhattacharya (2019) have prescribed the following strategy of group formation.

Let Λ be the prior correlation matrix of $\boldsymbol{\xi}$. Let the (i, j)-th element of Λ be λ_{ij}. We first consider the correlations between the i-th and j-th parameters, with $i < j$, and obtain a desired percentile λ of these quantities. Then, in G_i we include only those

indices $j\ (\neq i)$ such that $\lambda_{ij} \geq \lambda$. Thus, the i-th group contains indices of the parameters that are highly correlated with the i-th parameter. If there exists no index j such that $\lambda_{ij} \geq \lambda$, then $G_i = \{i\}$. In our applications, we have considered λ to be the 95-*th* percentile, which is seen to have yielded good results.

Once the prior associated with the model is decided and well chosen, the G_is as defined above will also be fixed and would lead to reliable results. In case the prior information on the correlation structure of the parameters is weak, Λ can be considered as the posterior correlation matrix of the parameters. Groups formed on the basis of the true correlation give the best result as expected. However, groups formed on the basis of posterior correlation significantly improve the performance (Chandra and Bhattacharya 2019). In Sect. 6, the groups are formed on the basis of posterior correlation and the strategy has outperformed some popular existing multiple testing methods in a variable selection context.

Notably for large samples, Bayesian methods are usually robust with respect to prior choice and there is a huge literature formalizing this aspect. For example Schwartz (1965), Ghosal et al. (2000) discussed that Bayesian models are asymptotically consistent given that the priors satisfy certain regularity conditions. In the same vein, we study the asymptotic properties of the Bayesian non-marginal method in this article and show that the procedure is asymptotically robust with respect to the choice of group structure later in Sect. 2.4. In the same section, we provide sufficient conditions for the asymptotic consistency of the non-marginal method. For illustrative purposes, we show that the conditions hold under a very general class of prior distributions in a time-varying covariate selection problem where the response variables possess inherent autocorrelation structure for any proper prior distribution over the parameter space.

1.3 Existing and new error measures in multiple testing

Storey (2003) advocated *positive False Discovery Rate* (pFDR) as a measure of type-I error in multiple testing. Let $\delta_{\mathcal{M}}(\boldsymbol{d}|\boldsymbol{X}_n)$ be the probability of choosing $\boldsymbol{d}$ as the optimal decision configuration given data $\boldsymbol{X}_n$ when a multiple testing method $\mathcal{M}$ is employed. Then, pFDR is defined as:

$$\text{pFDR} = E_{\boldsymbol{X}_n}\left[\sum_{\boldsymbol{d}\in\mathbb{D}} \frac{\sum_{i=1}^{m} d_i(1-r_i)}{\sum_{i=1}^{m} d_i}\delta_{\mathcal{M}}(\boldsymbol{d}|\boldsymbol{X}_n)|\delta_{\mathcal{M}}(\boldsymbol{d}=0|\boldsymbol{X}_n)=0\right].$$

Analogous to type-II error, the *positive False Non-discovery Rate* (pFNR) is defined as

$$\text{pFNR} = E_{\boldsymbol{X}_n}\left[\sum_{\boldsymbol{d}\in\mathbb{D}} \frac{\sum_{i=1}^{m}(1-d_i)r_i}{\sum_{i=1}^{m}(1-d_i)}\delta_{\mathcal{M}}\left(\boldsymbol{d}|\boldsymbol{X}_n\right)\Big|\delta_{\mathcal{M}}\left(\boldsymbol{d}=\boldsymbol{1}|\boldsymbol{X}_n\right)=0\right].$$

Under prior $\pi(\cdot)$, Sarkar et al. (2008) defined posterior FDR and FNR. The measures are given as follows:

$$\textit{posterior}\ \text{FDR} = E_{\xi|\boldsymbol{X}_n}\left[\sum_{\boldsymbol{d}\in\mathbb{D}}\frac{\sum_{i=1}^m d_i(1-r_i)}{\sum_{i=1}^m d_i\vee 1}\delta_{\mathcal{M}}(\boldsymbol{d}|\boldsymbol{X}_n)\right] = \sum_{\boldsymbol{d}\in\mathbb{D}}\frac{\sum_{i=1}^m d_i(1-v_{in})}{\sum_{i=1}^m d_i\vee 1}\delta_{\mathcal{M}}(\boldsymbol{d}|\boldsymbol{X}_n);$$

$$\textit{posterior}\ \text{FNR} = E_{\xi|\boldsymbol{X}_n}\left[\sum_{\boldsymbol{d}\in\mathbb{D}}\frac{\sum_{i=1}^m (1-d_i)r_i}{\sum_{i=1}^m (1-d_i)\vee 1}\delta_{\mathcal{M}}(\boldsymbol{d}|\boldsymbol{X}_n)\right] = \sum_{\boldsymbol{d}\in\mathbb{D}}\frac{\sum_{i=1}^m (1-d_i)v_{in}}{\sum_{i=1}^m (1-d_i)\vee 1}\delta_{\mathcal{M}}(\boldsymbol{d}|\boldsymbol{X}_n),$$

where $v_{in} = P_{\xi|\boldsymbol{X}_n}(\Theta_{1i})$. Also under any non-randomized decision rule $\mathcal{M}$, $\delta_{\mathcal{M}}(\boldsymbol{d}|\boldsymbol{X}_n)$ is either 1 or 0 depending on data $\boldsymbol{X}_n$. Given $\boldsymbol{X}_n$, we denote these posterior error measures by $\text{FDR}_{\boldsymbol{X}_n}$ and $\text{FNR}_{\boldsymbol{X}_n}$, respectively. With respect to the new notions of errors in (2) and (3), $\text{FDR}_{\boldsymbol{X}_n}$ is modified as

$$\begin{aligned}\textit{modified}\ \text{FDR}_{\boldsymbol{X}_n} &= E_{\xi|\boldsymbol{X}_n}\left[\sum_{\boldsymbol{d}\in\mathbb{D}}\frac{\sum_{i=1}^m d_i(1-r_iz_i)}{\sum_{i=1}^m d_i\vee 1}\delta_{\mathcal{M}}(\boldsymbol{d}|\boldsymbol{X}_n)\right]\\ &= \sum_{\boldsymbol{d}\in\mathbb{D}}\frac{\sum_{i=1}^m d_i(1-w_{in}(\boldsymbol{d}))}{\sum_{i=1}^m d_i\vee 1}\delta_{\mathcal{M}}(\boldsymbol{d}|\boldsymbol{X}_n).\end{aligned}$$

We denote *modified* $\text{FDR}_{\boldsymbol{X}_n}$ by $\text{mFDR}_{\boldsymbol{X}_n}$. Notably, the expectations of $\text{FDR}_{\boldsymbol{X}_n}$ and $\text{FNR}_{\boldsymbol{X}_n}$ with respect to $\boldsymbol{X}_n$, conditioned on the event that their respective denominators are positive, yield the *positive Bayesian* FDR (pBFDR) and FNR (pBFNR), respectively. The same expectation over $\text{mFDR}_{\boldsymbol{X}_n}$ yields *modified positive* BFDR (mpBFDR).

We advocate the posterior error measures $\text{mFDR}_{\boldsymbol{X}_n}$, $\text{FDR}_{\boldsymbol{X}_n}$ and $\text{FNR}_{\boldsymbol{X}_n}$ as multiple testing error controlling measures in Bayesian multiple testing. These measures give the performance of the employed multiple testing procedure given the data and hence most appropriate from the Bayesian perspective. In particular, wisdom gained from the traditional debate between the classical and Bayesian paradigms suggests that avoiding expectation with respect to the data in the error measures can help avoid possible paradoxes analogous to examples such as the Welch's paradox (Welch 1939). Not only that the posterior error measures are readily estimable in practical situations, however, complicated the dependent structure may be, without any assumption. In Sect. 2, we show that the asymptotic convergence rates of these measures are associated with the KL divergence between the true data-generating process and the selected model. As regards $\text{mFDR}_{\boldsymbol{X}_n}$, it takes into account the joint dependence structure between parameters through the z_i terms. As will be shown subsequently, this joint dependence structure manifests itself through the convergence rate of $\text{mFDR}_{\boldsymbol{X}_n}$. Chandra and Bhattacharya (2019) also showed that $\text{mFDR}_{\boldsymbol{X}_n}$ can be interpreted as the posterior probability of an incorrect decisions within each group. Extensive simulation studies indicated that controlling the $\text{mFDR}_{\boldsymbol{X}_n}$ gives better protection against the type-II error in dependent cases.

Müller et al. (2004) considered the following additive loss function

$$L(\boldsymbol{d}, \boldsymbol{\xi}) = c \sum_{i=1}^{m} d_i(1 - r_i) + \sum_{i=1}^{m} (1 - d_i) r_i, \tag{5}$$

where c is a positive constant. The decision rule that minimizes the posterior risk of the above loss is $d_i = I\left(v_i > \frac{c}{1+c}\right)$ for all $i = 1, \cdots, m$, where $I(\cdot)$ is the indication function.

This loss function has been widely used in the Bayesian multiple testing setups and also in frequentist decision theoretic approaches (Sun and Cai 2009; Xie et al. 2011). Notably, the non-marginal method boils down to this additive loss function-based approach when $G_i = \{i\}$, that is, when the information regarding dependence between hypotheses is not available or overlooked. Hence, the convergence properties of the additive loss function-based methods can be easily derived from our theories. We discuss this subsequently later in this article.

It is to be seen that multiple testing problems can be regarded as model selection problems where the task is to choose the correct specification for the parameters under consideration. Even if one decision is taken incorrectly, the model gets misspecified. Shalizi (2009) considered asymptotic behaviour of misspecified models under very general conditions. We adopt his basic assumptions and some of his convergence results to build a general asymptotic theory for our multiple testing method.

In Sect. 2, we provide the setup, assumptions and the main result which we adopt for our purpose. In the same section, we investigate consistency of the non-marginal multiple testing procedure. In Sect. 3, we study the rates of convergence of different versions of FDRs and FNRs and asymptotic comparison between them. We then investigate, in Sect. 4, the asymptotic properties of FNRs when the multiple testing methods are adjusted so that mpBFDR and pBFDR controlled at some level α, for some $\alpha \in (0, a)$, where $a \leq 1$. Indeed, as we show, any value of $\alpha \in (0, 1)$ is not permissible asymptotically. We further show that FNRs tend to zero at a faster rate compared to the situations where α-control is not exercised. In Sect. 5, we illustrate the asymptotic properties of the non-marginal method in a time-varying covariate selection problem where the response variables possess inherent autocorrelation structure. In Sect. 6, we compare the performance of this method with some popular existing Bayesian multiple testing methods. In Sect. 7, we apply our non-marginal method to a variable selection problem in a real, maize data with 7389 covariates representing SNP (single-nucleotide polymorphism) markers, concerning linear regression of "days to anthesis male flowering time" on the covariates. Excellent fit is the outcome, once the significant variables have been selected by our Bayesian multiple testing method. Finally, in Sect. 8 we summarize our contributions and provide concluding remarks.

2 Consistency of the non-marginal procedure and other procedures based on additive loss

2.1 Preliminaries for ensuring posterior convergence under general setup

We consider a probability space $(\Omega, \mathcal{F}, P)$ and a sequence of random variables $X_1, X_2, \ldots$, taking values in some measurable space $(\Xi, \mathcal{X})$, whose infinite-dimensional distribution is P. The natural filtration of this process is $\sigma(\boldsymbol{X}_n)$. We denote the distributions of processes adapted to $\sigma(\boldsymbol{X}_n)$ by $P_{X_n|\xi}$, where $\boldsymbol{\xi}$ is associated with a measurable space $(\boldsymbol{\Xi}, \mathcal{T})$ and is generally infinite-dimensional. For the sake of convenience, we assume, as defined by Shalizi (2009), that P and all the $P_{X_n|\xi}$ are dominated by a common reference measure, with respective densities p and f_ξ. The usual assumptions that $P \in \boldsymbol{\Xi}$ or even P lies in the support of the prior on $\boldsymbol{\Xi}$ are not required for Shalizi's result, rendering it very general indeed. We put the prior distribution $\pi(\cdot)$ on the parameter space $\boldsymbol{\Xi}$. Following Shalizi, we first define some notations: Consider the following likelihood ratio:

$$R_n(\boldsymbol{\xi}) = \frac{f_\xi(\boldsymbol{X}_n)}{p(\boldsymbol{X}_n)}.$$

For every $\boldsymbol{\xi} \in \Theta$, the KL divergence rate $h(\boldsymbol{\xi})$ is defined as

$$h(\boldsymbol{\xi}) = \lim_{n\to\infty} \frac{1}{n} E\left(\log \frac{p(\boldsymbol{X}_n)}{f_\xi(\boldsymbol{X}_n)}\right), \tag{6}$$

given that the above limit exists. For $A \subseteq \boldsymbol{\Xi}$, let

$$h(A) = \operatorname*{ess\,inf}_{\xi\in A} h(\boldsymbol{\xi}); J(\boldsymbol{\xi}) = h(\boldsymbol{\xi}) - h(\boldsymbol{\Xi}); J(A) = \operatorname*{ess\,inf}_{\xi\in A} J(\boldsymbol{\xi}). \tag{7}$$

We have stated the assumptions (S1)-(S7) considered by Shalizi in Section S-1. Under those assumptions, the following theorem can be seen to hold:

Theorem 1 (Shalizi 2009) *Consider assumptions* (S1)–(S7) *and any set* $A \in \boldsymbol{\Xi}$ *with* $\pi(A) > 0$. *If* $\varsigma > 2h(A)$, *where* ς *is given in* (S-3) *under assumption* (S5), *then*

$$\lim_{n\to\infty} \frac{1}{n} \log P_{\xi|X_n}(A|\boldsymbol{X}_n) = -J(A).$$

We shall frequently make use of this theorem for our purpose. Also throughout this article, we show consistency results for general models satisfying (S1)–(S7). For all our results, we assume these conditions.

2.2 Some requisite notations for the non-marginal method

It is very interesting that we need not assume that the true data-generating process P is in the class of postulated model $F_\xi; \boldsymbol{\xi} \in \boldsymbol{\Xi}$. However, asymptotic consistency of

the non-marginal procedure can still be achieved in the sense that with increasing sample size the model with minimal misspecification is selected. Note that depending on $d_j = 0$ or 1, $\Theta_{d_j j}$ is the specification corresponding to θ_j directed by d_j. Now for all possible decision configurations the parameter space $\boldsymbol{\Xi}$ can have the following partition

$$\boldsymbol{\Xi}(\boldsymbol{d}) = \prod_{i=1}^{m} \Theta_{d_i i} \times \prod_{i=m+1}^{M} \Theta_i.$$

Note that $J(\boldsymbol{\Xi})$ is the minimal KL divergence between the true data-generating process P and the class of all postulated models. Among all possible decision configurations $\boldsymbol{d} \in \mathbb{D}$, let $\boldsymbol{d}^t$ be such that $J(\boldsymbol{\Xi}(\boldsymbol{d}^t)) = J(\boldsymbol{\Xi})$. Note that $\boldsymbol{d}^t$ minimizes the KL divergence between the true data-generating model P among all possible decision configurations. We regard $\boldsymbol{d}^t$ as the true decision configuration. We will show that with increasing sample size the non-marginal procedure will choose $\boldsymbol{d}^t$ as the optimal decision rule almost surely (*a.s.*). We now define some notations required for further advancements. Let

$$\boldsymbol{\Xi}_{\boldsymbol{d},i} = \left\{ \theta_i \in \Theta_{1i}, \theta_j \in \Theta_{d_j j} \; \forall \, j \neq i \; \& \; j \in G_i \right\}.$$

Then, $\boldsymbol{\Xi}_{\boldsymbol{d},i}$ is the joint parameter space for the parameters in G_i directed by $\boldsymbol{d}$. Notably, $\boldsymbol{\Xi}_{\boldsymbol{d},i}$ is not the same as $\Theta_{d_i i}$; it concerns (possibly) multiple parameters in G_i, whereas the latter is only concerned with the ith parameter and the corresponding decision. For any decision configuration $\boldsymbol{d}$ and group G, let $\boldsymbol{d}_G = \{d_j : j \in G\}$. Define

$$\mathbb{D}_i = \left\{ \boldsymbol{d} : \text{ all decisions in } \boldsymbol{d}_{G_i} \text{ are correct} \right\}.$$

Here $\mathbb{D}_i$ is the set of all decision configurations where the decisions corresponding to the hypotheses in G_i are at least correct. Clearly, $\mathbb{D}_i$ contains $\boldsymbol{d}^t$ for all i. Hence, $\mathbb{D}_i^c = \left\{ \boldsymbol{d} : \text{ at least one decision in } \boldsymbol{d}_{G_i} \text{ is incorrect} \right\}$. By Theorem 1, for any $\epsilon > 0$, there exists $n_0(\epsilon)$ such that for each $i = 1, \ldots, m$, for $n \geq n_0(\epsilon)$,

$$\exp\left[-n\left(J\left(\boldsymbol{\Xi}_{\boldsymbol{d},i}\right) + \epsilon\right)\right] < w_{in}(\boldsymbol{d}) < \exp\left[-n\left(J\left(\boldsymbol{\Xi}_{\boldsymbol{d},i}\right) - \epsilon\right)\right] \text{ if } \boldsymbol{d} \in \mathbb{D}_i^c, \tag{8}$$

$$\exp\left[-n(J(\boldsymbol{\Xi}^c_{\boldsymbol{d}^t,i}) + \epsilon)\right] < 1 - w_{in}(\boldsymbol{d}) < \exp\left[-n(J(\boldsymbol{\Xi}^c_{\boldsymbol{d}^t,i}) - \epsilon)\right] \text{ if } \boldsymbol{d} \in \mathbb{D}_i. \tag{9}$$

Also, for $i = 1, \ldots, m$, and for $n \geq n_0(\epsilon)$,

$$\exp\left[-n\left(J\left(H_{1i}\right) + \epsilon\right)\right] < v_{in} < \exp\left[-n\left(J\left(H_{1i}\right) - \epsilon\right)\right], \text{ if } d_i^t = 0, \tag{10}$$

$$1 - \exp\left[-n\left(J\left(H_{0i}\right) - \epsilon\right)\right] < v_{in} < 1 - \exp\left[-n\left(J\left(H_{0i}\right) + \epsilon\right)\right] \text{ if } d_i^t = 1 \tag{11}$$

where $J(\boldsymbol{\Xi}_{\boldsymbol{d},i}) = \operatorname*{ess\,inf}_{\xi \in \Psi_{id}} J(\boldsymbol{\xi})$; $J(H_{ki}) = \operatorname*{ess\,inf}_{\xi \in Y_{ki}} J(\boldsymbol{\xi})$,

$$\Psi_{i\boldsymbol{d}} = \{\theta_i \in \Theta_{1i}, \theta_j \in \Theta_{d_j j} \,\forall\, j \neq i \,\&\, j \in G_i, \theta_k \in \Theta_k \,\forall\, k \in G_i^c\} \text{ and}$$
$$\Upsilon_{ki} = \{\theta_i \in \Theta_{ki}, \theta_j \in \Theta_j \,\forall\, j \neq i\},\ k = 0, 1.$$

Note that in $\Psi_{i\boldsymbol{d}}$, θ_k lies in its whole parameter space for all $k \in G_i^c$, irrespective of the fact that d_k might be incorrect. Hence, it corresponds to a model where only $\{\theta_k : k \in G_i\}$ may be misspecified. $J(\Xi_{\boldsymbol{d},i})$ gives the KL divergence rate (defined in (7)) between the true model and this model. Also $J(\Theta_{i\boldsymbol{d}}) > 0$ if $\boldsymbol{d} \in \mathbb{D}_i^c$, $J(H_{1i}) > 0$ if $d_i^t = 0$ and $J(H_{0i}) > 0$ if $d_i^t = 1$.

It is important to observe that, in Eqs. (8)–(11), we have referred to the same ϵ and the same $n_0(\epsilon)$ for every $i = 1, \ldots, m$. Due to the finiteness of m, taking the same $n_0(\epsilon)$ is possible here.

2.3 The basic consistency theory for multiple testing with application to the non-marginal and additive loss-based procedures

With the above notations, in this section we show that the non-marginal procedure is asymptotically consistent under any general dependent model satisfying the conditions in Section S-1. As a simple corollary, we show that other existing multiple testing procedures based on additive loss are also consistent. Let us first formally define what we mean by asymptotic consistency of a multiple testing procedure.

Definition 2 Let $\boldsymbol{d}^t$ be the true decision configuration among all possible decision configurations as defined in Sect. 2.2. Then a multiple testing method $\mathcal{M}$ is said to be asymptotically consistent if almost surely

$$\lim_{n\to\infty} \delta_{\mathcal{M}}(\boldsymbol{d}^t|\boldsymbol{X}_n) = 1.$$

We now state the requisite conditions for NMD to be asymptotically consistent.

(A1) We assume that the sequence β_n is neither too small nor too large, that is,

$$\underline{\beta} = \liminf_{n\geq 1} \beta_n > 0; \tag{12}$$

$$\overline{\beta} = \limsup_{n\geq 1} \beta_n < 1. \tag{13}$$

(A2) We assume that neither all the null hypotheses are true and nor all of them are false, that is, $\boldsymbol{d}^t \neq \boldsymbol{0}$ and $\boldsymbol{d}^t \neq \boldsymbol{1}$, where $\boldsymbol{0}$ and $\boldsymbol{1}$ are vectors of 0's and 1's, respectively.

Recall the constant β_n in (4), which is the penalizing constant between the error E and true positives TP and (A1) ensures a fine balance between these two. It is necessary for the asymptotic consistency of both the non-marginal method and additive loss function-based method. Notably, (A2) is not required for the consistency results. Its role is to ensure that the denominator terms in the multiple testing error measures (defined in Sect. 1.3) do not become 0 by ruling out two

very extreme situations where none/all of the null hypotheses are false. It is also important in the asymptotic studies of different versions of FDR and FNR that we consider. With these conditions, we propose and prove the following results.

Theorem 2 *Let $\delta_{\mathcal{NM}}(\cdot|\boldsymbol{X}_n)$ denote the non-marginal decision rule given data $\boldsymbol{X}_n$. Assume condition (A1) on β_n. Then, the non-marginal decision procedure is asymptotically consistent.*

Remark 1 It is important to note that in the proof of Theorem 2, we do not require any assumption on how the groups should be formed. The theorem is valid even if the groups are implicitly dependent on the observed data. This shows that, in case the prior information on the correlation structure of the parameters is weak, the non-marginal method is also valid when the groups are formed on the basis of posterior correlation or by other data adaptive methods.

We have already mentioned that the optimal decision rules corresponding to the loss function in (5) is a special case of the non-marginal method when dependence among the hypotheses is ignored. As we have not considered any particular structure of G_i's in Theorem 2, consistency of the additive loss function-based method can also be obtained from the previous theorem.

Corollary 1 *Assuming condition (A1), the optimal decision rule corresponding to the additive loss function (5) is asymptotically consistent.*

2.4 Asymptotic robustness with respect to group choice

Note that in Theorem 2 no specification on group formation is required. Generally for large samples, Bayesian methods are robust with respect to the prior choice given that the prior distribution follows some regularity conditions. Theorem 2 entails that the non-marginal method is consistent for any group choice given that the model and prior distributions satisfy the conditions of Section S-1. Hence, we see that the non-marginal method is asymptotically robust with respect to the choice of groups.

3 Asymptotic analyses of multiple testing error rates

3.1 Asymptotic properties of versions of FDR

First, we study the convergence properties of $\text{mFDR}_{\boldsymbol{X}_n}$ and $\text{FDR}_{\boldsymbol{X}_n}$ in this section. We show that the convergence rates of the posterior error measures are directly associated with the KL divergence from the true model.

Theorem 3 *Assume conditions* (A1), (A2). *Let* $J_{\min} = \min_{i:d_i^t=1} J(\Xi^c_{\boldsymbol{d}^t,i})$ *and* $H_{\min} = \min_{i:d_i^t=1} J(H_{0i})$. *Then, for the non-marginal multiple testing procedure the following hold almost surely*:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathrm{mFDR}_{\boldsymbol{X}_n} = -J_{\min}; \quad \lim_{n\to\infty} \frac{1}{n} \log \mathrm{FDR}_{\boldsymbol{X}_n} = -H_{\min}.$$

Notably, both $J_{\min}$ and $H_{\min}$ are positive, and hence, the posterior FDR along with its modified version converges to 0 at an exponential rate with increasing sample size. Interestingly, the convergence rate is in terms of the KL divergence between the true data-generating process P and the class of postulated models $F_{\boldsymbol{\xi}}$. We see that the posterior error measures have this very interesting property where they truly indicate the divergence from the true data-generating process.

Remark 2 Even though NMD is asymptotically consistent for data-dependent group formations, asymptotic convergence rate of $\mathrm{mFDR}_{\boldsymbol{X}_n}$ may not hold in such case. For increasing sample size, the group structures may change, resulting in ambiguity in the definition of $\mathrm{mFDR}_{\boldsymbol{X}_n}$. Therefore, in Theorem 3 we assume that the group structures are known *a priori*.

So far we have investigated the asymptotic properties of $\mathrm{mFDR}_{\boldsymbol{X}_n}$ and $\mathrm{FDR}_{\boldsymbol{X}_n}$, which is a valid exercise from the Bayesian perspective, as the data are conditioned upon in these error measures. We now study the asymptotic properties of mpBFDR and pBFDR. mpBFDR is defined as

$$\mathrm{mpBFDR} = E_{\boldsymbol{X}_n}\left[\mathrm{mFDR}_{\boldsymbol{X}_n} | \delta_{\mathcal{M}}(\mathbf{0}|\boldsymbol{X}_n) = 0\right], \tag{14}$$

where $\mathbf{0}$ is the decision configuration that no null hypothesis is rejected. pBFDR is where the expectation in (14) is of $\mathrm{FDR}_{\boldsymbol{X}_n}$. Indeed, expectations of the error measures are traditionally more popular in multiple testing. The following theorem provides the asymptotic results of mpBFDR and pBFDR.

Corollary 2 *Under conditions* (*A1*), (*A2*)*, for the non-marginal procedure, we have*

$$\lim_{n\to\infty} \mathrm{mpBFDR} = 0; \quad \lim_{n\to\infty} \mathrm{pBFDR} = 0.$$

It is important to remark that although the aforementioned expected error measures converge to zero as shown by Corollary 2, it does not seem to be possible to obtain the rates of convergence to zero in general, as in Theorem 3 associated with the corresponding posterior versions.

As discussed in Sect. 2.4, the non-marginal method is robust in the sense that it is consistent for any group structure. However, the convergence rate of the $\mathrm{mFDR}_{\boldsymbol{X}_n}$ shows that it takes into account the dependence among hypotheses through the group structures. Hence, it may lose its effectiveness over $\mathrm{FDR}_{\boldsymbol{X}_n}$ in case the group choice is injudicious. In practical situations, where sample size is fixed, thoughtful choice of groups is very important.

3.2 Asymptotic properties of versions of FNR

As in the case of FDR, similar results can also be derived for different versions of FNR. We state the result in the following theorem.

Theorem 4 *Assume conditions* (A1) *and* (A2). *Let* $\tilde{H}_{\min} = \min_{i:d_i'=0} J(H_{1i})$. *Then, for the non-marginal multiple testing procedure*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathrm{FNR}_{\boldsymbol{X}_n} = -\tilde{H}_{\min}. \tag{15}$$

Thus we see that for the non-marginal method, $\mathrm{FNR}_{\boldsymbol{X}_n}$ do diminish to zero exponentially fast with the convergence rate being directly proportional to the KL divergence from the true data-generating model. pBFNR is defined as follows:

$$\mathrm{pBFNR} = E_{\boldsymbol{X}_n}\left[\mathrm{FNR}_{\boldsymbol{X}_n} | \delta_{\mathcal{M}}(\mathbf{1}|\boldsymbol{X}_n) = 0\right],$$

where $\mathbf{1}$ is the decision configuration that no null hypothesis is accepted. The following asymptotic result holds for pBFNR as well.

Corollary 3 *Under conditions* (*A1*), (*A2*), *for the non-marginal procedure we have*

$$\lim_{n\to\infty} \mathrm{pBFNR} = 0.$$

Although Corollary 3 asserts convergence of the relevant versions of BFNR to zero, it does not seem to be possible to provide their rates of convergence, as in $\mathrm{FNR}_{\boldsymbol{X}_n}$. This issue is in keeping with the corresponding versions of BFDR.

Remark 3 It is proper to envisage possible modification of FNR with respect to the new notions of errors. In Section S-2, we show that, under a mild assumption, the asymptotic convergence rates of $\mathrm{FNR}_{\boldsymbol{X}_n}$ and its modified counterpart are equal. Therefore, in this main article, we continue the relevant discussions with respect to the existing versions of FNR only.

4 Convergence of $\mathrm{FNR}_{\boldsymbol{X}_n}$ and BFNR when versions of BFDR are α -controlled

We now enforce asymptotic control over mpBFDR and pBFDR in the sense that they converge to $\alpha \in (0, a)$, instead of zero, for some $0 < a \leq 1$ and study the asymptotic behaviour of pBFNR. Here it is important to point out that Chandra and Bhattacharya (2019) proved that for both NMD and additive loss function-based methods, mpBFDR and pBFDR are continuous and non-increasing in β, and therefore, β can be tuned to set the type-I errors at any desired size α. However, as we show in the asymptotic case, it is not possible to incur too high type-I error, that is, a cannot be arbitrarily close to 1. This is not unexpected, since consistent methods cannot commit arbitrarily large errors asymptotically. Naturally the question arises

whether α-control of versions of BFDR is at all necessary. The answer is that since it is a standard practice in multiple testing to exercise α-control on versions of FDR in order to incur lesser type-II error, it is important to investigate what would be the feasible range of values of α to attain in large or even moderately large samples and for such α's how the type-II error would behave. We attempt to address these questions with respect to the non-marginal procedure and additive loss function-based method.

4.1 Convergence of mpBFDR and pBFDR to α for NMD

We begin with the following theorem that provides the bound for the maximum mpBFDR that can be incurred asymptotically.

Theorem 5 *In addition to* (*A1*), (*A2*), *assume the following:*

(B1) *Let each group of a particular set of* m_1 $(< m)$ *groups out of the total* m *groups be associated with at least one false null hypothesis and that all the null hypotheses associated with the remaining* $m - m_1$ *groups be true. Let us further assume that the latter* $m - m_1$ *groups do not have any overlap with the remaining* m_1 *groups. Without loss of generality, assume that* $G_1, \ldots, G_{m_1}$ *are the groups each consisting of at least one false null and* $G_{m_1+1}, G_{m_1+2}, \cdots, G_m$ *are the groups where all the null hypotheses are true.*

Then, the maximum mpBFDR *that can be incurred asymptotically lies in* $\left(\frac{1}{\sum_{i=1}^{m} d_i^t+1}, \frac{m-m_1}{\sum_{i=1}^{m} d_i^t+m-m_1}\right)$.

Remark 4 The proof of Theorem 5 crucially uses the result that mpBFDR is non-increasing with β. It can be easily seen that this monotonicity with respect to β holds for $\mathrm{mFDR}_{\boldsymbol{X}_n}$ as well. Hence, Theorem 5 is also valid for $\mathrm{mFDR}_{\boldsymbol{X}_n}$.

Remark 5 Theorem 5 holds when $G_i \subset \{1, \ldots, m\}$ for at least one $i \in \{1, \ldots, m\}$. But if $G_i = \{1, \ldots, m\}$ for $i = 1, \ldots, m$, then mpBFDR $\to 0$ as $n \to \infty$, for any sequence $\beta_n \in [0, 1]$. This is because in this case there does not exist any $\boldsymbol{d} \neq \boldsymbol{d}^t$ such that

$$P\left(\sum_{i=1}^{m} d_i w_{in}(\boldsymbol{d}) - \sum_{i=1}^{m} d_i^t w_{in}(\boldsymbol{d}^t) > \beta_n\left(\sum_{i=1}^{m} d_i - \sum_{i=1}^{m} d_i^t\right)\right) > 0,$$

as $n \to \infty$.

Theorem 5 also clarifies that for any arbitrary configuration of groups, it is not possible to commit arbitrarily large error when the sample size is large enough. The joint structure provides a safeguard against incurring large errors. However, in practical situations dealing with real-life data, it is common practice to control type-I error at some pre-specified level α (> 0) both in single and multiple

hypothesis testing problems, which renders the very important task of investigating the feasible range of α.

In this regard, (B1) is the condition under which possible values of type-I error to be controlled are available, at least for large n. Note that to incur type-I error it is required to reject some true null hypotheses. As the grouping structures prevent from committing arbitrary error by the non-marginal procedure, (B1) is required. By virtue of this condition, there are some true null hypotheses in isolation which can be rejected. In the following theorem, we provide an asymptotic bound on the maximum type-I error that can be incurred.

Theorem 6 *Assume condition (B1), and let* mpBFDR_β *denote the procured* mpBFDR *in the non-marginal procedure where the penalizing constant is* β. *Suppose*

$$\lim_{n\to\infty} \text{mpBFDR}_{\beta=0} = E. \tag{16}$$

Then, for any $\alpha < E$ *and* $\alpha \in \left(\frac{1}{\sum_{i=1}^m d_i^t + 1}, \frac{m-m_1}{\sum_{i=1}^m d_i^t + m - m_1} \right)$, *there exists a sequence* $\beta_n \to 0$ *such that* $\text{mpBFDR}_{\beta_n} \to \alpha$ *as* $n \to \infty$.

Since mpBFDR is decreasing in β, β can be interpreted as a balance provider between type-I and type-II errors. Corollary 2 shows that mpBFDR decays to 0 when $\liminf_{n\to\infty} \beta_n > 0$ and Theorem 6 shows that for α-control, we must have $\lim_{n\to\infty} \beta_n = 0$. Since mpBFDR is decreasing in β, it intuitively indicates that in the case of α-control of mpBFDR the sequence $\{\beta_n\}$ has to be dominated by any $\{\beta_n\}$ sequence for which Corollary 2 holds. Theorem 6 formalizes this intuition and shows that a smaller sequence of β_n has to be taken for α-control.

From the proofs of Theorem 5 and 6, it can be seen that replacing $w_{in}(\hat{\boldsymbol{d}})$ by v_{in} does not affect the results. Hence, we state the following corollary.

Corollary 4 *Assume condition (B1) and let* pBFDR_β *denote the procured* pBFDR *in the non-marginal procedure where the penalizing constant is* β. *Suppose*

$$\lim_{n\to\infty} \text{pBFDR}_{\beta=0} = E'.$$

Then, for any $\alpha < E'$ *and* $\alpha \in \left(\frac{1}{\sum_{i=1}^m d_i^t + 1}, \frac{m-m_1}{\sum_{i=1}^m d_i^t + m - m_1} \right)$, *there exists a sequence* $\beta_n \to 0$ *such that* $\text{pBFDR}_{\beta_n} \to \alpha$ *as* $n \to \infty$.

We now investigate, as special cases of the above results, the situations where $G_i = \{i\}$ for all i. Recall that in this case the additive loss function-based methods are special cases of the non-marginal procedure. In such cases, mpBFDR also boils down to pBFDR. The following theorem gives the result for asymptotic α -control of pBFDR in this situation.

Theorem 7 *Let m_0 $(< m)$ be the number of true null hypotheses. Then, for any $0 < \alpha < \frac{m_0}{m}$, there exists a sequence $\beta_n \to 0$ as $n \to \infty$ such that for the additive loss function-based methods*

$$\lim_{n\to\infty} \text{pBFDR}_{\beta_n} = \alpha.$$

In the above, we have noted that mpBFDR reduces to pBFDR when $G_i = \{i\}$ for all i. However, for any additive loss function-based multiple testing method, we may still envisage the measure mpBFDR where the definition of mpBFDR considers the adequate dependent structure by means of non-singleton G_i's. This has the advantage of yielding non-marginal decisions even though the actual criterion to be optimized is a sum of loss functions associated with individual parameters and decisions. In the following theorem, we show that the same asymptotic result as Theorem 7 also holds for mpBFDR in the case of additive loss functions, without assumption (B1).

Theorem 8 *Let α be the desired level of significance where $0 < \alpha < \frac{m_0}{m}$, m_0 $(< m)$ being the number of true null hypotheses. Then, there exists a sequence $\beta_n \to 0$ as $n \to \infty$ such that for the additive loss function-based method*

$$\lim_{n\to\infty} \text{mpBFDR}_{\beta_n} = \alpha.$$

It is interesting that for the additive loss function-based method, Theorem 8 holds without condition (B1). This condition is an added imposition to study the theoretical properties of the non-marginal procedure when mpBFDR is controlled at level α. (B1) ensures that there are some isolated groups of hypotheses. Although there is no notion of grouping in the additive loss function, as we pointed out above, mpBFDR does correspond to groups that are not singletons. However, $\text{mpBFDR}(\mathcal{M}) \geq \text{pBFDR}(\mathcal{M})$ for any multiple testing method $\mathcal{M}$, for arbitrary sample size, and this crucially ensures that the result asserted by Theorem 8 goes through even without (B1).

Remark 6 Note that Theorems 6–8 and Corollary 4 use continuity of the expected versions of FDR with respect to β, in addition to their non-increasing nature with respect to β. The continuity property need not be satisfied by the corresponding Bayesian versions given the data, and hence, we cannot assert that the aforementioned results continue to hold for the corresponding Bayesian versions of FDR (conditional on the data).

Remark 7 We have already discussed in the context of Theorems 5 and 6 that condition (B1) is crucial for α-control for the non-marginal method, and without the assumption, mpBFDR would diminish to zero asymptotically. This signifies that it is difficult to commit errors by the non-marginal method, thanks to its dependence structure, so that extra assumption is needed for positive α-control. On the other hand, Theorems 7 and 8 show that for other multiple testing methods based on additive loss, α-control is possible without (B1), not only with respect to pBFDR but also with respect to mpBFDR, which includes the dependence structure in its

definition. It certifies that if the underlying multiple testing procedure does not consider dependence, then, however, sensible the underlying model is, the errors can be larger compared to the non-marginal procedure.

4.2 Asymptotic properties of type-II errors when mpBFDR and pBFDR are asymptotically controlled at α

Theorem 9 *Assume condition* (*B1*). *Then, for asymptotic* α*-control of* mpBFDR *in the non-marginal procedure the following holds almost surely*:

$$\limsup_{n\to\infty} \mathrm{FNR}_{\boldsymbol{X}_n} \leq -\tilde{H}_{\min}.$$

Corollary 5 *Assume condition* (*B1*). *Then, for asymptotic* α*-control of* mpBFDR *in the non-marginal procedure, the following holds*:

$$\lim_{n\to\infty} \mathrm{pBFNR} = 0.$$

Thus, we see that pBFNR also goes to 0 with increasing sample size when type-I error is asymptotically controlled at α. In fact, the posterior type-II error, that is, $\mathrm{FNR}_{\boldsymbol{X}_n}$, converges to zero at a rate faster than or equal to that compared to the case when α control is not imposed. In other words, allowing asymptotically non-negligible type-I error may result in lower type-II error.

5 Illustration of consistency of NMD in time-varying covariate selection in autoregressive process

Let the true model P stand for the following $AR(1)$ model consisting of time-varying covariates:

$$x_t = \rho_0 x_{t-1} + \sum_{i=0}^{m} \beta_{i0} z_{it} + \epsilon_t,\ t = 1, 2, \ldots, \tag{17}$$

where $x_0 \equiv 0$, $|\rho_0| < 1$ and $\epsilon_t \overset{iid}{\sim} N(0, \sigma_0^2)$, for $t = 1, 2, \ldots$. We further assume that for $i = 1, \ldots, m$, the time-varying covariates $\{z_{it} : t = 1, 2, \ldots\}$ are realizations of some asymptotically stationary stochastic process. We set $z_{0t} \equiv 1$ for all t.

Now let the data be modelled by the same model as P but with ρ_0, β_{i0} and σ_0^2 be replaced with the unknown quantities ρ, β_i and σ^2, respectively, that is,

$$x_t = \rho x_{t-1} + \sum_{i=0}^{m} \beta_i z_{it} + \epsilon_t,\ t = 1, 2, \ldots, \tag{18}$$

where we set $x_0 \equiv 0$, $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$, for $t = 1, 2, \ldots$. As in P, we assume that for $i = 1, \ldots, m$, the time-varying covariates are realizations of some asymptotically stationary stochastic process. For notational convenience, we define $\boldsymbol{z}_t = (z_{0t}, z_{1t}, \ldots, z_{mt})'$, $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{10}, \ldots, \beta_{m0})'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)'$.

For our asymptotic theories regarding the multiple testing methods that we consider in our main manuscript, we must verify the assumptions of Shalizi for the modelling setups (17) and (18), with $\boldsymbol{\xi} = (\rho, \beta_0, \beta_1, \ldots, \beta_m, \sigma)$. As regards the parameter space, let $\rho \in \mathbb{R}$, where $\mathbb{R}$ is the real line, $\boldsymbol{\beta} \in \mathbb{R}^m$ and $\sigma \in \mathbb{R}^+$, where $\mathbb{R}^+$ is the positive part of the real line. Thus, $\boldsymbol{\Xi} = \mathbb{R}^{m+1} \times \mathbb{R}^+$, is the parameter space. We denote the true data-generating value of $\boldsymbol{\xi}$ by $\boldsymbol{\xi}_0$. We consider any prior on $\boldsymbol{\Xi}$ that is dominated by the Lebesgue measure, with mild condition on the moments.

With respect to the above setup, we consider the following multiple-testing framework:

$$\begin{aligned} &H_{01} : |\rho| < 1 \text{ versus } H_{11} : |\rho| \geq 1 \text{ and} \\ &H_{0i} : \beta_i \in \mathcal{N}_0 \text{ versus } H_{1i} : \beta_i \in \mathcal{N}_0^c, \text{ for } i = 2, \ldots, m+1, \end{aligned} \tag{19}$$

where $\mathcal{N}_0$ is some neighbourhood of zero and $\mathcal{N}_0^c$ is the complement of the neighbourhood in the corresponding parameter space.

Verification of consistency of our non-marginal procedure amounts to verification of assumptions (S1)–(S7) for the above setup. In this regard, we make the following assumptions regarding the true model and prior distribution:

(C1) As $n \to \infty$ the following hold

$$\begin{aligned} &\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{z}_t \to \mathbf{0}; \\ &\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{z}_{t+k}\boldsymbol{z}_t' \to \mathbf{0} \text{ (null matrix), for any } k \geq 1; \\ &\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{z}_t\boldsymbol{z}_t' \to \boldsymbol{\Sigma}_z. \end{aligned} \tag{20}$$

(C2) $\sup_{t \geq 1} |\boldsymbol{z}_t'\boldsymbol{\beta}_0| < C$, for some $C > 0$.

(C3) $\boldsymbol{\xi}_0$ is an interior point of $\boldsymbol{\Xi}$.

With these model assumptions, we have to verify the seven assumptions in Section S-1 in order to show consistency. Theorem 1 essentially tells that under certain model and prior assumptions, the posterior distribution asymptotically concentrates around the true data-generating process. In this problem, we need to show that the posterior distribution concentrates around $\boldsymbol{\xi}_0$.

An important concept related to the posterior convergence theory is the asymptotic equipartition property, which needs to hold for this model. This is ensured by conditions (S1)–(S3). (S4) fortifies that the class of postulated models are not completely orthogonal to the true data-generating process. The sequence of sets $\{\mathcal{G}_n\}_{n=1}^{\infty}$ in condition (S5) is analogous to the method of sieves (Geman and Hwang 1982) which ensures that the behaviour of the posterior distribution on the full parameter space is dominated by its behaviour on the sieves. (S6), together with (S5), makes sure that the prior probability mass outside the sieve is exponentially small with the

decay rate large enough so that the posterior probability mass outside it also goes to zero. Using the analogy to the sieve again, the interpretation of the assumption is that the convergence of the log-likelihood ratio is sufficiently fast and eventually the convergence is uniform, almost surely.

To show that Bayesian multiple testing methods are consistent for model (17), we need to verify the conditions in Section S-1. These are shown in Section S-6 which leads to the following theorem.

Theorem 10 *Under model assumptions (C1)–(C3), the non-marginal multiple testing procedure for the hypothesis testing problem in* (19) *is consistent.*

Needless to mention, all the results regarding the asymptotic convergence rate of different multiple testing error measures will also continue to hold for this setup.

As an aside, from the above results, we also get a method for variable selection problem from the multiple testing approach. We do not require any restriction on the choice of prior distribution, except that it has to be a proper probability distribution. We have proved the results for dependent data making it quite general.

6 Simulation study

In this section, we compare the performance of the non-marginal procedure (NMD) with the widely used Bayesian multiple testing methods of Müller et al. (2004) (MPR) and Sarkar et al. (2008) (SZG). With increasing sample sizes, we study the convergence rates of these methods. We elaborate the simulation design in the following section.

6.1 True data-generating mechanism

In the simulation study, we take $\rho_0 = -0.5$, $\sigma_0^2 = 1$ and $m = 150$. As regards the m-dimensional true regression vector $\boldsymbol{\beta}_0$, we take 10 randomly chosen components to be nonzero and the rest to be zero. We generate the covariates as the following

$$z_1, \ldots, z_t \overset{iid}{\sim} \mathcal{MN}(\mathbf{0}, \boldsymbol{\Phi}), \tag{21}$$

where $\mathcal{MN}(\mathbf{0}, \boldsymbol{\Phi})$ denotes a multivariate distribution with mean vector $\mathbf{0}$ and dispersion matrix $\boldsymbol{\Phi}$. In this study, $\boldsymbol{\Phi}$ is a known positive definite matrix. With these covariates and true set of parameters $\boldsymbol{\xi}_0 = (\boldsymbol{\beta}_0, \sigma_0, \rho_0)$, we generate the observation $x_1, \ldots, x_t$ following the model in (17).

6.2 The postulated Bayesian model

Since most of the true β_{0i}s are zero, we consider the following global local shrinkage prior similar to Ishwaran and Rao (2005) over the β_is:

$$\begin{aligned}
\beta_i|\gamma_i &\overset{iid}{\sim} \gamma_i N(0,\tau_i^2) + (1-\gamma_i)N(0,v\tau_i^2),\\
\tau_i &\overset{iid}{\sim} IG(a_0,b_0),\\
v &\sim\ C^+(0,1),\\
\gamma_i|p &\overset{iid}{\sim} Bernoulli(p),\\
p &\sim\ Beta(a_1,b_1),
\end{aligned}$$

where $C^+(0,1)$ is the Cauchy distribution restricted on the positive real line and $IG(\cdot,\cdot)$ denotes a *Inverse-gamma* distribution. Similar prior has previously been considered in variable selection problem from a multiple testing perspective by Ghosh et al. (2006). Here γ_is are the allocation variables signifying whether the i-th variable is included in the model or not. It is a common practice to work with the allocation variables in Bayesian variable selection problems (Narisetty and He 2014), and therefore, we reframe the hypothesis testing problem in (19) as follows:

$$H_{0i} : \gamma_i = 0 \text{ versus } H_{1i} : \gamma_i = 1, \quad \text{for } i = 2,\ldots,m+1.$$

τ_is are positive numbers taking into account the uncertainty of β_is being nonzero when $\gamma_i = 1$. v is a very small quantity allocating very high probability around 0 when $\gamma_i = 0$. We have adjusted a_1 and b_1 such that the mode of the prior Beta distribution of p is 0.1. As regards σ and ρ, we consider the following distributions as prior for these parameters:

$$\sigma^2 \sim IG(a_2,b_2), \quad \rho \sim N(0,1).$$

s Here a_2 and b_2 are adjusted such that the mode of the prior distribution is 1 and variance 100. For all the three methods, the same prior distribution is considered.

For implementation of the NMD method in this simulation study, groups are formed according to the strategy in Sect. 1.2 where Λ is taken to be the posterior correlation matrix of $\boldsymbol{\beta}$ computed the MCMC samples. With these groups, we implement the non-marginal method.

6.3 Criteria for comparing different multiple testing methods in this study

Different multiple testing methods are expected to yield different decision configurations for the same given dataset. We adopt three different criteria for comparing the performances of the competing multiple testing procedures, which we briefly discuss below.

Let $\boldsymbol{d}_{\mathcal{M}}$ be the decision configuration obtained by a multiple testing method $\mathcal{M}$. We compute the Jaccard similarity coefficient (Jaccard 1901, 1908) between the true decision configuration $\boldsymbol{d}_0$ and $\boldsymbol{d}_{\mathcal{M}}$ for each of three multiple testing methods and compare their performances.

Let $\boldsymbol{\beta}_{\mathcal{M}}$ and $\hat{\rho}$ be the mode of the posterior distributions of $\boldsymbol{\beta}$ and ρ, respectively, given the data. We also compute the Euclidean distance between $(\boldsymbol{\beta}_0, \rho_0)$ and $(\boldsymbol{\beta}_{\mathcal{M}}, \hat{\rho})$. In this context, note once the multiple testing procedure identifies the

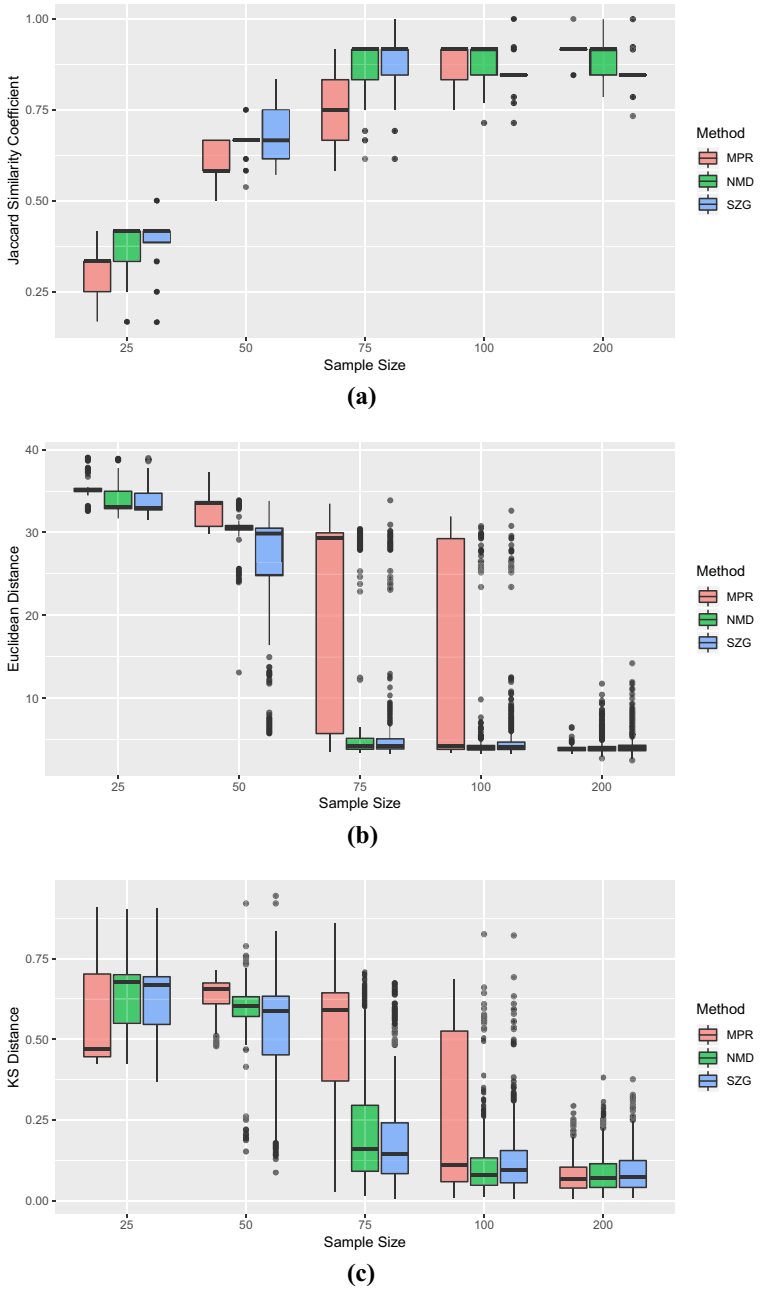

Fig. 1 Performance comparison via boxplots for increasing sample sizes; the X-axis plots the sample sizes: panel **a** shows the Jaccard similarity coefficient; panel **b** shows the Euclidean distance between the estimated model parameters and the true parameters; panel **c** shows the Kolmogorov–Smirnov distance between the true data generative distribution and posterior predictive distribution for the competing methods across different sample sizes. Consistent to the asymptotic theories, all the panels exhibit improvement in performance with increasing sample size

significant covariates, we no longer consider the shrinkage prior for β_i for computing the posterior distributions of $\boldsymbol{\xi}$ and ρ, but set $\beta_i \overset{iid}{\sim} N(0, \tau^2)$.

With the significant covariates and a future covariate z_{t+1}, we compute the posterior predictive distribution of x_{t+1} and compute the Kolmogorov–Smirnov (KS) distance from the true predictive distribution of x_{t+1}. Again, we consider $\beta_i \overset{iid}{\sim} N(0, \tau^2)$.

In other words, we compare the performance and accuracy of the three competing Bayesian multiple testing methods by means of the Jaccard similarity coefficient, Euclidean distance and KS distance. For five different sample sizes, we replicate our simulation experiments 750 times and compare the boxplots. For all the three competing Bayesian multiple testing methods, FDR_{X_n} is controlled at level 0.05.

6.4 Comparison of the results

From Fig. 1a, we see that the Jaccard Similarity Coefficients have stabilized near 1 sample size 75 onward indicating that the asymptotic theory indeed takes precedence for all the methods, when the sample size gets sufficiently large. Interestingly, the NMD method has the fastest convergence rate with respect to sample size in terms of accurately detecting the truly significant covariates and also exhibits the best performance when the sample sizes are small. Similar behaviour can be observed with respect to the Euclidean distance from the true parameter values (see Fig. 1b). As regards the KS distances depicted in Fig. 1c, we can see that the results of the NMD are the most stable for every sample size, and with moderately large sample size this method gives the best performance. In this study, greater accuracy of the NMD method, particularly for small sample size, indicates that in practical multiple hypothesis testing applications where the sample size is generally much smaller as compared to the number of parameters, incorporating the dependence structure in the multiple testing method indeed boosts accuracy.

Observe that variability is much higher in the Euclidean and KS distances compared to Jaccard similarity coefficients. Figure 1a indicates that as we are observing more and more samples the right regressors are getting selected with increasing precision. Nonetheless, incorrect decision regarding some regressors, even with moderately high regression coefficient, would contribute significantly to the Euclidean and KS distances. This is reflected in Figs. 1b and c.

6.5 Empirical studies on model misspecification

In this section, we study the effect of model misspecification on multiple testing methods. We generate data from the $AR(1)$ model in (17) for varying values of ρ_0. To allow model misspecification, we ignore the autoregressive part while fitting the data and perform variable selection according to the global local shrinkage prior in Sect. 6.2. The true values of the parameters are same as we have considered in Sect. 6.1 with a sample size of $n = 100$. Different values of ρ_0 are provided in the x-axis of different panels in Fig. 2. We compute the Jaccard similarity coefficient, Euclidean norm and KS distance in the same way as described in Sect. 6.3.

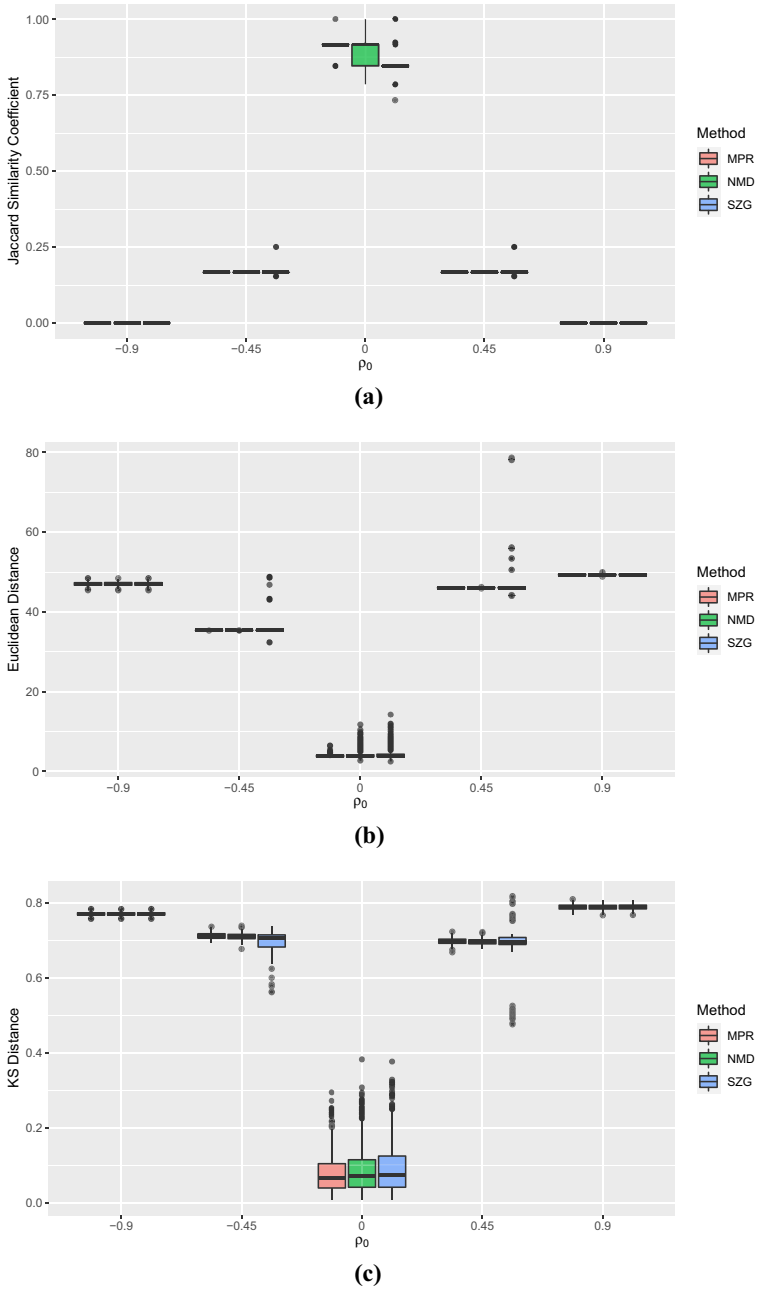

Fig. 2 Effect of model misspecification on multiple testing methods for varying autoregressive parameter with sample size of $n = 100$; the X-axis plots the true ρ_0 in the generative model: panel (a) shows the Jaccard similarity coefficient; panel (b) shows the Euclidean distance between the estimated model parameters and the true parameters; panel (c) shows the Kolmogorov–Smirnov distance between the true generative distribution and posterior predictive distribution for the competing methods across different values of ρ_0. All the panels show increased degree of misspecifications as ρ_0 deviates from fitted $\rho = 0$

Note that for $\rho_0 = 0$ all the methods perform quite accurately. In this case, there is no autoregressive component in the true data-generating model. Also Fig. 1 shows that asymptotics is taking precedence from sample size 75 onward. As the performance of all the three competing methods depends upon appropriate posterior probabilities, accurate results are quite expected for $\rho_0 = 0$. Variability in the Euclidean norms and KS distances is much lesser here compared to Figs. 1b and c for $n = 100$. The added precision is not surprising as we do not have the autoregressive component to model here.

However, the performance of all the methods deteriorates with the increase in model misspecification. The posterior probabilities of events may not properly showcase the uncertainty in case the class of postulated models have a high KL divergence from the true data-generating process. As ρ_0 deviates from zero the extent of misspecification increases (see Lemma S-6.1). Apparently from Fig. 2, the Bayesian multiple testing methods under consideration, being based on posterior probabilities fail to perform adequately. This study highlights that with misspecified models inadequate for explaining the variability in the data, it is indeed difficult to extract meaningful inference.

7 Real data analysis

We now consider variable selection using our Bayesian non-marginal multiple testing method in a real data context. The data, available at https://www4.stat.ncsu.edu/~boos/var.select/maize.html, obtained from Buckler et al. (2009), are regarding 25 crosses (also called families or populations) of maize flowers, each with about 200 observations on recombinant inbred lines (RILs). There are 7389 independent variables (covariates) representing the SNP markers, and the response variable is "days to anthesis male flowering time" (dtoa). In all, there are 4981 observations for the 25 crosses (excluding the missing values). Our aim to apply the Bayesian non-marginal multiple testing procedure to select the influential marker variables from the total of 7389, in a linear regression context, for each of the 25 crosses, each having about 200 observed values.

We consider the same Bayesian model as in Sect. 6.2 for this variable selection problem and subsequently employ our multiple testing procedure to select the relevant SNP markers. With the selected markers, we compute the corresponding fitted values for each of different populations. Figure 3, displaying the observed versus fitted dtoa values for each of different populations, indicates that the data variability is adequately explained by our model and methodologies. Due to space constraints, we show the plots of 12 populations in the main article and the rest in Section S-7. In the same latter section, we also report the causal SNPs for some of the populations.

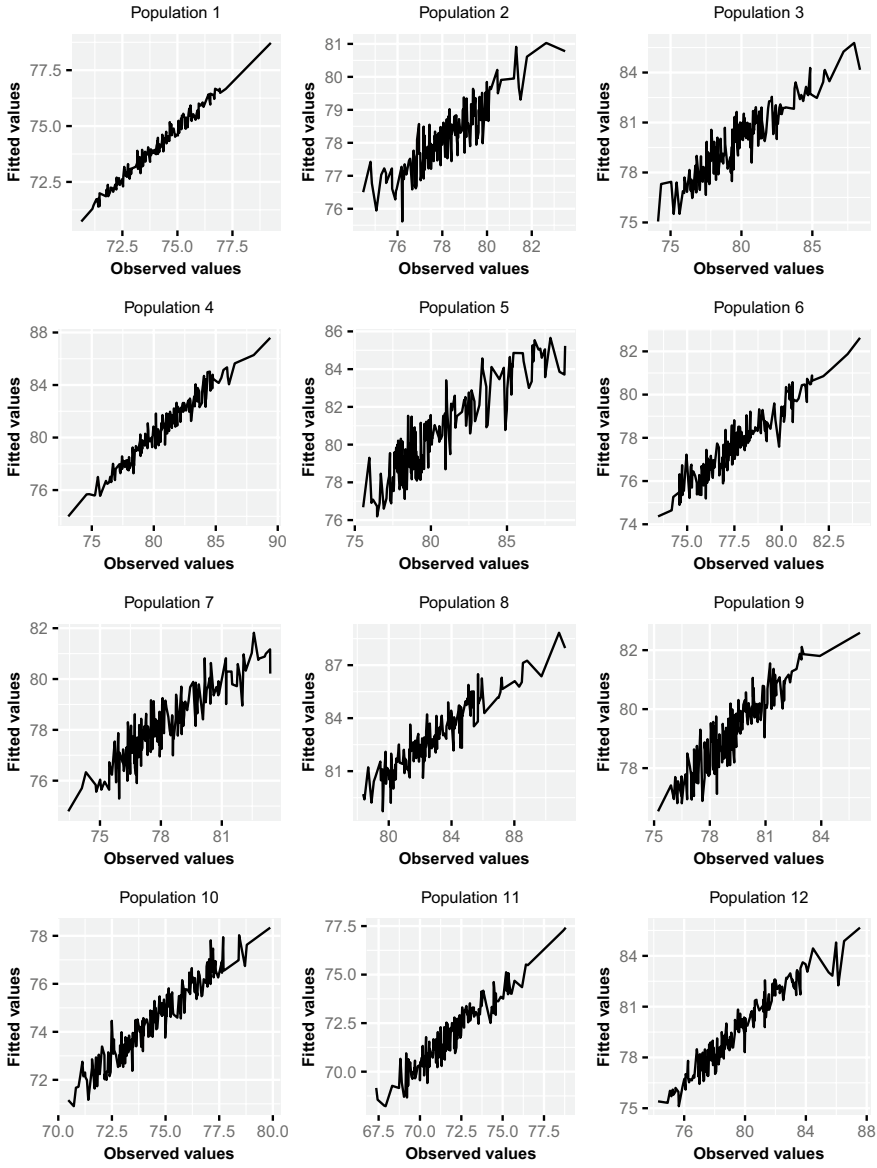

Fig. 3 Observed versus fitted *days to anthesis male flowering time* for the significant markers corresponding to the first 12 populations

8 Summary and conclusion

In this article, we have investigated asymptotic properties of Bayesian multiple testing procedures. We have shown strong consistency of the non-marginal Bayesian procedure under general dependence structure. As a corollary, we have shown that

additive loss function-based approaches are also consistent.

We have also studied asymptotic properties of multiple testing error rates. We have shown that the posterior versions of the error rates, namely $\text{FDR}_{\boldsymbol{X}_n}$ and $\text{FNR}_{\boldsymbol{X}_n}$, are directly associated with the entropy rate of the true data-generating model. Hence, from the Bayesian perspective, we advocate the posterior versions of error rates conditioned on the data. In the light of the dependence structure associated with the hypotheses, we introduce $\text{mFDR}_{\boldsymbol{X}_n}$ - a modified version of $\text{FDR}_{\boldsymbol{X}_n}$, the modification being with respect to the dependence among the parameters. The modified version is seen to be associated with a smaller entropy compared to its existing counterpart.

For α-control of type-I errors in the non-marginal procedure, a mild, but still an extra assumption of existence of disjoint groups of hypothesis where the nulls are true, is required. However, as we elucidated, this condition indeed indicates that grouping dependent hypotheses pools information across them and provides an extra safeguard against committing error. Importantly, as we have shown, for large sample sizes, α cannot take any value in (0, 1); in particular, we have provided lower bounds to the maximum possible values of mpBFDR and pBFDR and have shown that these lower bounds are significantly bounded away from 1, so that setting large values of α is not possible for large samples. Hence, for large samples, the practitioner must choose α carefully. As regards type-II error, we have shown that, with α-control of type-I error rates, pBFNR is likely to converge to zero at a faster rate than that without α-control of the type-I errors. Thus, the usual expectation of statisticians that controlling type-I error yields smaller type-II error in single hypothesis testing is expected to hold in our multiple testing framework.

We draw attention to the fact that most of our asymptotic results crucially hinge on the assumptions considered in Section S-1. In this regard, we have illustrated these assumptions in a variable selection problem with autoregressive response variables from a multiple testing perspective, along with the test for stationarity. In this problem, we show that the assumptions hold for any choice of proper prior over the general, non-compact parameter space, entailing strong consistency of Bayesian multiple testing methods. We have also discussed how verification of these assumptions is implicitly related to showing consistency of the maximum likelihood estimator. Indeed, proving strong consistency of Bayesian posterior distributions or maximum likelihood estimators is certainly quite challenging for non-compact parameter spaces and dependent setups, and our approach is probably of independent interest in this respect.

We have backed up our theoretical investigations with extensive simulation studies, comparing the performance of our NMD method with two other Bayesian multiple testing procedures for sample sizes ranging from small to moderately large. The results indicate clear superiority of the NMD method, particularly for small sample sizes. This is quite encouraging, since in practice, sample sizes are expected to be small compared to the number of available covariates. The message underlying the superior performance of NMD is that it exploits the dependence structure in a more wholesome way compared to the existing methods.

The empirical studies on misspecified models are particularly important. These studies show that multiple testing methods relying on inadequate models would

suffer. The results by Shalizi (2009) show that asymptotically the model with the minimum KL divergence from the true data-generating process would be preferred, however, that preferred model can be quite bad. Methods reckoning on the uncertainty delivered by the posterior probabilities suffer in such cases.

Application of our multiple testing procedure to a real maize data concerning selection of influential marker variables from a total of 7389 variables, yielded quite encouraging results. Since variable selection from among many variables is an important real problem, our results seem to indicate the importance of our multiple testing procedure.

In this article, we have assumed m, the number of hypotheses, to be fixed. But it is also important to investigate the asymptotic theory when m also grows with the sample size n, particularly because of its relevance in practical problems. As Shalizi's framework is valid for infinite-dimensional models, it is not too difficult to extend our consistency results in the high-dimensional setup. It is also worth studying the asymptotic behaviour of the error rates in such scenario. Chandra and Bhattacharya (2020) made some progress in these directions.

Acknowledgements We sincerely express our gratitude to the Editor, the Associate Editor, and the referees for their responsible handling of our paper and providing valuable comments that led to significant improvement in the presentation and readability of our paper.

References

Benjamini, Y., Heller, R. (2007). False discovery rates for spatial signals. *Journal of the American Statistical Association*, *102*(480), 1272–1281.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, *57*(1), 289–300.

Benjamini, Y., Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998.

Berry, D. A., Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, *82*(1), 215–227.

Brown, A., Lazar, N. A., Dutta, G. S., Jang, W., McDowell, J. E. (2014). Incorporating spatial dependence into bayesian multiple testing of statistical parametric maps in functional Neuroimaging. *NeuroImage*, *84*(1), 97–112.

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., Ersoz, E., et al. (2009). The genetic architecture of maize flowering time. *Science*, *325*(5941), 714–718. https://doi.org/10.1126/science.1174276.

Chandra, N. K., Bhattacharya, S. (2019). Non-marginal decisions: A novel Bayesian multiple testing procedure. *Electronic Journal of Statistics*, *13*(1), 489–535. https://doi.org/10.1214/19-EJS1535.

Chandra, N. K., Bhattacharya, S. (2020). High-dimensional asymptotic theory of Bayesian multiple testing procedures under general dependent setup and possible misspecification. arXiv preprint arXiv :2005.00066.

Chandra, N. K., Singh, R., Bhattacharya, S. (2019). A novel Bayesian multiple testing approach to deregulated miRNA discovery harnessing positional clustering. *Biometrics*, *75*(1), 202–209. https://doi.org/10.1111/biom.12967.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, *102*(477), 93–103.

Fan, J., Han, X., Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*, *107*(499), 1019–1035. https://doi.org/10.1080/01621459.2012.7204784.

Finner, H., Roters, M. (2002). Multiple hypotheses testing and expected number of type I. Errors. *The Annals of Statistics*, *30*(1), 220–238. https://doi.org/10.1214/aos/1015362191.

Finner, H., Dickhaus, T., Roters, M. (2007). Dependency and false discovery rate: Asymptotics. *The Annals of Statistics*, *35*(4), 1432–1455. https://doi.org/10.1214/009053607000000046.

Finner, H., Dickhaus, T., Roters, M. (2009). On the false discovery rate and an asymptotically optimal rejection curve. *The Annals of Statistics*, *37*(2), 596–618. https://doi.org/10.1214/07-AOS569.

Geman, S., Hwang, C. R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, *10*(2), 401–414. https://doi.org/10.1214/aos/1176345782.

Ghosal, S., Ghosh, J. K., van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, *28*(2), 500–531. https://doi.org/10.1214/aos/1016218228.

Ghosh, D., Chen, W., Raghunathan, T. (2006). The false discovery rate: A variable selection perspective. *Journal of Statistical Planning and Inference*, *136*(8), 2668–2684. https://doi.org/10.1016/j.jspi.2004.10.024.

Ishwaran, H., Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773. https://doi.org/10.1214/009053604000001147.

Jaccard, P. (1901). Étude Comparative de la Distribution Florale dans une Portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, *37*, 547–579.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles*, *44*, 223–270.

Jensen, S. T., Erkan, I., Arnardottir, E. S., Small, D. S. (2009). Bayesian testing of many hypotheses × many genes: A study of sleep apnea. *The Annals of Applied Statistics*, *3*(3), 1080–1101.

Liu, Y., Sarkar, S. K., Zhao, Z. (2016). A new approach to multiple testing of grouped hypotheses. *Journal of Statistical Planning and Inference*, *179*, 1–14. https://doi.org/10.1016/j.jspi.2016.07.004.

Müller, P., Parmigiani, G., Robert, C., Rousseau, J. (2004). Optimal sample size for multiple testing: The case of gene expression microarrays. *Journal of the American Statistical Association*, *99*(468), 990–1001.

Narisetty, N. N., He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, *42*(2), 789–817. https://doi.org/10.1214/14-AOS1207.

Risser, M. D., Paciorek, C. J., Stone, D. A. (2019). Spatially dependent multiple testing under model misspecification, with application to detection of anthropogenic influence on extreme climate events. *Journal of the American Statistical Association*, *114*(525), 61–78.

Sarkar, S. K., Zhou, T., Ghosh, D. (2008). A general decision theoretic formulation of procedures controlling FDR and FNR from a Bayesian perspective. *Statistica Sinica*, *18*(3), 925–945.

Schwartz, L. (1965). On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *4*(1), 10–26.

Schwartzman, A., Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika*, *98*(1), 199–214.

Scott, J. G. (2009). Nonparametric Bayesian multiple testing for longitudinal performance stratification. *The Annals of Applied Statistics*, *3*(4), 1655–1674.

Scott, J. G., Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, *38*(5), 2587–2619. https://doi.org/10.1214/10-AOS792.

Shalizi, C. R. (2009). Dynamics of Bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, *3*, 1039–1074. https://doi.org/10.1214/09-EJS485.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics*, *31*(6), 2013–2035. https://doi.org/10.1214/aos/1074290335.

Sun, W., Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, *102*(479), 901–912.

Sun, W., Cai, T. T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *71*(2), 393–424.

Sun, W., Reich, B. J., Tony Cai, T., Guindani, M., Schwartzman, A. (2015). False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *77*(1), 59–83. https://doi.org/10.1111/rssb.12064.

Welch, B. L. (1939). On confidence limits and sufficiency, and particular reference to parameters of location. *Annals of Mathematical Statistics*, *10*, 58–69.

Xie, J., Cai, T. T., Maris, J., Li, H. (2011). Optimal false discovery rate control for dependent data. *Statistics and Its Interface*, *4*(4), 417.

Zhang, C., Fan, J., Yu, T. (2011). Multiple testing via FDR_l for large scale imaging data. *The Annals of Statistics*, *39*(1), 613–642. https://doi.org/10.1214/10-AOS848.