TEST
https://doi.org/10.1007/s11749-023-00886-7

ORIGINAL PAPER

Selective inference for false discovery proportion in a hidden Markov model

Marie Perrot-Dockès[1,3,4] · Gilles Blanchard[2] · Pierre Neuvial[3] · Etienne Roquain[4]

Received: 7 July 2022 / Accepted: 24 July 2023

Abstract
We address the multiple testing problem under the assumption that the true/false hypotheses are driven by a hidden Markov model (HMM), which is recognized as a fundamental setting to model multiple testing under dependence since the seminal work of Sun and Cai (J R Stat Soc Ser B (Stat Methodol) 71:393–424, 2009). While previous work has concentrated on deriving specific procedures with a controlled false discovery rate under this model, following a recent trend in selective inference, we consider the problem of establishing confidence bounds on the false discovery proportion, for a user-selected set of hypotheses that can depend on the observed data in an arbitrary way. We develop a methodology to construct such confidence bounds first when the HMM model is known, then when its parameters are unknown and estimated, including the data distribution under the null and the alternative, using a nonparametric approach. In the latter case, we propose a bootstrap-based methodology to take into account the effect of parameter estimation error. We show that taking advantage of the assumed HMM structure allows for a substantial improvement of confidence

✉ Marie Perrot-Dockès
marie.perrot-dockees@u-paris.fr

Gilles Blanchard
gilles.blanchard@universite-paris-saclay.fr

Pierre Neuvial
pierre.neuvial@math.univ-toulouse.fr

Etienne Roquain
etienne.roquain@upmc.fr

1 CNRS MAP5 UMR 8145, Université de Paris, Paris, France

2 Inria Laboratoire de mathématiques d'Orsay, CNRS, Université Paris-Saclay, 91405 Orsay, France

3 Institut de Mathématiques de Toulouse UMR 5219, CNRS UPS, Université de Toulouse, 31062 Toulouse Cedex 9, France

4 Laboratoire de Probabilités, Statistique et Modélisation, CNRS, Sorbonne Université, Université de Paris, Paris, France

Published online: 14 September 2023

bound sharpness over existing agnostic (structure-free) methods, as witnessed both via numerical experiments and real data examples.

Keywords Post hoc bounds · Hidden Markov model · False discovery proportion · Posterior distribution · Bootstrap

Mathematics Subject Classification 62J15

1 Introduction

1.1 Context and motivation

To analyze large, heterogeneous and complex data, the analyst often adopts a *post hoc* approach, by choosing methods, formulating questions, selecting models or variables, on the basis of the data set. In particular, performing statistical inference after selection is a flourishing research field, often named as *selective inference*. The main challenge is to avoid the selection bias, by properly calibrating the error probabilities or risks.

Observations arising from applications are generally not independent. Hidden Markov Models (HMMs) are a tool of choice to model stochastic processes with temporal or spatial dependence, and they have been widely successfully used in various areas including signal processing (Gales and Young 2008), economics (Kim et al. 1999), or computational biology (Koski 2001). This paper is motivated by a specific use case in genomics: the differential analysis of DNA copy number alterations (CNA) in cancer cells. Cancer cells are characterized by structural changes in the number of gene copies along the genome, and modern biotechnologies such as microarrays and sequencing are commonly used to quantify such changes at high resolution (Albertson et al. 2003). DNA copy number (CN) profiles are generally modeled as piece-wise constant signals, and HMMs have been extensively used for this purpose (see, e.g., Fridlyand et al. 2004; Shah et al. 2009; Zhang 2010). Given the observation of CN profiles along the genome for individuals classified in two groups corresponding to different types of cancer, differential analysis aims at identifying regions of the genome for which the CN profiles differ "significantly" between the two cancer types. As a case in point, we consider a study of ovarian cancers where a differential analysis of 117 patients with or without endometriosis is performed (Okamoto et al. 2015), see Sect. 5. This study comprises CNA measurements for 236,385 genomic locations (loci) for 63 patients without endometriosis and 54 patients with endometriosis. Figure 1 displays two-sided Wilcoxon test statistics of the null hypothesis of no signal difference between the two cancer types, for 4799 loci on chromosome 7.

We model these test statistics as an HMM with two states, corresponding to the null and alternative hypotheses for the test. Loci highlighted in orange correspond to a specific data-driven selection of regions. The methods developed in this paper make it possible to construct confidence bounds on the false positives in such data-driven selections. They are both more informative than approaches based on the control of local False Discovery Rates (Sun and Cai 2009), which only provide posterior point estimates, and more powerful than agnostic post-selection bounds (Goeman and Solari 2011; Blanchard et al. 2020) which do not take the dependency structure into account

in the inference. While the advantage of our approach will be highlighted specifically with this data set (see Sect. 5), comprehensive numerical experiments will show that this superiority holds in a wide spectrum of cases, see Sects. 4 and 5: in general, in any situation where assuming an HMM structure is reasonable.

1.2 Selective inference

In very general terms, given an observation $X \in \mathcal{X}$, statistical inference is deemed *selective* if it concerns a quantity of interest $\xi(\theta, R) \in \mathbb{R}$, depending on a vector of unknown parameters θ and on a data-dependent selection set $R = S(X)$ taking values in some space $\mathcal{R}$, which generally represents a subset of coordinates of θ. We call a function $S(.) : \mathcal{X} \to \mathcal{R}$ a *selection policy*. The standard setting is to assume a statistical model indexed by the unknown parameter θ, $\{P_\theta, \theta \in \Theta\}$, and to aim at constructing a confidence interval $I_\alpha(X)$, satisfying

$$\forall \theta \in \Theta : \qquad \mathbb{P}_\theta(\xi(\theta, S(X)) \in I_\alpha(X)) \geq 1 - \alpha. \tag{1}$$

To this end, a first approach, the *conditional approach*, is to assume the selection policy $S(.)$ known and to construct a family of intervals $(I^S_\alpha(X, R))_{R \in \mathcal{R}}$ such that

$$\forall \theta \in \Theta, \forall R \in \mathcal{R} : \qquad \mathbb{P}_\theta(\xi(\theta, R) \in I^S_\alpha(X, R) | S(X) = R) \geq 1 - \alpha. \tag{2}$$

Another approach, the *agnostic approach*, does not assume the selection policy to be known but only its realization $S(X)$, and aims at constructing $(I_\alpha(X, R))_{R \in \mathcal{R}}$ such that

$$\forall \theta \in \Theta : \qquad \mathbb{P}_\theta(\forall R \in \mathcal{R} : \xi(\theta, R) \in I_\alpha(X, R)) \geq 1 - \alpha. \tag{3}$$

Observe that both (2) and (3) imply (1), but the interval family $I_\alpha(X, R)$ in (3) is valid for any selection policy, while the interval family $I^S(X, R)$ in (2) is tuned to a specific selection policy $S(.)$.

In the linear regression model and various error measures concerning inference, the conditional approach has been advocated, e.g., by Lee et al. (2016) (more specifically for selection policies related to the LASSO) and Tibshirani et al. (2018); while the

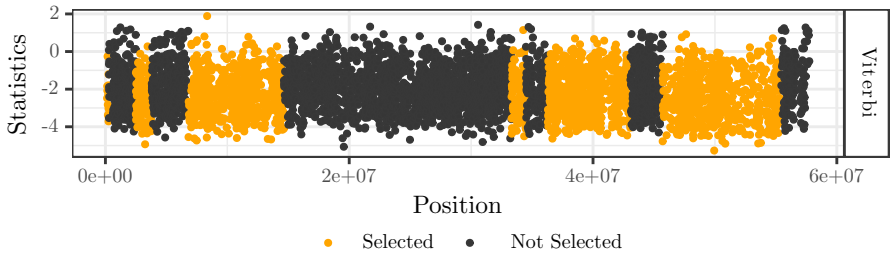

Fig. 1 Wilcoxon test statistics for 13,239 loci on chromosome 7 in the Okamoto data set (Okamoto et al. 2015). Loci highlighted in orange correspond to a specific data-driven selection of regions

agnostic approach has been studied, e.g., by Scheffé (1959), Berk et al. (2013), Bachoc et al. (2018, 2019).

In the model under scrutiny here, we assume the truth is carried by a *configuration vector* $\theta \in \Theta = \{0, 1\}^m$ for which $\theta_i = 0$ if and only if the i-th variable is not active (in the example of Sect. 1.1, $\theta_i = 1$ then means chromosomal aberration at position i). Given the configuration θ, we observe $X \in \mathbb{R}^m$ with independent coordinates $X_i \sim P_{\theta_i}$, where P_0, P_1 are two distributions on $\mathbb{R}$ with densities f_0, f_1. Then, for a data-dependent selection set $S(X) = R \subset \mathbb{N}_m := \{1, \ldots, m\}$ of "detected" active variables, define the following quantity of interest

$$\mathrm{FDP}(\theta, R) = \frac{\sum_{i \in R} \mathbb{1}\{\theta_i = 0\}}{|R| \vee 1},$$

which is called the false discovery proportion for region R. Since its introduction by Benjamini and Hochberg (1995), it has become a classical way of accounting for the errors made by the selection R.

Confidence bounds for the FDP which are "agnostic" in the sense of (3) have been considered by Genovese and Wasserman (2006), Goeman and Solari (2011), Blanchard et al. (2020). A counterpart of the uniform guarantee granted by the agnostic approach is conservativeness, that is, the obtained bound can be far from the true value $\mathrm{FDP}(\theta, S(X))$ for a given particular selection policy $S(\cdot)$. Some improvements have been proposed in the literature by adding some local structure on the null hypotheses (Durand et al. 2020) or by considering subset restricted to a "path" of procedures of interest (Katsevich and Ramdas 2020).

Let us also mention that suitable confidence interval adjustments after selection have been proposed by Benjamini and Yekutieli (2005), Benjamini and Bogomolov (2014), Weinstein and Ramdas (2019) in a general multiple testing context.

1.3 Selective inference in latent variables models

Following the motivation in Sect. 1.1, we depart from assuming a statistical model indexed by the configuration θ and assume instead that the configuration vector θ is a latent random variable in a HMM model for the observations. This model will be specified in detail in Sect. 2. For the general considerations in the remainder of the present introduction section though, it is sufficient to assume that we have a model parametrized by Γ for the joint distribution of (θ, X), denoted by P_Γ. The parameter Γ specifies on the one hand the marginal distribution of the configuration vector θ as a Markov chain, and on the other hand, the conditional distribution of the observation X conditionally on θ, via the distributions P_0, P_1 as in the previous section. The distribution on the underlying probability space is denoted by $\mathbb{P}_\Gamma$.

Modeling θ as a latent variable is in line with empirical Bayes methods for multiple testing, and in particular, of the widely used two-group model and so-called local FDR method, see Efron et al. (2001), Efron (2008), Sun and Stephens (2018), Jin and Cai (2007), Sun and Cai (2009), Cai and Sun (2009), Cai and Jin (2010), Heller and Yekutieli (2014), Cai et al. (2019), Rebafka et al. (2019), among others (more details about the relation to these references can be found in Sects. 1.5 and 1.6).

Under that setting, we consider the following aim: for any selection policy $S(\cdot) : x \in \mathcal{X} \mapsto S(x) \subset \mathbb{N}_m$, build a functional $I_\alpha(X, S(\cdot))$ valued in the intervals of $[0, 1]$, such that

$$\mathbb{P}_\Gamma\left(\mathrm{FDP}(\theta, S(X)) \in I_\alpha(X, S(\cdot))\right) \geq 1 - \alpha. \tag{4}$$

Sometimes, the post-selection interval $I_\alpha(X, S(\cdot))$ only requires knowledge of the selected region $S(X)$ on the observed data, in which case we will denote it by $I_\alpha(X, S(X))$ with an overload of notation.

As a particular case, intervals of the form $I_\alpha(X, S(\cdot)) = [0, \overline{\mathrm{FDP}}_\alpha(X, S(\cdot))]$ can be considered, in which case (4) reduces to a post-selection upper-bound $\overline{\mathrm{FDP}}_\alpha(X, S(\cdot))$ on $\mathrm{FDP}(\theta, S(X))$. However, (4) also allows for considering post-selection lower-bounds, which can also be informative in practice, see Sect. 5.

1.4 Contribution: posterior post-selection bounds

The standard approach (1), in the considered model, amounts to derive a bound which holds conditionally on any value θ of the latent true null hypothesis configuration, i.e.,

$$\forall \theta \in \{0, 1\}^m : \qquad \mathbb{P}_\Gamma\left(\mathrm{FDP}(\theta, S(X)) \in I_\alpha(X, S(\cdot)) \mid \theta\right) \geq 1 - \alpha. \tag{5}$$

While (5) implies (4), conditioning on θ amounts to disregarding the assumed model on the latent variables and defeats the purpose of the HMM modeling. To exploit this structural assumption, we consider instead tackling (4) via conditioning with respect to X and solving

$$\mathbb{P}_\Gamma\left(\mathrm{FDP}(\theta, S(X)) \in I_\alpha(X, S(\cdot)) \mid X\right) \geq 1 - \alpha, \qquad \mathbb{P}_\Gamma - \text{ a.s.}, \tag{6}$$

which can be achieved by considering the *posterior distribution*, that is, the distribution of θ conditionally on X under $\mathbb{P}_\Gamma$.

If the parameter Γ governing the distribution of the latent variables is known, one can build a posterior interval only depending on $S(X)$ that fulfills (4): denoting $R = S(X)$, and provided $|R| > 0$, let

$$I_\alpha(X, R) = [L_{\alpha\gamma}(X, R; \Gamma), U_{\alpha(1-\gamma)}(X, R; \Gamma)]; \tag{7}$$

$$U_\beta(X, R; \Gamma) = |R|^{-1} \min\left\{n \in \{0, \ldots, m\} \,:\, \mathbb{P}_\Gamma\left(\sum_{i \in R}(1 - \theta_i) \leq n \,\middle|\, X\right) \geq 1 - \beta\right\}; \tag{8}$$

$$L_\beta(X, R; \Gamma) = |R|^{-1} \max\left\{n \in \{0, \ldots, m\} \,:\, \mathbb{P}_\Gamma\left(\sum_{i \in R}(1 - \theta_i) \geq n \,\middle|\, X\right) \geq 1 - \beta\right\}, \tag{9}$$

for some $\gamma \in (0, 1)$ balancing the errors between the upper and lower bounds. We will sometimes drop the X in the notation for brevity.

This interval accounts for the particular HMM modeling via the posterior distribution and thus, is expected to be considerably sharper than the other intervals described

in Sect. 1.2 and/or based on (5), that ignore this structure. Unfortunately, the functionals $L_\beta(X, S(X); \Gamma)$ and $U_\beta(X, S(X); \Gamma)$ are not directly accessible because the model parameter Γ is typically unknown.

We propose the following approximations of the oracle bound $U_\beta(S(X); \Gamma)$ (similar for $L_\beta(S(X); \Gamma)$):

- Plug-in: $U_\beta^{\mathrm{PI}}(X, S(X)) = U_\beta(X, S(X); \widehat{\Gamma})$, where $\widehat{\Gamma}$ is an estimator of Γ. We will consider an estimator $\widehat{\Gamma}$ based upon an iterative EM-type algorithm, see Sect. 2.2;
- Bootstrap 1: $U_\beta^{\mathrm{boot1}}(X, S(\cdot))$, correcting the fluctuations in $\widehat{\Gamma}$ and $S(X)$ of the above plug-in bounds by using a bootstrap approach generating resampled data X^* under $P_{\widehat{\Gamma}}$ (thus, computing repeatedly $\widehat{\Gamma}^*$ and $S(X^*)$), see Sect. 3.3.
- Bootstrap 2: $U_\beta^{\mathrm{boot2}}(X, S(X))$, a heuristic approximation of $U_\beta^{\mathrm{boot1}}(X, S(\cdot))$, which follows the same scheme, except that only $\widehat{\Gamma}^*$ is recomputed from each bootstrap sample; in particular, the selection set $S(X)$ is kept fixed during the bootstrap process, see Sect. 3.3.
- Bootstrap 3: $U_\beta^{\mathrm{boot3}}(X, S(\cdot))$, a bootstrap bound based on generating resampled data (θ^*, X^*) under $P_{\widehat{\Gamma}}$ to approximate the distribution of $\mathrm{FDP}(\theta, S(X))$, recentered with the plug-in bound $U_\beta(X, S(X); \widehat{\Gamma})$, see Sect. 3.3.

Remark 1.1 Bootstrap 1 and 3 require the knowledge of the whole selection policy $S(\cdot)$. Bootstrap 2 only needs $R = S(X)$, the value of the selection policy for the observation X. For Bootstrap 1 and 3, we need to be able to simulate jointly $(X^*, R^*) = (X^*, S(X^*))$, while for Bootstrap 2, we only have to simulate X^* marginally and to observe the realization of $R = S(X)$ on the original data. In practice, it means that for Bootstrap 2, the full selection $S(\cdot)$ could be kept private from the statistician, and only its value on the given data is communicated (with the guarantee that it is indeed a function of X only).

The coverage of these bounds is evaluated via extensive numerical experiments in Sect. 4, which typically reflect the impact of the estimation error of Γ in the different bounds. Our conclusion is that while the existing approaches, ignoring the latent HMM structure, are over-pessimistic, the plug-in approach can be slightly over-optimistic. A good trade-off is provided by bootstrap-based strategies, which ensure a correct coverage while taking full advantage of the HMM structure. This is shown in Fig. 2 that compares two of the proposed bounds (Oracle and Bootstrap 3) to some of the bounds from the literature Goeman and Solari (2011) ("Simes"), Blanchard et al. (2020) ("BNR", with the so-called β-template defined therein), Katsevich and Ramdas (2020) ("KR") which do not take the HMM structure into account. The difference Δ between the actual FDP and the estimated bound is plotted for different selection policies. The percentage of simulation runs for which $\Delta < 0$ is displayed within rectangles; it should not exceed the target risk $\beta = 10\%$. The bounds taking into account the HMM structure are closer to the actual FDP than structure-agnostic bounds. More details about the scenario of simulation and the bounds are provided in Sect. 4.

These bounds have been implemented in the R package `SansSouci.HMM`, which is available from https://github.com/Marie-PerrotDockes/sanssouci.hmm.

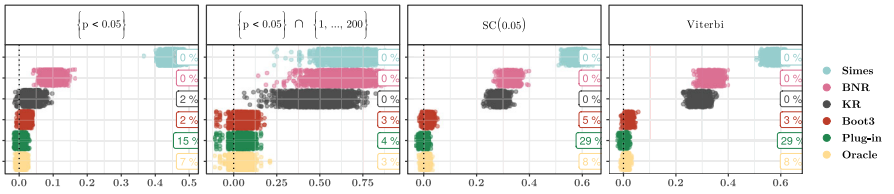

Fig. 2 Results for 300 simulation runs with the model parameters described in Sect. 4.2. Difference $\Delta = U(\cdot) - \mathrm{FDP}(\theta, S(X))$ for different upper bounds (rows) and selection policies $S(X)$ (columns), with the empirical violation probability (proportion of $\Delta < 0$) of the bound displayed within rectangles. Target risk $\beta = 10\%$

1.5 Relation to posterior point estimates of the FDR

In the previous literature, under a joint latent configuration/observation variable model P_Γ such as the one considered here, a commonly considered goal is to focus on the FDR of a given selection policy $S(\cdot) : x \in \mathcal{X} \mapsto S(x) \subset \mathbb{N}_m$, that is,

$$\mathrm{FDR}(S(\cdot), \Gamma) = \mathbb{E}_{(\theta,X)\sim P_\Gamma}(\mathrm{FDP}(\theta, S(X))).$$

Observe that, since $\mathrm{FDR}(S(\cdot), \Gamma) = \mathbb{E}_\Gamma(\mathbb{E}_\Gamma(\mathrm{FDP}(\theta, S(X)) \mid X))$, the quantity

$$\widehat{\mathrm{FDR}}(S(X), \Gamma) := \mathbb{E}_\Gamma(\mathrm{FDP}(\theta, S(X)) \mid X) = |S(X)|^{-1} \sum_{i \in S(X)} \mathbb{P}_\Gamma(\theta_i = 0 \mid X) \quad (10)$$

is an unbiased estimator for $\mathrm{FDR}(S(\cdot), \Gamma)$ (assuming Γ known for now) for any selection policy $S(\cdot)$. Since this holds for any selection policy $S(\cdot)$, $\widehat{\mathrm{FDR}}(S(X), \Gamma)$ can be considered as a post-selection estimate.

The marginal conditional probabilities $\mathbb{P}_\Gamma(\theta_i = 0 \mid X)$, $1 \le i \le m$ appearing in (10), are generally referred to in the literature as the local FDR (Efron 2008) or the ℓ-values (Castillo and Roquain 2020), but without the explicit purpose of being applied to any selection policy to our knowledge. In fact, the estimate (10) can be used for several purposes:

- for *any given* selection policy $S(\cdot)$, to compute an unbiased estimate of the FDR. Observe that this estimate only depends on the selected region $S(X)$, and not on the full selection policy $S(\cdot)$.
- to *design* a selection policy $S(\cdot)$ such that $\mathrm{FDR}(S(\cdot), \Gamma)$ is equal to a specified level α. The simplest way to guarantee this is to construct $S(X)$ so that $\widehat{\mathrm{FDR}}(S(X), \Gamma)$ is constant equal to α. This is, for instance, the method proposed by Sun and Cai (2007, 2009), Cai and Sun (2009), Cai et al. (2019), where selections take the form of level sets of a relevant statistic T: $S_t(X) = \{i \in \mathbb{N}_m : T_i(X) \ge t\}$, and the threshold $t = t(X)$ is chosen also depending on X as the smallest value such that $\widehat{\mathrm{FDR}}(S_{t(X)}(X), \Gamma) = \alpha$.

In practice, the parameter Γ is not known, so after introducing some suitable estimator $\widehat{\Gamma}$ of Γ (which can be built via an empirical Bayes approach), the plug-in quantity

$\widehat{\mathrm{FDR}}(S(X), \widehat{\Gamma})$ is used. In general, theoretical studies then ensure that the above properties hold in a suitable asymptotical sense, see Sun and Cai (2007, 2009), Cai and Sun (2009), Cai et al. (2019).

However, this approach only provides an (asymptotically) unbiased point estimate, and not a confidence interval for the FDP. The starting point of the present work is that we can use the same principle to derive confidence bounds for the FDP as in (7)-(8)-(9), which are more informative for practice, see Fig. 5 in Sect. 5 and Figure S9 in supplementary material. Observe that, analogously to the two points above, the derived bounds can be used both to evaluate and design a policy.

1.6 Relation to other work

1.6.1 Multiple testing via a latent structure

Latent variables are widely used in statistics to design models with specific structures. In multiple testing, they have been used to model external "confounding factors" or "systemic effects" that induce dependencies between the tests, see Leek and Storey (2008), Friguet et al. (2009), Fan and Han (2017), Fan et al. (2019) and references of Sun et al. (2012) for genetic or genomic applications. In the present paper, the philosophy is rather different: latent variables are used to model dependencies between the null hypothesis configurations (here, between the coordinates of the configuration vector θ). While the unstructured case where the θ_i are i.i.d. can be traced back to the two-group model (Efron et al. 2001; Sun and Cai 2007), the structured HMM case has been shown to improve over the independent case by Sun and Cai (2009), with a substantial gain in power. More recent examples of structures include two-sample sparsity (Cai et al. 2019) or stochastic block models for graph-structured nulls (Rebafka et al. 2019), for which substantial improvements are also shown with respect to the unstructured case.

1.6.2 Empirical Bayes FDR methods

Our method, as all methods cited above, has a Bayesian flavor, which is the case in general for all latent-based multiple testing based on the use of "local FDR", that is, the posterior probability that an item was generated under the null. While oracle versions (with a known model parameter) of such methods are known to be optimal in some way (see Sun and Cai 2007 or more recently Heller and Rosset 2021), a challenging part is to evaluate how the methods can handle parameter estimation, which is generally shown in an asymptotic manner. As already underlined in Sect. 1.5, all these methods have been developed for the FDR metric, not for FDP confidence intervals. Since the FDP metric considered in our post-selection bounds is considerably more difficult to analyze than the FDR one, obtaining a theoretical consistency result for the coverage of our bounds is out of scope here. It is left as a challenge for future investigations, see Sect. 6.3 for a discussion on this issue.

1.6.3 Nonparametric inference in HMM

As said above, the feasibility of parameter estimation plays a crucial role in empirical Bayes multiple testing. The quality of estimation heavily relies on the structural assumptions made for the latent variable distribution and for the considered model for the observations conditionally on the latent variables. Adding structure allows to increase performance. For an HMM with nonparametric emission densities, the situation is more favorable than in the unstructured two-group model, see Gassiat et al. (2016), Alexandrovich et al. (2016), with provable consistency guarantees for the "local FDR" values (De Castro et al. 2017; Abraham et al. 2021a).

1.6.4 Null estimation

Estimating the null density is both crucial and difficult in the two group model, as shown empirically by Efron (2004, 2007, 2008, 2009) and further studied by Schwartzman (2010), Azriel and Schwartzman (2015), Stephens (2017), Sun and Stephens (2018). In particular, the recent work of Roquain and Verzelen (2020) theoretically shows that some sparsity is needed to obtain valid FDR inference when the null is estimated. By contrast, and as said above, the HMM assumption made in our setting makes this estimation much easier. In order to stick with the usual multiple testing line of research, we will make a distinction between the situation where the null distribution is assumed to be known (which is studied in the core of this paper as it is the most standard case), and the case where it is not (which is studied more specifically in supplementary material Section S5). The numerical experiments of Sect. 4 confirm the validity of our bounds even in the latter case.

1.6.5 Bayesian conditional selective inference

We emphasize that our guarantees hold conditionally on X, see (6). This is markedly different from the following conditional control used in Yekutieli (2012) and Panigrahi et al. (2020): for all $R \subset \{1, \dots, m\}$,

$$\mathbb{E}_\theta \left[\mathbb{P}_{X|\theta} \left(\mathrm{FDP}(\theta, R) \in I_\alpha(X, S(\cdot)) \mid S(X) = R \right) \right] \geq 1 - \alpha. \tag{11}$$

The latter criterion has completely different interpretation and would require a substantially different methodology. For more details on this issue, we refer the reader to Yekutieli (2012), to the recent paper of Panigrahi et al. (2020) and references therein.

2 Hidden Markov modeling

In this section, we define the HMM modeling of our problem and add some more material that will be useful to deal with the post-selection bounds.

2.1 Model, notation and posterior distribution

The model we use here is essentially the one considered by Sun and Cai (2009). Let us consider a hidden Markov process $(\theta, X) = ((\theta_i, X_i)_{1 \leq i \leq m})$, where

- $\theta = (\theta_1, \ldots, \theta_m) \in \{0, 1\}^m$ is a unobserved latent variable sequence following a stationary Markov chain with transition matrix $A = (a_{q,\ell})_{q,\ell \in \{0,1\}}$ where $a_{q,\ell} \in (0, 1)$ for $q, \ell \in \{0, 1\}$, $a_{q,0} + a_{q,1} = 1$ for $q \in \{0, 1\}$, and $a_{0,0} \neq a_{1,0}$ to ensure that A is full rank.
- conditionally on θ, the observation $X = (X_i)_{1 \leq i \leq m} \in \mathbb{R}^m$ has independent coordinates and for all $i \in \{1, \ldots, m\}$, $X_i \,|\, \theta_i \sim f_0$ if $\theta_i = 0$ and $X_i \,|\, \theta_i \sim f_1$ if $\theta_i = 1$, where f_0 and f_1 are two distinct densities on $\mathbb{R}$ (with respect to the Lebesgue measure).

While Sun and Cai (2009) assumed f_0 known and modeled f_1 via a mixture of Gaussians, we consider a fully nonparametric model here. Namely, following Gassiat et al. (2016), the above model is identifiable under the stated assumptions (which is a remarkable result since f_0, f_1 can be arbitrary densities).

Observe that, since the sequence θ is assumed to be stationary, this implicitly means that θ_1 is generated according to the marginal stationary distribution π on $\{0, 1\}$ associated with A. Also, since we consider a testing paradigm where f_0 is the null distribution, whereas f_1 is the alternative distribution, the roles of f_0 and f_1 are not exchangeable. We will consider both of the following cases:

- f_0 is known, in which case the parameter is $\Gamma = (A, f_1)$. In that case, the multiple testing task aims at finding observations X_i that depart from the distribution f_0;
- f_0 is unknown, in which case the parameter is $\Gamma = (A, f_0, f_1)$. In this case, since we do not know the distribution of the nulls, the task is detect 'outliers', that are X_i's with abnormal behavior. To distinguish f_0 from f_1, a classical assumption is $\mathbb{P}(\theta_1 = 0) > \mathbb{P}(\theta_1 = 1)$, that is, f_0 is the predominant class in the sample X.

The distribution of (θ, X) is denoted by P_Γ. For convenience, we will also sometimes write $X \sim P_\Gamma$ when the variable θ is not needed. In our model, the distribution of θ conditionally on X (the posterior distribution) is well known.

Proposition 2.1 (Proposition 3.3.2 of Cappé et al. 2006) *In the model described above, the distribution of $\theta \mid X$ is a heterogeneous Markov chain.*

The initial distribution and transition probabilities $\Pi_i(\Gamma)$, $2 \leq i \leq m$, of this Markov chain can be computed via the standard forward-backward algorithm, see Section S1 in the supplementary material.

2.2 Parameter estimation

Since our posterior bounds depend on the unknown parameter Γ, building an estimator $\widehat{\Gamma} = (\widehat{A}, \widehat{f_1})$ of $\Gamma = (A, f_1)$ is crucial to obtain a computable bound. In this section, we only deal with the case where f_0 is known, the other case being similar and postponed to supplementary material Section S5.

For this, we make use of a pseudo-Expectation-Maximization (EM) algorithm, combined with a weighted kernel estimator of f_1, as proposed by Robin et al. (2007) and studied by Nguyen and Matias (2014). In our framework, this estimator can be written as follows: starting from an initial guess $\widehat{\Gamma}^{(t-1)}$ of Γ, let

$$\widehat{f}_1^{(t)}(x) = \sum_{i=1}^{m} \boldsymbol{\ell}_{i,1}(\widehat{\Gamma}^{(t-1)}) \frac{K((x-X_i)/h)}{h} \Big/ \sum_{i=1}^{m} \boldsymbol{\ell}_{i,1}(\widehat{\Gamma}^{(t-1)}), \tag{12}$$

where K is a kernel function and $\boldsymbol{\ell}_{i,q}(\Gamma) = \mathbb{P}_\Gamma(\theta_i = q \mid X)$, for $1 \le i \le m$ and $q \in \{0, 1\}$, denote the ℓ-values. We will typically use the Gaussian kernel $K(u) = (2\pi)^{-1/2} e^{-u^2/2}$, $u \in \mathbb{R}$. The rationale behind this is that weighting by $\boldsymbol{\ell}_{i,1}(\widehat{\Gamma}^{(t-1)})$ will put more weights to the observations X_i for which θ_i is likely to be equal to 1. Doing so, $\widehat{f}_1^{(t)}(x)$ is close to an ideal standard kernel estimator of the density of an i.i.d. sample that would be based on the observations X_i under the alternative only. This gives rise to Algorithm 1 detailed in supplementary material (Section S3).

3 Posterior post-selection confidence bounds

We first present the oracle bound, that is, the one using the true model parameter Γ. Then, we introduce several options for estimating this oracle, firstly based on a rough "plug-in" via the estimator $\widehat{\Gamma}$ and then, refinements based on bootstrap approaches. Also, we only present formulae for the upper bounds U_β for brevity. The corresponding lower bounds L_β are similar and quickly described in supplementary material (Section S4).

3.1 Oracle bound

In this section, we fix any non-empty selected set $R = S(X) = \{j_1, \ldots, j_s\} \subset \mathbb{N}_m$, for some indices $1 \le j_1 < \cdots < j_s \le m$ and $s = |R|$. While the case where R consists of contiguous indices is typical, we consider the more general situation where R is arbitrary. For notation brevity, we let $\theta_t^R = \theta_{j_t}$ for $t \in \{1, \ldots, s\}$.

According to (8), we let

$$U_\beta(X, R; \Gamma) = s^{-1} \min \left\{ n \in \{0, \ldots, m\} \,:\, \mathbb{P}_\Gamma \left(\sum_{t=1}^{s} (1 - \theta_t^R) \le n \,\middle|\, X \right) \ge 1 - \beta \right\}. \tag{13}$$

Proposition 2.1 ensures that, conditionally on X, $(\theta_t^R)_{1 \le t \le s}$ follows a heterogeneous Markov chain with initial probability $\boldsymbol{\ell}_{j_1,1}(\Gamma)$ and transition matrices

$$\Pi_t^R(\Gamma) = \prod_{i=j_{t-1}+1}^{j_t} \Pi_i(\Gamma), \quad t \in \{2, \ldots, s\}. \tag{14}$$

As such, $U_\beta(X, R; \Gamma)$ is not explicit. In the sequel, we provide an algorithm to compute $U_\beta(X, R; \Gamma)$. For this, let for $1 \leq k \leq s$, $0 \leq \ell \leq s$,

$$B_{k,\ell,0} = \mathbb{P}_\Gamma\left(\sum_{t=1}^{k}(1-\theta_t^R) \leq \ell,\ \theta_k^R = 0 \,\middle|\, X\right);$$
$$B_{k,\ell,1} = \mathbb{P}_\Gamma\left(\sum_{t=1}^{k}(1-\theta_t^R) \leq \ell,\ \theta_k^R = 1 \,\middle|\, X\right). \tag{15}$$

In words, $B_{k,\ell,0}$ is the posterior probability that there are at most ℓ zero-occurrences in $\theta_{1:k}^R$, with a zero in the last position. Similarly, $B_{k,\ell,1}$ is the posterior probability that there are at most ℓ zero-occurrences in $\theta_{1:k}^R$, with a one in the last position. Since $B_{s,\ell,0} + B_{s,\ell,1}$ is the posterior probability that at most ℓ zero-occurrences occurs in the whole sequence $\theta^R = \theta_{1:s}^R$, the following holds.

Proposition 3.1 *The quantity $U_\beta(X, R; \Gamma)$ defined by (13) can be computed as*

$$U_\beta(X, R; \Gamma) = s^{-1} \min\left\{n \in \{0, \ldots, m\} \ :\ B_{s,n,0} + B_{s,n,1} \geq 1 - \beta\right\}, \tag{16}$$

where $B_{k,\ell,0}$ and $B_{k,\ell,1}$ are defined by (15), respectively.

Algorithm 2 in supplementary material (Section S3) provides an explicit recursive computation of the quantities $B_{k,\ell,0}$ and $B_{k,\ell,1}$.

3.2 Plug-in bound

The first non-oracle bound that is proposed is simply obtained by plugging the estimator derived in Sect. 2.2 into the oracle bound, that is,

$$U_\beta^{\mathrm{PI}}(X, S(X)) := U_\beta(X, S(X); \widehat{\Gamma}), \tag{17}$$

where $\widehat{\Gamma}$ comes from Algorithm 1 and $U_\beta(X, R; \Gamma)$ is given by (13) and (16). Since the oracle bound is based on the conditional distribution of the latent variable given the observation, the above bound can be interpreted as an "empirical Bayes credible set" for the FDP.

Unfortunately, this plug-in bound can be anti-conservative, meaning that it can violate (4), as we will see in the simulations of Sect. 4. An intuitive explanation is that $U_\beta(X, S(X); \widehat{\Gamma})$ is expected to fluctuate on both sides of $U_\beta(X, S(X); \Gamma)$, and thus, the event $\mathrm{FDP}(\theta, S(X)) \leq U_\beta(X, S(X); \widehat{\Gamma})$ may not be true when $U_\beta(X, S(X); \widehat{\Gamma})$ is smaller than $U_\beta(X, S(X); \Gamma)$.

3.3 Bootstrap bounds

To correct for the additional statistical error due to parameter estimation, we propose different bounds based on a bootstrap approach. The supplementary material provides

full details for these bounds (rationale of the construction, comments) in Section S2 and implementation in Section S3 (Algorithm 3). In what follows, we only report a summary. Let $\delta \in [0, 1]$ be a parameter fixed by the user.

The first bootstrap bound splits β between the plug-in bound and the additional statistical error due to parameter estimation, which itself is estimated via bootstrap:

$$U^{\text{boot1}}_{\beta,\delta}(X, S(\cdot)) := U_{\beta(1-\delta)}(X, S(X); \widehat{\Gamma}) + \tilde{q}^{(B)}_{1,\beta\delta}(\beta(1-\delta), S(\cdot); \widehat{\Gamma}), \tag{18}$$

where $\tilde{q}^{(B)}_{1,\beta\delta}(\beta(1-\delta), S(\cdot); \widehat{\Gamma})$ is a Monte Carlo estimator of $q_{1,\beta\delta}(\beta(1-\delta), S(\cdot); \widehat{\Gamma})$:

$$q_{1,\gamma}(\beta, S(\cdot); \Gamma) := \min\left\{x \in \mathbb{R} \,:\, \mathbb{P}_\Gamma\left(U_\beta(X, S(X); \Gamma) - U_\beta(X, S(X); \widehat{\Gamma}) \le x\right) \ge 1-\gamma\right\}.$$

The second bootstrap bound is a heuristic approximation of the first one, for which only $\widehat{\Gamma}^*$ is recomputed from each bootstrap sample, while the selection set $S(X)$ is kept fixed. It mimics a situation where parameter estimation and region selection would be performed on independent data:

$$U^{\text{boot2}}_{\beta,\delta}(X, S(X)) := U_{\beta(1-\delta)}(X, S(X); \widehat{\Gamma}) + \tilde{q}^{(B)}_{2,\beta\delta}(\beta(1-\delta), S(X); \widehat{\Gamma}), \tag{19}$$

where $\tilde{q}^{(B)}_{2,\beta\delta}(\beta(1-\delta), S(X); \widehat{\Gamma})$ is a Monte Carlo estimator of $q_{2,\beta\delta}(\beta(1-\delta), S(X); \widehat{\Gamma})$:

$$\begin{aligned} &q_{2,\gamma}(\beta, S(X); \Gamma) \\ &:= \min\left\{x \in \mathbb{R} \,:\, \mathbb{P}_{Y\sim P_\Gamma}\left(U_{\beta(1-\delta)}(X, S(X); \Gamma) - U_{\beta(1-\delta)}(X, S(X); \widehat{\Gamma}(Y)) \right.\right. \\ &\qquad \left.\left. \le x \,\middle|\, X\right) \ge 1-\gamma\right\}. \end{aligned}$$

The third bootstrap bound is based on a different idea. First, we consider the naive bound

$$U^{\text{naive}}_{\beta}(S(\cdot)) := \tilde{q}^{(B)}_{\text{naive},\beta}(S(\cdot); \widehat{\Gamma}) \tag{20}$$

where $\tilde{q}^{(B)}_{\text{naive},\beta}(S(\cdot); \widehat{\Gamma})$ is as a Monte Carlo bootstrap estimator of

$$q_{\text{naive},\beta}(S(\cdot); \Gamma) := \min\left\{x \in \mathbb{R} \,:\, \mathbb{P}_\Gamma\left(\text{FDP}(\theta, S(X)) \le x\right) \ge 1-\beta\right\}.$$

The latter accounts for the (unconditional) variations of the underlying true FDP. As demonstrated by the numerical experiments in Sect. 4, using U^{naive}_{β} is generally too conservative. This can be improved by a proper re-centering by the plug-in bound $U_\beta(X, S(X); \widehat{\Gamma})$ given in (17) that acts like a stabilization. This is done in the third bootstrap bound which corrects the plug-in bound by a bootstrap quantile between the plug-in bound and the true value of the FDP:

$$U^{\text{boot3}}_{\beta}(X, S(\cdot)) := U_\beta(X, S(X); \widehat{\Gamma}) + \tilde{q}^{(B)}_{3,\beta}(\beta, S(\cdot); \widehat{\Gamma}), \tag{21}$$

where $\tilde{q}^{(B)}_{3,\beta}(\beta, S(\cdot); \widehat{\Gamma})$ is a Monte Carlo estimator of $q_{3,\beta}(\beta, S(\cdot); \widehat{\Gamma})$:

$$q_{3,\beta}(\beta, S(\cdot); \Gamma) := \min \Big\{ x \in \mathbb{R} \ : \ \mathbb{P}_{(\theta,X)\sim P_\Gamma} \left(\mathrm{FDP}(\theta, S(X)) - U_\beta(X, S(X); \widehat{\Gamma}(X)) \right. \\ \left. \leq x \right) \geq 1 - \beta \Big\}.$$

Note that, if the plug-in bound was perfect, the latter quantile should be 0. Hence, this quantile accounts for the errors done by the plug-in bound. A notable advantage of this bound compared to the previous ones is that splitting β into $\beta\delta$ and $\beta(1-\delta)$ is not required.

4 Numerical experiments

This section summarizes numerical experiments performed in order to assess the quality of the bounds introduced in Sect. 3.

4.1 Setting

4.1.1 Post-selection bounds

We have performed numerical experiments using the proposed bounds (Table 1), as well as structure-agnostic bounds Goeman and Solari (2011), Blanchard et al. (2020), Katsevich and Ramdas (2020). Recall

$$U^{\mathrm{Simes}}_\gamma(X, R) = \min_{1\leq k\leq m} \left\{ \sum_{i\in R} \mathbb{1}\{p_i > \gamma k/m\} + k - 1 \right\}$$

$$U^{\mathrm{KR}}_\gamma(X, R) = \min_{1\leq k\leq m} \left\{ \sum_{i\in R} \mathbb{1}\{p_i > p_{(k)}\} + \left\lceil \frac{\log(1/\gamma)}{\log(1+\log(1/\gamma))}(1 + m p_{(k)}) \right\rceil - 1 \right\},$$

where the p-values p_i are based either on f_0 (if unknown, we use $\hat{f}_0$ see Sect. 4.3.3). However, as illustrated by Fig. 2, the latter are by construction far more conservative than our proposed bounds that take into account the latent structure of the model.

To ensure that the bootstrap bounds are never less conservative than the Plug-in, these bounds have been slightly modified and $\tilde{q}_\beta$ has been replaced by its positive part $\tilde{q}^+_\beta$. This modification generally has no effect on the upper bounds in practice, but does affect the corresponding lower bounds, for which we have observed a tendency of the bootstrap toward overcompensation.

Throughout this section, the target risk level is set to $\beta = 0.1$. Therefore, our proposed upper bounds are expected to satisfy $\mathbb{P}(U_\beta(X, S(X); \Gamma) < \mathrm{FDP}(\theta, S(X))) \leq 0.1$ and the corresponding lower bounds are expected to satisfy $\mathbb{P}(L_\beta(X, S(X); \Gamma) > \mathrm{FDP}(\theta, S(X))) \leq 0.1$.

Table 1 (Upper) bounds considered in the numerical experiments

Name	Definition	References
Oracle	$U_\beta(S(X), \Gamma)$	(16)
Plug-in	$U_\beta(S(X), \widehat{\Gamma})$	(17)
Boot1	$U^{\text{boot1}}_{\beta,\delta}(S(\cdot), \widehat{\Gamma}), \delta \in \{0.1; 0.5; 0.9\}$	(18)
Boot2	$U^{\text{boot2}}_{\beta,\delta}(S(X), \widehat{\Gamma}), \delta = 0.5$	(19)
Naive	$U^{\text{naive}}_{\beta}(S(\cdot))$	(20)
Boot3	$U^{\text{boot3}}_{\beta}(S(X), \widehat{\Gamma})$	(21)

4.1.2 Selection policies

We consider three different selection policies $S(\cdot)$, labeled as follows in the figures:

- "$S(X) = \{p_i < 0.05\}$": the p-value level set associated with the threshold 0.05.
- "$S(X) = SC(0.05)$": items selected by the procedure introduced by Sun and Cai (2009) in the HMM model (see Sect. 1.5) to control FDR at level 0.05 using the parameter estimators $\widehat{\Gamma}$ of Sect. 2.2.
- "$S(X) =$ Viterbi": items i selected by the Viterbi algorithm, that is, such that $\widehat{\theta}_i = 1$, where $\widehat{\theta}$ is the estimation of θ using the Viterbi algorithm with the parameter estimators $\widehat{\Gamma}$ of Sect. 2.2.

Section 4.2 illustrates the behavior of the considered bounds in a setting where our assumptions are met. The robustness of the method is then studied in Sect. 4.3, where we report the results of numerical experiments in the case where the HMM is not or poorly identifiable, or where the selected set $S(X)$ depends on prior knowledge not included in X. Finally, in Sect. 4.4 we report the results of further numerical experiments on DNA copy number data from a cancer study with known ground truth. Another application to the influenza-like illness is provided in supplementary material (Section S8).

4.2 Results in a typical within-model scenario

The Markov chain $(\theta_i)_{1 \le i \le m}$ is generated from transition matrix: $A = \begin{pmatrix} 0.95 & 0.05 \\ 0.2 & 0.8 \end{pmatrix}$, with the stationary distribution (0.8, 0.2) as initial state. The X variables are generated such that $X_i|\theta_i = 0 \sim \mathcal{N}(0, 1)$ and $X_i|\theta_i = 1 \sim P_1$. Here, $P_1 = \mathcal{N}(3, 1)$. Here, we generate $m = 3200$ variables. The experimental setting is the same as in Fig. 2, where it is patent that the KR, BNR and Simes methods (which are agnostic with respect to the HMM structure and do not exploit it) underperform. For this reason, we did not include them in the more detailed Fig. 3 aimed at a more fined-grained comparison.

Figure 3A displays the value of the difference Δ between the upper bound $U(\cdot)$ and the true FDP value $\text{FDP}(\theta, S(X))$ for each upper bound (in rows) and selection policy (in columns), for 300 simulation runs. The empirical violation probability of the bound, that is, the proportion of simulation runs for which a given bound is lower

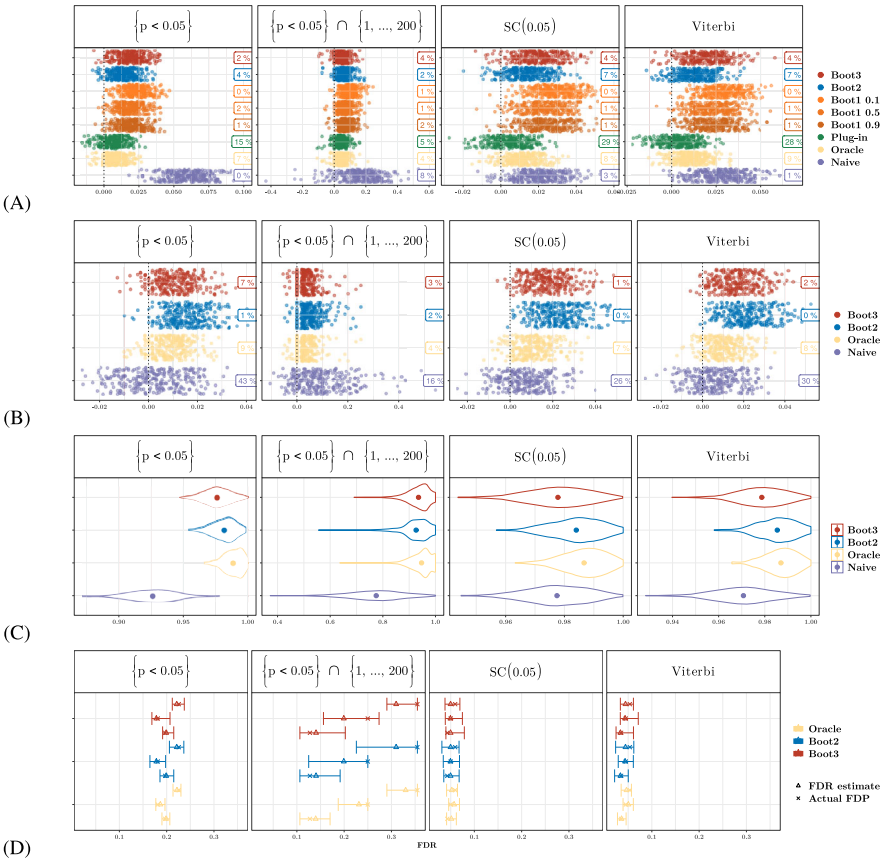

Fig. 3 **A**–**C** Results for 300 simulation runs with the model parameters described in Sect. 4.2 ($m = 3200$). **A** Difference $\Delta = U(\cdot) - \mathrm{FDP}(\theta, S(X))$ for each upper bound (rows) and each selection policy $S(X)$ (columns), with the empirical violation probability of the bound displayed within rectangles (Target risk $\beta = 10\%$). **B** Difference $\Delta' = \mathrm{FDP}(\theta, S(X)) - L(\cdot)$ for the corresponding lower bounds. **C** Power of the different bounds. **D** Realizations of 80% FDP post-selection intervals for three illustrative simulation runs

than the true FDP (i.e., $\Delta < 0$), is displayed within rectangles. This proportion is expected to be lower than $\beta = 10\%$.

As stated earlier, the Plug-in bound is an estimator of the oracle bound and not an upper bound of it; hence, it is not surprising that its empirical violation probability exceeds the target. All of the other bounds have empirical violation probability below the target in this setting. This is expected for the bootstrap-corrected bounds Boot1, Boot2, and Boot3. The fact that the naive bound is also below the target risk was not guaranteed, because (similarly to $U(\beta, S(X); \widehat{\Gamma})$) this bound is only an estimator of $U(\beta, S(X); \Gamma)$. We observed that the impact of the choice of δ in Boot1 seems moderate. Nevertheless, Boot3 seems always less conservative than Boot1, which could be explained by the fact that Boot3 does not use such split of the confidence budget. Overall, neither of Boot2 or Boot3 uniformly dominates the other one, which is well expected because they come from two different bootstrap strategies. For the sake

of readability of the illustrations, Boot1, which is too conservative, and the Plug-in bound, which is anti-conservative, will not be displayed in the remainder of the paper.

Figure 3B displays, for the corresponding lower bound the value of the difference Δ' between the FDP and the lower bound $L(\cdot)$. Similarly to (A), valid lower bounds are expected to be above the true FDP (i.e., $\Delta' < 0$) in less than 10% of the simulations runs. This is the case for all the proposed bounds, expect for the naive one. Figure 3C displays the power of the different bounds, which is defined consistently with Blanchard et al. (2020):

$$\text{Power} = \mathbb{E}\left(\frac{|S(X)| - U(X)}{|S(X) \cap \mathcal{H}_1|} \,\middle|\, |S(X) \cap \mathcal{H}_1| \neq 0, |S(X) \cap \mathcal{H}_0| \leq U(X)\right) \quad (22)$$

In this setting, the bounds are almost as powerful as the oracle. Finally, in order to emphasize the interest of FDP post-selection interval compared to pointwise FDR estimate $\widehat{\text{FDR}}(S(X), \widehat{\Gamma})$, see (10), we have displayed both in Fig. 3D for three arbitrarily chosen simulation runs. The FDP post-selection intervals are clearly more informative than the corresponding point estimates as they quantify its uncertainty, which can be widely different according to the scenario: this is reflected in the interval lengths. In two of the three displayed simulation runs, the estimated FDR is quite far from the true value, whereas the true value still lies in the post-selection interval.

4.3 Challenging model assumptions

This section briefly summarizes further numerical experiments carried out in order to test the robustness of the proposed bounds either to violations of the model assumptions, or to departures from a mild scenario toward more challenging settings. The corresponding illustrations are postponed to supplementary material (Section S7).

4.3.1 Invalid selection policies

One of the assumptions of the method is that the selected set $S(X)$ cannot depend on an additional prior knowledge not included in the observation X. In terms of the modeling via the HMM, violating this assumption would correspond to a selection policy having access to "insider information" about the latent configuration vector θ under one form or another. To assess how important this constraint is for the validity of the method, we have considered three selection policies that include full knowledge of $\mathcal{H}_0$ (this of course a very unrealistic situation, only considered here to illustrate the point). The results are displayed in Figure S1 and show that the oracle bound (as well as the second bootstrap bound which tries to mimic the oracle) is too liberal in this case. The third bootstrap bound, which corrects the plug-in bound to try to match the FDP, respects the risk. However, it requires knowledge of the full selection policy $S(.)$. Depending on the situation, this might be realistic (e.g., if the additional information stems from an ancillary statistic that can also be simulated) or not (e.g., if the additional information comes from some vaguely defined "insider expert knowledge"), see also Sect. 6.5 for a discussion.

4.3.2 Identifiability issues

When the latent variables are independent, the model is not identifiable (Alexandrovich et al. 2016; Gassiat et al. 2016). When the latent variables are close to independence (that is, when $\det(A)$ gets close to 0), the model is close to singular. The bounds obtained for transition matrices A with various values of $\det(A)$ are displayed in Figure S2. As expected, the bounds are too liberal in the independent case (that is, $\det(A) = 0$). However, they appear to be valid even for small determinants.

4.3.3 Unknown f_0

Finally, we have also computed the bounds in a case where f_0 is unknown. We have considered two options for initializing the estimation of f_0 in the corresponding EM-type algorithm, see Algorithm 4: (i) using the true f_0, and (ii) estimating f_0 using local FDR algorithm (Efron 2004). In both cases, the proposed bounds remain below the target risk, as illustrated in Figure S3.

4.3.4 Small number of hypotheses

Our results are based on the estimation of A and f_1. When the number of hypothesis is small, this estimation may not be accurate, and our estimator may not fulfill the guarantee. In Figure S4, we study the behavior of our estimators for different values of m from 50 to 5000. We can see that both Boot1 and Boot2 are valid for $m \geq 500$.

4.3.5 Stationarity

To challenge the stationarity hypothesis of HMM setting, we simulate a dataset in the same way as in Sect. 4.2 but forcing $\theta_1 = 1$. The results are displayed in Figure S5; both Boot2 and Boot3 are still valid in this scenario.

4.4 Semi-simulated data based on DNA copy numbers

In order to test the robustness of our methodology, we have considered a more realistic scenario, which does not rely on a probabilistic simulation model. Using the R package `jointseg` (Pierre-Jean et al. 2019), we have generated synthetic copy number profiles as proposed (and further described) in Pierre-Jean et al. (2015): given a number m of loci and a number K of regions, we draw uniformly $K - 1$ breakpoint positions, thus defining K regions. Then, we draw K region labels, corresponding to the number of DNA copies for each parent (a.k.a. parental CN). Finally, for a region of size m_k, we draw m_k samples by resampling from a real DNA copy number data corresponding to this type of region. These data are available in the R package `acnr` (Pierre-Jean and Neuvial 2017), which contain annotated CN profiles from several cancer data sets. Importantly, these data sets correspond to dilution series where tumor and normal cells are mixed in known proportion. Therefore, signal to noise ratio of the corresponding CN profile is implicitly controlled by the fraction of tumor cells.

We proceed as follows to obtain two groups of samples with known differential regions. First, we generate $n_1 = 50$ CN profiles as described in the preceding paragraph, with the same regions (here, 10 regions). Then, we generate $n_2 = 50$ samples from the same regions, but modify the label of two regions in such a way that this group of samples only has 8 different regions instead of 10. In this setting, $\theta_i = 1$ if the position i belongs to one of the two modified regions and $\theta_i = 0$ otherwise. We then compute Wilcoxon statistics and scale them using their limit law. More precisely, to compare a group 1 of size n_1 to a group 2 of size n_2 at position i we use

$$X_i = \left(W_i - \frac{n_1 n_2}{2}\right) \Bigg/ \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}, \tag{23}$$

where W_i is the classical Wilcoxon statistic, the sum of the ranks of group 1. This process is illustrated by Figure S6 in the supplementary material. In this particular example, the proportion of tumor cells has been set to 70%, corresponding to a moderate to high signal to noise ratio. By construction of these data sets, the distribution of the test statistics is the same for all the genomic regions for which $\theta = 0$, but may differ across regions for which $\theta = 1$, as illustrated by the bottom plot of Figure S7 in the supplementary material. Since the distribution of $X_i|\theta$ depends on the underlying true copy number regions, the assumption of conditional independence of the HMM introduced in Sect. 2.1 is not fulfilled here. As such, the numerical experiments reported in this section make it possible to assess the robustness of the proposed bounds to some violation of this assumption.

The results of these numerical experiments are summarized in Fig. 4, where the difference Δ between the FDP upper bound and the true FDP is displayed for each of 100 simulation runs. The lines in this figure correspond to increasing values for the SNR (as governed by the fraction of tumor cells), while the columns correspond to the selection policies introduced at the beginning of Sect. 4.

All the bounds appear to be valid in this settings, since their empirical violation probability is less than the target level $\beta = 10\%$ (i.e., $\Delta < 0$ in less than 10% of the simulation runs). Overall, the bounds are quite conservative in all settings, with empirical violation probabilities often closer to 0 than 10%. This conservativeness is partly explained by the fact that for high SNR values, the problem is so easy that the upper bounds often match the true FDP exactly, as illustrated by the presence of a mode at 0. The Simes and KR bounds are the most conservative, with null empirical violation probability in all but one scenario.

As SNR decreases, we observe for all bounds a shift of the empirical distribution of Δ toward positive values. The Simes bounds are highly conservative already at tumor fraction 0.7, and Δ even concentrates at 1 for three of the four selection policies at tumor fractions 0.3 and 0.5: this corresponds to a true FDP close to 0 and an upper bound close to 1, meaning that the Simes bound is not informative at all in these settings. The KR bound is marginally closer to the true value of the FDP but is not really informative either for the scenario with a low tumor fraction (0.5 and 0.3). The first bootstrap bound has the same tendency, but to a lesser extent. The second and third bootstrap bounds show a remarkable behavior, with Δ remaining very close to 0 for the

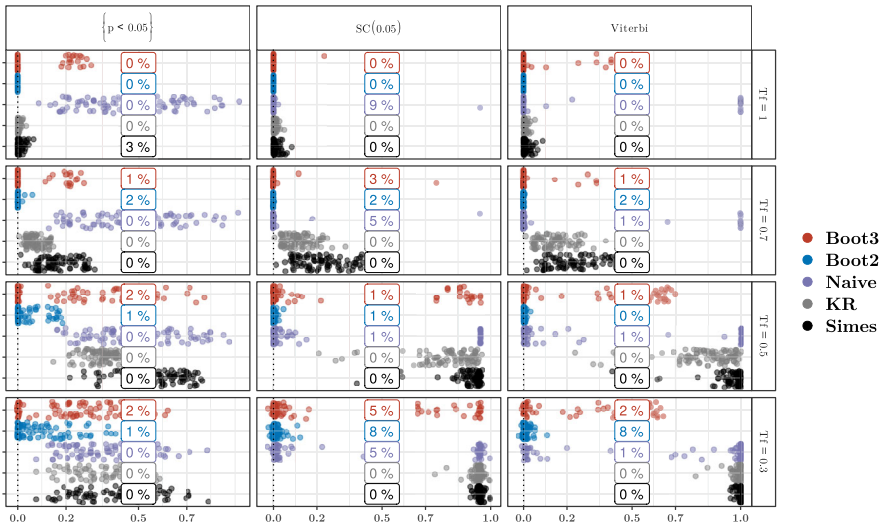

Fig. 4 Summary of 100 numerical experiments with semi-simulated CN data. Each dot corresponds to a realization of the difference $\Delta = U(\cdot) - \mathrm{FDP}(\theta, S(X))$ for each upper bound (colors) and each selection policy $S(X)$ (columns) for different values of the SNR (rows). The empirical violation probability of the bound is displayed within a rectangle

vast majority of simulation runs for all selection policies, even for small SNR values. In addition, Boot2 and Boot3 have also a high power (22), as shown in Figure S8 (supplementary material Section S7.7). These results strengthen our confidence in the applicability of our methods and suggest that even in practical applications, our proposed bounds (and especially the second and third bootstrap bounds) will be able to evaluate the amount of true signal after selection with accuracy and correct coverage.

5 Application to differential copy number analysis

In this section, we propose an application of our approach to Copy Number Alterations Associated with Endometriosis in Ovarian Clear Cell Adenocarcinoma. Our proposed bounds are applied to a study of 117 ovarian cancer patients briefly mentioned in the introduction (Okamoto et al. 2015). We focus on the part of the study that aims at comparing, for 13,239 loci located on chromosome 7, the DNA copy numbers of 54 patients with endometriosis (a common gynecological disorder characterized by ectopic growth of endometrial glands and stroma) to that of 63 patients without endometriosis. For such CNA data, HMM modeling is one of the recommended approaches, see, e.g., Fridlyand et al. (2004), Shah et al. (2009), Zhang (2010), Luo (2019). As in the semi-simulated setting described in Sect. 4.4, we compare the two groups using Wilcoxon tests and scale the statistics using (23). The obtained statistics for the different positions of the genome are displayed in the left part of Fig. 5A for the full chromosome 7 and (B) for the "short arm" of chromosome 7 (corresponding to the first 4,799 loci).

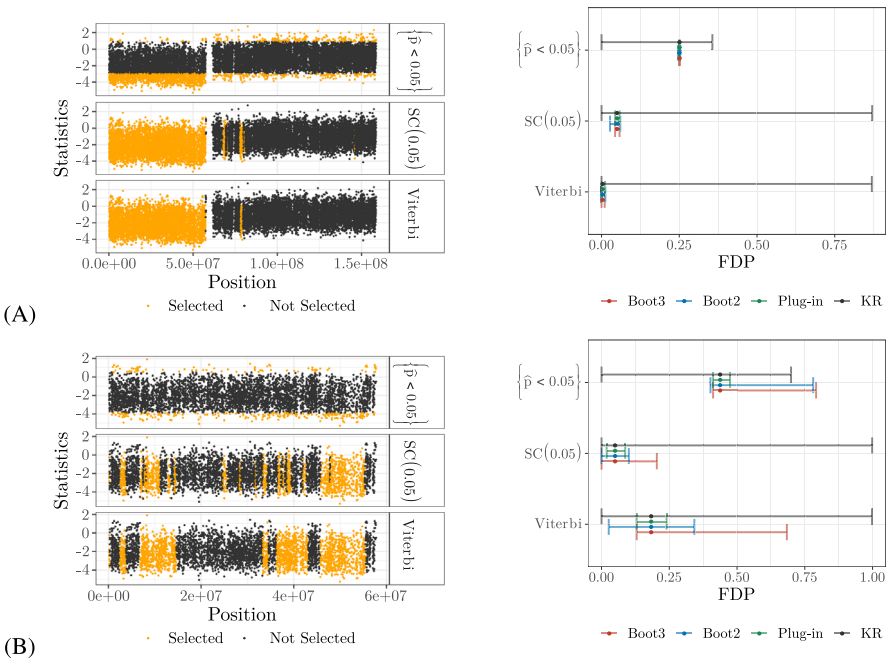

Fig. 5 Left: Scaled Wilcoxon test statistic X_i (see (23)) comparing the copy numbers between patients with and without endometriosis at different loci on chromosome 7 for (**A**) and 7p for (**B**). Right: 80% FDP post-selection interval different selection policies. Since f_0 is unknown, empirical p-values defined in supplementary material (Section S7.3) are used

A first important observation is that the test statistics are not centered at 0 but globally shifted toward negative values across the entire chromosome 7, and more prominently so in chromosome 7p. This indicates that patients with endometriosis have on average a larger CN than the others, which may be due to an increase prevalence of trisomy 7 in these patients, or to a larger proportion of tumor cells in the biological samples corresponding to these patients. Here, our goal is not to detect such macroscopic changes but to pinpoint chromosomal regions that deviate from the rest of the chromosome. Indeed, such regions could indicate the presence of "driver" genes more actively or more early implicated in the tumor process compared to other "passenger" regions.

Therefore, we chose to estimate f_0 from the data, as described in supplementary material (Section S5). Note that $\widehat{f_0}$ will not estimate the distribution of the null hypothesis "the two groups have the same number of DNA copy at this position," but the distribution of the most frequent type of difference between the two groups. For instance, if all patients without endometriosis have two copies of chromosome 7 (not affected by their ovarian cancer) and patients with endometriosis have a third copy of chromosome 7 and in some rare position they have a fourth copy, the estimated f_0 will be the law of the statistics comparing copy numbers two and three. Because of the marked shift between the short (left) and long (right) arms of chromosome 7, the estimation of f_0 is quite different depending on whether the entire chromosome

or only one arm is analyzed. Therefore, we have carried out an analysis of the entire chromosome (Fig. 5A) and an analysis of its short arm, chromosome 7p (Fig. 5B):

Full chromosome 7: By construction, the selection policies that take into account the position mostly select regions on chromosome 7p. The post-selection intervals corresponding to each selection policy are represented in the right panel of Fig. 5A. In this scenario, the post-selection intervals are very tight around the pointwise FDR estimate, which reflects the fact that $\hat{f}_0$ and $\hat{f}_1$ are quite distinct here.

Chromosome 7p: When focusing on chromosome 7p, $\hat{f}_0$ and $\hat{f}_1$ are much closer to each other, resulting in wider FDP post-selection intervals than for the full chromosome. For instance, in the sets SC(0.05) the FDR is controlled at 5% but the upper bounds goes up to 20%. Even the plug-in bound, which has been shown in the simulation not to be conservative enough, goes up to 10%.

These examples emphasize the added value of post-selection intervals on the FDP over FDR point estimates, since a given value of the point estimate will be interpreted very differently depending on the width of the interval.

Also note that other structure-agnostic bounds by Goeman and Solari (2011) ("Simes") and Katsevich and Ramdas (2020) ("KR") (plugging in estimated p-values and ignoring the fluctuations linked to this estimation) are outperformed by the bounds considered here, i.e., return larger confidence intervals, so are omitted.

Remark 5.1 As discussed in Sects. 4.3.2 and 6.2, it is appropriate to validate empirically our bounds by ensuring that the determinant of the estimated transition matrix is not too close to zero. In the two considered scenarios above, the determinants are both close to one (0.99 and 0.98, respectively). Moreover, the number of hypotheses is large. Hence, we are in a situation where we expect our estimators to be sharp enough and thus, our bound to be valid.

6 Discussion

6.1 Choosing among the bootstrap approaches

In practice, we endorse the bootstrap approaches over the plug-in approach. Here, we discuss how to choose between the three bootstrap options Boot 1, 2 or 3. Given Remark 1.1 and the numerical experiments of Sect. 4, we propose the following guidance: if only $S(X)$ is available and not the whole policy $S(\cdot)$, Boot 2 is the only usable method. Otherwise, choose Boot 3 when the model fits well (Boot 3 more powerful than Boot 2) or Boot 2 if the user has some doubt on the validity of the model (Boot 2 more robust than Boot 3). We discard Boot 1 because it has been shown to be too conservative.

6.2 Detecting close to singular scenarios

The HMM model becomes singular in the situation where $f_0 = f_1$, or when the transition matrix A is of rank 1 (i.e., the coordinates of the configuration vector θ are

drawn i.i.d. from a Bernoulli distribution), in which case the model is not identifiable: the same data distribution for the observable X can be obtained for several sets of parameters, in particular different joint distributions of (θ, X) so that the notion of ground truth for the FDP is questionable.

The behavior of our approach in a situation close to singular is discussed in Sect. 4.3; in a truly singular situation, almost surely the estimator $\widehat{A}$ is not of rank 1, so that the estimation error (estimated at $\widehat{\Gamma}$) will surely be underestimated by the bootstrap procedure. It would be therefore in principle necessary to have a test for the singular case and stop the procedure if that test is not rejected (using an agnostic multiple testing procedure would then be more appropriate; see also the discussion on null estimation in the independent case, Sect. 1.6). While we do not cover the precise design of such a test in the present work, a suitable heuristic procedure in practice can be to monitor if $|\det \widehat{A}|$ is too close to 0, or $\widehat{f_1}$ is too close to f_0.

6.3 Asymptotic consistency of plug-in

An important theoretical insight for FDR control methods using empirical Bayes-type approaches is that asymptotic consistency is usually granted for plug-in methods (Sun and Cai 2009). In the case of nonparametric HMMs, consistent estimation with quantitative guarantees for the HMM parameters and the local FDR have been obtained recently, see the discussion in Sect. 1.6. In the setting considered here, since we are considering in principle arbitrary policies and therefore, selection sets $S(X)$, it seems we would need a stronger convergence property for parameter estimation, e.g., in the sense of total variation distance for the full joint distribution of (θ, X). If such a convergence holds, we can guarantee the asymptotic of the plug-in approach for FDP bounds (see supplementary material Section S6.)

However, obtaining such a consistency result for convergence of the full estimated joint distribution seems a tall order, since the asymptotics would be in the increasing size of the observation space, which will also make TV distance larger. (It might be more plausible for theory to assume that the parameter is estimated using an observation sequence of longer size than the one it is used to perform multiple testing on.) Furthermore, our experiments show that the simple plug-in approach is generally unsatisfactory in practice and that the proposed bootstrap-based bounds are more appropriate.

6.4 Validity of the bootstrap

The bootstrap used here is a type of 'smoothed bootstrap,' see Hall et al. (1989). In contrast with Efron's original bootstrap, which uses the empirical distribution, the generating distribution for the smoothed bootstrap samples comes from a (nonparametric) estimator. The properties of the smoothed bootstrap for dependent processes and time series have been only scarcely studied in the literature, and we mention few relevant references: in an autoregressive model, a consistency result for bootstrap confidence bands has been obtained in Franke et al. (2002), for which the bootstrap sample is generated from the kernel estimate of the autoregression function. Maybe closer to our

framework, the smoothed bootstrap has been used for a Markov process over a continuous space in Horowitz (2003), for which the bootstrap sample is generated from the nonparametric estimator of the transition density. The latter work establishes the consistency of the smoothed bootstrap for the mean of the stationary distribution marginal. It is also argued that the smoothed bootstrap for time series enjoys better properties than the "block resampling" bootstrap approach (which closer is in spirit to Efron's bootstrap for i.i.d. data). Obtaining comparable theoretical results in our context is significantly more challenging: first, the HMM uses an additional hidden layer which should be considered. Second, since the selection policy may depend on the whole observation sequence, the quantity of interest is a function of the whole joint distribution and not only the mean or some other regular function of the stationary marginal. Thus, the whole sequence should be taken into account in the bootstrap approximation. Certainly, deriving consistency for the coverage of our bootstrap bounds is a key point, but we believe it deserves an entire devoted study and is left for future investigations.

6.5 Bayesian versus frequentist selective inference

The posterior (oracle) bound (6) can accommodate absolutely any selection policy $S(\cdot)$. Thus, it seems that the bound holds in fact for any selected set $R = S(X)$. This can appear surprising at first, since in a frequentist setting, uniform agnostic post hoc bounds (holding for all subsets R and over all latent parameter configurations θ) generally have a price for complexity (e.g., a union bound over a reference family of candidate rejection sets), while the Bayesian approach seems to offer a free lunch in that regard. A key to understand this apparent conundrum is to insist that

1. strictly speaking, the guarantees only hold if the assumed structural model (prior) on the latent variable θ is correct. Hence, "the Bayesian solution is perfectly sensible so long as the prior is taken seriously", to quote Dawid (1994), see also Senn (2008), Efron (2011), in relation to apparent paradoxes of Bayesian post-selection inference.
2. the posterior bound (6) is *not* a uniform statement over all possible rejection sets, but should be interpreted as conditioning with respect to X and for an arbitrary but given selection policy $S(\cdot)$ that must *only depend on the observation* X.

The latter point excludes to use any form of "insider" or "leaked" information on the latent variable configuration to be used for to determine the selected set, such as ancillary statistics conveying additional information or "expert knowledge" that would not be incorporated in the prior. As discussed in Sect. 4.3, violation of these model assumptions can result in the oracle bound to be incorrect. From a Bayesian point of view, the prior structural distribution assumption on θ should reflect the entirety of the available prior information and not be used merely as a convenient default.

Connected to this, it would be interesting to study whether "frequentist Bayes" approaches based on posterior concentration to the true latent parameter (see, e.g., Castillo and Roquain 2020, for a recent review) would be applicable here. While recent progress has been achieved to analyze the FDR from this perspective (Castillo and

Roquain 2020; Abraham et al. 2021a, b), the same question relative to post-selection bounds has not been explored yet up to our knowledge.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11749-023-00886-7.

Acknowledgements The authors would like to thank an associate editor and the two referees, whose insightful comments led to considerable improvements to this paper. This work has been supported by ANR-16-CE40-0019 (SansSouci), ANR-17-CE40-0001 (BASICS), ANR-19-CHIA-0021-01 (BiSCottE), ANR-21-CE23-0035 (ASCAI), the UPSaclay Excellency Chair REC-2019-044, the DFG CRC 1294 - 318763901 'Data Assimilation', and by the GDR ISIS through the "projets exploratoires" program (project TASTY).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abraham K, Castillo I, Gassiat E (2021a) Multiple testing in nonparametric hidden Markov models: an empirical Bayes approach. arXiv:2101.03838

Abraham K, Castillo I, Roquain E (2021b) Empirical Bayes cumulative ℓ-value multiple testing procedure for sparse sequences

Albertson DG, Collins C, McCormick F, Gray JW (2003) Chromosome aberrations in solid tumors. Nat Genet 34:369–376

Alexandrovich G, Holzmann H, Leister A (2016) Nonparametric identification and maximum likelihood estimation for hidden Markov models. Biometrika 103:423–434

Azriel D, Schwartzman A (2015) The empirical distribution of a large number of correlated normal variables. J Am Stat Assoc 110:1217–1228. https://doi.org/10.1080/01621459.2014.958156

Bachoc F, Blanchard G, Neuvial P (2018) On the post selection inference constant under restricted isometry properties. Electron J Stat 12:3736–3757. https://doi.org/10.1214/18-EJS1490

Bachoc F, Leeb H, Pötscher BM (2019) Valid confidence intervals for post-model-selection predictors. Ann Stat 47:1475–1504. https://doi.org/10.1214/18-AOS1721

Benjamini Y, Bogomolov M (2014) Selective inference on multiple families of hypotheses. J R Stat Soc Ser B (Stat Methodol) 76:297–318

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

Benjamini Y, Yekutieli D (2005) False discovery rate-adjusted multiple confidence intervals for selected parameters. J Am Stat Assoc 100:71–81

Berk R, Brown L, Buja A, Zhang K, Zhao L (2013) Valid post-selection inference. Ann Stat 41:802–837. https://doi.org/10.1214/12-AOS1077

Blanchard G, Neuvial P, Roquain E (2020) Post hoc confidence bounds on false positives using reference families. Ann Stat 48:1281–1303. https://doi.org/10.1214/19-AOS1847

Cai TT, Jin J (2010) Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. Ann Stat 38:100–145. https://doi.org/10.1214/09-AOS696

Cai TT, Sun W (2009) Simultaneous testing of grouped hypotheses: finding needles in multiple haystacks. J Am Stat Assoc 104:1467–1481. https://doi.org/10.1198/jasa.2009.tm08415

Cai TT, Sun W, Wang W (2019) Covariate-assisted ranking and screening for large-scale two-sample inference. J R Stat Soc Ser B (Stat Methodol) 81:187–234. https://doi.org/10.1111/rssb.12304

Cappé O, Moulines E, Rydén T (2006) Inference in hidden Markov models. Springer, Berlin

Castillo I, Roquain E (2020) On spike and slab empirical Bayes multiple testing. Ann Stat 48:2548–2574

Dawid AP (1994) Selection paradoxes of Bayesian inference. Lect Notes Monogr Ser 24:211–220

De Castro Y, Gassiat E, Le Corff S (2017) Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. IEEE Trans Inf Theory 63:4758–4777

Durand G, Blanchard G, Neuvial P, Roquain E (2020) Post hoc false positive control for structured hypotheses. Scand J Stat 47:1114–1148. https://doi.org/10.1111/sjos.12453
Efron B (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Am Stat Assoc 99:96–104. https://doi.org/10.1198/016214504000000089
Efron B (2007) Doing thousands of hypothesis tests at the same time. Metron Int J Stat LXV:3–21
Efron B (2008) Microarrays, empirical Bayes and the two-groups model. Stat Sci 23:1–22. https://doi.org/10.1214/07-STS236
Efron B (2009) Empirical Bayes estimates for large-scale prediction problems. J Am Stat Assoc 104:1015–1028. https://doi.org/10.1198/jasa.2009.tm08523
Efron B (2011) Tweedie's formula and selection bias. J Am Stat Assoc 106:1602–1614
Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96:1151–1160
Fan J, Han X (2017) Estimation of the false discovery proportion with unknown dependence. J R Stat Soc Ser B (Stat Methodol) 79:1143–1164
Fan J, Ke Y, Sun Q, Zhou W-X (2019) Farmtest: factor-adjusted robust multiple testing with approximate false discovery control. J Am Stat Assoc 1–29
Franke J, Kreiss J-P, Mammen E, Neumann MH (2002) Properties of the nonparametric autoregressive bootstrap. J Time Ser Anal 23:555–585
Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN (2004) Hidden Markov models approach to the analysis of array CGH data. J Multivar Anal 90:132–153
Friguet C, Kloareg M, Causeur D (2009) A factor model approach to multiple testing under dependence. J Am Stat Assoc 104:1406–1415
Gales M, Young S (2008) The application of hidden Markov models in speech recognition. Now Publishers Inc, Hanover
Gassiat É, Cleynen A, Robin S (2016) Inference in finite state space non parametric hidden Markov models and applications. Stat Comput 26:61–71
Genovese CR, Wasserman L (2006) Exceedance control of the false discovery proportion. J Am Stat Assoc 101:1408–1417
Goeman JJ, Solari A (2011) Multiple testing for exploratory research. Stat Sci 26:584–597. https://doi.org/10.1214/11-STS356
Hall P, DiCiccio TJ, Romano JP (1989) On smoothing and the bootstrap. Ann Stat 17:692–704
Heller R, Rosset S (2021) Optimal control of false discovery criteria in the two-group model. J R Stat Soc Ser B (Stat Methodol) 83:133–155
Heller R, Yekutieli D (2014) Replicability analysis for genome-wide association studies. Ann Appl Stat 8:481–498. https://doi.org/10.1214/13-AOAS697
Horowitz JL (2003) Bootstrap methods for Markov processes. Econometrica 71:1049–1082
Jin J, Cai TT (2007) Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. J Am Stat Assoc 102:495–506. https://doi.org/10.1198/016214507000000167
Katsevich E, Ramdas A (2020) Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. Ann Stat 48:3465–3487. https://doi.org/10.1214/19-AOS1938
Kim C-J, Nelson CR et al (1999) State-space models with regime switching: classical and Gibbs-sampling approaches with applications, vol 1. The MIT press, Cambridge
Koski T (2001) Hidden Markov models for bioinformatics, vol 2. Springer, Berlin
Lee JD, Sun DL, Sun Y, Taylor JE et al (2016) Exact post-selection inference, with application to the lasso. Ann Stat 44:907–927
Leek JT, Storey JD (2008) A general framework for multiple testing dependence. Proc Natl Acad Sci 105:18718–18723
Luo F (2019) A systematic evaluation of copy number alterations detection methods on real SNP array and deep sequencing data. BMC Bioinform 20:1–16
Nguyen VH, 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 PS 18:584–612. https://doi.org/10.1051/ps/2013041
Okamoto A, Sehouli J, Yanaihara N, Hirata Y, Braicu I, Kim B-G, Takakura S, Saito M, Yanagida S, Takenaka M et al (2015) Somatic copy number alterations associated with Japanese or endometriosis in ovarian clear cell adenocarcinoma. PLoS ONE 10:e0116977

Panigrahi S, Taylor J, Weinstein A (2020) Integrative methods for post-selection inference under convex constraints
Pierre-Jean M, Neuvial P (2017) acnr: annotated copy-number regions R package version 1.0.0
Pierre-Jean M, Rigaill G, Neuvial P (2015) Performance evaluation of DNA copy number segmentation methods. Brief Bioinform 16:600–615
Pierre-Jean M, Rigaill G, Neuvial P (2019) jointseg: Joint segmentation of multivariate (copy number) signals R package version 1.0.2
Rebafka T, Roquain E, Villers F (2019) Graph inference with clustering and false discovery rate control
Robin S, Bar-Hen A, Daudin J-J, Pierre L (2007) A semi-parametric approach for mixture models: application to local false discovery rate estimation. Comput Stat Data Anal 51:5483–5493
Roquain E, Verzelen N (2020) False discovery rate control with unknown null distribution: is it possible to mimic the oracle?
Scheffé H (1959) The analysis of variance. Chapman & Hall Ltd, London, p 0116429
Schwartzman A (2010) Comment: correlated z-values and the accuracy of large-scale statistical estimates. J Am Stat Assoc 105:1059–1063. https://doi.org/10.1198/jasa.2010.tm10237
Senn S (2008) A note concerning a selection "paradox" of Dawid's. Am Stat 62:206–210
Shah SP, Cheung K-J Jr, Johnson NA, Alain G, Gascoyne RD, Horsman DE, Ng RT, Murphy KP (2009) Model-based clustering of array CGH data. Bioinformatics 25:i30–i38
Stephens M (2017) False discovery rates: a new deal. Biostatistics 18:275–294
Sun W, Cai TT (2007) Oracle and adaptive compound decision rules for false discovery rate control. J Am Stat Assoc 102:901–912. https://doi.org/10.1198/016214507000000545
Sun W, Cai TT (2009) Large-scale multiple testing under dependence. J R Stat Soc Ser B (Stat Methodol) 71:393–424
Sun L, Stephens M (2018) Solving the empirical Bayes normal means problem with correlated noise
Sun Y, Zhang NR, Owen AB (2012) Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data. Ann Appl Stat 6:1664–1688
Tibshirani RJ, Rinaldo A, Tibshirani R, Wasserman L (2018) Uniform asymptotic inference and the bootstrap after model selection. Ann Stat 46:1255–1287
Weinstein A, Ramdas A (2019) Online control of the false coverage rate and false sign rate
Yekutieli D (2012) Adjusted Bayesian inference for selected parameters. J R Stat Soc Ser B (Stat Methodol) 74:515–541
Zhang NR (2010) DNA copy number profiling in normal and tumor genomes. In: Feng J, Fu W, Sun F (eds) Frontiers in computational and systems biology. Springer, Berlin, pp 259–281. https://doi.org/10.1007/978-1-84996-196-7_14