On spike and slab empirical Bayes multiple testing

Ismael Castillo, Etienne Roquain

On spike and slab empirical Bayes multiple testing

Ismaël Castillo and Étienne Roquain

Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4, Place Jussieu, 75252 Paris cedex 05, France
e-mail: ismael.castillo@upmc.fr; etienne.roquain@upmc.fr

Abstract: This paper explores a connection between empirical Bayes posterior distributions and false discovery rate (FDR) control. In the Gaussian sequence model, this work shows that empirical Bayes-calibrated spike and slab posterior distributions allow a correct FDR control under sparsity. Doing so, it offers a frequentist theoretical validation of empirical Bayes methods in the context of multiple testing. Our theoretical results are illustrated with numerical experiments.

AMS 2000 subject classifications: Primary 62C12, 62G10.
Keywords and phrases: Frequentist properties of Bayesian procedures, False discovery rate, sparsity, multiple testing.

1. Introduction

1.1. Context

In modern high dimensional statistical models, several aims are typically pursued, often at the same time: *testing* of hypotheses on the parameters of interest, *estimation* and *uncertainty* quantification, among others. Due to their flexibility, in particular in the choice of the prior, Bayesian posterior distributions are routinely used to provide solutions to a variety of such inference problems. However, although practitioners may often directly read off quantities such as the posterior mean or credible sets once they have simulated posterior draws, the question of mathematical justification of the use of such quantities, in particular from a frequentist perspective, has recently attracted a lot of attention. While the seminal papers [21], [34] set the stage for the study of posterior estimation rates in general models, the case of estimation in high dimensional models has been considered only recently from the point of view of estimation, see [24], [14], [41] among others, while results on frequentist coverage of credible sets are just starting to emerge, see e.g. [3], [40]. Some of the previous approaches rely on automatic data-driven calibration of the prior parameters, following the so-called *empirical Bayes* approach, notably [24], estimating the proportion of significant parameters, and [23], where the full distribution function of the unknowns is estimated.

*Work partly supported by grants of the ANR, projects ANR-16-CE40-0019 (SansSouci) and ANR-17-CE40-0001 (BASICS)

Our interest here is on the issue of *multiple testing* of hypotheses. Typically, the problem is to identify the active variables among a large number of candidates. This task appears in a wide variety of applied fields as genomics, neuro-imaging, astrophysics, among others. Such data typically involve more than thousands of variables with only a small part of them being significant (sparsity).

In this context, a typical aim is to control the false discovery rate (FDR), see (8) below, that is, to find a selection rule that ensures that the averaged proportion of errors among the selected variables is smaller than some prescribed level α. This multiple testing type I error rate, introduced in [4], became quickly popular with the development of high-throughput technologies because it is "scalable" with respect to the dimension: the more rejections are possible, the more false positives are allowed. A common way to achieve this goal is to compute the p-values (probability under the null that the test statistic is larger than the observed value) and to run the Benjamini-Hochberg (BH) procedure [4], which is often considered as a benchmark procedure. In the last decades, an extensive literature aimed at studying the BH method, by showing that it (or versions of it) controls the FDR in various frameworks, see [6, 5, 31, 19], among others.

In a fundamental work [1], Abramovich, Benjamini, Donoho and Johnstone proved that a certain hard thresholding rule deduced from the BH procedure – keeping only observations with significant p-values – satisfies remarkable risk properties: it is minimax adaptive simultaneously for a range of losses and sparsity classes over a broad range of sparsity parameters. In addition, similar results hold true for the misclassification risks, see [7, 29]. These results in particular suggest a link between FDR controlling procedures and adaptation to sparsity. Further results in this direction include [9], [36] for the SLOPE estimate. Here, we shall follow a questioning that can be seen as 'dual' to the former one: starting from a commonly used Bayesian procedure that is known to optimally adapt to the sparsity in terms of risk over a broad range of sparsity classes (and even, under appropriate self-similarity type conditions, to produce adaptive confidence sets), we ask whether a uniform FDR control can be guaranteed.

1.2. Setting

In this paper, we consider the Gaussian sequence model. One observes, for $1 \leq i \leq n$,

$$X_i = \theta_{0,i} + \varepsilon_i, \tag{1}$$

for an unknown n-dimensional vector $\theta_0 = (\theta_{0,i})_{1\leq i\leq n} \in \mathbb{R}^n$ and ε_i i.i.d. $\mathcal{N}(0,1)$. This model can be seen as a stylized version of an high-dimensional model. The problem is to test

$$H_{0,i} : \text{“}\theta_{0,i} = 0\text{”} \text{ against } H_{1,i} : \text{“}\theta_{0,i} \neq 0\text{”},$$

simultaneously over $i \in \{1, \dots, n\}$. We also introduce the assumption that the vector θ_0 is s_n-sparse, that is, is supposed to belong to the set

$$\ell_0[s_n] = \{\theta \in \mathbb{R}^n \ : \ \#\{1 \leq i \leq n \ : \ \theta_i \neq 0\} \leq s_n\}, \tag{2}$$

for some sequence $s_n \in \{0, 1, \dots, n\}$, typically much smaller than n, measuring the sparsity of the vector.

1.3. Bayesian multiple testing methodology

From the point of view of posterior distributions, one natural approach for testing is simply based on comparing posterior probabilities of the hypotheses under consideration. Yet, to do so, a choice of prior needs to be made, and for this reason it is important to carefully design a prior that is flexible enough to adapt to the unknown underlying structure (and, here, sparsity) of the model. This is one of the reasons behind the use of *empirical Bayes* approaches, that aim at calibrating the prior in a fully automatic, data-driven, way. Empirical Bayes methods for multiple testing have been in particular advocated by Efron (see e.g. [17] and references therein) in a series of works over the last 10-15 years, reporting excellent behaviour of such procedures – we describe two of them in more detail in the next paragraphs – in practice. Fully Bayes methods, that bring added flexibility by putting prior on sensible hyperparameters, are another alternative. In the sequel *Bayesian multiple testing procedures* will be referred to as BMT for brevity.

Several popular BMT procedures rely on two quantities that can be seen as possible Bayesian counterparts of standard p-values:

- the ℓ-value: the probability that the null is true conditionally on the fact that the test statistics is *equal* to the observed value, see e.g. [18];
- the q-value: the probability that the null is true conditionally on the fact that the test statistics is *larger* than the observed value, introduced in [35].

(Note that the ℓ-value is usually called "local FDR". Here, we used another terminology to avoid any confusion between the procedure and the FDR.) Obviously, these quantities are well defined only if the trueness/falseness of a null hypothesis is random, which is obtained by introducing an appropriate prior distribution.

Once the prior is calibrated (in a data-driven way or not), the q-values (resp. ℓ-values) can be computed and combined to produce BMT procedures. For instance, existing strategies reject null hypotheses with:

- a ℓ-value smaller than a fixed cutoff $t = 0.2$ [16];
- a q-value smaller than the nominal level α [17];
- averaged ℓ-values smaller than the nominal level α [28, 37, 38].

For alternatives see, e.g., [32]. In particular, one popular fact is that the use of Bayesian quantities "automatically corrects for the multiplicity of the tests", see, e.g., [35]; while using p-values requires to use a cutoff t that decreases with the dimension n, using ℓ-values/q-values can be used with a cutoff t close to the nominal level α, without any further correction. This is well known to be valid from a decision theoretic perspective for the Bayes FDR, that is, for the FDR integrated w.r.t. the prior distribution, as we recall in Proposition 1 below.

When the hyper-parameters are estimated from the data within the BMT, the Bayes FDR is still controlled to some extent, as proved in [37, 38]. However, controlling the Bayes FDR does not give theoretical guarantees for the usual frequentist FDR, that is, for the FDR at the true value of the parameter, as the pointwise FDR may deviate from an integrated version thereof.

1.4. Frequentist control of BMT

In this paper, our main aim is to study whether BMT procedures have valid frequentist multiple testing properties.

A first hint has already been given in [35, 17]: it turns out that the BH procedure can loosely be seen as a "plug-in version" of the procedure rejecting the q-values smaller than α (namely, the theoretical c.d.f. of the p-values is estimated by its empirical counterpart). Since the BH procedure controls the (frequentist) FDR, this might suggest a possible connection between BMT and successful frequentist multiple testing procedures.

In regard to the rapidly increasing literature on frequentist validity of Bayesian procedures from the *estimation* perspective, the multiple testing question for BMT procedures has been less studied so far from the theoretical, frequentist, point of view. This is despite a number of very encouraging simulation performance results, see e.g. [28, 10, 22, 26]. A recent exception is the interesting preprint [30] that shows a frequentist FDR control for a BMT based on a continuous shrinkage prior; yet, this control holds under a certain signal-strength assumption only. One main question we ask in the present work is whether a fully *uniform* control (over sparse vectors) of the frequentist FDR is possible for some posterior-based BMT procedures. Also, we would like to clarify whether the final FDR control is made at, or close to, the required level α.

1.5. Spike and slab prior distributions and sparse priors

Let $w \in (0,1)$ be a fixed hyper-parameter. Let us define the prior distribution $\Pi = \Pi_{w,\gamma}$ on $\mathbb{R}^n$ as

$$\Pi_{w,\gamma} = ((1-w)\delta_0 + w\mathcal{G})^{\otimes n}, \tag{3}$$

where $\mathcal{G}$ is a distribution with a symmetric density γ on $\mathbb{R}$. Such a prior is a tensor product of a mixture of a Dirac mass at 0 (spike), that reflects the sparsity assumption, and of an absolutely continuous distribution (slab), that models nonzero coefficients. This is arguably one of the most natural priors on sparse vectors and has been considered in many key contributions on Bayesian sparse estimation and model selection, see, e.g., [27], [20].

Of course, an important question is that of the choice of w and γ. A popular choice of w is data-driven and based on a marginal maximum likelihood empirical Bayes method (to be described in more details below). The idea is to make the procedure learn the intrinsic sparsity while also incorporating some automatic multiplicity correction, as discussed e.g. in [33, 8]. Following such

an approach in a fundamental paper, Johnstone and Silverman [24] show that, provided γ has tails at least as heavy as Laplace, the posterior median of the empirical Bayes posterior is rate adaptive for a wide range of sparsity parameters and classes, is fast to compute and enjoys excellent behaviour in simulations (the corresponding R–package `EBayesThresh` [25] is widely used). Namely, if $\|\cdot\|$ denotes the euclidian norm and $\hat{\theta} = \hat{\theta}(X)$ is the coordinate-wise median of the empirical Bayes posterior distribution, there exists $c_1 > 0$ such that

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \|\hat{\theta} - \theta_0\|^2 \leq c_1 s_n \log(n/s_n). \tag{4}$$

Thus, asymptotically (in the regime $s_n, n \to \infty$, $s_n/n \to 0$), it matches up to a constant the minimax risk for this problem ([15]). In the recent work [11], the convergence of the empirical Bayes full posterior distribution (not only aspects such as median or mean) is considered, and similar results can be obtained, under stronger conditions on the tails of γ (for instance γ Cauchy works). More precisely, for $\Pi(\cdot \,|\, X) = \hat{\Pi}(\cdot \,|\, X)$ the empirical Bayes posterior, one can find a constant $C_1 > 0$ such that

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi(\theta \,|\, X) \leq C_1 s_n \log(n/s_n). \tag{5}$$

Further, under some conditions, one can show that certain credible sets from the posterior distributions are also adaptive confidence sets in the frequentist sense [13]. Alternatively, one can also follow a hierarchical approach and put a prior on w. The paper [14] obtains adaptive rates for such a fully Bayes procedure over a variety of sparsity classes, and presents a polynomial time algorithm to compute certain aspects of the posterior.

Empirical Bayes approaches have also been successfully applied to a variety of different sparse priors such as empirically recentered Gaussian slabs as in [3, 2], or the horseshoe [39, 40], both studied in terms of estimation and the possibility to construct adaptive confidence sets. In [23], an empirical Bayes approach based on the 'empirical' cdf of the θs is shown to allow for optimal adaptive estimation over various sparsity classes. For an overview on the rapidly growing literature on sparse priors, we refer to the discussion paper [40].

Yet, most of the previous results are concerned with estimation or confidence sets, although a few of them report empirical false discoveries, e.g. [40], Figure 7, without theoretical analysis though.

1.6. Aim and results of the paper

Here we wish to find – if this is at all possible – a posterior-based procedure using a prior Π (possibly an empirical Bayes one i.e. $\Pi = \hat{\Pi}$), that can perform *simultaneous inference* in that a) it behaves optimally up to constants in terms of the quadratic risk in the sense of (4) (or (5)), b) its frequentist FDR at *any* sparse vector is bounded from above by (a constant times) a given nominal level. More precisely, given a nominal level $t \in (0, 1)$ and φ_t a multiple testing procedure

deduced from Π (ℓ-values or q-values procedure, as listed in Section 1.3) we want to *validate* its use in terms of a uniform control of its false discovery rate $\mathrm{FDR}(\theta_0, \varphi_t)$, see (8) below, over the whole parameter space. That is, we ask whether we can find $C_2 > 0$ independent of t such that, for n large enough,

$$\sup_{\theta_0 \in \ell_0[s_n]} \mathrm{FDR}(\theta_0, \varphi_t) \leq C_2\, t. \tag{6}$$

Our main results are as follows: for a sparsity $s_n = O(n^\upsilon)$ with $\upsilon \in (0,1)$,

- Theorem 1 shows that (6) holds with C_2 arbitrary small for the BMT procedure rejecting the nulls whenever the corresponding ℓ-value is smaller than t.
- Theorem 2 shows that (6) holds for some $C_2 > 0$ for the BMT procedure rejecting the nulls whenever the corresponding q-value is smaller than t (with a slight modification if only few signals are detected).

These results hold for spike and slab priors, for γ being Laplace or Cauchy, or even for slightly more general heavy-tailed distributions. The hyperparameter $\hat{w}$ is chosen according to a certain empirical Bayes approach to be specified below (with minor modifications with respect to the choice of [24]).

In addition, from a pure multiple testing perspective, it is important to evaluate the amplitude of $C_2 > 0$ in (6). Our numerical experiments support the fact that, roughly, $C_2 = 1$. Furthermore, Theorem 3 shows that for some subset $\mathcal{L}_0[s_n] \subset \ell_0[s_n]$ (containing strong signals), we have for the q-value BMT, for any (sequence) $\theta_0 \in \mathcal{L}_0[s_n]$,

$$\lim_n \ \mathrm{FDR}(\theta_0, \varphi_t) = t\,, \tag{7}$$

so the FDR control is exactly achieved asymptotically in that case.

It follows from these results (combined with previous results of [24, 11]) that the posterior distribution associated to a spike and slab prior, with γ Cauchy and a suitably empirical Bayes–calibrated w, is appropriate to perform several tasks: (6) (multiple testing), (5)–(4) (posterior concentration in L^2-distance). The posterior can also be used to build honest adaptive confidence sets ([13]). The present work, focusing on the multiple testing aspect, then completes the inference picture for spike and slab empirical Bayes posteriors, confirming their excellent behaviour in simulations.

1.7. Organisation of the paper

In Section 2, we introduce Bayesian multiple testing procedures associated to spike and slab posterior distributions as well as the considered empirical Bayes choice of w. In Section 3, our main results are stated, while Section 4 contains numerical experiments and Section 5 a short discussion. Preliminaries for the proofs are given in Section 6, while the proof of the main results can be found in Section 7. The supplementary file [12] gathers a number of useful lemmas

used in the proofs, as well as the proofs of Propositions 1–2 and Theorem 3. The sections and equations of this supplement are referred to with an additional symbol "S-" in the numbering.

1.8. Notation

In this paper, we use the following notation:

- for F a cdf, we set $\overline{F} = 1 - F$
- $\phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$ and $\Phi(x) = \int_{-\infty}^{x} \phi(u)du$
- $u_n \asymp v_n$ means that there exists constants $C, C' > 0$ such that $|v_n|c \leq |u_n| \leq C|v_n|$ for n large enough;
- $u_n \lesssim v_n$ means that there exists constants $C > 0$ such that $|u_n| \leq C|v_n|$ for n large enough;
- $f(y) \asymp g(y)$, for $y \in A$ means that there exists constants $C, C' > 0$ such that for all $y \in A$, $c|g(y)| \leq |f(y)| \leq C|g(y)|$;
- $f(y) \asymp g(y)$, as $y \to \infty$ means that there exists constants $C, C' > 0$ such that $c|g(y)| \leq |f(y)| \leq C|g(y)|$ for y large enough;
- $u_n \sim v_n$ means $u_n - v_n = o(u_n)$.

Also, for $\tau \in \mathbb{R}^n$, the symbols E_τ (resp. P_τ) denotes the expectation (resp. probability) under $\theta_0 = \tau$ in the model (1). The support of $\theta_0 \in \mathbb{R}^n$ is denoted by $S_{\theta_0} = \{i \,:\, \theta_{0,i} \neq 0\}$ or sometimes S_0 for simplicity. The cardinality of the support S_{θ_0} is denoted by $\sigma_0 = |S_0|$.

2. Preliminaries

2.1. Procedure and FDR

A multiple testing procedure is a measurable function of the form $\varphi(X) = (\varphi_i(X))_{1 \leq i \leq n} \in \{0,1\}^n$, where each $\varphi_i(X) = 0$ (resp. $\varphi_i(X) = 1$) codes for accepting $H_{0,i}$ (resp. rejecting $H_{0,i}$). For any such procedure φ, we let

$$\mathrm{FDR}(\theta_0, \varphi) = E_{\theta_0}\left[\frac{\sum_{i=1}^n \mathbf{1}\{\theta_{0,i} = 0\}\varphi_i(X)}{1 \vee \sum_{i=1}^n \varphi_i(X)}\right]. \tag{8}$$

A procedure φ is said to control the FDR at level α if $\mathrm{FDR}(\theta_0, \varphi) \leq \alpha$ for any θ_0 in $\mathbb{R}^n$. Note that under $\theta_0 = 0$, we have $\mathrm{FDR}(\theta_0, \varphi) = P_{\theta_0 = 0}(\exists i \,:\, \varphi_i(X) = 1)$, which means that an α–FDR controlling procedure provides in particular a (single) test of level α of the full null "$\theta_{0,i} = 0$ for all i". As already mentioned, in the framework of this paper, our goal is a control of the FDR around the pre-specified target level, as in (6) or (7) (where $t = \alpha$).

2.2. Prior, posterior, ℓ-values and q-values

Recall the definition of the prior distribution $\Pi = \Pi_{w,\gamma}$ from (3) and let

$$g(x) = \int \gamma(x - u)\phi(u)du. \tag{9}$$

The posterior distribution $\Pi[\cdot \mid X]$ of θ is then explicitly given by

$$\theta \mid X \sim \bigotimes_{i=1}^{n} \ell_i(X)\,\delta_0 + (1-\ell_i(X))\,\mathcal{G}_{X_i} \tag{10}$$

where $\mathcal{G}_x$ is the distribution with density $\gamma_x(u) := \phi(x-u)\gamma(u)/g(x)$ and

$$\ell_i(X) = \ell(X_i; w, g); \tag{11}$$

$$\ell(x; w, g) = \Pi(\theta_1 = 0 \mid X_1 = x) = \frac{(1-w)\phi(x)}{(1-w)\phi(x) + wg(x)}. \tag{12}$$

The quantities $\ell_i(X)$, $1 \le i \le n$, given by (11) are called the *ℓ-values*. Note that, although we do not emphasize it in the notation for short, the ℓ-values depend also on w and g. The ℓ-value measures locally, for a given observation X_i, the probability that the latter comes from pure noise. This is why it is sometimes called 'local-FDR', see [18].

If one has in mind a range of values –i.e. those that exceed a given amplitude–, a different measure is given by the *q-values* defined by

$$q_i(X) = q(X_i; w, g); \tag{13}$$

$$q(x; w, g) = \Pi(\theta_1 = 0 \mid |X_1| \ge |x|) = \frac{(1-w)\overline{\Phi}(|x|)}{(1-w)\overline{\Phi}(|x|) + w\,\overline{G}(|x|)}; \tag{14}$$

$$\overline{G}(s) = \int_s^{+\infty} g(x)dx. \tag{15}$$

The identity (14) relating the q-value to $\overline{\Phi}, \overline{G}$ is proved in Section S-3.

2.3. Assumptions

We follow throughout the paper assumptions similar to those of [24]. The prior γ is assumed to be unimodal, symmetric and so that

$$|\log\gamma(x) - \log\gamma(y)| \le \Lambda|x-y|, \quad x, y \in \mathbb{R}; \tag{16}$$

$$\gamma(y)^{-1}\int_y^{\infty} \gamma(u)du \asymp y^{\kappa-1}, \quad \text{as } y \to \infty,\ \kappa \in [1,2]; \tag{17}$$

$$y \in \mathbb{R} \to y^2\gamma(y) \text{ is bounded.} \tag{18}$$

Conditions (16), (17) and (18) above are for instance true when γ is Cauchy ($\kappa = 2$, $\Lambda = 1$) or Laplace ($\kappa = 1$, Λ is the scaling parameter). As we show in Remark S-11, explicit expressions exist for g, see (9), in the Laplace case. In the Cauchy case, the integral is not explicit, but in practice (to avoid approximating the integral) one can work with the quasi-Cauchy prior, see [25], that satisfies the above conditions and corresponds to

$$\gamma(x) = (2\pi)^{-1/2}(1 - |x|\overline{\Phi}(x)/\phi(x)); \tag{19}$$

$$g(x) = (2\pi)^{-1/2}x^{-2}(1 - e^{-x^2/2}). \tag{20}$$

2.4. Bayesian Multiple Testing procedures (BMT)

We define the multiple procedures defined from the ℓ-values/q-values in the following way:

$$\varphi_i^{\ell\text{-val}}(t; w, g) = \mathbb{1}_{\{\ell_i(X) \le t\}}, \ \ 1 \le i \le n; \tag{21}$$
$$\varphi_i^{q\text{-val}}(t; w, g) = \mathbb{1}_{\{q_i(X) \le t\}}, \ \ 1 \le i \le n, \tag{22}$$

where $t \in (0,1)$ is some threshold, that possibly depends on X. As we will see in Section 6.2, these two procedures, denoted $\varphi^{\ell\text{-val}}(t)$, $\varphi^{q\text{-val}}(t)$ for brevity, simply correspond to (hard) thresholding procedures that select the $|X_i|$'s larger than some (*random*) threshold. The value of the threshold is driven by the posterior distribution in a very specific way: it depends on γ, t, and on the whole data vector X through the empirical Bayes choice of the hyper-parameter w, that automatically "scales" the procedure according to the sparsity of the data.

2.5. Controlling the Bayes FDR

If the aim is to control the FDR at some level α, a first result indicates that choosing $t = \alpha$ in $\varphi^{\ell\text{-val}}(t)$ and $\varphi^{q\text{-val}}(t)$ may be appropriate, because the corresponding procedures control the Bayes FDR, that is, the FDR where the parameter θ has been integrated with respect to the prior distribution (see, e.g., [32]). More formally, for any multiple testing procedure φ, and hyper-parameters w and γ, define

$$\mathrm{BFDR}(\varphi; w, \gamma) = \int_{\mathbb{R}^n} \mathrm{FDR}(\theta, \varphi) d\Pi_{w,\gamma}(\theta). \tag{23}$$

Then the following result holds.

Proposition 1. *Let $\alpha \in (0,1)$ and $w \in (0,1)$ and consider any density γ satisfying the assumptions of Section 2.3. Let $\varphi^\ell = \varphi^{\ell\text{-val}}(\alpha; w, g)$ as defined in (21) and $\varphi^q = \varphi^{q\text{-val}}(\alpha; w, g)$ as defined in (22). Then we have*

$$BFDR(\varphi^\ell; w, \gamma) \le \alpha \, P(\exists i \ : \ \ell_i(X) \le \alpha) \tag{24}$$
$$\le \alpha \, P(\exists i \ : \ q_i(X) \le \alpha) = BFDR(\varphi^q; w, \gamma) \le \alpha. \tag{25}$$

This result can be certainly considered as well known, as (24) (resp. (25)) is similar in essence to Theorem 4 of [38] (resp. Theorem 1 of [35]). It is essentially a consequence of Fubini's theorem, see Section S-2.1 for a proof. While Proposition 1 justifies the use of ℓ/q-values from the purely Bayesian perspective, it does not bring any information about $\mathrm{FDR}(\theta_0, \varphi^\ell)$ and $\mathrm{FDR}(\theta_0, \varphi^q)$ at an arbitrary sparse vector $\theta_0 \in \mathbb{R}^n$.

2.6. Marginal maximum likelihood

In order to choose the hyper-parameter w, we explore now the choice made in [24], following the popular marginal maximum likelihood method. Let us

introduce the auxiliary functions

$$\beta(x) = \frac{g}{\phi}(x) - 1; \quad \beta(x, w) = \frac{\beta(x)}{1 + w\beta(x)}. \tag{26}$$

A useful property is that β is increasing on $[0, \infty)$ from $\beta(0) \in (-1, 0)$ to infinity, see Section 6.1. The marginal likelihood for w is by definition the marginal density of X, given w, in the Bayesian setting. Its logarithm is equal to

$$L(w) = \sum_{i=1}^{n} \log \phi(X_i) + \sum_{i=1}^{n} \log \left(1 + w\beta(X_i)\right),$$

which is a differentiable function on $[0, 1]$. The derivative S of L, the score function, can be written as

$$S(w) = \sum_{i=1}^{n} \frac{\beta(X_i)}{1 + w\beta(X_i)} = \sum_{i=1}^{n} \beta(X_i, w). \tag{27}$$

The function $w \in [0, 1] \to S(w)$ is (a.s.) decreasing and thus $w \in [0, 1] \to L(w)$ is (a.s.) strictly concave. Hence, almost surely, the maximum of the function L on a compact interval exists, is unique, and we can define the marginal maximum likelihood estimator $\hat{w}$ by

$$\hat{w} \;=\; \underset{w \in \left[\frac{1}{n}, 1\right]}{\operatorname{argmax}} L(w) \quad \text{(a.s.)}. \tag{28}$$

This choice of $\hat{w}$ is close to the one in [24]. The only difference is in the lower bound, here $1/n$, of the maximisation interval, which differs from the choice in [24] by a slowly varying term. This difference is important for multiple testing in case of weak or zero signal (in contrast to the estimation task, for which this different choice does not modify the results). Another slightly different choice of interval, still close to $[1/n, 1]$, will also be of interest below.

In addition, if $\hat{w} \in (1/n, 1)$, it solves the equation $S(w) = 0$ in w. However, note that the maximiser $\hat{w}$ can be at the boundary and thus may not be a zero of S.

3. Main results

Let us first describe the ℓ-value algorithm.

Algorithm `EBayesL`

`Input:` $X_1, \ldots, X_n$, slab prior γ, target confidence t
`Output:` BMT procedure $\varphi^{\ell\text{-val}}$

1. Find the maximiser $\hat{w}$ given by (28);
2. Compute $\hat{\ell}_i(X) = \ell(X_i; \hat{w}, g)$ given by (12);
3. Return, for $1 \le i \le n$,

$$\varphi_i^{\ell\text{-val}} = \mathbf{1}\{\hat{\ell}_i(X) \le t\}. \tag{29}$$

Theorem 1. *Consider the parameter space $\ell_0[s_n]$ given by (2) with sparsity $s_n \leq n^\upsilon$ for some $\upsilon \in (0,1)$. Let γ be a unimodal symmetric slab density that satisfies (16)–(18) with κ as in (17). Then the algorithm* `EBayesL` *produces as output the BMT $\varphi^{\ell\text{-}val}$ defined in (29) that satisfies the following: there exists a constant $C = C(\gamma, \upsilon)$ such that for any $t \leq 3/4$, there exists an integer $N_0 = N_0(\gamma, \upsilon, t)$ such that, for any $n \geq N_0$,*

$$\sup_{\theta_0 \in \ell_0[s_n]} FDR(\theta_0, \varphi^{\ell\text{-}val}) \leq C \frac{\log\log n}{(\log n)^{\kappa/2}}. \tag{30}$$

Theorem 1 is proved in Section 7. The proof relies mainly on two different arguments: first, a careful analysis of the concentration of $\hat{w}$, which requires to distinguish between two regimes (signal weak/moderate or strong, basically). Second, study the FDR behavior of the ℓ-value procedure taken at some sparsity parameter w (not random but depending on n) in each of these two regimes. This requires to analyse the mathematical behavior of a number of functions of w, θ_0, uniformly over a wide range of possible sparsities, which is one main technical difficulty of our results. In particular, the concentration of $\hat{w}$ is obtained uniformly over all sparse vectors with polynomial sparsity, without any strong-signal or self-similarity-type assumption, as would typically be the case for obtaining adaptive confidence sets. Such assumptions would of course simplify the analysis significantly, but the point here is precisely that a uniform FDR control is possible for rate-adaptive procedures without any assumption on the true sparse signal. The uniform concentration of $\hat{w}$ is expressed implicitly and requires sharp estimates, contrary to rate results for which a concentration in a range of values is typically sufficient. In particular, some of our lemmas in the supplementary file [12] are refined versions of lemmas in [24].

As a corollary, (30) entails

$$\overline{\lim_n} \sup_{\theta_0 \in \ell_0[s_n]} \mathrm{FDR}(\theta_0, \varphi^{\ell\text{-val}}) = 0,$$

and this for any chosen threshold $t \in (0,1)$ in $\varphi^{\ell\text{-val}}$. From a pure α-FDR controlling point of view, while making a vanishing small proportion of errors is obviously desirable, it implies that $\varphi^{\ell\text{-val}}$ is, as far as the FDR is concerned, somewhat conservative, in the sense that it does not spend all the allowed type I errors (0 instead of α) and thus will make too few (true) discoveries at the end. It turns out that in the present setting ℓ-values are not quite on the "exact" scale for FDR control. An alternative is to consider the q-value scale, as we now describe.

Algorithm `EBayesq`

`Input:` $X_1, \dots, X_n$, slab prior γ, target confidence t
`Output:` BMT procedure $\varphi^{q\text{-val}}$

1. Find the maximiser $\hat{w}$ given by (28).
2. Compute $\hat{q}_i(X) = q(X_i; \hat{w}, g)$
3. Return, for $1 \le i \le n$,

$$\varphi_i^{q\text{-val}} = \mathbf{1}\{\hat{q}_i(X) \le t\} \tag{31}$$

We also consider the following variant of the procedure `EBayesq`, which is mostly the same, except that it does not allow for too small estimated weight $\hat{w}$. Set, for L_n tending slowly to infinity,

$$\omega_n = \frac{L_n}{n\overline{G}(\sqrt{2.1\log n})}. \tag{32}$$

For instance, for γ Cauchy or quasi-Cauchy, we have $\omega_n \asymp (L_n/n)\sqrt{\log n}$ while for γ Laplace(1) we have $\omega_n \asymp (L_n/n)\exp\{C\sqrt{\log n}\}$.

Algorithm `EBayesq.0`

`Input:` $X_1, \dots, X_n$, slab prior γ, target confidence t, sequence L_n
`Output:` BMT procedure $\varphi^{q\text{-val.0}}$

1.-2. Same as for `EBayesq`, returning $\hat{q}_i(X)$
3. Return, for $1 \le i \le n$, and ω_n as in (32),

$$\varphi_i^{q\text{-val.0}} = \mathbf{1}\{\hat{q}_i(X) \le t\}\mathbf{1}\{\hat{w} > \omega_n\} \tag{33}$$

Theorem 2. *Consider the same setting as Theorem 1. Then the algorithm* `EBayesq` *produces the BMT procedure* $\varphi^{q\text{-val}}$ *in* (31) *that satisfies the following: there exists a constant* $C = C(\gamma, \upsilon)$ *such that for any* $t \le 3/4$*, there exists an integer* $N_0 = N_0(\gamma, \upsilon, t)$ *such that, for any* $n \ge N_0$*,*

$$\sup_{\theta_0 \in \ell_0[s_n]} FDR(\theta_0, \varphi^{q\text{-val}}) \le Ct\log(1/t). \tag{34}$$

In addition, the algorithm `EBayesq.0` *produces the BMT procedure* $\varphi^{q\text{-val.0}}$ *in* (33) *that satisfies, for* ω_n *as in* (32) *with* $L_n \to \infty$*,* $L_n \le \log n$*,* $t \le 3/4$ *and* C, N_0 *as before (but with possibly different numerical values), for any* $n \ge N_0$*,*

$$\sup_{\theta_0 \in \ell_0[s_n]} FDR(\theta_0, \varphi^{q\text{-val.0}}) \le Ct. \tag{35}$$

The proof of Theorem 2 is technically close to that of Theorem 1 and is given in Section 7, see also Section 7.2 for an informal heuristic that serves as

guidelines for the proof. The statements of Theorem 2 are however of different nature, because the q-value threshold t appears explicitly in the bounds (34)-(35), that do not vanish as $n \to \infty$.

The two bounds (34) and (35) differ from a $\log(1/t)$ term, which may become significant for small t. This term appears in the case where the signal is weak (only few rejected nulls), for which the calibration $\hat{w}$ is slightly too large. This may not be the case using a different type of sparsity–adaptation, or a different estimate $\hat{w}$. Indeed, this phenomenon disappears when using `EBayesq.0`, since $\hat{w}$ is then set to 0 when it is not large enough, in which case the FDR control is shown to be guaranteed, and we retrieve a dependence in terms of a constant times the target level t.

A consequence of Theorem 2 is that an α–FDR control can be achieved with `EBayesq`/`EBayesq.0` procedures by taking $t = t(\alpha)$ sufficiently small (although not tending to zero). Again, it is important to known how small the constant $C > 0$ can be taken in (34) and (35). When the signal is strong enough, the following result shows that $C = 1$ and the $\log(1/t)$ factor can be removed in (34).

Let us first introduce a set $\mathcal{L}_0[s_n]$ of 'large' signals, for arbitrary $a > 1$,

$$\mathcal{L}_0[s_n] = \Big\{ \theta \in \ell_0[s_n] \,:\, |\theta_i| \geq a\sqrt{2\log(n/s_n)} \text{ for } i \in S_\theta, \quad |S_\theta| = s_n \Big\}. \tag{36}$$

Theorem 3. *Consider $\mathcal{L}_0[s_n] = \mathcal{L}_0[s_n; a]$ defined by (36) with an arbitrary $a > 1$, for $s_n \to \infty$ and $s_n \leq n^\upsilon$ for some $\upsilon \in (0,1)$. Assume that γ is a unimodal symmetric slab density that satisfies (16)–(18) with κ as in (17). Then, for any pre-specified level $t \in (0,1)$, `EBayesq` produces the BMT procedure $\varphi^{q\text{-}val}$ in (31) such that*

$$\lim_n \sup_{\theta_0 \in \mathcal{L}_0[s_n]} FDR(\theta_0, \varphi^{q\text{-}val}) = \lim_n \inf_{\theta_0 \in \mathcal{L}_0[s_n]} FDR(\theta_0, \varphi^{q\text{-}val}) = t\,. \tag{37}$$

In addition, `EBayesq.0` with $L_n \to \infty$, satisfies the same property whenever $s_n/n \geq 2\omega_n$, for ω_n as in (32), which is in particular the case if s_n grows faster than a given power of n and $L_n \leq \log n$.

Theorem 3, although focused on a specific regime, shows that empirical Bayes procedures are able to produce an asymptotically exact FDR control. Again, this may look surprising at first, as the prior slab density γ is not particularly linked to the true value of the parameter $\theta_0 \in \mathcal{L}_0[s_n]$ in (37). This puts forward a strong adaptive property of the spike and slab prior for multiple testing.

Remark 4. *Our three results can be extended to the case where g is not of the form (9) (that is, not necessarily of the form of a convolution with the standard gaussian), but satisfies some weaker properties, see Section 6.1. This extended setting corresponds to a "quasi-Bayesian" approach where the ℓ-values (resp. q-values) are directly given by the formulas (11) (resp. (13)), without specifying a slab prior γ.*

4. Numerical experiments

In this section, our theoretical findings are illustrated via numerical experiments. A motivation here is also to evaluate how the parameters s_n, $\theta_0 \in \ell_0[s_n]$, and the hyper-parameter γ (or g) affect the FDR control, in particular the value of the constant in the bound of Theorem 2.

For this we consider $n = 10^4$, $s_n \in \{10, 10^2, 10^3\}$ and the following two possible scenarios for $\theta_0 \in \ell_0[s_n]$:

- constant alternatives: $\theta_{0,i} = \mu$ if $1 \leq i \leq s_n$ and 0 otherwise;
- randomized alternatives: $\theta_{0,i}$ i.i.d. uniformly distributed on $(0, 2\mu)$ if $1 \leq i \leq s_n$ and 0 otherwise.

The parameter range for μ is taken equal to $\{0.01, 0.5, 1, 2, \ldots, 10\}$. The marginal likelihood estimator $\hat{w}$ given by (28) is computed by using a modification of the function `wfromx` of the package `EbayesThresh` [25], that accommodates the lower bound $1/n$ in our definition (instead of $w_n = \zeta^{-1}(\sqrt{2\log n})$, see (49), in the original version). The parameter γ is either given by the quasi-Cauchy prior (19)-(20) or by the Laplace prior of scaling parameter $a = 1/2$ (see Remark S-11 for more details). For any of the above parameter combinations, the FDR of the procedures `EBayesL`, `EBayesq` (defined in Section 3) is evaluated empirically via 500 replications.

Figure 1 displays the FDR of the procedures `EBayesL` (ℓ-values) and `EBayesq` (q-values). Concerning `EBayesL`, in all situations, the FDR is small while not exactly equal to the value 0, which seems to indicate that the bound found in Theorem 1 is not too conservative. Moreover, the quasi-Cauchy version seems more conservative than the Laplace version, which corroborates our theoretical findings (in our bound (30), we have the factor $(\log n)^{-1}$ for quasi-Cauchy and $(\log n)^{-1/2}$ for Laplace). As for `EBayesq`, when the signal is large, the FDR curves are markedly close to the threshold value t when s_n/n is small, which is in line with Theorem 3. However, for a weak sparsity $s_n/n = 0.1$, the FDR values are slightly inflated (above the threshold t), which seems to indicate that the asymptotical regime is not yet reached for this value. Looking now at the whole range of signal strengths, one notices the presence of a 'bump' in the regime of intermediate values of μ, especially for the Laplace prior. However, this bump seems to disappear when s_n/n decreases. We do not known presently whether this bump is vanishing with n or if this corresponds to a necessary additional constant $C = C(\gamma, \upsilon) > 1$ (or $\log(1/t)$) in the achieved FDR level, but we suspect that this is related to the fact that the intermediate regime was the most challenging part of our proofs. Overall, the Cauchy slab prior seems to have a particularly suitable behavior. This was not totally surprising for us as it already showed more stability than the Laplace prior in the context of estimation with the full empirical Bayes posterior distribution, as seen in [11].

Finally, we provide additional experiments in the supplement, see Section S-7. The findings can be summarized as follows:

- the curves behave qualitatively similarly for randomized alternatives (second scenario);

- at least in the considered framework, the classical BMT procedure based on averaged ℓ-values [37] seems to fail to control the frequentist FDR for *large* signals, which is markedly different from `EBayesq`;
- the procedure `EBayesq.0` (with $L_n = \log\log n$) has a global behavior similar to `EBayesq`, with more conservativeness for weak signal (as expected).
- it is possible to uniformly improve `EBayesq.0` by considering the following modification (named `EBayesq.hybrid` below): if $w \leq \omega_n$, instead of rejecting no null, `EBayesq.hybrid` performs a standard Bonferroni correction, that is, rejects the $H_{0,i}$'s such that $p_i(X) \leq t/n$. Note that a careful inspection of the proof of Theorem 2 (`EBayesq.0` part) shows that the bound (35) is still valid for `EBayesq.hybrid`.

5. Discussion

Our results show that spike and slab priors produce posterior distributions with particularly suitable multiple testing properties. One main challenge in deriving the results was to build bounds that are uniform over sparse vectors. We demonstrate that such a uniform control is possible up to a constant term away from the target control level. This constant is very close to 1 in simulations, and can even be shown to be 1 asymptotically for some subclass of sparse vectors.

The results of the paper are meant as a theoretical validation of the common practical use of posterior-based quantities for (frequentist) FDR control. While the main purpose here was validation, and as such the proposed procedures were not particularly meant to improve upon the classical FDR-controlling procedures (which specifically target the FDR control, while here the starting point is the posterior, which is rather a global inference object over the whole vector θ), it is remarkable that a uniform control of the FDR very close to the target level can be obtained for the spike and slab BMT procedure in the present unstructured sparse high-dimensional model.

While many studies focused on controlling the Bayes FDR with Bayesian multiple testing procedures, this work paves the way for a frequentist FDR analysis of Bayesian multiple testing procedures in different settings. In our study, the perhaps most surprising fact is how well marginal maximum likelihood estimation combines with FDR control under sparsity: as shown in our proof (and summarized in our heuristic) the score function is linked to a peculiar equation that makes perfectly the link between the numerator and the denominator in the FDR of the q-value–based multiple testing procedure. This phenomenon has not been noticed before to the best of our knowledge. We suspect that this link is only part of a more general picture, in which the concentration of the score process in general sparse high dimensional models plays a central role. While this exceeds the scope of this paper, generalizing our results to such settings is a very interesting direction for future work.

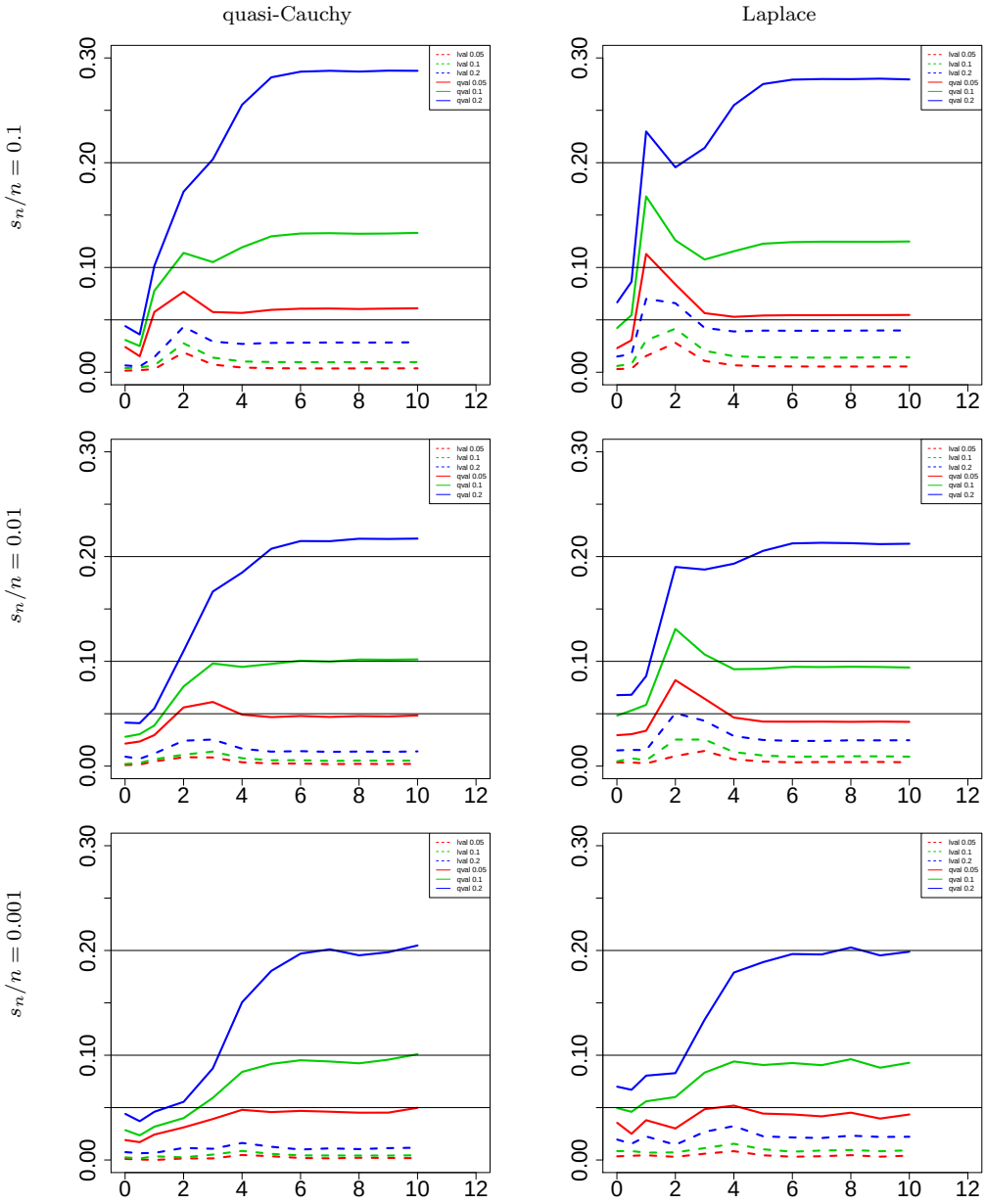

FIG 1. *FDR of* `EBayesL` *and* `EBayesq` *procedures with threshold* $t \in \{0.05, 0.1, 0.2\}$. $\alpha = 0.2$; $n = 10,000$; 500 *replications; alternative all equal to* μ *(on the* X*-axis).*

6. Preliminaries for the proofs

6.1. Working with general g

As noted in Remark 4, the results of Theorems 1, 2 and 3 are also true under slightly more general assumptions, that do not impose that g is coming from a γ by a convolution product. Namely, let us assume

$$\begin{array}{c} g \text{ is a positive, symmetric, differentiable density} \\ \text{that decreases on a vicinity of } +\infty \end{array} \tag{38}$$

(g decreasing on a vicinity of $+\infty$ means that $x \to g(x)$ is decreasing for $x > M$, for a suitably large constant $M = M(g)$). Assume moreover that

$$|(\log g)'(y)| \leq \Lambda, \text{ for all } y \in \mathbb{R}, \ \Lambda > 0; \tag{39}$$

$$\overline{G}(y) \asymp g(y)\, y^{\kappa-1}, \quad \text{as } y \to \infty, \quad \text{for some } \kappa \in [1,2]; \tag{40}$$

$$y \in \mathbb{R} \to (1+y^2)g(y) \text{ is bounded}; \tag{41}$$

$$g/\phi \text{ is increasing on } [0,\infty) \text{ from } (g/\phi)(0) < 1 \text{ to } \infty; \tag{42}$$

By Lemma S-9, it is worth to note that (42) implies

$$\overline{G}/\overline{\Phi} \text{ is increasing on } [0,\infty) \text{ from 1 to } \infty. \tag{43}$$

In the case where g is of the form of a convolution with γ, see (9), conditions (39), (40) and (41) are easy consequences of the fact $g(y) \asymp \gamma(y)$ when $y \to \infty$ and condition (42) follows from the fact that for all fixed $u > 0$, the function $x \in [0,\infty) \to (\phi(x+u) + \phi(x-u))/\phi(x)$ is increasing, see Lemma 1 of [24] for a detailed derivation.

A consequence of (39) is that g and $\overline{G}$ have at least Laplace tails

$$g(y) \geq g(0)e^{-\Lambda y}, \quad y \geq 0; \tag{44}$$

$$\overline{G}(y) \geq g(0)\Lambda^{-1}e^{-\Lambda y}, \quad y \geq 0. \tag{45}$$

6.2. BMT as thresholding-based procedures

Recall the definitions (21) and (22). Let, for any w and t in $[0,1)$,

$$r(w,t) = \frac{wt}{(1-w)(1-t)}. \tag{46}$$

The following quantity plays the role of threshold for ℓ–values,

$$\xi = (\phi/g)^{-1} : (0, (\phi/g)(0)] \to [0,\infty), \tag{47}$$

i.e. ξ is the decreasing continuous invert of ϕ/g (that exists thanks to (42)). Simple algebra shows that for $w, t \in [0,1)$ with $r(w,t) \leq \phi(0)/g(0)$,

$$\ell_i(X) \leq t \ \Leftrightarrow \ |X_i| \geq \xi(r(w,t)). \tag{48}$$

When u becomes small, the order magnitude of $\xi(u)$ is given in Lemma S-13: $\xi(u)$ slightly exceeds $(-2\log u)^{1/2}$ but not by much, which comes from the fact that g has heavy tails.

Another quantity close to ξ we shall use in the sequel is the threshold ζ introduced in [24] and defined as, for any $w \in (0,1]$,

$$\zeta(w) = \beta^{-1}(w^{-1}). \tag{49}$$

Combining the definitions leads, see (S-8) for details, to $\zeta(w) = \xi(w/(1+w))$ and $\xi(w) \leq \zeta(w)$. Similarly, let us introduce a threshold for q–values as

$$\chi = (\overline{\Phi}/\overline{G})^{-1} : (0,1] \to [0,\infty), \tag{50}$$

which is the decreasing continuous invert of $\overline{\Phi}/\overline{G}$ (that exists thanks to (43)). For all $w \in [0,1)$ and $t \in [0,1)$ with $r(w,t) \leq 1$,

$$q_i(X) \leq t \;\Leftrightarrow\; |X_i| \geq \chi(r(w,t)). \tag{51}$$

Lemma S-14 shows that, for small u, the order of magnitude of $\chi(u)$ is slightly more than $\overline{\Phi}^{-1}(u)$ but not by much, which comes from the fact that $\overline{G}$ has heavy tails. Also, Lemma S-10 together with (48)-(51) imply

$$\chi(u) \leq \xi(u), \qquad \text{for } u \leq 1. \tag{52}$$

6.3. Single type I error rates

The single type I error rates of our procedures are evaluated by the following result (proved in Section S-2.2).

Proposition 2. *Consider any function g satisfying the assumptions of Section 6.1. Consider $r(\cdot,\cdot)$ as in (46), ξ as in (47) and χ as in (50). Then the following bounds hold. For all t, w such that $r(w,t) \leq (\phi/g)(0)$,*

$$P_{\theta_0=0}(\ell_i(X) \leq t) \leq 2r(w,t)\frac{g(\xi(r(w,t)))}{\xi(r(w,t))}; \tag{53}$$

$$P_{\theta_0=0}(\ell_i(X) \leq t) \geq r(w,t)\frac{g(\xi(r(w,t)))}{\xi(r(w,t))} \text{ if } r(w,t) \text{ is small enough.} \tag{54}$$

For q-values, we have, for all t, w such that $r(w,t) \leq 1$,

$$P_{\theta_0=0}(q_i(X) \leq t) = r(w,t)\, 2\overline{G}\left(\chi(r(w,t))\right). \tag{55}$$

As a result, for a fixed w, we see that the more g is heavily tailed, the more the type I error rate is large. This is well–expected, as the heavier the tails of g, the more mass the prior puts on large values.

7. Proof of of the main results

7.1. Notation

The following moments are useful when studying the score function S. Let us set

$$\tilde{m}(w) = -E_0\beta(X,w) = \int_{-\infty}^{\infty} \beta(t,w)\phi(t)dt \tag{56}$$

and further denote

$$m_1(\tau,w) = E_\tau[\beta(X,w)] = \int_{-\infty}^{\infty} \beta(t,w)\phi(t-\tau)dt. \tag{57}$$

$$m_2(\tau,w) = E_\tau[\beta(X,w)^2] = \int_{-\infty}^{\infty} (\beta(t,w))^2\phi(t-\tau)dt. \tag{58}$$

These expectations are well defined and studied in detail in Appendix S-5, refining previous results established in [24].

In order to study the FDR of a procedure φ, we introduce the notation

$$V(\varphi) = \sum_{i:\ \theta_{0,i}=0} \varphi_i\ , \qquad S(\varphi) = \sum_{i:\ \theta_{0,i}\neq 0} \varphi_i, \tag{59}$$

counting for φ the number of false and true discoveries, respectively.

7.2. Heuristic

Why should the marginal empirical Bayes choice of w lead to a correct control of the FDR? Here is an informal argument that will give a direction for our proofs. We consider the case of $\varphi^{q\text{-val}}$ here as it is expected to reject more nulls than $\varphi^{\ell\text{-val}}$ and thus to have a larger FDR.

First, let us note that, when there is enough signal, one can expect $\hat{w}$ to be approximately equal to the solution $w^\star$ of the score equation in expectation $E_{\theta_0}(S(w^\star)) = 0$, that is, by using (27),

$$\sum_{i:\theta_{0,i}\neq 0} m_1(\theta_{0,i},w^\star) = (n-s_n)\tilde{m}(w^\star),$$

where $\tilde{m}$ and m_1 are defined by (56) and (57), respectively, if there θ_0 has exactly s_n nonzero coordinates. As seen in Section S-5, up to log-terms,

$$\sum_{i:\theta_{0,i}\neq 0} m_1(\theta_{0,i},w^\star) \approx \sum_{i:\theta_{0,i}\neq 0} \frac{\overline{\Phi}(\zeta(w^\star)-\theta_{0,i}) + \overline{\Phi}(\zeta(w^\star)+\theta_{0,i})}{w^\star};$$
$$\tilde{m}(w^\star) \approx 2\overline{G}(\zeta(w^\star)).$$

Now consider the FDR and assume that all quantities are well concentrated (in particular, take the expectation both in the numerator and denominator in (8)). Then, by using (55), we have

$$\begin{aligned}\mathrm{FDR}(\theta_0, \varphi^{q\text{-val}}(\alpha; \hat{w}, g)) &\approx \mathrm{FDR}(\theta_0, \varphi^{q\text{-val}}(\alpha; w^\star, g)) \\ &\approx \frac{\sum_{i:\theta_{0,i}=0} P_{\theta_{0,i}}(q_i^\star(X) \le \alpha)}{\sum_{i:\theta_{0,i}=0} P_{\theta_{0,i}}(q_i^\star(X) \le \alpha) + \sum_{i:\theta_{0,i}\neq 0} P_{\theta_{0,i}}(q_i^\star(X) \le \alpha)} \\ &\approx \frac{(n-s_n) r(w^\star, \alpha)\, 2\overline{G}\left(\zeta(w^\star)\right)}{(n-s_n) r(w^\star, \alpha)\, 2\overline{G}\left(\zeta(w^\star)\right) + \sum_{i:\theta_{0,i}\neq 0} P_{\theta_{0,i}}(q_i^\star(X) \le \alpha)},\end{aligned}$$

where we denoted $q_i^\star(X) = q(X_i; w^\star, g)$ and we used that $\chi(r(w^\star, t))$ is close to $\zeta(w^\star)$, as seen in Section S-4. Now, by using the definition of $q_i^\star(X)$,

$$\begin{aligned}\sum_{i:\theta_{0,i}\neq 0} P_{\theta_{0,i}}(q_i^\star(X) \le \alpha) &= \sum_{i:\theta_{0,i}\neq 0} \overline{\Phi}\left(\chi(r(w^\star, \alpha)) - \theta_{0,i}\right) + \overline{\Phi}\left(\chi(r(w^\star, \alpha)) + \theta_{0,i}\right) \\ &\approx \sum_{i:\theta_{0,i}\neq 0} \overline{\Phi}\left(\zeta(w^\star) - \theta_{0,i}\right) + \overline{\Phi}\left(\zeta(w^\star) + \theta_{0,i}\right),\end{aligned}$$

where we used again $\chi(r(w^\star, t)) \approx \zeta(w^\star)$. Now using the above properties of $w^\star$, the latter is

$$\approx w^\star \sum_{i:\theta_{0,i}\neq 0} m_1(\theta_{0,i}, w^\star) = (n-s_n) w^\star \tilde{m}(w^\star) \approx (n-s_n) w^\star 2\overline{G}(\zeta(w^\star)).$$

Putting the previous estimates together yields

$$\begin{aligned}\mathrm{FDR}(\theta_0, \varphi^{q\text{-val}}(\alpha; \hat{w}, g)) &\approx \frac{(n-s_n) r(w^\star, \alpha)\, 2\overline{G}\left(\zeta(w^\star)\right)}{(n-s_n) r(w^\star, \alpha)\, 2\overline{G}\left(\zeta(w^\star)\right) + (n-s_n) w^\star 2\overline{G}(\zeta(w^\star))} \\ &= \frac{r(w^\star, \alpha)}{r(w^\star, \alpha) + w^\star} = \frac{\frac{w^\star}{1-w^\star}\frac{\alpha}{1-\alpha}}{\frac{w^\star}{1-w^\star}\frac{\alpha}{1-\alpha} + w^\star} \approx \frac{\frac{\alpha}{1-\alpha}}{\frac{\alpha}{1-\alpha} + 1} = \alpha.\end{aligned}$$

We will see that this heuristic holds, up to some constant terms that may come in factor of the target level α.

7.3. *Proof of Theorems 1 and 2*

We prove results for ℓ- and q-values together. The proof for `EBayesq.0` is given at the end of this section. First, let w_0 be the solution of the equation,

$$n w_0 \tilde{m}(w_0) = M, \tag{60}$$

for M to be chosen below in the range $[1, \log n]$ (more precisely, equal to either $C \log(1/t)$ or $C t^{-1} \log\log n$ for a constant C independent of t and large enough; both bounds belong to the previous interval for n large enough). For

any $M \in [1, \log n]$, this equation has always a unique solution, as $\tilde{m}$ is continuous increasing (see Lemma S-22) so the map $w \to w\tilde{m}(w)$ increases from 0 at $w = 0$ to a constant at $w = 1$, and in particular has a continuous inverse. This implies that w_0 goes to 0 with n, which we use freely in the sequel. Also, we note that w_0 is larger than $1/n$ for C in the choice of M large enough. Indeed, $w_0 \geq \tilde{m}(1)^{-1}M/n$ by monotonicity of $\tilde{m}$. But $\tilde{m}(1)$ is at most a constant, so, provided M is large enough, $w_0 \geq 1/n$. Thus w_0 is always inside the interval $[n^{-1}, 1]$ over which the maximiser $\hat{w}$ is defined.

Let $\nu \in (0, 1)$ and $\theta_0 \in \ell_0[s_n]$. Recall that S_0 denotes the support of θ_0 and that $\sigma_0 = |S_0|$ denotes the exact number of nonzero coefficients of θ_0, so that $0 \leq \sigma_0 \leq s_n$. The next equation, depending on the configuration θ_0, and on the just defined w_0, plays a key role in the proof:

$$\sum_{i \in S_0} m_1(\theta_{0,i}, w) = (1 - \nu)(n - \sigma_0)\tilde{m}(w), \quad w \in [w_0, 1). \tag{61}$$

This equation may or may not have a solution, depending on the true θ_0 and the values of n and ν. We will now assume $n \geq N_0$ for some universal constant N_0 to be determined below.

7.3.1. Case 1: (61) has no solution

For a given value of n, let us consider the case where (61) has no solution in $w \in [w_0, 1)$.

First, the maps $w \in [0, 1] \to \tilde{m}(w)$ and $w \in [0, 1] \to m_1(\mu, w)$ ($\mu \in \mathbb{R}$) are continuous, see Lemmas S-22 and S-24 and, for any $\mu \in \mathbb{R}$,

$$|m_1(\mu, 1)| \leq \int \left| \frac{\beta(x)}{1 + \beta(x)} \right| \phi(x - \mu)dx \leq \max_{x \in \mathbb{R}} \left| \frac{\beta(x)}{1 + \beta(x)} \right|,$$

so thatD $\sum_{i \in S_0} m_1(\theta_{0,i}, 1) \leq C\sigma_0 < (1 - \nu)(n - \sigma_0)\tilde{m}(1)$ for $n \geq N_0$, where we use $\sigma_0 \leq s_n \leq dn^{\upsilon}$ and $\tilde{m}(1) > 0$ and $N_0 = N_0(g, \upsilon)$. This means

$$\sum_{i \in S_0} m_1(\theta_{0,i}, w) < (1 - \nu)(n - \sigma_0)\tilde{m}(w), \quad \text{for } w \in [w_0, 1), \tag{62}$$

as otherwise by the intermediate value theorem the graphs of the functions on the two sides of the previous inequality would have to cross on $[w_0, 1)$ and (61) would have a solution. Lemma S-3 shows that, under (62), we have

$$P_{\theta_0}(\hat{w} > w_0) \leq e^{-C_0\nu^2 M}, \tag{63}$$

for some constant $C_0 = C_0(g, \upsilon)$. Now consider φ being either $\varphi^{\ell\text{-val}}$ or $\varphi^{q\text{-val}}$ and upper-bound the FDR by the so-called family-wise error rate by distinguishing the two cases $\hat{w} \leq w_0$ and $\hat{w} > w_0$ as follows:

$$\begin{aligned} \mathrm{FDR}(\theta_0, \varphi(t; \hat{w}, g)) &\leq P_{\theta_0}(\exists i \ : \ \theta_{0,i} = 0, \ \varphi_i(t; \hat{w}, g) = 1) \\ &\leq P_{\theta_0}(\exists i \ : \ \theta_{0,i} = 0, \ \varphi_i(t; w_0, g) = 1) + P_{\theta_0}(\hat{w} > w_0) \\ &\leq (n - \sigma_0)P_{\theta_{0,i}=0}(\varphi_i(t; w_0, g) = 1) + e^{-C_0\nu^2 M}, \end{aligned} \tag{64}$$

where we use that $w \to \varphi_i(t; w, g)$ is nondecreasing, see Lemma S-7, together with a union bound.

ℓ-value part Let $\xi_0 = \xi(r(w_0, t))$ and $\zeta_0 = \zeta(w_0)$, hen (53) leads to (provided $r(w_0, t) \le (\phi/g)(0)$, which holds for e.g. $t \le 3/4$ and $w_0 \le 1/4$)

$$\mathrm{FDR}(\theta_0, \varphi^{\ell\text{-val}}(t; \hat{w}, g)) \le 2\frac{nw_0}{1-w_0}\frac{t}{1-t}\frac{g(\xi_0)}{\xi_0} + e^{-C_0\nu^2 M}.$$

Combining the definition of w_0 and Lemma S-24, taking n large enough so that w_0 is appropriately small, with $t \le 3/4$,

$$\mathrm{FDR}(\theta_0, \varphi^{\ell\text{-val}}(t; \hat{w}, g)) \le \frac{5M}{\xi_0}\frac{g(\xi_0)}{\overline{G}(\zeta_0)}t + e^{-C_0\nu^2 M}.$$

Noting that $|\xi_0 - \zeta_0| \lesssim 1$, $g(\xi_0) \le Dg(\zeta_0)$ and $\overline{G}(\zeta_0) \asymp \zeta_0^{\kappa-1} g(\zeta_0)$ by Lemma S-17 and S-24, one obtains

$$\mathrm{FDR}(\theta_0, \varphi^{\ell\text{-val}}(t; \hat{w}, g)) \le \frac{C(g)M}{\zeta_0^{\kappa}}t + e^{-C_0\nu^2 M}. \tag{65}$$

q-value part For the q-value case, we come back to (64) and use (55) instead of (53) to get, setting $\chi_0 = \chi(r(w_0, t))$,

$$\mathrm{FDR}(\theta_0, \varphi^{q\text{-val}}(t; \hat{w}, g)) \le 2\frac{nw_0}{1-w_0}\frac{t}{1-t}\overline{G}(\chi_0) + e^{-C_0\nu^2 M}.$$

As a result, by (60) and Lemma S-24, one gets for n large enough, $t \le 3/4$,

$$\mathrm{FDR}(\theta_0, \varphi^{q\text{-val}}(t; \hat{w}, g)) \le 5Mt\frac{\overline{G}(\chi_0)}{\overline{G}(\zeta_0)} + e^{-C_0\nu^2 M}.$$

Now, by the last assertion of Lemma S-17, the ratio in the last display is bounded by 2 (say) provided n is large enough, which gives

$$\mathrm{FDR}(\theta_0, \varphi^{q\text{-val}}(t; \hat{w}, g)) \le 10Mt + e^{-C_0\nu^2 M}. \tag{66}$$

7.3.2. Case 2: (61) has a solution

In this case we denote the solution by $w_1 \in [w_0, 1)$, so that one can write

$$\sum_{i \in S_0} m_1(\theta_{0,i}, w_1) = (1-\nu)(n-\sigma_0)\tilde{m}(w_1). \tag{67}$$

Now consider the slightly different equation in w

$$\sum_{i \in S_0} m_1(\theta_{0,i}, w) = (1+\nu)(n-\sigma_0)\tilde{m}(w), \quad w \in [0, 1). \tag{68}$$

Equation (68) always has a (unique) solution $w_2 \in [0, w_1)$. To see this, first note that the case $\theta_0 = 0$ is excluded from (67), as $m_1(0, w) = -\tilde{m}(w) < 0$ if $w \neq 0$. By Lemma S-22, $w \to m_1(\mu, w)$ and $w \to \tilde{m}(w)$ are continuous and respectively decreasing and increasing (both strictly), and $\tilde{m}(0) = 0$, while it can be seen that $m_1(\mu, 0) > 0$ if $\mu \neq 0$, see Lemma S-22. On the other hand, the value at $w = 1$ of the left hand side of (68) is at most $\sigma_0 C/w \lesssim \sigma_0$, and so is of smaller order than $(1 + \nu)(n - \sigma_0)\tilde{m}(1) \asymp n$.

The purpose of w_1, w_2 is to provide (implicit) deterministic upper and lower bounds for the random $\hat{w}$: this is the content of Lemma S-4. Additionally, the key Lemma S-5 shows that, in case where the solution w_1 of (67) exists, we have $w_1 \asymp w_2$; that is, the bounds are of the same order.

q-value part Recall the notation (59). We focus on the case of q-values first. We come back to the case of ℓ-values at the end, its proof being similar. For simplicity, we write $V_q(w) = V(\varphi^{q-val}(t; w, g))$ and $S_q(w) = S(\varphi^{q-val}(t; w, g))$. By definition of the FDR,

$$\begin{aligned}\mathrm{FDR}(\theta_0, \varphi^{q\text{-val}}(t; \hat{w}, g)) &= E_{\theta_0}\left[\frac{V_q(\hat{w})}{(V_q(\hat{w}) + S_q(\hat{w})) \vee 1}\right] \\ &\leq E_{\theta_0}\left[\frac{V_q(\hat{w})}{(V_q(\hat{w}) + S_q(\hat{w})) \vee 1}\mathbf{1}\{w_2 \leq \hat{w} \leq w_1\}\right] + P_{\theta_0}\left[\hat{w} \notin [w_2, w_1]\right].\end{aligned}$$

The last expectation in the previous display is now bounded by, using first the monotonicity of the maps $w \to V_q(w)$, $w \to S_q(w)$, $x \to x/(1 + x)$ and $x \to 1/(1+x)$, then bounding the indicator variable by 1, and finally combining with Lemma S-37 applied to the *independent* variables $U = V_q(w_1)$ and $T = S_q(w_2)$,

$$\begin{aligned}E_{\theta_0}\left[\frac{V_q(\hat{w})}{(V_q(\hat{w}) + S_q(\hat{w})) \vee 1}\mathbf{1}\{w_2 \leq \hat{w} \leq w_1\}\right] &\leq E_{\theta_0}\left[\frac{V_q(w_1)}{(V_q(w_1) + S_q(w_2)) \vee 1}\right] \\ &\leq \exp\{-E_{\theta_0} S_q(w_2)\} + 12\frac{E_{\theta_0} V_q(w_1)}{E_{\theta_0} S_q(w_2)}.\end{aligned}$$

Next, by using the definition of V_q, one writes

$$E_{\theta_0} V_q(w_1) = \sum_{i:\ \theta_{0,i}=0} 2\overline{\Phi}(\chi(r(w_1, t))) = 2(n - \sigma_0)\overline{\Phi}(\chi(r(w_1, t))).$$

Using the definition of χ, we have $\overline{\Phi}(\chi(u)) = \overline{G}(\chi(u))u$ for $u \in (0, 1)$, so

$$\overline{\Phi}(\chi(r(w_1, t))) = r(w_1, t)\overline{G}(\chi(r(w_1, t))).$$

Then (S-21) in Lemma S-17 implies, for small enough w_1,

$$\overline{G}(\chi(r(w_1, t))) \leq 2\overline{G}(\zeta(w_1)).$$

Combining (S-3) in Lemma S-5, that is $w_1/C \leq w_2 \leq w_1$, for a constant $C = C(\nu, \upsilon, g) > 0$ and Lemma S-19, we have (with, say, $c_1 = 1/2$),

$$(1/2)\overline{G}(\zeta(w_1)) \leq \overline{G}(\zeta(w_1/C)) \leq \overline{G}(\zeta(w_2)).$$

Next using Lemma S-24, one obtains $\overline{G}(\chi(r(w_1,t))) \leq 3\,\tilde{m}(w_2)$, so that

$$\begin{aligned} E_{\theta_0} V_q(w_1) &\leq 3(n-\sigma_0)\frac{w_1}{1-w_1}\tilde{m}(w_2)\frac{t}{1-t} \\ &\leq 3C(n-\sigma_0)\frac{w_2}{1-Cw_2}\tilde{m}(w_2)\frac{t}{1-t} \\ &\leq C^*(n-\sigma_0)w_2\tilde{m}(w_2)t, \end{aligned}$$

because $t \leq 3/4$ for some constant $C^* = C^*(\nu, \upsilon, g) > 0$. On the other hand, by definition of S_q, one can write

$$E_{\theta_0} S_q(w_2) = \sum_{i:\theta_{0,i}\neq 0} \overline{\Phi}\left(\chi(r(w_2,t)) - \theta_{0,i}\right) + \overline{\Phi}\left(\chi(r(w_2,t)) + \theta_{0,i}\right).$$

Let us introduce the set of indices, for $K_1 = 2/(1-\upsilon)$,

$$\mathcal{C}_0(w, K_1) = \left\{1 \leq i \leq n \ : \ |\theta_{0,i}| \geq \frac{\zeta(w)}{K_1}\right\}. \tag{69}$$

Moreover, $\chi(r(w_2,t)) \leq \zeta(w_2)$ by Lemma S-16. Hence,

$$\begin{aligned} E_{\theta_0} S_q(w_2) &\geq \sum_{i\in\mathcal{C}_0(w_2,K_1)} \overline{\Phi}\left(\zeta(w_2) - \theta_{0,i}\right) + \overline{\Phi}\left(\zeta(w_2) + \theta_{0,i}\right) \\ &\geq \sum_{i\in\mathcal{C}_0(w_2,K_1)} \overline{\Phi}\left(\zeta(w_2) - |\theta_{0,i}|\right). \end{aligned} \tag{70}$$

First, we apply Corollary S-26 with $K = K_1$, $w = w_2$ to bound each term in the sum in terms of m_1, noting that $|\theta_{0,i}| \geq \zeta(w_2)/K_1$ by definition of the set $\mathcal{C}_0(w_2, K_1)$. Next, one uses Lemma S-31 restricting the suprema to $w = w_2$ (which is in the prescribed interval by Lemmas S-1, S-2 and S-5) and $K = K_1$, to get for n large enough and constants $C = C(\upsilon, g) > 0$, $C' = C'(\upsilon, g) > 0$, $D = D(\upsilon, g) \in (0,1)$,

$$\begin{aligned} &\sum_{i\in\mathcal{C}_0(w_2,K_1)} \overline{\Phi}\left(\zeta(w_2) - |\theta_{0,i}|\right) \geq Cw_2 \sum_{i\in\mathcal{C}_0(w_2,K_1)} m_1(\theta_{0,i}, w_2) \\ &\geq Cw_2\Big\{\sum_{i\in S_0} m_1(\theta_{0,i}, w_2) - Cn^{1-D}\tilde{m}(w_2)\Big\} \\ &= Cw_2\Big\{(1+\nu)(n-\sigma_0)\tilde{m}(w_2) - C'n^{1-D}\tilde{m}(w_2)\Big\}, \end{aligned}$$

where the last equality comes from (68). As a consequence, for n large enough, for a positive constant $C_* = C_*(\upsilon, g) > 0$, we have

$$E_{\theta_0} S_q(w_2) \geq C_*(n-\sigma_0)w_2\tilde{m}(w_2).$$

Combining the previous bounds leads to

$$E_{\theta_0}\left[\frac{V_q(\hat{w})}{V_q(\hat{w}) + S_q(\hat{w}) \vee 1}\mathbf{1}\{w_2 \leq \hat{w} \leq w_1\}\right] \leq e^{-C_*(n-\sigma_0)w_2\tilde{m}(w_2)} + 12\frac{C^*}{C_*}t.$$

As $w \to w\tilde{m}(w)$ is increasing, and $w_1/C \leq w_2$ by Lemma S-5, we have $w_2\tilde{m}(w_2) \geq (w_1/C)\tilde{m}(w_1/C)$. Recall that $w_1 \geq w_0$ by definition, so Lemma S-24 together with (S-24) of Lemma S-19 imply

$$\tilde{m}(w_1/C) \geq (1/2)\tilde{m}(w_1) \geq (1/2)\tilde{m}(w_0).$$

Combining the obtained inequalities leads to

$$(n-\sigma_0)w_2\tilde{m}(w_2) \geq C'(n-\sigma_0)w_0\tilde{m}(w_0) \geq C'M, \tag{71}$$

where the last inequality follows from the definition of w_0. Now turning to a bound on the FDR, Lemma S-4 and the above inequality imply, with $\nu = 1$,

$$P_{\theta_0}[\hat{w} \notin [w_1, w_2]] \leq 2e^{-C_1\nu^2 n w_2\tilde{m}(w_2)} \leq 2e^{-CM}, \tag{72}$$

for some $C = C(\upsilon, g) > 0$. Conclude that in the considered case, for some constants $c_1 = c_1(\upsilon, g), c_2 = c_2(\upsilon, g) > 0$,

$$\mathrm{FDR}(\theta_0, \varphi^{q\text{-val}}(t; \hat{w}, g)) \leq c_2 t + 3e^{-c_1 M}. \tag{73}$$

ℓ-value part In the case of ℓ-values, one can follow a similar argument. We write $V_\ell(w) = V(\varphi^{\ell\text{-val}}(t; w, g))$ and $S_\ell(w) = S(\varphi^{\ell\text{-val}}(t; w, g))$. Again, the maps $w \to V_\ell(w)$ and $w \to S_\ell(w)$ are monotone. As above for q-values, one deduces

$$\mathrm{FDR}(\theta_0, \varphi^{\ell\text{-val}}(t; \hat{w}, g)) \leq \exp\{-E_{\theta_0}S_\ell(w_2)\} + 12\frac{E_{\theta_0}V_\ell(w_1)}{E_{\theta_0}S_\ell(w_2)}.$$

By definition of V_ℓ and ξ, one can write

$$E_{\theta_0}V_\ell(w_1) = 2(n-\sigma_0)\overline{\Phi}(\xi(r(w_1, t))).$$

The bound $\overline{\Phi}(u) \leq \phi(u)/u$ for $u > 0$ (see Lemma S-33), combined with the definition of ξ and that $|\xi(r(w_1, t)) - \zeta(w_1)| \lesssim 1$ by Lemma S-17 leads to

$$E_{\theta_0}V_\ell(w_1) \leq 3(n-\sigma_0)\zeta(w_1)^{-1}r(w_1, t)g(\xi(r(w_1, t))).$$

Lemma S-17 then implies $g(\xi(r(w_1, t))) \leq 2g(\zeta(w_1))$ (say), for n large enough. Using $w_1/C \leq w_2 \leq w_1$, and (S-24) in Lemma S-19, we have

$$(1/2)g(\zeta(w_1)) \leq g(\zeta(w_1/C)) \leq g(\zeta(w_2)).$$

Next using the relation $\zeta^{\kappa-1}g(\zeta) \asymp \tilde{m}(w)$ from Lemma S-24, one obtains $g(\xi(r(w_1, t))) \lesssim \zeta(w_2)^{1-\kappa}\tilde{m}(w_2) \lesssim \zeta(w_1)^{1-\kappa}\tilde{m}(w_2)$, so that

$$\begin{aligned} E_{\theta_0}V_\ell(w_1) &\leq Ct(n-\sigma_0)w_1\tilde{m}(w_2)\zeta(w_1)^{-\kappa} \\ &\leq c^*t(n-\sigma_0)w_2\tilde{m}(w_2)\zeta(w_1)^{-\kappa}, \end{aligned}$$

for a constant $c^* = c^*(\upsilon, g) > 0$. On the other hand, by definition of S_ℓ, one can write

$$E_{\theta_0}S_\ell(w_2) = \sum_{i:\theta_{0,i}\neq 0} \overline{\Phi}\left(\xi(r(w_2, t)) - \theta_{0,i}\right) + \overline{\Phi}\left(\xi(r(w_2, t)) + \theta_{0,i}\right).$$

Lemma S-18 now enables to bound from below the two terms in the previous display in terms of $\zeta(w_2)$, and further restricting the sum to the set of indices $\mathcal{C}_0(w_2, K_1)$ defined by (69) with the same choice of K_1 leads to

$$E_{\theta_0} S_\ell(w_2) \geq Ct \sum_{i \in \mathcal{C}_0(w_2, K_1)} \overline{\Phi}\left(\zeta(w_2) - |\theta_{0,i}|\right).$$

Appart from the Ct term in factor, it is the same bound as for q-values, see (70). Hence, using the bound obtained above, for n large enough and $c_* = c_*(\upsilon, g) > 0$,

$$E_{\theta_0} S_\ell(w_2) \geq c_* t (n - \sigma_0) w_2 \tilde{m}(w_2).$$

Combining the previous bounds leads to

$$E_{\theta_0}\left[\frac{V_\ell(\hat{w})}{V_\ell(\hat{w}) + S_\ell(\hat{w}) \vee 1} \mathbf{1}\{w_2 \leq \hat{w} \leq w_1\}\right] \leq e^{-c_* t (n-\sigma_0) w_2 \tilde{m}(w_2)} + 12 \frac{c^*}{c_*} \frac{1}{\zeta(w_1)^\kappa}.$$

As in (71), we have $(n - \sigma_0) w_2 \tilde{m}(w_2) \geq C'(n - \sigma_0) w_0 \tilde{m}(w_0) \geq C' M$. One concludes that, in Case 2, for some constants $d_1 = d_1(\upsilon, g), d_2 = d_2(\upsilon, g) > 0$ and taking $\nu = 1$, setting $\zeta(w_1) = \zeta_1$,

$$\begin{aligned} \mathrm{FDR}(\theta_0, \varphi^{\ell\text{-val}}(t; \hat{w}, g)) &\leq d_2 \zeta_1^{-\kappa} t + e^{-C' M c_* t} + 2e^{-CM} \\ &\leq d_2 \zeta_1^{-\kappa} t + 3e^{-d_1 M t}. \end{aligned} \tag{74}$$

7.3.3. Combining cases 1 and 2

For q-values, for $\nu = 1$ and $t \leq 3/4$, we get by combining (66) and (73)

$$\mathrm{FDR}(\theta_0, \varphi^{q\text{-val}}(t; \hat{w}, g)) \leq \max\left\{10Mt + e^{-C_0 M}, c_2 t + 3e^{-c_1 M}\right\}$$

Taking $M = (C_0 \wedge c_1)^{-1} \log(1/t)$ gives the upper bound

$$\mathrm{FDR}(\theta_0, \varphi^{q\text{-val}}(t; \hat{w}, g)) \leq \max\{C' t \log(1/t) + e^{-\log(1/t)}, c_2 t + 3e^{-\log 1/t}\},$$

which is smaller than $Ct \log(1/t)$, giving the result for q-values.

In the ℓ-values case, with $\zeta_1 \leq \zeta_0$ and setting $\nu = 1$, we get by combining (65) and (74)

$$\begin{aligned} \mathrm{FDR}(\theta_0, \varphi^{\ell\text{-val}}(t; \hat{w}, g)) &\leq \max\left\{CM\zeta_0^{-\kappa} t + e^{-C_0 M}, d_2 \zeta_1^{-\kappa} t + 3e^{-d_1 t M}\right\} \\ &\leq d_3\{(M+1)\zeta_1^{-\kappa} t + e^{-d_4 t M}\}. \end{aligned}$$

The announced bound is obtained upon setting $M = t^{-1} d_4^{-1} \log(\zeta_1^\kappa)$ and noting that $\zeta_1^2 \lesssim \log(1/w_1) \lesssim \log n$ and $\zeta_1^2 \gtrsim \log(1/w_1) \gtrsim \log n$ by using Lemmas S-1, S-2 to bound w_1 and Lemma S-15 to bound $\zeta(w_1)$. This concludes the proof of Theorem 1 for ℓ-values and Theorem 2 for q–values.

7.3.4. Proof for EBayesq.0

First notice that

$$\begin{aligned}\mathrm{FDR}(\theta_0, \varphi^{q\text{-val}.0}(t;\hat{w},g)) &= E_{\theta_0}\left[\frac{\sum_{i=1}^n \mathbf{1}\{\theta_{0,i}=0\}\varphi^{q\text{-val}.0}(t;\hat{w},g)}{1\vee\sum_{i=1}^n \varphi^{q\text{-val}.0}(t;\hat{w},g)}\right] \\ &= E_{\theta_0}\left[\frac{\sum_{i=1}^n \mathbf{1}\{\theta_{0,i}=0\}\varphi^{q\text{-val}}(t;\hat{w},g)}{1\vee\sum_{i=1}^n \varphi^{q\text{-val}}(t;\hat{w},g)}\mathbf{1}\{\hat{w}>\omega_n\}\right],\end{aligned} \tag{75}$$

by definition of algorithm EBayesq.0 . The strategy of proof is similar to the q-value case. Let us take M in the definition (60) of w_0 equal to L_n from the statement of Theorem 2, see (32), and suppose $L_n \in [1, \log n]$. Let us show for n large enough,

$$\omega_n \geq w_0. \tag{76}$$

As $\zeta(w_0) \leq \zeta(1/n) \leq \sqrt{2.1\log n}$ for n large enough by Lemmas S-1, S-15,

$$\omega_n = \frac{L_n}{n\overline{G}(\sqrt{2.1\log n})} \geq \frac{L_n}{n\overline{G}(\zeta(1/n))} \geq \frac{L_n}{n\overline{G}(\zeta(w_0))}.$$

Now, by using Lemma S-24, for n large enough,

$$\frac{L_n}{n\overline{G}(\zeta(w_0))} \geq 0.9\frac{2L_n}{n\tilde{m}(w_0)} \geq \frac{L_n}{n\tilde{m}(w_0)} = w_0,$$

leading to (76). Next, on the one hand, in Case 1, the FDR is bounded by

$$\mathrm{FDR}(\theta_0, \varphi^{q\text{-val}.0}(t;\hat{w},g)) \leq P_{\theta_0}(\hat{w}>\omega_n) \leq P_{\theta_0}(\hat{w}>w_0).$$

By using (63), the last display is at most $e^{-C_0\nu^2 L_n}$. On the other hand, in Case 2, we simply use that by (75),

$$\mathrm{FDR}(\theta_0, \varphi^{q\text{-val}.0}(t;\hat{w},g)) \leq \mathrm{FDR}(\theta_0, \varphi^{q\text{-val}}(t;\hat{w},g)) \leq c_2 t + 3e^{-c_1 L_n},$$

which concludes the proof.

Acknowledgments

This work has been supported by ANR-16-CE40-0019 (SansSouci) and ANR-17-CE40-0001 (BASICS).

References

[1] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653, 2006.

[2] E. Belitser and S. Ghosal. Empirical bayes oracle uncertainty quantification for regression. preprint, available on authors' webpage.
[3] E. Belitser and N. Nurushev. Needles and straw in a haystack: robust empirical Bayes confidence for possibly sparse sequences. *ArXiv e-prints*, Nov. 2015.
[4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995.
[5] Y. Benjamini, A. M. Krieger, and D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.
[6] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188, 2001.
[7] M. Bogdan, A. Chakrabarti, F. Frommlet, and J. K. Ghosh. Asymptotic bayes-optimality under sparsity of some multiple testing procedures. *Ann. Statist.*, 39(3):1551–1579, 2011.
[8] M. g. Bogdan, J. K. Ghosh, and S. T. Tokdar. A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, volume 1 of *Inst. Math. Stat. (IMS) Collect.*, pages 211–230. Inst. Math. Statist., Beachwood, OH, 2008.
[9] M. g. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès. SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103–1140, 2015.
[10] J. Cao, X.-J. Xie, S. Zhang, A. Whitehurst, and M. A White. Bayesian optimal discovery procedure for simultaneous significance testing. *BMC Bioinformatics*, 10(1):5, Jan 2009.
[11] I. Castillo and R. Mismer. Empirical Bayes analysis of Spike and Slab posterior distributions. 2017. preprint arXiv:1801.01696.
[12] I. Castillo and E. Roquain. Supplement to "On spike and slab empirical Bayes multiple testing". 2018.
[13] I. Castillo and B. Szabó. Spike and Slab Empirical Bayes sparse credible sets. 2018. in preparation.
[14] I. Castillo and A. W. van der Vaart. Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *Ann. Statist.*, 40(4):2069–2101, 2012.
[15] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern. Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B*, 54(1):41–81, 1992. With discussion and a reply by the authors.
[16] B. Efron. Size, power and false discovery rates. *Ann. Statist.*, 35(4):1351–1377, 2007.
[17] B. Efron. Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23(1):1–22, 2008.
[18] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96(456):1151–1160, 2001.

[19] H. Finner, T. Dickhaus, and M. Roters. Dependency and false discovery rate: asymptotics. *Ann. Statist.*, 35(4):1432–1455, 2007.
[20] E. I. George and D. P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000.
[21] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.
[22] M. Guindani, P. Müller, and S. Zhang. A Bayesian discovery procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(5):905–925, 2009.
[23] W. Jiang and C.-H. Zhang. General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.*, 37(4):1647–1684, 2009.
[24] I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, 32(4):1594–1649, 2004.
[25] I. M. Johnstone and B. W. Silverman. EbayesThresh: R Programs for Empirical Bayes Thresholding. *Journal of Statistical Software*, 12(8), 2005.
[26] R. Martin and S. Tokdar. A nonparametric Empirical Bayes framework for Large-scale significance testing. *Biostatistics*, 13:427–439, 2012.
[27] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.*, 83(404):1023–1036, 1988. With comments by James Berger and C. L. Mallows and with a reply by the authors.
[28] P. Müller, G. Parmigiani, C. Robert, and J. Rousseau. Optimal sample size for multiple testing: the case of gene expression microarrays. *J. Amer. Statist. Assoc.*, 99(468):990–1001, 2004.
[29] P. Neuvial and E. Roquain. On false discovery rate thresholding for classification under sparsity. *Ann. Statist.*, 40(5):2572–2600, 2012.
[30] J.-B. Salomond. Risk quantification for the thresholding rule for multiple testing using gaussian scale mixtures. 2017. preprint arXiv:1711.08705.
[31] S. K. Sarkar. Stepup procedures controlling generalized FWER and generalized FDR. *Ann. Statist.*, 35(6):2405–2420, 2007.
[32] S. K. Sarkar, T. Zhou, and D. Ghosh. A general decision theoretic formulation of procedures controlling FDR and FNR from a Bayesian perspective. *Statist. Sinica*, 18(3):925–945, 2008.
[33] J. G. Scott and J. O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.*, 38(5):2587–2619, 2010.
[34] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714, 2001.
[35] J. D. Storey. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.*, 31(6):2013–2035, 2003.
[36] W. Su and E. Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.*, 44(3):1038–1068, 2016.
[37] W. Sun and T. T. Cai. Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.*, 102(479):901–912, 2007.
[38] W. Sun and T. T. Cai. Large-scale multiple testing under dependence. *J.*

R. Stat. Soc. Ser. B Stat. Methodol., 71(2):393–424, 2009.

[39] S. van der Pas, B. Szabó, and A. van der Vaart. Adaptive posterior contraction rates for the horseshoe. *Electron. J. Stat.*, 11(2):3196–3225, 2017.

[40] S. van der Pas, B. Szabó, and A. van der Vaart. Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.*, 12(4):1221–1274, 2017. With a rejoinder by the authors.

[41] S. L. van der Pas, B. J. K. Kleijn, and A. W. van der Vaart. The horseshoe estimator: posterior concentration around nearly black vectors. *Electron. J. Stat.*, 8(2):2585–2618, 2014.