Journal of Machine Learning Research 23 (2022) 1-57 Submitted 1/21; Revised 2/22; Published 4/22

Multiple Testing in Nonparametric Hidden Markov Models: An Empirical Bayes Approach

Kweku Abraham LKWA2@CAM.AC.UK
University of Cambridge
Statistical Laboratory
Wilberforce Road, Cambridge CB3 0WB, UK

Ismaël Castillo ISMAEL.CASTILLO@SORBONNE-UNIVERSITE.FR
Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation,
4 Place Jussieu, 75005 Paris, France

Elisabeth Gassiat ELISABETH.GASSIAT@UNIVERSITE-PARIS-SACLAY.FR
Université Paris-Saclay, CNRS, Laboratoire de Mathématiques d'Orsay,
91405 Orsay, France

Editor: Animashree Anandkumar

Abstract

Given a nonparametric Hidden Markov Model (HMM) with two states, the question of constructing efficient multiple testing procedures is considered, treating the states as unknown null and alternative hypotheses. A procedure is introduced, based on nonparametric empirical Bayes ideas, that controls the False Discovery Rate (FDR) at a user-specified level. Guarantees on power are also provided, in the form of a control of the true positive rate. One of the key steps in the construction requires supremum-norm convergence of preliminary estimators of the emission densities of the HMM. We provide the existence of such estimators, with convergence at the optimal minimax rate, for the case of a HMM with $J \geq 2$ states, which is of independent interest.

Keywords: efficient multiple testing, hidden Markov models, false discovery rate, true discovery rate, minimax supremum norm estimation

1. Introduction

1.1 Aim of the Paper

We consider the problem of multiple testing in a hidden Markov model (HMM) setting. Given data $(X_i : i \leq N)$ whose distribution is governed by an unobserved categorical variable $\theta = (\theta_i : i \leq N) \in \{0,1\}^N$ drawn from a Markov chain with unknown parameters, for each $i \leq N$ one seeks to test the null hypothesis $H_{0,i} : \theta_i = 0$ against the alternative $H_{1,i} : \theta_i \neq 0$, where the number of tests N is "large".

To make the problem concrete, we highlight an example given in Sun and Cai (2009). The index i tracks the passage of time, and the variable X_i denotes the recorded cases of an influenza-like illness in some location. When $\theta_i = 0$ one sees typical disease levels, and when $\theta_i = 1$ there is an atypical outbreak. Such outbreaks tend to cluster temporally, so that placing a Markov structure on θ is natural. Procedures which ignore this dependence structure cannot be optimal: high levels of recorded cases are more likely to be outliers if recorded only very briefly than if sustained for a period, and independence-based methods do not account for this. Here, as is typical in multiple testing settings, an optimal test is defined as one which maximises the True Discovery Rate (TDR) while controlling the False Discovery Rate (FDR) at some specified level t; see Section 2.1 for definitions.

When the model parameters are known, classical decision theory arguments show that procedures which threshold based on the probabilities of the θ_i's being zero conditional on the observations $X_1, \dots, X_N$ are optimal (e.g. see Lemma 22). These conditional probabilities are simply posterior probabilities in the Bayesian world, and smoothing probabilities in the latent variables vocabulary. They will (mainly) be called ℓ-values in this work. Here we make the realistic assumption that the model parameters are unknown, hence we replace such 'oracle' thresholding procedures (so called because the ℓ-values depend on the model parameters), with procedures which plug in estimates of the parameters in the chosen modelling: the 'empirical' Bayes method. The optimality of the empirical Bayes method in the HMM setting with parametric modelling of the distributions of $X_i \mid \theta = j$, $j \in \{0, 1\}$ was addressed in Sun and Cai (2009).

Here we consider instead modelling these distributions *nonparametrically.* Parametric modelling of HMMs can lead poor results in case of misspecification, as discussed for example in Yau et al. (2011). We draw attention also to the extensive simulations conducted and discussed in Wang et al. (2019) for real valued observations, and in Su and Wang (2020) for count data. These latter two works demonstrate empirically that the FDR and TDR are badly impacted by parametric modelling in case of misspecification, while nonparametric empirical Bayes methods appear to closely match the optimal behaviour of oracle ℓ-value procedures.

The goal of this paper is to prove this last fact: that the discussed thresholding procedures still (asymptotically) maintain multiple testing optimality properties when the parameters are estimated. Note that the plug-in operation must be addressed more delicately in the current nonparametric framework compared to the (well-specified) parametric framework considered in Sun and Cai (2009). Our key theorems can be summarised as follows.

- Our first main results, Theorems 2 and 3, show theoretically that in the nonparametric HMM setting an empirical Bayesian procedure attains the target FDR level and enjoys TDR optimality. The proofs of these two theorems are partly based on a result in De Castro et al. (2017), which shows how control of plug-in estimators propagates to give control of ℓ-value errors. A key step is to have good *supremum-norm* estimators, in contrast to the L^2-norm estimators previously found in the literature.

- Our second main results, which are both key to obtaining the first and also of independent interest, concern supremum-norm estimation of emission densities in nonparametric HMMs. We provide estimators, and prove in Theorems 4 and 5 that the supremum-norm risk of these estimators achieves the parametric convergence rate $N^{-1/2}$ for discrete observations (where the set of possible values X_1 can take is countable), and the convergence rate $(N/\log N)^{-s/(2s+1)}$, familiar from the classical i.i.d. density estimation setting and also proved to be optimal in the HMM context (see Proposition 6), for Hölder densities with regularity s.

Let us remark that a further advantage of modelling the HMM densities nonparametrically is that it ensures our results allow for fairly arbitrary distributions under the null hypothesis. In contrast, many common multiple testing procedures – including the original Benjamini–Hochberg procedure – assume that the null distribution is known. One can of course adjust such procedures to use an estimated null hypothesis, but there are so far only a few settings in which it has been proved that this plug-in step has no negative effect on the desired properties of the procedures. We refer to the recent work by Roquain and Verzelen (2020) for more discussion concerning this issue.

Finally, we note that as well as enabling the plug-in results which yield control of the FDR, estimating the emission densities in terms of the supremum norm is useful in its own right. Indeed, practically speaking, results of this type justify that plots of density estimators will be visually close to the original density. Such estimators can also be helpful for identifying change points, estimating level sets, and constructing confidence bands for uncertainty quantification.

1.2 Context

Let us place these results in the broader multiple testing and HMM contexts. See also Section 1.3 where links to frequentist-Bayesian literature are given.

Multiple testing. The problem of identifying relevant variables among a large number of possible candidates is ubiquitous with high dimensional data: indeed, multiple testing methods are very popular in the analysis of genomic data, in astrostatistics, and in imaging, to name just a few practical applications. Since the seminal work of Benjamini and Hochberg (1995), controlling the FDR has been the goal of much of the extensive literature on the subject.

Early works tended to assume i.i.d. data. Efron (2007b) noted that ignoring dependence and using methods designed for FDR control with independent data could result in either too conservative or too liberal procedures, showing that dependence must carefully being taken into account. A number of works, including those of Benjamini and Yekutieli (2001), Farcomeni (2007), Finner et al. (2007) and Wu (2008), have shown that under certain assumptions on the dependence structure, some multiple testing procedures designed for independent case (such as the step-up Benjamini–Hochberg procedure) still control the FDR below a given target level. Such procedures, although having guaranteed FDR even under dependence, may suffer from being too conservative. Another line of work considers so-called knockoff methods, designed initially in the independent case in Candès et al. (2018) and extended to the hidden Markov setting to be considered herein in Sesia et al. (2019). Such methods again focus on controlling the FDR, saying little about the power.

The control of power in dependent data settings is less developed. Some works in this direction include those of Xie et al. (2011) and of Heller and Rosset (2021) which consider the 'general two group model', wherein the θ_i's are independent and identically distributed, but for each i the distribution of X_i given θ may depend on the whole vector θ and not only on θ_i. In some settings, such as with genetic data, allowing for the θ_i's themselves to be dependent can however be more natural, and the HMM model for X considered here allows for a natural local structure of θ – while still remaining tractable – by modelling it as a Markov chain. Let us note that other structures can also lead to tractable modelling, for example the stochastic block model considered in Rebafka et al. (2019).

Hidden Markov models. HMMs have been widely used for applications as varied as speech modelling, computational finance and gene prediction since works of Baum, Petrie and coauthors introduced practical algorithms and proved parametric estimation rates in a discrete data setting (Petrie, 1967; Baum and Petrie, 1966; Baum et al., 1970). Later works, including those of Bickel et al. (1998) and of Douc and Matias (2001), extended these proofs to allow parametric modelling of the emission distributions.

Recently, Gassiat et al. (2016) opened the possibility that consistency holds also when the emission densities are modelled nonparametrically by proving identifiability under mild conditions. Anandkumar et al. (2012) introduced in the parametric case a spectral method which was then generalised in De Castro et al. (2016) and Lehéricy (2018) to indeed give consistency at a usual rate in the nonparametric setting. These nonparametric works however focus on L^2-estimation, and do not immediately generalise to give rate-optimal supremum norm estimation: indeed, attempting to apply a typical wavelet method of estimating individual coefficients at a parametric rate and aggregating, one runs into an alignment issue arising from the fact that the emission densities are identifiable only up to a permutation. An insight of the current work is that returning to the spectral method and making sure to *simultaneously* diagonalise matrices bypasses these issues; in particular we do this for a kernel-based estimator.

Finally, note that recent works have considered HMM type settings with *non-stationary* data: see Section 4.5 for some examples.

1.3 Setting

Consider a hidden Markov model (HMM), in which the observations $X = (X_n)_{n\leq N}$ satisfy

$$\begin{aligned} X_n \mid \theta \sim f_{\theta_n}, \quad 1 \leq n \leq N, \\ \theta = (\theta_n)_{n\leq N} \sim \mathrm{Markov}(\pi, Q), \end{aligned} \tag{1}$$

and, conditional on θ, the entries of X are independent. The vector θ of 'hidden states' takes values in $\{0,1\}^N$ (we will later also consider the case where θ takes values in $\{1,\dots,J\}^N$ for some $J \geq 2$) and $\mathrm{Markov}(\pi, Q)$ denotes a Markov chain of initial distribution $\pi = (\pi_0, \pi_1)$, and 2×2 transition matrix Q. The 'emission densities' f_0, f_1 are probability densities with respect to some common dominating measure μ on a measurable space $\mathcal{X}$. For simplicity we will assume that μ is either Lebesgue measure on $\mathbb{R}$ or counting measure on $\mathbb{Z} \subset \mathbb{R}$; our results adapt straightforwardly to the d-dimensional setting, and in principle versions should hold for more general measure spaces (see the discussion in Section 4.4). We use $H = \{Q, \pi, f_0, f_1\}$ to denote a generic set of parameters for the HMM. We denote by Π_H the law of (X, θ) in (1), and by extension also the marginal laws of X and θ. We write E_H to denote the expectation operator associated to Π_H. Let us underline that the observations consist of a single sequence $X_1, \dots, X_N$ of length N and that in particular the sequence $\theta_1, \dots, \theta_N$ is not observed; moreover, all parameters comprising H are unknown.

The goal of multiple testing is to provide a procedure $\varphi = \varphi(X)$ which identifies well for which i we have signal ($\theta_i \neq 0$). Testing errors in the multiple testing sense, to be defined next, are measured collectively through all hypotheses $i = 1, \dots, N$ simultaneously rather than by considering a fixed single coordinate i. We will measure the performance of φ through the false discovery rate (FDR) and the true discovery rate (TDR). Defining the false discovery proportion (FDP) at θ as

$$\mathrm{FDP}_\theta(\varphi) := \frac{\sum_{i=1}^N \mathbb{1}\{\theta_i = 0, \varphi_i = 1\}}{1 \vee \left(\sum_{i=1}^N \varphi_i\right)}, \tag{2}$$

the FDR at θ is given by

$$\mathrm{FDR}_\theta(\varphi) := E_H[\mathrm{FDP}_\theta(\varphi(X)) \mid \theta]. \tag{3}$$

We consider the average false discovery rate for θ generated according to the "prior" law Π_H:

$$\mathrm{FDR}_H(\varphi) := E_{\theta\sim\Pi_H} \mathrm{FDR}_\theta(\varphi) \equiv E_H \mathrm{FDP}_\theta(\varphi), \tag{4}$$

and we define the 'posterior FDR' as the average FDP obtained when θ is drawn from its posterior,

$$\mathrm{postFDR}_H(\varphi) = \mathrm{postFDR}_H(\varphi; X) := E_H[\mathrm{FDP}_\theta(\varphi) \mid X]. \tag{5}$$

The true discovery rate is defined as the expected proportion of signals which are detected by a procedure:

$$\mathrm{TDR}_H(\varphi) = E_H\Big[\frac{\sum_{i=1}^N \mathbb{1}\{\theta_i = 1, \varphi_i = 1\}}{1 \vee (\sum_{i=1}^N \theta_i)}\Big]. \tag{6}$$

Bayesian formulation and latent variable formulation. Let P_0 denote the "true" distribution of the data X arising from model (1). If, in (1), the distribution of θ is interpreted as "prior" distribution (it is of course an "oracle prior", as π, Q are components of the unknown "true" parameter $H = (\pi, Q, f_0, f_1)$), the distribution of $X = (X_n)_{n\leq N}$ in the (oracle) Bayesian setting is simply the true distribution P_0. Of course, one may also avoid the Bayesian vocabulary and simply view model (1) as a latent variable model: under such point of view, ℓ-values are known as smoothing probabilities and $\theta \,|\, X$ is simply a conditional distribution. We find it convenient to nevertheless use Bayesian terminology (in contrast to previous works such as Sun and Cai 2007, 2009). Partially this is in accordance with classical decision theory, wherein Bayesian terminology is commonly used for describing optimal classifiers; indeed, as Storey (2003) observed, "classical classification theory seems to be a bridge between Bayesian modeling and hypothesis testing". It is also helpful preparation for considering a setting where θ is fixed and non-random, as dicussed next.

Connection with frequentist analysis of Bayesian procedures. Recent years have seen notable progress on providing frequentist validations of the use of posterior distributions for inference, with most results concerning the estimation task, and more recently also uncertainty quantification and confidence sets (see e.g. Ghosal and van der Vaart, 2017, for a summary). One can consider using the HMM model (1) not because one believes θ is genuinely random with a Markov structure, but rather as a way to model some block structure of a fixed true θ, wherein neighbour coordinates of X have a higher chance of coming from the same distribution. The first results in this spirit in a multiple testing setting were obtained recently for sparse sequences in Castillo and Roquain (2020), and in Abraham et al. (2021) for the posterior-based procedure considered here, in both cases without block structure. Investigating the Bayesian procedure studied in this paper for structured sequences of fixed θ where the HMM modeling will then be a Bayesian prior seems to be a interesting (but difficult) open problem.

We also note that the results we obtain below still constitute a (partial) frequentist Bayes validation, in the following sense. Consider a standard Bayesian approach where θ is viewed as parameter and given a Markov prior, but not the other parameters (f_0, f_1, π, Q), which are estimated separately. Then Theorems 2 and 3 below prove that if the true (frequentist) data generating distribution is some nonparametric HMM, then the empirical Bayes procedure derived from the posterior on θ behaves consistently from the multiple testing viewpoint: its FDR is controlled with optimality guarantees on the TDR. It is a less strong frequentist analysis than under an arbitrary fixed θ_0, but it validates the frequentist use of the procedure assuming that the data comes from some (fairly arbitrary) non-parametric HMM: this still allows one to capture many typical signals with varied latent densities.

1.4 Outline of the Paper

In Section 2 we introduce our multiple testing procedure and establish its asymptotic performance in Theorems 2 and 3. Section 3 is devoted to the estimation of the emission densities, with asymptotic supremum norm control established in Theorems 4 and 5. We also give in Proposition 6 a lower bound for the estimation of Hölder emission densities with regularity s in the HMM context. Finally, Proposition 7 gives examples of conditions under which one can overcome the 'label switching' issue, present in the HMM setting as for mixture models, in order to know which estimator corresponds to the null state and which to the alternative. This allows us to avoid the assumption, common to many multiple testing methods, that the distribution of the data under the null is known.

In Section 4, we provide a detailed discussion of our assumptions and comparisons of our results with the literature. We also explain the extent to which the rates of convergence of our emission densities estimators are uniform in the parameters.

Proofs of the main theorems are given in Section 5. Intermediate results useful for these proofs are given in Appendices A and B. Appendix C gives a proof of a minimax lower bound. For the reader's convenience, the notation introduced throughout the paper is gathered in Appendix D.

2. The Empirical Bayesian Procedure

2.1 Definition

We analyse an empirical Bayesian approach to the multiple testing problem, based on thresholding by the posterior (smoothing) probabilities, here called the 'ℓ-values' and also known in the literature as the 'local indices of significance' (Efron et al., 2001; Efron, 2007a; Sun and Cai, 2009):

$$\ell_i(X) \equiv \ell_{i,H}(X) = \Pi_H(\theta_i = 0 \mid X). \tag{7}$$

In the 'oracle' setting (where the parameter H is known), it is well known that the optimal (weighted) classification procedure is an ℓ-value thresholding procedure; that is, it is $\varphi_{\lambda,H}$ for some λ, where

$$\varphi_{\lambda,H}(X) = (\mathbb{1}\{\ell_{i,H}(X) < \lambda\})_{i \le N}. \tag{8}$$

It has been shown in Sun and Cai (2009) that this class of procedures (possibly with data-driven thresholds) is also optimal in a multiple testing sense, in that a procedure making false discoveries at a pre-specified rate and maximising a suitable notion of the multiple testing power is necessarily an ℓ-value thresholding procedure (see also Lemma 22).

The FDR is the expectation of the posterior FDR, so that using the latter (which is observable) to choose the threshold is a natural approach. When the parameter H is unobserved, we use an estimator $\hat{H} = (\hat{Q}, \hat{\pi}, \hat{f}_0, \hat{f}_1)$ instead (to be constructed later), and so we are led to the procedure $\varphi_{\hat{\lambda},\hat{H}}$, where

$$\hat{\lambda} = \hat{\lambda}(\hat{H}, t) := \sup\{\lambda : \mathrm{postFDR}_{\hat{H}}(\varphi_{\lambda,\hat{H}}) \le t\}. \tag{9}$$

We also note an alternative characterisation of the threshold $\hat{\lambda}$. In view of the definitions (5) and (7), we have the following expression for the posterior FDR:

$$\mathrm{postFDR}_H(\varphi) = \frac{\sum_{i=1}^N \ell_{i,H}\varphi_i}{1 \vee (\sum_{i=1}^n \varphi_i)}. \tag{10}$$

That is, the posterior FDR of a procedure φ is the average of the selected ℓ-values. Consequently, the procedure $\varphi_{\hat{\lambda},\hat{H}}$ must threshold at one of the "empirical ℓ-values" (i.e. at some $\hat{\ell}_i = \ell_{i,\hat{H}}$), as $\mathrm{postFDR}_{\hat{H}}(\varphi_{\lambda,\hat{H}})$ only changes when λ crosses such a threshold. The threshold $\hat{\lambda}$ can therefore equivalently be expressed, as in Sun and Cai (2009), as $\hat{\lambda} = \hat{\ell}_{(\hat{K}+1)}$, with $\hat{\ell}_{(i)}$ denoting the ith order statistic[1] of $\{\ell_{i,\hat{H}} : 1 \le i \le N\}$, where $\hat{K}$ is defined by

$$\frac{1}{\hat{K}} \sum_{i=1}^{\hat{K}} \hat{\ell}_{(i)} \le t < \frac{1}{\hat{K}+1} \sum_{i=1}^{\hat{K}+1} \hat{\ell}_{(i)}. \tag{11}$$

[By convention the left inequality automatically holds in the case $\hat{K} = 0$, and we define $\hat{\ell}_{(N+1)} := \infty$ so that the right inequality automatically holds in the case $\hat{K} = N$.] Note that $\hat{K}$ is well defined and unique, by monotonicity of the average of nondecreasing numbers. This monotonicity also makes clear the following dichotomy:

$$\mathrm{postFDR}_{\hat{H}}(\varphi_{\lambda,\hat{H}}) \le t \iff \lambda \le \hat{\lambda}. \tag{12}$$

If there are no ties, the procedure $\varphi_{\hat{\lambda},\hat{H}}$ necessarily rejects $\hat{K}$ of the null hypotheses. In the case of ties, it may reject fewer, and to avoid potential conservativity, we therefore consider a slightly adjusted procedure $\hat{\varphi}$.

Definition 1. *Define $\hat{\varphi} = \hat{\varphi}^{(t)}$ to be a procedure rejecting exactly $\hat{K}$ of the hypotheses with the smallest $\hat{\ell}_i$ values, choosing arbitrarily in case of ties, where $\hat{K}$ is defined by* (11)*. We write $\hat{S}_0$ for the rejection set*

$$\hat{S}_0 = \{i \le N : \hat{\varphi}_i = 1\},$$

and we note that by construction we have $|\hat{S}_0| = \hat{K}$ and

$$\{i : \hat{\ell}_i(X) < \hat{\lambda}\} \subseteq \hat{S}_0 \subseteq \{i : \hat{\ell}_i(X) \le \hat{\lambda}\}.$$

[1] We define the order statistics so that repeats are allowed: the order statistics are defined by the fact that $\{\ell_i, i \le N\} = \{\ell_{(j)}, j \le N\}$ as a multiset ($\forall x \in \mathbb{R}$, $\#\{i : \ell_i = x\} = \#\{i : \ell_{(i)} = x\}$) and $\ell_{(1)} \le \ell_{(2)} \le \cdots \le \ell_{(N)}$.

We make the following assumptions on the parameters. The assumptions are not particularly restrictive, and are discussed in detail in Section 4.1. The procedures we construct do not require the quantities involved in these assumptions to be known.

Assumption A. There exists $x^* \in \mathbb{R} \cup \{\pm\infty\}$ such that either

$$\begin{aligned} f_1(x)/f_0(x) &\to \infty, \quad \text{as } x \uparrow x^*, \quad \text{or} \\ f_1(x)/f_0(x) &\to \infty, \quad \text{as } x \downarrow x^*, \end{aligned}$$

where we take the conventions that $1/0 = \infty$, $0/0 = 0$. [In the case where μ is counting measure on $\mathbb{Z}$ and $x^* \notin \{\pm\infty\}$, the limits are interpreted to mean that $f_1(x^*) > 0$ and $f_0(x^*) = 0$.]

Assumption B. There exists a constant $\nu > 0$ such that

$$\max_{j=0,1} E_{X\sim f_j}(|X|^\nu) < \infty.$$

Assumption C. 1. The matrix Q has full rank (i.e. its two rows are distinct), and

$$\delta := \min_{i,j} Q_{i,j} > 0.$$

2. The Markov chain is stationary: the initial distribution $\pi = (\pi_0, \pi_1)$ is the invariant distribution for Q.

Throughout we will write

$$f_\pi(x) = \pi_0 f_0(x) + \pi_1 f_1(x) \tag{13}$$

for the marginal distribution of each X_i, $i \leq N$, under Assumption C; note that necessarily $\min(\pi_0, \pi_1) \geq \delta$ under the assumption. We note the following illustrative examples of pairs of densities with respect to the Lebesgue measure $\mu = \mathrm{d}x$ which satisfy both Assumption A and Assumption B.

Examples. i. $f_j(x) = \phi(x - \mu_j)$, where ϕ is the density of a standard normal random variable and $\mu_1 \neq \mu_2$, with $\nu > 0$ arbitrary and $x^* = \pm\infty$.

ii. f_0 is the density of any normal random variable, and f_1 is the density of any Cauchy random variable (or indeed any other distribution with polynomial tails, for ν adjusted appropriately), with $0 < \nu < 1$ and $x^* = \pm\infty$.

iii. f_0, f_1 are compactly supported densities, and the support of f_1 is not a subset of the support of f_0, with any $\nu > 0$ and any x^* in the support of f_1 but not of f_0.

iv. f_0, f_1 are the densities of Beta random variables, $f_j(x) = c_j x^{\alpha_j - 1}(1-x)^{\beta_j - 1}\mathbb{1}\{x \in [0,1]\}$ for a normalising constant c_j, and $\alpha_0 > \alpha_1$ or $\beta_0 > \beta_1$ (or both), with any $\nu > 0$ and $x^* = 0$ if $\alpha_0 > \alpha_1$ or $x^* = 1$ if $\beta_0 > \beta_1$.

2.2 Theoretical guarantees

Our main result shows that for suitably chosen $\hat{H} = (\hat{Q}, \hat{\pi}, \hat{f}_0, \hat{f}_1)$, the procedure $\hat{\varphi}$ achieves an FDR upper bounded by the level t chosen by the user, at least asymptotically. The existence of estimators with suitable consistency properties is shown in the next section under mild further assumptions. Here $\|\cdot\|$ denotes the usual Euclidean norm for vectors (and later also the corresponding operator norm for matrices), $\|\cdot\|_F$ denotes the Frobenius matrix norm $\|A\|_F^2 = \sum_{ij} A_{ij}^2$, and $\|\cdot\|_\infty$ denotes the L^∞ (supremum) norm on functions taking values in $\mathbb{R}$.

Theorem 2. *Grant Assumptions A to C. Suppose that for some $u > 1+\nu^{-1}$ and some sequence ε_N such that $\varepsilon_N(\log N)^u \to 0$, the estimators $\hat{Q}, \hat{\pi}$ and $\hat{f}_j,\ j=0,1$ satisfy*

$$\Pi_H(\max\{\|\hat{Q}-Q\|_F, \|\hat{\pi}-\pi\|, \|\hat{f}_0-f_0\|_\infty, \|\hat{f}_1-f_1\|_\infty\} > \varepsilon_N) \to 0, \quad \text{as } N\to\infty. \tag{14}$$

Then for $\hat{\varphi}$ the multiple testing procedure of Definition 1 we have

$$\mathrm{FDR}_H(\hat{\varphi}) \to \min(t, \pi_0).$$

As alluded to, the construction of $\hat{\varphi}$ suggests it should have close to optimal power, and the following result shows that this is indeed true under an extra condition on the distribution of $(f_1/f_0)(X_1)$. The extra condition cannot hold if μ is the counting measure, but is only used to prove a property of the limiting ℓ-values, so that a version of Theorem 3 may also hold in the discrete setting – see the discussion in Section 4.4. As is common in the literature (again see Section 4.4), the precise notion of power is given by the marginal true discovery rate (mTDR), the average proportion of true signals which a testing procedure discovers (note that the denominator is necessarily non-zero under Assumption C):

$$\mathrm{mTDR}_H(\varphi) = \frac{E_H\#\{i : \theta_i = 1, \varphi_i = 1\}}{E_H\#\{i : \theta_i = 1\}}. \tag{15}$$

The marginal FDR is defined correspondingly:

$$\mathrm{mFDR}_H(\varphi) = \frac{E_H\#\{i : \theta_i = 0, \varphi_i = 1\}}{E_H\#\{i : \varphi_i = 1\}}, \tag{16}$$

with the convention that $0/0=0$. These ‘marginal’ quantities are, by concentration results, close to the original quantities $\mathrm{TDR}_H(\varphi)$, $\mathrm{FDR}_H(\varphi)$ for many procedures, including $\hat{\varphi}$ (as is implied by ideas in the proof of the following result; see also the discussion in Section 4.4).

Theorem 3. *In the setting of Theorem 2, additionally grant that the distribution function of the random variable $(f_1/f_0)(X_1)$ (i.e. the function $t\mapsto \Pi_H\big((f_1/f_0)(X_1)\le t\big)$) is continuous and strictly increasing. Then the procedure $\hat{\varphi}$ of Theorem 2 satisfies the following as $N\to\infty$:*

$$\begin{aligned}\mathrm{mTDR}_H(\hat{\varphi}) &= \sup\{\mathrm{mTDR}_H(\psi) : \mathrm{mFDR}_H(\psi) \le \mathrm{mFDR}_H(\hat{\varphi})\} + o(1)\\ &= \sup\{\mathrm{mTDR}_H(\psi) : \mathrm{mFDR}_H(\psi) \le t\} + o(1).\end{aligned}$$

The suprema are over all multiple testing procedures ψ satisfying the bound on their mFDR, including oracle procedures allowed knowledge of the parameters H.

The essence of the proof of Theorem 2 is to show that $\hat{\ell}_i \approx \ell_i$ for most $i \le N$ (see Lemma 9, in Section 5.1) and that consequently $\mathrm{postFDR}_H(\hat{\varphi})$ is close to $\mathrm{postFDR}_{\hat{H}}(\hat{\varphi})$. The latter, thanks to our definition of $\hat{\lambda}$, is close t.

In proving Theorem 3, there is no *a priori* control of the power analogous to the bound $\mathrm{postFDR}_{\hat{H}}(\hat{\varphi}) \le t$, hence we cannot simply argue by symmetry. Instead, one shows that $\hat{\lambda}$ concentrates around some $\lambda^* \in (0,1]$: see Lemma 10 in Section 5.1. Then, again using that $\hat{\ell}_i \approx \ell_i$, it follows that $\mathrm{mTDR}_H(\hat{\varphi}) \approx \mathrm{mTDR}_H(\varphi_{\lambda^*,\hat{H}}) \approx \mathrm{mTDR}_H(\varphi_{\lambda^*,H})$ and similarly that $\mathrm{mFDR}_H(\hat{\varphi}) \approx \mathrm{mFDR}_H(\varphi_{\lambda^*,H}) \approx t$. Known optimality results for the class $(\varphi_{\lambda,H} : \lambda \ge 0)$ mean that one is able to show that $\mathrm{mTDR}_H(\varphi_{\lambda^*,H})$ is the largest of procedures with mFDR at most $\mathrm{mFDR}_H(\varphi_{\lambda^*,H}) \approx t$ (see Lemma 22), so that the same is approximately true of $\mathrm{mTDR}_H(\hat{\varphi})$.

See Section 5.1 for the proofs.

3. Supremum Norm Estimation of Emission Densities

Of course, Theorems 2 and 3 are only useful if one can estimate H at an appropriate rate in the specified norms, and the results of this section ensure that this is indeed possible in a wide range of nonparametric settings. Estimation is possible not only in the two-state setting, and since estimation results are of independent interest we assume in this section that the data X is drawn from model (1) for Q a $J \times J$ matrix and π a distribution on $\{1, \dots, J\}$, with the state vector θ taking values in $\{1, \dots, J\}^N$, for some known $J \geq 2$.

Assumptions A and C are designed with the particular FDR context in mind. In the J-state estimation setting we instead use the following conditions, designed to ensure a spectral estimation method works. We will still require the moment condition Assumption B for some results.

Assumption C'. The matrix Q is full rank, the J-state Markov chain $(\theta_n)_{n\in\mathbb{N}}$ is irreducible and aperiodic, and θ_1 follows the invariant distribution. [This is weaker than Assumption C in general, but equivalent in the two-state setting.]

Assumption D. The density functions $f_1, \dots f_J$ are linearly independent. [In the two-state setting it suffices to assume $f_0 \neq f_1$, which is implied by Assumption A.]

Under these assumptions, in a parametric setting a variant of a typical regularity condition suffices to show that estimation is possible at a parametric rate, so that our theorems offer a new proof of the results of Sun and Cai (2009): see Section 4.2. Of greater interest here, though, is that Theorem 2 also allows for a nonparametric setting. As noted already, this is a major improvement for applications – see for instance Yau et al. (2011), Wang et al. (2019) and Su and Wang (2020). Estimating the Markov parameters Q and π consistently up to a permutation at a polynomial rate has already been proved possible (see De Castro et al., 2017, Appendix C), and we therefore focus on estimation, in the supremum norm, of the emissions densities themselves. Note first of all that in a discrete setting estimation is possible at a parametric rate.

Theorem 4. *Assume that the dominating measure μ is the counting measure on $\mathbb{Z}$. Let M_N be a sequence tending to infinity, arbitrarily slowly. Under Assumptions C' and D, there exist estimators $\hat{f}_1, \dots, \hat{f}_J$ and a permutation τ such that*

$$\Pi_H(\|\hat{f}_j - f_{\tau(j)}\|_\infty \geq M_N N^{-1/2}) \to 0.$$

The proof is a simplification of that of Theorem 5 (to follow) and so is sketched only: see Appendix B.4.

For the remainder of this section we assume that the functions $f_1, \dots, f_J$ are densities with respect to the Lebesgue measure on $\mathbb{R}$, $\mu = \mathrm{d}x$. We demonstrate that consistent estimation of these densities in the supremum norm is indeed possible at a near-minimax rate in the nonparametric setting, under the following typical smoothness condition.

Assumption E. $f_1, \dots f_J$ belong to $C^s(\mathbb{R})$ for some $s > 0$, where for $C^0(\mathbb{R})$ denoting all bounded continuous functions from $\mathbb{R}$ to itself (equipped with the usual supremum norm $\|\cdot\|_\infty$) and writing $j = \lfloor s \rfloor$ for the integer part of s, $C^s(\mathbb{R})$ denotes the usual space of (locally) Hölder-continuous functions

$$C^s(\mathbb{R}) = \{f : f^{(j)} \in C^{s-j}(\mathbb{R})\}, \qquad s \geq 1$$
$$C^s(\mathbb{R}) = \{f \in C^0(\mathbb{R}) : \sup_{0<|x-y|\leq 1} \Big(\frac{|f(x) - f(y)|}{|x-y|^s} \Big) < \infty\} \quad s \in (0,1),$$

equipped with the usual norm

$$\|f\|_{C^s} = \|f^{(\lfloor s\rfloor)}\|_{C^{s-\lfloor s\rfloor}} + \sum_{0\leq i<\lfloor s\rfloor} \|f^{(i)}\|_\infty, \quad s\geq 1$$
$$\|f\|_{C^s} = \|f\|_\infty + \sup_{0<|x-y|\leq 1} \frac{|f(y)-f(x)|}{|y-x|^s}, \quad 0<s<1.$$

The results also extend in the usual way to Besov spaces, e.g. using results from Giné and Nickl (2016, Chapter 4).

Theorem 5. *Grant Assumptions B, C', D and E. Suppose $L_0\to\infty$ as $N\to\infty$, and $L_0^{\max(5,(J+3)/2)}r_N\to 0$, where $r_N=(N/\log N)^{-s/(1+2s)}$. Then there exist estimators $\hat{f}_j$, $1\leq j\leq J$ (continuous so that the supremum below is measurable) and a permutation τ such that, for some $C>0$,*

$$\Pi_H\Big(\|\hat{f}_j - f_{\tau(j)}\|_\infty \geq CL_0^5 r_N\Big) \to 0. \tag{17}$$

Convergence in expectation also holds: for some $C'>0$,

$$E_H\|\hat{f}_j - f_{\tau(j)}\|_\infty \leq C' L_0^5 r_n. \tag{18}$$

The proof is given in Section 5.2. The parameter L_0 has the interpretation of the dimension of a matrix used in the contruction of the estimators and it can be chosen to diverge arbitrarily slowly (or even, under slightly strengthened versions of the assumptions, to take the fixed value J – see Algorithm 1 and the remarks thereafter), so that the upper bound is arbitrarily close to the following lower bound. Such a lower bound is familiar from the i.i.d. setting, but does not automatically apply in the current setting. Indeed, the mixture components in a nonparametric mixture model are not identifiable, so that our assumptions necessarily exclude the i.i.d. subcase of a HMM. The content of the following proposition is that these assumptions do not, however, make estimation easier than having i.i.d. samples from each of the emission densities. We refer to Appendix C for a formal statement and proof. Let us just mention that the idea consists of identifying a broad class of parameters over which the upper bound results hold uniformly – some details on this can already be found in Section 4.3 below – and proving the corresponding lower bound over this class of parameters.

Proposition 6 (informal statement)**.** *The rate $r_N=(N/\log N)^{-s/(1+2s)}$ is a lower bound for the minimax supremum-norm estimation rate for the emission densities in a J-state nonparametric HMM.*

The algorithm solving Theorem 5 uses a 'spectral' method similar to those of Anandkumar et al. (2012) De Castro et al. (2017) and Lehéricy (2018). However, De Castro et al. (2017) and Lehéricy (2018) expand in terms of orthonormal basis functions, and use particular properties of L^2-projections which do not straightforwardly adapt to the L^∞ setting. In developing herein a spectral *kernel density estimator* we are forced to bypass the need for these projection properties (note though that one could apply similar ideas to basis function based estimators). See Algorithm 1 for a description of the estimating procedure.

Finally, note that Theorems 4 and 5 only show that one may estimate the parameters consistently *up to a permutation.* While this is generally sufficient for estimation purposes, since the labelling of the states is usually of no relevance, any multiple testing procedure targeting FDR control necessarily treats the null and the alternative differently, so it is essential that we can identify which of our estimators corresponds to the null state. We will therefore also require the following condition.

Condition F. There exist estimators $\hat{f}_1,\dots,\hat{f}_J$ in Theorem 5 (or Theorem 4) for which the permutation τ is the identity.

It suffices that there exist $\{\hat{f}_1, \dots, \hat{f}_J, \tau\}$ as in Theorem 5 for which the permutation τ can be estimated consistently by some $\hat{\tau}$, since we can define $\check{f}_j = \hat{f}_{\hat{\tau}(j)}$. We give two illustrative assumptions, each plausible in the original two-state FDR setting, under which Condition F holds. A version of the following proposition also holds under such assumptions in the discrete setting using Theorem 4 in place of Theorem 5 in the proof, which can be found at the end of Section 5.2.

Proposition 7. *In the setting of Theorem 2 grant also Assumption E (and recall that Assumption D automatically holds). Then Condition F is verified, and there exist estimators $\hat{Q}, \hat{\pi}, \hat{f}_0, \hat{f}_1$ satisfying* (14) *for any rate ε_N slower than $r_N = (N/\log N)^{-s/(1+2s)}$, under either of the following assumptions:*

1. *For some* known $x^* \in \mathbb{R} \cup \{+\infty\}$, $f_1(x)/f_0(x) \to \infty$ *as* $x \uparrow x^*$.
2. π_0 *is known to be greater than* π_1.

4. Discussion

4.1 Applicability of the Results

Generality of the assumptions. Assumptions A to E and Condition F are not restrictive, so that Theorems 2, 4 and 5 hold in typical nonparametric settings (we discuss the extra assumption of Theorem 3 in Section 4.4).

Assumption A is a signal strength assumption, without which the proofs (in particular the proof of Lemma 14) remain valid only for large enough values of t. It is known that weak signals are a case requiring special attention for multiple testing, discussed for example in a different setting in Heller and Rosset (2021).

The moment condition Assumption B is used for Lemmas 12 and 31. A different proof of Theorem 5 might bypass the need for this condition since kernel methods have been known to work in density estimation in the absence of tail conditions.

The full rank assumption on Q in Assumption C' is necessary even for identifiability up to a permutation in the two-state case (with nonparametric emission densities) since otherwise the HMM reduces to an i.i.d. nonparametric mixture model, whose non-identifiability is easily seen. For $J > 2$ states it is not known whether identifiability holds without this assumption, and full rank is assumed in all papers know to the authors concerning nonparametric inference of HMM parameters. Irreducibility is essential to ensure all hidden states genuinely influence the data. Aperiodicity is assumed to allow typical Markov chain convergence and concentration results to apply, but in principle it should be possible to avoid this assumption at the expense of requiring specially tailored proofs, since the proofs use empirical averages as a building block.

Consistent estimation of the HMM parameters is possible upon replacing Assumption D by the weaker assumption that the emission densities are all distinct, see Alexandrovich et al. (2016) and Lehéricy (2018). Proving rates under this weaker assumption is much harder and no results exist yet.

Implementing the method. Our proposed method for estimating the emission densities can be implemented through Algorithm 1. Then, given estimators of the parameters, efficient computations of ℓ-values is easily done using the forward-backward algorithm for HMMs. Indeed the empirical Bayes multiple testing procedure is implemented in Sun and Cai (2009), Wang et al. (2019) and Su and Wang (2020). [These works use mixture models with unknown number of components to estimate the emission densities, either via fully Bayesian methods or via some model selection method.]

4.2 The Parametric Setting

Bickel et al. (1998) prove a central limit theorem for the maximum likelihood estimator of the model parameter (which we denote, say, by h) under standard regularity conditions, so that it may be estimated at a parametric rate up to label switching. To these, adding the condition that the parametrisation map $h \mapsto (f_{1,h}, \dots f_{J,h})$ is Lipschitz continuous with respect to the Euclidean norm and the supremum norm (at least on a neighbourhood of the true parameter), we arrive at the following.

Proposition 8. *In a parametric model satisfying mild regularity conditions, Assumptions C' and D are enough to ensure that there exist estimators $\hat{Q}, \hat{\pi}, \hat{f}_1, \dots, \hat{f}_J$ such that for some permutation τ and any $M_N \to \infty$,*

$$\max\bigl(\|\hat{Q} - Q\|_F, \|\hat{\pi} - \pi\|, \|\hat{f}_1 - f_{\tau(1)}\|_\infty, \dots, \|\hat{f}_J - f_{\tau(J)}\|_\infty\bigr) < M_N N^{-1/2},$$

with probability tending to 1.

We note that many common parametric families, including Gaussian models, exponential models and Poisson models, satisfy a suitable regularity condition (this can be seen by using standard formulae for exponential families to calculate the derivative of the parametrisation map and bounding).

Under an assumption akin to those of Proposition 7 to ensure that a version of Condition F holds, we see that Theorems 2 and 3 apply in a parametric setting. Except perhaps for the regularity condition, our assumptions are weaker than those of Sun and Cai (2009) (after adapting Theorem 3 slightly – see Section 4.4), so that we slightly generalise their main results even in the parametric setting.

4.3 Uniformity in the Parameters

The constants of Theorem 5 depend only on quantitative measures (as listed below) of the degree to which Assumptions B, C', D and E hold, so that a uniform version of (18),

$$\sup_{H \in \mathcal{H}} E_H \|\hat{f}_j - f_{\tau(j)}\|_\infty \leq C' L_0^5 r_n,$$

holds if the following bounds are satisfied on the set $\mathcal{H}$ (and similarly for (17)). The estimators $\hat{f}_1, \dots \hat{f}_J$ do not depend on knowledge of the bound $M < \infty$, so the result is adaptive in these quantities (though recall that the smoothness s is assumed known – see also the discussion of adaptation in Section 4.4).

- $\sup_{H \in \mathcal{H}} (\max_j E_{X \sim f_j} |X|^{1/M}) \leq M$.
- $\sup_{H \in \mathcal{H}} (\kappa(Q)) \leq M$, where $\kappa(Q) = \|Q\| \|Q^{-1}\|$, the condition number, measures how far Q is from having less than full rank.
- $\inf_{H \in \mathcal{H}} \gamma_{\text{ps}} \geq M^{-1}$ where γ_{ps} denotes the pseudo spectral gap of the matrix Q as defined in Paulin (2015). This bound quantitatively measures how far the chain θ is from being reducible or periodic, and is only used to control the mixing time of the chain θ. It can therefore be replaced by any assumption ensuring a uniform bound on the mixing time; in particular, in the two-state case of Section 2, the chain θ is necessarily reversible and it suffices to assume a uniform lower bound on the absolute spectral gap γ^*, defined by

$$\gamma^* = \begin{cases} 1 - \sup\{|\lambda| : \lambda \text{ an eigenvalue of } Q,\ \lambda \neq 1\} & \text{the eigenvalue 1 of } Q \text{ has multiplicity 1,} \\ 0 & \text{otherwise.} \end{cases}$$

- $\inf_{H \in \mathcal{H}} \min_j(\pi_j) > M^{-1}$. This too measures how far the chain is from being reducible.

- $\sup_{H\in\mathcal{H}} \max_j \|f_j\|_{C^s} \leq M$.
- $\sup_{H\in\mathcal{H}} \max(\underline{L}, 1/C) \leq M$, where $(C, \underline{L})$ are the constants, depending on H, from Lemma 24 in Appendix B. Denoting by $\sigma_J(A)$ the Jth largest singular value of a matrix A, these constants control $\sigma_J(O^{L_0})$ where $O^{L_0} = (E[h_l(X_1) \mid \theta_1 = j])_{l\leq L_0, j\leq J}$ for some suitably chosen functions h_l, $l \leq L_0$. The lemma shows that $h_1, \dots, h_{L_0}$ can be chosen in a universal way such that $\max(\underline{L}, 1/C) < \infty$ whenever $f_1, \dots f_J$ are linearly independent, so these constants quantitively measure the linear independence of these functions. In the case $J = 2$, a sufficient (but not necessary) condition for such a uniform bound to hold is that $\inf_H |P_{X\sim f_0}(X \in A) - P_{X\sim f_1}(X \in A)| > 0$ for some *known* set A: one then constructs the estimators $\hat{f}_0$, $\hat{f}_1$ using, in Algorithm 1, $L_0 = 2$, $h_1 = 1$, $h_2 = \mathbb{1}_A$.
- $\inf_{H\in\mathcal{H}} c \geq M^{-1}$ where $c = c(H)$ is the constant of Lemma 26 in Appendix B. The lemma shows that this constant is positive whenever $f_1, \dots, f_J$ are distinct and so it provides a quantitative measure of the degree of distinctness of these functions. In view of the proof, a sufficient (but not necessary) condition for such a uniform bound to hold is that $f_1, \dots f_J$ can uniformly be separated at a point, i.e. that the set $\mathcal{H}$ is such that
$$\inf_{H\in\mathcal{H}} \sup_{x\in\mathbb{R}} \min_{j\neq j'} |f_j(x) - f_{j'}(x)| > M^{-1}.$$

In what follows, we use for example $C = C(\mathcal{H})$ to denote any constant which depends only on the above bounds (i.e. on $M < \infty$). We will also allow such a constant to depend on the kernel K, the functions $h_1, \dots, h_{L_0}$ and the sets $\mathbb{D}_N$ of Algorithm 1 since these can be chosen independently of the parameter $H \in \mathcal{H}$. We note that the set $\mathcal{H}$ over which the upper bound is uniform (under the sufficient conditions of the last two items, with $A = [-1, 1]$) includes the set over which the lower bound Proposition 6 is proved in Appendix C, so that the estimation result Theorem 5 can genuinely be viewed as a minimax result.

The FDR result Theorem 2 is uniform over a large subset $\mathcal{I} \subset \mathcal{H}$. In particular, in addition to the above constraints, one needs to add the following conditions.

- $\inf_{H\in\mathcal{I}} \Pi_H((f_1/f_0)(X_1) > u) > 0$ for each $u > 0$.
- Condition F holds in a uniform way for $H \in \mathcal{I}$.
- $\inf_{H\in\mathcal{I}} \min_{i,j} Q_{ij} > 0$. [This is in fact implied already by the bounds on the π_j and on the pseudo-spectral gap, since for Theorem 2 we are in the two-state setting.]

We write $C = C(\mathcal{I})$ to denote any constant which depends only on $\mathcal{H}$ and these quantities.

4.4 Extensions of the Theorems

Weakening the assumption of Theorem 3. Theorem 3 remains true if we replace the assumption on $(f_1/f_0)(X_1)$ with the following; see Lemma 17 for a proof that this new condition holds under the assumptions of Theorem 3.

Condition G. Viewing the sample $(X_n : 1 \leq n \leq N)$ as coming from a bi-infinite HMM $(X_n : n \in \mathbb{Z})$, grant that the distribution function of
$$\ell_i^\infty(X) := \Pi_H(\theta_i = 0 \mid (X_n)_{n\in\mathbb{Z}}) \tag{19}$$
is continuous and strictly increasing on $[0, 1]$.

This condition is weaker than the 'monotone ratio condition' assumed in Sun and Cai (2009), since the latter implicitly assumes that the distribution function of ℓ_i^∞ has a strictly positive derivative. In the discrete context (that is, when the X_i's take discrete values), understanding when the distribution of the variables ℓ_i^∞ has a density with respect to Lebesgue measure is known to be hard, since it is mostly still an open problem for the closely related stationary filter $\Phi_i^\infty(X) := \Pi_H(\theta_i = 0 \mid (X_n)_{n\leq i})$, see Blackwell (1957), Bárány and Kolossváry (2015) and references therein.

Of particular interest, though, is the fact that this new condition is only about the continuity of the distribution function, not about its absolute continuity. Continuity is a weaker property that could be easier to understand and could hold in much more generality, so that Condition G opens up the possiblity that a version of Theorem 3 may hold even in certain discrete settings. Indeed, simulations in Su and Wang (2020) are suggestive that the conclusions of the theorem hold. They compare various multiple testing procedures and provide empirical evidence that the TDR of the empirical Bayes multiple testing method using nonparametric modeling of HMMs roughly matches that of an oracle thresholding procedure and is the best among the multiple testing procedures they compare.

Use of marginal FDR and TDR in Theorem 3. The proof of Theorem 3 in fact shows, after some minor adjustments, that

$$\mathrm{TDR}_H(\hat\varphi) \geq \mathrm{TDR}_H(\varphi_{\lambda_{\max},H}) - o(1),$$

where $\lambda_{\max} = \lambda_{\max}(t,H)$ is chosen maximal such that $\mathrm{FDR}_H(\varphi_{\lambda_{\max},H}) \leq t$, so that $\hat\varphi$ is (asymptotically) optimal for the TDR when restricting to the class of procedures whose TDR and FDR asymptotically coincide with their marginal equivalents. Heller and Rosset (2021) show in a non-Markovian setting that the procedure maximising the TDR among all procedures with controlled FDR is not in this class, but their results leave open the possiblity that Theorem 3 remains true with the full FDR and TDR. Indeed, a main conclusion of their work is that the class $(\varphi_{\lambda,H} : \lambda \geq 0)$ (or rather, the equivalent of this class for their setting) is optimal for the problem of maximing TDR with controlled FDR provided one allows data-driven thresholds – such as $\hat\lambda$ – whereas the current proof of Theorem 3 uses that for mTDR optimality with mFDR control it suffices to consider the class for non-random thresholds. Furthermore, the difference between the FDR and TDR of the optimal procedure and their marginal versions in the setting of Heller and Rosset (2021) manifests itself for weak signals, so that our signal strength assumption may suffice to rule out any such difference.

Adaptation. The estimator we construct for Theorem 5 uses knowledge of the smoothness s. One can adjust the arguments of Lehéricy (2018) to show that a careful application of Lepskii's method allows adaptation up to a maximum smoothness $s_{\max} < \infty$, and indeed state-by-state adaptation, wherein each state is estimated at a rate adapting to its smoothness parameter s_j, rather than requiring $s_j = s$ for all j. As usual, the rough idea is to construct estimators $\hat f_j^L$, $j \leq J$ for each $L \leq L_{\max}$ and use $\|\hat f_j^L - \hat f_j^{L_{\max}}\|_\infty$ as a proxy for the bias, so that one can make a suitable bias-variance tradeoff. In the HMM setting, as noted in Lehéricy (2018), one must also use $\hat f_j^{L_{\max}}$ to "align" the estimators $\hat f_j^L$ up to a single permutation τ rather than needing a different permutation τ_L for each level L; one can show using the triangle inequality that this alignment is successful for all large enough $L \leq L_{\max}$ with probability tending to 1.

General measure spaces. The proofs of Theorems 2 and 3 essentially only use the assumption that μ is Lebesgue measure on $\mathbb{R}$ or counting measure on $\mathbb{Z}$ in showing Lemma 12, so that versions of these theorems continue to hold on general (metric) measure spaces after adjusting Assumption B appropriately. Theorem 4 readily generalises to μ being any discrete measure of known support. The proof of Theorem 5 uses kernel density estimation techniques, and in principle it should be possible to prove a version of this result in any setting where kernel-type estimators with suitable properties exist – for example, using results from Cleanthous et al. (2020), on manifolds.

4.5 Other HMM settings

Wei et al. (2009) adapt HMM methods to more realistically model non-homogeneities of genetic data. More generally, the methodology of empirical Bayes multiple testing could be applied to other hidden Markov settings, such as seasonal hidden Markov models where nonparametric identifiability is proved in Touron (2019), or hidden Markov models with covariates as described in Zucchini et al. (2016, Chapter 10) (see also the references therein for applications). In such extensions, theoretical grounding would rely on a good control of the estimators, and on the propagation of errors in the posterior probabilities when plugging estimates of the transition probabilities and of the emission densities. Note also that the same posterior-based procedure as described herein has been investigated through simulations for hidden Markov random fields in Shu et al. (2015), showing better performances than usual multiple testing procedures.

5. Proofs

5.1 Proofs: FDR Control and TDR Optimality

The following lemma isolates part of the proof of Theorems 2 and 3, showing that $\hat{\ell}_i(X)$ converges to $\ell_i(X)$ at a rate slightly slower than the convergence rate ε_N of the estimators $\hat{H}$. Recalling the sketch proofs in Section 2.2, this lemma will be essential in obtaining bounds on the FDR and TDR.

Lemma 9. *In the setting of Theorem 2 define $\varepsilon'_N = \varepsilon_N(\log N)^u$, and recall that by definition $u > 1+\nu^{-1}$ and by assumption $\varepsilon'_N \to 0$, where ν is the parameter of Assumption B. Then*

$$\max_{i\le N} \Pi_H(|\hat{\ell}_i(X)-\ell_i(X)| > \varepsilon'_N) \to 0, \quad \text{as } N\to\infty. \tag{20}$$

Consequently, there exists $\delta_N \to 0$ such that

$$\Pi_H(\#\{i\le N : |\hat{\ell}_i(X)-\ell_i(X)| > \varepsilon'_N\} > N\delta_N) \to 0.$$

Proof We begin by showing that $\Pi_H(|\hat{\ell}_i(X)-\ell_i(X)| > M\varepsilon'_N) \to 0$ for each fixed i, for some constant $M = M(\mathcal{I})$. [Recall that a constant $M(\mathcal{I})$ depends only on certain bounds for the parameter $H = (Q,\pi,f_0,f_1)$ as described in Section 4.3.]

Let $(E_N)_N$ be a sequence of events with probability tending 1 on which

$$\max\Big(\|\hat{Q}-Q\|_F, \|\hat{\pi}-\pi\|, \max_{j\in\{0,1\}}\|\hat{f}_j - f_j\|_\infty\Big) \le \varepsilon_N,$$

and define

$$\delta = \min_{i,j} Q_{i,j}, \quad \rho = (1-2\delta)/(1-\delta),$$
$$\hat{\delta} = \min_{i,j} \hat{Q}_{i,j}, \quad \hat{\rho} = (1-2\hat{\delta})/(1-\hat{\delta}).$$

Then Proposition 2.2 of De Castro et al. (2017) yields that for some C depending only on a lower bound for δ,

$$\begin{aligned}|\hat{\ell}_i(X)-\ell_i(X)| \le C\Big\{&\rho^{i-1}\|\hat{\pi}-\pi\| + \big[(1-\rho)^{-1} + (1-\hat{\rho})^{-1}\big]\|\hat{Q}-Q\|_F + \\ &\sum_{n=1}^N ((\hat{\rho}\vee\rho)^{|n-i|}/f_\pi(X_n)) \max_{j=0,1}|\hat{f}_j(X_n)-f_j(X_n)|\Big\}.\end{aligned} \tag{21}$$

[The proposition there is stated with $c_*(x) := \min_{j=0,1}\sum_k Q_{jk}f_k(x)$ in place of $f_\pi(x)$, but we note $c_*(x)$ so defined is lower bounded by $\delta f_\pi(x)$. Also note that the authors assume that f_0, f_1

are densities with respect to *Lebesgue* measure, but this assumption is not used in the proof of Proposition 2.2 therein.] Recalling we assumed that δ was (strictly) positive, we see that on E_N, for N large enough we have $\hat{\delta} > \tilde{\delta} = \delta/2$, $\hat{\rho} < \tilde{\rho} = (1+\rho)/2$, and we replace $\rho, \hat{\rho}$ and $\delta, \hat{\delta}$ in (21) by $\tilde{\rho} < 1$ and $\tilde{\delta} > 0$. On the event E_N, choosing the constant $M = M(\tilde{\delta}, \tilde{\rho}, C) = M(\mathcal{I})$ large enough we see by a union bound that

$$\Pi_H(|\hat{\ell}_i(X) - \ell_i(X)| > M\varepsilon'_N) \leq \Pi_H(E_N^c) + \Pi_H\Big(\varepsilon_N \sum_{n=1}^N \frac{\tilde{\rho}^{|n-i|}}{f_\pi(X_n)} > \varepsilon'_N\Big).$$

For $\kappa > 0$ to be chosen, define $S_{\kappa,i} = \{n \leq N : |n-i| \leq \kappa \log N\}$. We can split the terms in $S_{\kappa,i}$ from those in $S_{\kappa,i}^c$ to see, for $C' = 2\sum_{n=0}^\infty \tilde{\rho}^n < \infty$, that

$$\sum_{n \leq N} \frac{\tilde{\rho}^{|n-i|}}{f_\pi(X_n)} \leq C'\Big[\max_{n \in S_{\kappa,i}}\Big(\frac{1}{f_\pi(X_n)}\Big) + \tilde{\rho}^{\kappa \log N} \max_{n \leq N}\Big(\frac{1}{f_\pi(X_n)}\Big)\Big],$$

so that again appealing to a union bound, it suffices to show

$$\Pi_H\Big(\max_{n \in S_{\kappa,i}}\Big(\frac{1}{f_\pi(X_n)}\Big) > \frac{1}{2C'}(\varepsilon'_N/\varepsilon_N)\Big) \to 0, \text{ and} \tag{22}$$

$$\Pi_H\Big(\tilde{\rho}^{\kappa \log N} \max_{n \leq N}\Big(\frac{1}{f_\pi(X_n)}\Big) > \frac{1}{2C'}(\varepsilon'_N/\varepsilon_N)\Big) \to 0. \tag{23}$$

Lemma 12 (in Appendix A.1) tells us that for any $a > 1+\nu^{-1}$, with ν the constant of Assumption B, we have $\Pi_H(\max_{i \leq R} 1/f_\pi(X_i) > R^a) \to 0$ as $R \to \infty$. By stationarity of the process X, taking $R = |S_{\kappa,i}| \leq (2\kappa \log N + 1)$ we deduce that

$$\Pi_H\Big(\max_{n \in S_{\kappa,i}} \frac{1}{f_\pi(X_n)} > (2\kappa \log N + 1)^a\Big) \to 0.$$

Recalling that $\varepsilon'_N/\varepsilon_N > (\log N)^u$, we see that (22) holds for all κ if $u > a$. Next we apply Lemma 12 with $R = N$ to see

$$\Pi_H\Big(\max_{n \leq N}\Big(\frac{1}{f_\pi(X_n)}\Big) > N^a\Big) \to 0.$$

Noting that $\tilde{\rho}^{-\kappa \log N} = N^{\kappa \log(1/\tilde{\rho})}$ and choosing $\kappa > a(\log 1/\tilde{\rho})^{-1}$ yields (23). This concludes the proof that for some constant M and each $i \leq N$, $\Pi_H(|\hat{\ell}_i(X) - \ell_i(X)| > M\varepsilon'_N) \to 0$.

To see that $\max_i \Pi_H(|\hat{\ell}_i(X) - \ell_i(X)| > \varepsilon'_N) \to 0$, we note that by initially considering ε'_n defined for some $u' < u$ we can remove the constant M. Thanks to stationarity of the HMM X, we further note that

$$\max_{i \leq N} \Pi_H\Big(\max_{n \in S_{\kappa,i}} \frac{1}{f_\pi(X_n)} > (2\kappa \log N + 1)^a\Big) \to 0;$$

then, since the other terms in the calculations above do not depend on i, we deduce (20).

Finally, defining

$$\delta_N = \Big(\max_{i \leq N} \Pi_H(|\hat{\ell}_i(X) - \ell_i(X)| > \varepsilon'_N)\Big)^{1/2},$$

we appeal to Markov's inequality to see that

$$\begin{aligned}\Pi_H(\#\{i \leq N : |\hat{\ell}_i(X) - \ell_i(X)| > \varepsilon'_N\} > N\delta_N) &\leq \frac{1}{N\delta_N}\sum_{i=1}^N \Pi_H\big(|\hat{\ell}_i(X) - \ell_i(X)| > \varepsilon'_N\big) \\ &\leq \delta_N^{-1} \max_{i \leq N} \Pi_H(|\hat{\ell}_i(X) - \ell_i(X)| > \varepsilon'_N) = \delta_N,\end{aligned}$$

which tends to zero, concluding the proof. ■

Proof [Proof of Theorem 2] Write $\hat{t} = \operatorname{postFDR}_{\hat{H}}\hat{\varphi}$ and recall we write $\hat{S}_0$ for the rejection set of $\hat{\varphi}$. We have, for any sequences of positive numbers ε'_N and of events F_N,

$$\begin{aligned}|\mathrm{FDR}_H(\hat{\varphi}) - E_H\hat{t}| &= \left|E_{X\sim\Pi_H}[\operatorname{postFDR}_H(\hat{\varphi}) - \operatorname{postFDR}_{\hat{H}}(\hat{\varphi})]\right| \\ &\le E_H\Big[\frac{\sum_{i=1}^N |\ell_i(X) - \hat{\ell}_i(X)| \mathbb{1}\{i \in \hat{S}_0\}}{1 \vee |\hat{S}_0|}\Big] \\ &\le \varepsilon'_N + \Pi_H(F_N^c) + E_H\Big[\mathbb{1}_{F_N}\frac{\sum_{i=1}^N \mathbb{1}\{|\ell_i(X) - \hat{\ell}_i(X)| > \varepsilon'_N\}}{1 \vee |\hat{S}_0|}\Big],\end{aligned}$$

where we have used that $|\ell_i(X) - \hat{\ell}_i(X)| \le 1$ for all i. Lemma 13 in Appendix A.1 shows that $E_H[\hat{t}] \to \min(t, \pi_0)$, so that it is enough to show the right side tends to zero for suitable ε'_N and F_N.

Lemma 14 tells us that $\Pi_H(|\hat{S}_0| > aN) \to 1$ for some $a > 0$. Combining with Lemma 9 by a union bound, we deduce that for suitably chosen $\varepsilon'_N \to 0$, $\delta_N \to 0$ and $a > 0$, we have $\Pi_H(F_N^c) \to 0$ if we define

$$F_N = \big\{\#\{i \le N : |\hat{\ell}_i(X) - \ell_i(X)| > \varepsilon'_N\} \le N\delta_N\big\} \cap \big\{|\hat{S}_0| > aN\big\}.$$

Then

$$E_H\Big[\mathbb{1}_{F_N}\frac{\sum_{i=1}^N \mathbb{1}\{|\ell_i(X) - \hat{\ell}_i(X)| > \varepsilon'_N\}}{1 \vee |\hat{S}_0|}\Big] \le \frac{N\delta_N}{aN} \to 0,$$

yielding the result. ■

The following lemma, mentioned already in the sketch proof in Section 2.2, will help us in proving Theorem 3.

Lemma 10. *Under the assumptions of Theorem 3, define $\lambda^* \in (t, 1]$ implicitly by*

$$E_H[\ell_i^\infty(X) \mid \ell_i^\infty(X) < \lambda^*] = \min(t, \pi_0),$$

where ℓ_i^∞ is as in (19) *(by stationarity the conditional expectation does not depend on i).*

Such a λ^* exists; it satisfies, for $\varepsilon > 0$,

$$\begin{aligned}&E_H[\ell_i^\infty(X) \mid \ell_i^\infty(X) < \lambda^* - \varepsilon] < \min(t, \pi_0),\\ &E_H[\ell_i^\infty(X) \mid \ell_i^\infty(X) < \lambda^* + \varepsilon] > t \qquad\qquad \text{if } t < \pi_0;\end{aligned}$$

and we have

$$\hat{\lambda} \to \lambda^* \quad \text{in probability as } N \to \infty. \tag{24}$$

Proof Lemma 17 (in Appendix A.2) tells us that under the assumptions of Theorem 3, the distribution function of ℓ_i^∞ is continuous and strictly increasing. Lemma 18 then tells us that the same is true of the map $\lambda \mapsto E_H[\ell_i^\infty \mid \ell_i^\infty < \lambda]$, and that $E_H[\ell_i^\infty \mid \ell_i^\infty < t] < t$. Noting also that $E_H[\ell_i^\infty \mid \ell_i^\infty < 1] = E_H[\ell_i^\infty] = \pi_0$ (since $\ell_i^\infty < 1$ with probability 1), we deduce the existence of a unique solution $\lambda^* \in (t, 1]$ by the intermediate value theorem. Strict monotonicity of the conditional expectation implies the claimed inequalities when conditioning on $\ell_i^\infty < \lambda^* - \varepsilon$ and on $\ell_i^\infty < \lambda^* + \varepsilon$.

For the convergence in probability, we show for $\varepsilon > 0$ arbitrary that with probability tending to 1 we have $\operatorname{postFDR}_{\hat{H}}(\varphi_{\lambda^*-\varepsilon,\hat{H}}) < t$. We omit the almost identical proof that for $t < \pi_0$ we have $\operatorname{postFDR}_{\hat{H}}(\varphi_{\lambda^*+\varepsilon,\hat{H}}) > t$. From these two bounds one deduces that $\hat{\lambda} \in (\lambda^* - \varepsilon, \lambda^* + \varepsilon)$, implying (24).

By Lemma 19, there exist $\xi_N, \delta_N \to 0$, such that with probability tending to 1

$$\#\{i : 1 \leq i \leq N,\ |\hat{\ell}_i(X) - \ell_i^\infty(X)| > \xi_N\} \leq N\delta_N,$$

and we observe that

$$\begin{aligned}
\mathrm{postFDR}_{\hat{H}}(\varphi_{\lambda^*-\varepsilon,\hat{H}}) &= \frac{\sum \hat{\ell}_i \mathbb{1}\{\hat{\ell}_i < \lambda^* - \varepsilon\}}{1 \vee (\sum \mathbb{1}\{\hat{\ell}_i < \lambda^* - \varepsilon\})} \\
&\leq \frac{\sum \hat{\ell}_i \mathbb{1}\{\hat{\ell}_i < \lambda^* - \varepsilon, |\hat{\ell}_i - \ell_i^\infty| \leq \xi_N\}}{1 \vee (\sum \mathbb{1}\{\hat{\ell}_i < \lambda^* - \varepsilon, |\hat{\ell}_i - \ell_i^\infty| \leq \xi_N\})} + \frac{\#\{i : |\hat{\ell}_i - \ell_i^\infty| > \xi_N\}}{\#\{i : \hat{\ell}_i < \lambda^* - \varepsilon\}} \\
&\leq \frac{\sum \ell_i^\infty \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon + \xi_N\}}{1 \vee (\sum \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon - \xi_N, |\hat{\ell}_i - \ell_i^\infty| \leq \xi_N\}} + \xi_N + \frac{\#\{i : |\hat{\ell}_i - \ell_i^\infty| > \xi_N\}}{\#\{i : \hat{\ell}_i < \lambda^* - \varepsilon\}}.
\end{aligned}$$

Since $\lambda^* > t$ (the proof of) Lemma 14 implies that for some $c > 0$ and for $\varepsilon > 0$ small enough, $\#\{i : \hat{\ell}_i < \lambda^* - \varepsilon\} > cN$ with probability tending to 1. We also lower bound the denominator in the first term of the final line by $\#\{i : \ell_i^\infty < \lambda^* - \varepsilon - \xi_N\} - \#\{i : |\hat{\ell}_i - \ell_i^\infty| > \xi_N\}$; for ε, ξ_N, c' small enough note that $\#\{i : \ell_i^\infty < \lambda^* - \varepsilon - \xi_N\} > c'N$ with probability tending to 1 by ergodicity (i.e. applying Lemma 20 with $g(x) = \mathbb{1}\{x < \lambda^* - \varepsilon - \xi\}$ for some $\xi > \xi_N$), using that $\Pi_H(\ell_i^\infty < \lambda^* - \varepsilon - \xi_N) > 0$. It follows that for an event C_N of probability tending to 1, $\mathrm{postFDR}_{\hat{H}}(\varphi_{\lambda^*-\varepsilon,\hat{H}})$ is upper bounded by

$$\begin{aligned}
&\mathbb{1}_{C_N^c} + \frac{\sum \ell_i^\infty \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon + \xi_N\}}{\sum \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon - \xi_N\}} \Big(1 + O\Big(\frac{\#\{i : |\hat{\ell}_i - \ell_i^\infty| > \xi_N\}}{\#\{i : \ell_i^\infty < \lambda^* - \varepsilon - \xi_N\}}\Big)\Big) + \xi_N + \delta_N/c \\
\leq &\frac{\sum \ell_i^\infty \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon + \xi_N\}}{\sum \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon - \xi_N\}} + o_p(1).
\end{aligned}$$

Again using the ergodicity result Lemma 20, we have that, for fixed $\xi > 0$,

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \ell_i^\infty \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon + \xi\} &\to E_H[\ell_i^\infty \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon + \xi\}] \quad \text{in probability,} \\
\frac{1}{N}\sum_{i=1}^N \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon - \xi\} &\to \Pi_H(\ell_i^\infty < \lambda^* - \varepsilon - \xi) > 0 \quad \text{in probability,}
\end{aligned}$$

so that we may apply Slutsky's lemma (e.g. van der Vaart, 1998, Lemma 2.8) to deduce that for N large enough

$$\begin{aligned}
\frac{\sum_{i=1}^N \ell_i^\infty \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon + \xi_N\}}{\sum_{i=1}^N \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon - \xi_N\}} &\leq \frac{\sum_{i=1}^N \ell_i^\infty \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon + \xi\}}{\sum_{i=1}^N \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon - \xi\}} \\
&\leq \frac{E_H[\ell_i^\infty \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon + \xi\}]}{\Pi_H(\ell_i^\infty < \lambda^* - \varepsilon - \xi)} + o_p(1).
\end{aligned}$$

Finally we note that

$$\frac{E_H[\ell_i^\infty \mathbb{1}\{\ell_i^\infty < \lambda^* - \varepsilon + \xi\}]}{\Pi_H(\ell_i^\infty < \lambda^* - \varepsilon - \xi)} = E_H\big[\ell_i^\infty \mid \ell_i^\infty < \lambda^* - \varepsilon + \xi\big] \frac{\Pi_H(\ell_i^\infty < \lambda^* - \varepsilon + \xi)}{\Pi_H(\ell_i^\infty < \lambda^* - \varepsilon - \xi)}.$$

Uniformly in ξ satisfying $0 < \xi < \varepsilon/2$, we have by monotonicity

$$E_H[\ell_i^\infty \mid \ell_i^\infty < \lambda^* - \varepsilon + \xi] \leq E_H[\ell_i^\infty \mid \ell_i^\infty < \lambda^* - \varepsilon/2] < \min(t, \pi_0).$$

Observe also that $\Pi_H(\lambda^*-\varepsilon-\xi \leq \ell_i^\infty < \lambda^*-\varepsilon+\xi) \to 0$ as $\xi \to 0$ by the continuity of the distribution function of ℓ_i^∞, while $\Pi_H(\ell_i^\infty < \lambda^* - \varepsilon - \xi)$ is bounded away from zero for $\lambda^* - \varepsilon - \xi$ bounded away from zero. It follows that by choosing $\xi = \xi(\varepsilon)$ small enough we may ensure that

$$E_H[\ell_i^\infty \mid \ell_i^\infty < \lambda^* - \varepsilon + \xi] \frac{\Pi_H(\ell_i^\infty < \lambda^* - \varepsilon + \xi)}{\Pi_H(\ell_i^\infty < \lambda^* - \varepsilon - \xi)} < \min(t, \pi_0).$$

We conclude, as claimed, that $\mathrm{postFDR}_{\hat{H}}(\varphi_{\lambda^*-\varepsilon,\hat{H}}) < \min(t, \pi_0)$ with probability tending to 1. ■

Proof [Proof of Theorem 3] Define λ^* as in Lemma 10. In the case $\lambda^* = 1$, one shows that $\hat{\varphi}$ rejects all but $o_p(N)$ of the hypotheses. It follows that, asymptotically, its mTDR is close to that of the procedure which rejects all null hypotheses, which trivially has the best mTDR of any procedure. We omit the proof details in this case and henceforth assume that $\lambda^* < 1$, or equivalently (in view of Lemma 10) that $t < \pi_0$.

We compare $\hat{\varphi}$ to the 'oracle' procedure $\varphi_{\lambda^*,H}$, which we will argue has optimal multiple testing properties. For $\varepsilon_N > 0$ we may decompose

$$\mathbb{1}\{\ell_i < \lambda^*\} \leq \mathbb{1}\{\lambda^* - \varepsilon_N \leq \ell_i < \lambda^*\} + \mathbb{1}\{\hat{\ell}_i < \hat{\lambda}\} + \mathbb{1}\{\hat{\lambda} < \lambda^* - \varepsilon_N/2\} + \mathbb{1}\{\hat{\ell}_i - \ell_i > \varepsilon_N/2\}.$$

Lemma 10 tells us that $\hat{\lambda}$ tends to λ^* in probability, so that $\Pi_H(\hat{\lambda} < \lambda^* - \varepsilon_N/2) \to 0$ for ε_N tending to zero slowly enough, and Lemma 9 tells us that $\#\{i : |\hat{\ell}_i - \ell_i| > \varepsilon_N/2\}/N \to 0$ in probability, again for ε_N tending to zero slowly enough. Lemma 19 tells us that there exist $\xi_N \to 0$ such that $\#\{i : |\ell_i - \ell_i^\infty| > \xi_N\}/N \to 0$ in probability, and Lemma 17 tells us that under the conditions of Theorem 3 the distribution function of ℓ_i^∞ is continuous, so that as $N \to \infty$

$$\begin{aligned} E_H\#\{i : \lambda^* - \varepsilon_N \leq \ell_i < \lambda^*\}/N &\leq E_H\#\{i : |\ell_i - \ell_i^\infty| > \xi_N\}/N + \Pi_H(\lambda^* - \varepsilon_N - \xi_N \leq \ell_i^\infty < \lambda^* + \xi_N) \\ &\to 0. \end{aligned}$$

We deduce that

$$E_H[\#\{i : \theta_i = 1, \hat{\ell}_i < \hat{\lambda}\}] \geq E_H[\#\{i : \theta_i = 1, \ell_i < \lambda^*\}] - o(N),$$

so that, dividing each side by $E_H\#\{i : \theta_i = 1\} = N\pi_1$,

$$\mathrm{mTDR}_H(\hat{\varphi}) \geq \mathrm{mTDR}_H(\varphi_{\lambda^*,H}) - o(1).$$

Next we consider the mFDR. A similar decomposition to those above yields that

$$\begin{aligned} E_H[\#\{i : \theta_i = 0, \hat{\ell}_i < \hat{\lambda}\}] &\leq E_H[\#\{i : \theta_i = 0, \ell_i < \lambda^*\}] + o(N), \\ E_H[\#\{i : \hat{\ell}_i < \hat{\lambda}\}] &\geq E_H[\#\{i : \ell_i < \lambda^*\}] - o(N). \end{aligned}$$

One also has (by comparison to ℓ_i^∞ as above, or by adapting the proof of Lemma 14) that $E_H\#\{i : \ell_i < \lambda^* + \varepsilon_N\} \geq cN$ for some $c > 0$, so that a Taylor expansion yields

$$\mathrm{mFDR}_H(\hat{\varphi}) \leq \frac{E_H\#\{i : \theta_i = 0, \ell_i < \lambda^*\} + o(N)}{E_H\#\{i : \ell_i < \lambda^*\} - o(N)} \leq \mathrm{mFDR}_H(\varphi_{\lambda^*,H}) + o(1).$$

Define $g(x) = \sup\{\mathrm{mTDR}_H(\psi) : \mathrm{mFDR}_H(\psi) \leq x\}$. Trivially $\mathrm{mTDR}_H(\hat{\varphi}) \leq g(\mathrm{mFDR}_H(\hat{\varphi}))$, and hence the following chain of equalities (justified below) proves the first claim of the theorem:

$$\begin{aligned} \mathrm{mTDR}_H(\hat{\varphi}) &\geq \mathrm{mTDR}_H(\varphi_{\lambda^*,H}) - o(1) \\ &\geq g\big(\mathrm{mFDR}_H(\varphi_{\lambda^*,H})\big) - o(1) \\ &\geq g\big(\mathrm{mFDR}_H(\hat{\varphi}) - o(1)\big) - o(1) \\ &\geq g\big(\mathrm{mFDR}_H(\hat{\varphi})\big) - o(1). \end{aligned}$$

The first line was proved above. The second is a consequence of an optimality property for the class $(\varphi_{\lambda,H} : \lambda \in [0,1])$ given by Lemma 22 in Appendix A.2. The third line then follows from what was proved above, and the final line follows by a continuity-type result for g given by Lemma 23.

It remains to prove the second claim of the theorem. This will follow, with the same arguments as above, from proving that $\mathrm{mFDR}_H(\varphi_{\lambda^*,H}) \geq t - o(1)$. Observe that, using Lemma 19 as above, one can show

$$\begin{aligned}
E_H[\sum_{i\leq N} \ell_i \mathbb{1}\{\ell_i < \lambda^*\}] &= E_H \sum_{i\leq N} [\ell_i^\infty \mathbb{1}\{\ell_i^\infty < \lambda^*\}] + o(N)\\
E_H[\sum_{i\leq N} \mathbb{1}\{\ell_i < \lambda^*\}] &= E_H[\sum_{i\leq N} \mathbb{1}\{\ell_i^\infty < \lambda^*\}] + o(N).
\end{aligned}$$

Stationarity of the HMM implies that

$$\begin{aligned}
E_H[\sum_{i\leq N} \ell_i^\infty \mathbb{1}\{\ell_i^\infty < \lambda^*\}] &= N E_H[\ell_1^\infty \mid \ell_1 <^\infty \lambda^*] \Pi_H(\ell_1^\infty < \lambda^*),\\
E_H[\sum_{i\leq N} \mathbb{1}\{\ell_i^\infty < \lambda\}] &= N \Pi_H(\ell_1^\infty < \lambda^*),
\end{aligned}$$

and hence by definition of λ^* (recall we have assumed $t < \pi_0$)

$$E_H \sum_{i\leq N} (\ell_i^\infty - t) \mathbb{1}\{\ell_i^\infty < \lambda^*\} = 0.$$

Returning to the ℓ-values themselves and using also Lemma 17 to see that $\Pi_H(\ell_1^\infty < \lambda^*) > 0$, we deduce that

$$\begin{aligned}
N^{-1} E_H[\sum_{i\leq N} (\ell_i - t) \mathbb{1}\{\ell_i < \lambda^*\}] &\to 0,\\
N^{-1} E_H \sum_{i\leq N} \mathbb{1}\{\ell_i < \lambda^*\} &\to \Pi_H(\ell_1^\infty < \lambda^*) > 0,
\end{aligned}$$

and we may rearrange to see that $\mathrm{mFDR}_H(\varphi_{\lambda^*,H}) \geq t - o(1)$. ■

5.2 Proofs: Supremum Norm Estimation of HMM Parameters

We construct the estimators of Theorem 5 using a spectral kernel density estimation method. As usual, this involves approximating f by its convolution with a 'mollifier' K_L and estimating this convolution; the level $L = L_n$ governs how close the kernel K_L is to a Dirac mass and hence the tradeoff between the bias and the variance.

Let K be a bounded Lipschitz-continuous function, supported in $[-1, 1]$, such that if we define

$$\begin{aligned}
K_L(x,y) &= 2^L K(2^L(x-y)),\\
K_L[f](x) &= \int K_L(x,y) f(y) \,\mathrm{d}y,
\end{aligned} \tag{25}$$

then we have, for any $f \in C^s(\mathbb{R})$,

$$\|f - K_L[f]\|_\infty \leq C \|f\|_{C^s} 2^{-Ls}. \tag{26}$$

Note that such a function, a 'bounded convolution kernel of order s', exists, see Tsybakov (2009) (in particular, to ensure K is Lipschitz, one builds the kernel using a Gegenbauer basis with parameter $\alpha > 1$ as in Section 1.2.2 thereof). We also note here that for some $C = C(\mathcal{H})$,

$$\max_j \|K_L[f_j]\|_\infty \leq 2\|K\|_\infty \max_j \|f_j\|_\infty \leq C \tag{27}$$

since $\int_{-1}^{1} |K(x)| \,\mathrm{d}x \leq 2\|K\|_\infty$. [Recall that a constant $C = C(\mathcal{H})$ depends only on certain bounds for the parameter $H = (Q, \pi, f_1, \dots, f_J)$ as described in Section 4.3. Recall also that, as with the above, we allow such a constant to also depend on the kernel K since this kernel can be chosen independent of H, and similarly it may depend on the choice of functions $h_1, \dots, h_{L_0}$ and of sets $\mathbb{D}_N$ in Algorithm 1 below.]

To fix ideas, consider the case $J = 2$, suppose $s = 1$, and for simplicity of exposition imagine that we do not require K to be Lipschitz. Then it is straightforward to show that $K(x) = (1/2)\mathbb{1}\{-1 \leq x \leq 1\}$ defines a suitable kernel, satisfying (26) with $C = 1$, for which $K_L[f](x) = \int_{-1}^{1} (1/2) f(x + 2^{-L} z) \,\mathrm{d}z$ is a local average of f. For kernel density estimation with data $X_i \overset{iid}{\sim} f$, $i \leq N$ we would now estimate $K_L[f](x)$, and hence $f(x)$ itself, by $(1/2N)\sum_{i=1}^{N} \mathbb{1}\{x - 2^{-L} \leq X_i \leq x + 2^{-L}\}$. Here we instead adapt the spectral method from Anandkumar et al. (2012) and Lehéricy (2018) to perform this final estimation step.

Again to fix ideas, suppose that $P_{X\sim f_0}(X \in [-1,1]) \neq P_{X\sim f_1}(X \in [-1,1])$. We may straightforwardly estimate empirically the joint distributions, starting from stationarity, of X_1, X_2, X_3. Since X_1 and X_3 are independent conditional on X_2, it is convenient to focus on the distribution of X_2. Define 2×2 matrices M^x, $x \in \mathbb{R}$ and P as follows.

$$M^x = E_H \begin{pmatrix} K_L(x, X_2) & K_L(x, X_2)\mathbb{1}\{X_3 \in [-1,1]\} \\ K_L(x, X_2)\mathbb{1}\{X_1 \in [-1,1]\} & K_L(x, X_2)\mathbb{1}\{X_1 \in [-1,1], X_3 \in [-1,1]\} \end{pmatrix},$$

$$P = E_H \begin{pmatrix} 1 & \mathbb{1}\{X_3 \in [-1,1]\} \\ \mathbb{1}\{X_1 \in [-1,1]\} & \mathbb{1}\{X_1 \in [-1,1], X_3 \in [-1,1]\} \end{pmatrix}.$$

It is clear that M^x, P can be estimated by empirical averages. Since $P_{X\sim f_0}(X \in [-1,1]) \neq P_{X\sim f_1}(X \in [-1,1])$ one can show that the events $X_1 \in [-1,1]$, $X_3 \in [-1,1]$ are not independent, hence that P is invertible, so that one may define the matrix $B^x = P^{-1}M^x$. Ideas found in Anandkumar et al. (2012) and Lehéricy (2018) (see also the coming lemma) then reveal that the B^x, $x \in \mathbb{R}$ are simultaneously diagonalisable, with eigenvalues $K_L[f_j](x)$, $j = 0, 1$. The pairs of function values $\{f_0(x), f_1(x)\}$ can therefore be estimated by eigenvalues of an empirical version of B^x up to bias terms $|f_j(x) - \int_{-1}^{1} (1/2) f_j(x + 2^{-L})| \,\mathrm{d}z \leq 2^{-L}\|f_j\|_{C^1}$ which are small if $L = L_N$ is large, and variance terms whose magnitude depends on the errors in the empirical approximation of P, M^x and how those propagate to the eigenvalues; these variance terms vanish as N tends to infinity provided $L = L_N$ does not grow too fast. Finally, choosing a single matrix to *simultaneously* approximately diagonalise all B^x allows us to avoid any identifiability issues: if we instead diagonalised each empirical version of B^x individually, we would arrive at a uncountable collection $\{\{\lambda_1(x), \lambda_2(x)\} : x \in \mathbb{R}\}$ of pairs of eigenvalues which we would not necessarily know how to group into a single pair of estimators $\hat{f}_0$, $\hat{f}_1$ consistent up to a permutation, but using a single matrix to approximately diagonalise bypasses this issue.

The following lemma, which adapts ideas found in Anandkumar et al. (2012) and Lehéricy (2018), proves the simultaneous diagonalisability underpinning this spectral method. We write in a more general setting than the description above, allowing once more for $J \geq 2$. For P to be invertible it then becomes essential to consider multiple functions $h_1, \dots, h_{L_0}$ in place of the two functions $x \mapsto 1$ and $x \mapsto \mathbb{1}\{-1 \leq x \leq 1\}$ used in the description above: in principle $L_0 = J$ is sufficient (see the remarks after Algorithm 1), but we allow for $L_0 > J$, which requires the introduction of a matrix V to reduce the dimensions of M^x and P to $J \times J$.

Lemma 11. *For $L_0 \in \mathbb{N}$, let $h_1, \dots h_{L_0}$ be arbitrary functions. Define, for data X from the HMM (1),*

$$M^x \equiv M^{x,L_0,L} := (E_H[h_l(X_1)K_L(x,X_2)h_m(X_3)]_{l,m\le L_0}) \in \mathbb{R}^{L_0\times L_0}, \tag{28}$$

$$P \equiv P^{L_0} := (E_H[h_l(X_1)h_m(X_3)]_{l,m\le L_0}) \in \mathbb{R}^{L_0\times L_0}, \tag{29}$$

$$D^x \equiv D^{x,L} := \operatorname{diag}\big(K_L[f_j](x)_{j\le J}\big) \in \mathbb{R}^{J\times J}, \tag{30}$$

$$O \equiv O^{L_0} := (E_H[h_l(X_1) \mid \theta_1 = j]_{l\le L_0, j\le J}) \in \mathbb{R}^{L_0\times J}. \tag{31}$$

Then

$$M^x = O \operatorname{diag}(\pi) Q D^x Q O^\intercal, \quad \textit{and}$$
$$P = O \operatorname{diag}(\pi) Q^2 O^\intercal.$$

If $V \in \mathbb{R}^{L_0\times J}$ is such that $V^\intercal P V$ is invertible (it suffices to assume PV has rank J, which holds under the assumption that P has rank J if the columns of V consist of orthonormal right singular vectors of P, or any other orthonormal basis of the column space of P) then the matrix

$$B^x \equiv B^{x,L_0,L} := (V^\intercal P V)^{-1} V^\intercal M^x V \tag{32}$$

satisfies

$$B^x = (QO^\intercal V)^{-1} D^x (QO^\intercal V), \tag{33}$$

so that the matrices $(B^x : x \in \mathbb{R})$ are diagonalisable simultaneously*, with B^x having eigenvalues $(D^x_j : j \le J) = (K_L[f_j](x) : j \le J)$.*

Proof Conditioning on $(\theta_1, \theta_2, \theta_3)$, we see

$$\begin{aligned} M^x_{l,m} &= \sum_{a,b,c} \Pi_H(\theta_1 = a, \theta_2 = b, \theta_3 = c) E_H[h_l(X_1)K_L(x,X_2)h_m(X_3) \mid \theta_1 = a, \theta_2 = b, \theta_3 = c] \\ &= \sum_{a,b,c} \pi_a Q_{a,b} Q_{b,c} O_{l,a} O_{m,c} E_{X\sim f_b}[K_L(x,X)] \\ &= (O \operatorname{diag}(\pi) Q D^x Q O^\intercal)_{l,m}, \end{aligned}$$

and similarly we have

$$P = (\sum_{a,b,c} \pi_a Q_{a,b} Q_{b,c} O_{l,a} O_{m,c})_{l,m} = O \operatorname{diag}(\pi) Q^2 O^\intercal.$$

Next, note that if $V^\intercal P V$ is invertible then so is $QO^\intercal V$ (since $V^\intercal P V = V^\intercal O \operatorname{diag}(\pi) Q (QO^\intercal V)$, and a product AB of square matrices is invertible if and only if each of A and B are). The result (33) then follows from the expressions for P and M^x. ■

As discussed prior to the lemma, this result suggests that one estimates the eigenvalues $(K_L[f_j](x) : j \le J)$ of B^x (and hence, using (26), the function values $f_j(x)$, $j \le J$ themselves) by using empirical versions of V, P and M^x, an idea which is implemented in the following algorithm. The algorithm requires as inputs functions $h_1, \dots h_{L_0}$ and sets $\mathbb{D}_N$ with certain properties; the existence of suitable inputs is discussed in the remarks thereafter. We introduce notation for the "eigen-separation" of a diagonalisable matrix $B \in \mathbb{R}^{J\times J}$ with eigenvalues $\lambda_1, \dots, \lambda_J$:

$$\operatorname{sep}(B) = \min_{i\ne j} |\lambda_i - \lambda_j|. \tag{34}$$

Recall that $\sigma_J(B)$ denotes the Jth largest singular value of B.

Algorithm 1 Kernel density estimator

```
input
    • Data $(X_n : n \le N+2)$ drawn from the HMM (1).
    • Functions $h_1, \dots h_{L_0}$, uniformly bounded, such that
      $O = (E_H[h_l(X_1) \mid \theta_1 = j]_{l \le L_0, j \le J})$ is of rank $J$, with $\sigma_J(O)$ bounded
      away from 0 uniformly in $N$, at least for $N$ large enough.
    • Finite sets $\mathbb{D}_N \subseteq \{(a,u) \in \mathbb{R}^{J(J-1)/2} \times \mathbb{R}^{J(J-1)/2} : \sum |a_i| \le 1\}$ such that
      $\max_{(a,u)\in\mathbb{D}_N} \operatorname{sep}(B^{a,u})$ is bounded away from 0 uniformly in $N$, at least
      for $N$ large enough, where $B^{a,u} = \sum a_i B^{u_i}$ for $B^x$ as in Lemma 11 for
      some $V$.
estimate the matrices $P$, $(M^x,\ x \in \mathbb{R})$ of Lemma 11 by taking empirical averages: for
    $L$ such that $2^L \asymp (N/\log N)^{1/(1+2s)}$, define
        $\hat{P} = \hat{P}^{L_0} = (N^{-1} \sum_{n \le N} h_l(X_n) h_m(X_{n+2}))_{l,m \le L_0}$,
        $\hat{M}^x = \hat{M}^{x,L_0,L} = (N^{-1} \sum_{n \le N} h_l(X_n) K_L(x, X_{n+1}) h_m(X_{n+2}))_{l,m \le L_0}$.
    Let $\hat{V} = \hat{V}^{L_0} \in \mathbb{R}^{L_0 \times J}$ be a matrix of orthonormal right singular vectors of $\hat{P}$
    (fail if $\hat{P}$ is of rank less than $J$).
set, for $x \in \mathbb{R}$ and for $a, u \in \mathbb{R}^{J(J-1)/2}$
        $\hat{B}^x = \hat{B}^{x,L_0,L} := (\hat{V}^\intercal \hat{P} \hat{V})^{-1} \hat{V}^\intercal \hat{M}^x \hat{V}, \quad \hat{B}^{a,u} := \sum a_i \hat{B}^{u_i}$.
choose $\hat{R}$ of normalised columns diagonalising $\hat{B}^{\hat{a},\hat{u}}$, where $(\hat{a},\hat{u}) \in \operatorname{argmax}_{\mathbb{D}_N} \operatorname{sep}(\hat{B}^{a,u})$ (fail if
$\hat{B}^{\hat{a},\hat{u}}$ is not diagonalisable).
output $(\hat{f}_j : j \le J)$, where, defining
        $\tilde{f}_j^L(x) = (\hat{R}^{-1} \hat{B}^x \hat{R})_{jj}$,
    we set
        $\hat{f}_j(x) = \begin{cases} \tilde{f}_j^L(x) & |\tilde{f}_j^L(x)| \le N^\alpha \\ N^\alpha \operatorname{sign}(\tilde{f}_j^L(x)) & \text{otherwise,} \end{cases}$
    for $\alpha > 0$ arbitrary. [The in-probability result (17) also holds for $\tilde{f}_j^L$.]
```

Remarks. i. For notational convenience, we have considered observing $N+2$ data points $X_1, \dots, X_{N+2}$ so that we can form N triples of consecutive observations; the proofs go through for the original N data points by adjusting constants.

ii. Under Assumption D, $h_1, \dots h_{L_0}$ can be chosen without knowledge of the parameters, for example by letting L_0 tend to infinity arbitrarily slowly and taking the h_l to be indicator functions of the first L_0 of a countable collection of sets generating the Borel σ-algebra (see Lemma 24, in Appendix B.1). In principle, $L_0 = J$ is sufficient to achieve O of rank J, but without further assumptions, the appropriate functions $h_1, \dots h_J$ will necessarily depend on the unknown parameters. In the case $J = 2$, it suffices to assume in addition to the other conditions of Theorem 5 that $P_{X \sim f_1}(X \in A) \neq P_{X \sim f_2}(X \in A)$ for some known A, by taking $h_1 = 1, h_2 = \mathbb{1}_A$ (as in the example given before Lemma 11, where we took $A = [-1, 1]$).

iii. Lemma 11 implies that the condition on $\mathbb{D}_N$ is independent of V provided V is such that $V^\intercal P V$ is invertible. Lemma 26, the proof of which uses only that $f_1, \dots, f_J$ are distinct, shows that the choice $V = \hat{V}$ is suitable with probability tending to 1 and that $\mathbb{D}_N$ can be chosen independent of the parameters, for example by taking a cartesian product of increasing dyadic sets of rationals. In the case $J = 2$, the description of the algorithm simplifies, in that necessarily $\hat{a} = 1 \in \mathbb{R}^1$. A corresponding simplification also works in the general J state case if one is willing to assume that there exists $x_0 \in \mathbb{R}$ for which the values $f_j(x_0),\ j \leq J$ are all distinct, in that one may define $\hat{R}$ as diagonalising $\hat{B}^{\hat{x}}$ where $\hat{x}$ maximises $\operatorname{sep}(\hat{B}^x)$ over x in (some finite increasing sieve in) $\mathbb{R}$.

iv. Lemmas 25 and 27 prove that, with probability tending to 1, $\hat{P}$ has rank J and $\hat{B}^{\hat{a},\hat{u}}$ is diagonalisable, so that the outputs $\hat{f}_j$ are well-defined.

v. Note that the truncation step in defining $\hat{f}_j$ is not needed for bounds in probability. For bounds in expectation, if we have an a priori bound $\|f_j\|_\infty \leq C_j$, then we may define $\hat{f}_j$ by truncating $\tilde{f}_j$ at $\pm C_j$ rather than at N^α. The choice to truncate at N^α (with α arbitrary) avoids poor performance in expectation which could result from the errors being excessively large on an event of small probability. Since the bounds in probability hold without this truncation, we think in practice not truncating at all would be fine.

vi. Since the f_j are assumed Hölder continuous, and satisfy tail bounds, one could in fact calculate $\hat{f}_j(x)$ only for x in some finite set, then construct estimators $\check{f}_j$ via interpolation, in order to ease computation.

Proof [Proof of Theorem 5] Define M^x, P, D^x, O as in Lemma 11 and construct $\hat{f}_j, \tilde{f}_j^L$ using Algorithm 1. Continuity of the $\hat{f}_j, \tilde{f}_j$ follows from continuity of the map $x \mapsto \hat{B}^x$, which in turn follows from that of the map $x \mapsto \hat{M}^x$, proved in Lemma 25. Observe also that $\|f_j\|_\infty < \infty$ for all $j \leq J$, so that for any τ, for N large enough that $\|f_j\|_\infty \leq N^a$ we have

$$\|\hat{f}_j - f_{\tau(j)}\|_\infty \leq \|\tilde{f}_j^L - f_{\tau(j)}\|_\infty,$$

hence for the in-probability result it suffices to prove (17) with $\tilde{f}_j = \tilde{f}_j^L$ in place of $\hat{f}_j$.

Lemma 25 tells us that $\hat{P}, \hat{M}^X$ estimate P, M^x at the rate r_N and consequently both that the choice $V = \hat{V}$ is suitable in (32), and that the B so constructed is close to $\hat{B}$. Matrix perturbation arguments then yield the result.

Precisely, for a constant $c > 0$, define the event

$$\mathcal{A} = \{\|\hat{P} - P\| \leq cL_0 r_N,\ \|\hat{M}^x - M^x\| \leq cL_0^2 r_N\ \forall x \in \mathbb{R}\}. \tag{35}$$

This is indeed a measurable event, and for suitable $c = c(\kappa, \mathcal{H})$ it has probability at least $1 - N^{-\kappa}$, by Lemma 25, which also tells us that $\hat{V}^\intercal P \hat{V}$ is invertible on $\mathcal{A}$ and that, defining

$$\tilde{B}^x := (\hat{V}^\intercal P \hat{V})^{-1} \hat{V}^\intercal M^x \hat{V},$$

we have, for some C depending on $\mathcal{H}$ and on the constant c of event $\mathcal{A}$,

$$\mathbb{1}_{\mathcal{A}} \sup_{x \in \mathbb{R}} \|\tilde{B}^x - \hat{B}^x\| \leq CL_0^2 r_N.$$

Lemma 11 tells us (on $\mathcal{A}$) that $\tilde{B}^x = (QO^\intercal \hat{V})^{-1} D^x QO^\intercal \hat{V}$, and we write $\tilde{R}$ for a matrix whose columns are those of $QO^\intercal \hat{V}$ but scaled to have unit Euclidean norm, which thus diagonalises $\tilde{B}^x$ for all x. By Lemma 32 (and the remark thereafter), we may assume there exists a permutation τ such that $\|\hat{R} - \tilde{R}_\tau\| \leq CL_0^{7/2} r_N$ on $\mathcal{A}$, where $\tilde{R}_\tau$ is obtained by permuting the columns of $\tilde{R}$ according to τ. Next we apply Lemma 33 with $\mathcal{T} = \mathbb{R}$, $A_x = \tilde{B}^x$, $\hat{A}_x = \hat{B}^x$, $R = \tilde{R}$. Noting that $\|\tilde{R}^{-1}\| \leq C' L_0^{1/2}$ and $\kappa(\tilde{R}) := \|\tilde{R}\| \|\tilde{R}^{-1}\| \leq C' L_0$ for some constant $C' = C'(\mathcal{H})$ (see Lemma 35b), and that the constant $\lambda_{\max}$ of the lemma is bounded by a constant depending only on $\mathcal{H}$ (see (27)), we deduce that

$$\sup_x \max_j |\tilde{f}_j^L(x) - K_L[f_{\tau(j)}](x)| \leq c' L_0 [L_0^2 r_N + L_0^{1/2} L_0^{7/2} r_N] \leq c'' L_0^5 r_N,$$

for some constants c', c''. The in-probability result (17) follows, since the choice of L ensures by (26) that $\|f_{\tau(j)} - K_L[f_{\tau(j)}]\|_\infty \leq C'' r_N$ for some C'', so that for a suitable constant C,

$$\Pi_H(\|\tilde{f}_j - f_{\tau(j)}\|_\infty > C L_0^5 r_N) \leq \Pi_H(\mathcal{A}^c) \leq N^{-\kappa} \to 0.$$

For the in-expectation result (18), observe that by truncating at $\pm N^\alpha$ we have ensured that

$$E_H \|\hat{f}_j - f_{\tau(j)}\|_\infty \leq C L_0^5 r_N + 2N^\alpha \Pi_H(\mathcal{A}^c).$$

Choosing $c = c(\kappa, \mathcal{H})$ in the definition of the event $\mathcal{A}$ corresponding to some $\kappa \geq s/(1+2s) + \alpha$ concludes the proof. ■

Proof [Proof of Proposition 7] Let $\hat{f}_0, \hat{f}_1, \hat{Q}, \hat{\pi}$ be estimators which satisfy

$$\Pi_H(\|\hat{f}_0 - f_{\tau(0)}\| + \|\hat{f}_1 - f_{\tau(1)}\| + \|\hat{Q} - Q_{\sigma,\sigma}\|_F + \|\hat{\pi} - \pi_\sigma\| > C\varepsilon_N) \to 0 \tag{36}$$

for some permutations τ, σ and a constant $C > 0$, with $Q_{\sigma,\sigma}$ defined by permuting the rows and columns of Q, and π_σ defined similarly. The existence of suitable $\hat{f}_0, \hat{f}_1$ is given by Theorem 5, and the existence of suitable $\hat{Q}, \hat{\pi}$ is proved by results in De Castro et al. (2017, Appendix C) (and by arguments as in De Castro et al. 2016, Section 8.6 to accelerate the possibly slow rate to a near-parametric rate). Moreover, the estimators of De Castro et al. (2017) are constructed using a spectral method, so that one may in fact assume $\sigma = \tau$. [One could also "align" σ and τ by hand, by noting that by ergodicity the invariant density f_π can be estimated at the rate r_N using a standard kernel density estimator, and permuting rows and columns of $\hat{Q}$ and $\hat{\pi}$ so that $\sum \hat{\pi}_i \hat{f}_i$ is close to this kernel density estimator; linear independence of the f_i ensures that this alignment method works.]

Next, under the assumption $\pi_0 > \pi_1$, define $\check{f}_j = \hat{f}_{\hat{\tau}(j)}$, $\check{Q} = \hat{Q}_{\hat{\tau},\hat{\tau}}$ and $\check{\pi} = \hat{\pi}_{\hat{\tau}}$, where $\hat{\tau}(0) = 1 - \hat{\tau}(1) = \mathbb{1}\{\hat{\pi}_1 > \hat{\pi}_0\}$. Consistency of $\hat{\pi}$ implies that $\hat{\tau}$ consistently estimates the permutation $\tau = \sigma$ of (36), hence

$$\begin{aligned}
&\Pi_H(\|\check{f}_0 - f_0\| + \|\check{f}_1 - f_1\| + \|\check{Q} - Q\|_F + \|\check{\pi} - \pi\|) > C\varepsilon_N) \\
&\quad \leq \Pi_H(\hat{\tau} \neq \tau) + \Pi_H(\|\hat{f}_0 - f_{\tau(0)}\| + \|\hat{f}_1 - f_{\tau(1)}\| + \|\hat{Q} - Q_{\tau,\tau}\|_F + \|\hat{\pi} - \pi_\tau\| > C\varepsilon_N) \to 0.
\end{aligned}$$

For the other case, we want to define $\hat{\tau}(0) = \mathbb{1}\{\limsup_{x \uparrow x^*} (\hat{f}_0/\hat{f}_1)(x) > 1\}$ and proceed similarly, but the compact support of K means that $\hat{f}_1(x) = \hat{f}_0(x) = 0$ for $x > 2^{-L} + \max_k X_k$, and the right side may be strictly smaller than x^*. Instead, noting that necessarily $\Pi_H(X_1 \leq x^*) > 0$ and assuming without loss of generality that $x^* > 0$, we set $\tilde{X}_n = X_n \mathbb{1}\{X_n \leq x^*\}$ and define

$$\begin{aligned}
\hat{\tau}(0) = 1 - \hat{\tau}(1) &= \mathbb{1}\{\hat{f}_0(M_N) > \hat{f}_1(M_N)\}, \\
M_N &= \max_{i \leq \log N} (\tilde{X}_i);
\end{aligned}$$

note that by construction we have $\hat{f}_{\hat{\tau}(1)}(M_N) \geq \hat{f}_{\hat{\tau}(0)}(M_N)$. We show that $\|\hat{f}_{\hat{\tau}(1)} - f_0\|_\infty > C\varepsilon_N$ on an event A_N of probability tending to 1; it will follow from (36) that $\hat{\tau} \equiv \hat{\tau}^{-1} = \tau$ on A_N, and the result will follow.

The variables $\tilde{X}_i,\ i \leq N$ have a density with respect to the measure μ defined by adding an atom at 0 to Lebesgue measure. Let u be as in Theorem 2, so that $u > 1 + \nu^{-1}$ and $\varepsilon_N (\log N)^u \to 0$ for ν as in Assumption B. The proof of Lemma 12 shows that with probability tending to 1 we have

$f_1(M_N) \geq \min_{i \leq \log N}(f(\tilde{X}_i)) \geq (\log N)^{-u}$, hence $f_1(M_N) > 3C\varepsilon_N$. We also note that $M_N \uparrow x^*$ almost surely, so that $f_1(M_N) > 3f_0(M_N)$ for all N large enough.

Let A_N be an event of probability tending to 1 on which

$$f_1(M_N) > 3C\varepsilon_N, \quad f_1(M_N) > 3f_0(M_N), \quad \|\hat{f}_0 - f_{\tau(0)}\|_\infty \leq C\varepsilon_N, \quad \|\hat{f}_1 - f_{\tau(1)}\|_\infty \leq C\varepsilon_N,$$

whose existence we have just demonstrated. On A_N we have both $\hat{f}_1(M_N) \geq f_{\tau(1)}(M_N) - C\varepsilon_N$ and $\hat{f}_0(M_N) \geq f_{\tau(0)}(M_N) - C\varepsilon_N$ hence (for N large enough)

$$\begin{aligned}\hat{f}_{\hat{\tau}(1)}(M_N) = \max(\hat{f}_0(M_N), \hat{f}_1(M_N)) \geq \max_j(f_j(M_N) - C\varepsilon_N) = f_1(M_N) - C\varepsilon_N > \tfrac{1}{3}f_1(M_N) + C\varepsilon_N \\ > f_0(M_N) + C\varepsilon_N,\end{aligned}$$

so that $\|\hat{f}_{\hat{\tau}(1)} - f_0\|_\infty > C\varepsilon_N$ on A_N as claimed. ■

Acknowledgments

The authors would like to acknowledge Étienne Roquain, Gloria Buritica and Ramon van Handel for fruitful discussions about this work. We thank an action editor and two anonymous referees for their insightful comments, which helped us improve the paper. This work was mostly completed while K.A. was at Université Paris-Saclay, supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH. I.C. and E.G. would like to acknowledge support for this project from Institut Universitaire de France. I.C. is partly supported by ANR-17-CE40-0001 grant (BASICS).

Appendix A. Auxiliary Results for Section 2

A.1 Lemmas for Theorem 2

Recall $f_\pi = \pi_0 f_0 + \pi_1 f_1$ is the density of each X_i, $i \leq N$, in the HMM model (1).

Lemma 12. *Under Assumption B we have, for any $a > 1 + \nu^{-1}$,*

$$\Pi_H(\max_{i \leq R} 1/f_\pi(X_i) > R^a) \to 0 \quad \text{as } R \to \infty.$$

Proof For $A = R^a, B = R^b$ with $a, b > 0$ to be chosen, we have by a union bound and stationarity

$$\begin{aligned}\Pi_H\Big(\max_{i \leq R} \frac{1}{f_\pi(X_i)} > A\Big) &\leq R\Pi_H\big(f_\pi(X_1) < A^{-1}\big) \\ &\leq R\int_{-B}^{B} \mathbb{1}\big\{f_\pi(x) < A^{-1}\big\} f_\pi(x)\, \mathrm{d}\mu(x) + R\Pi_H\big(|X_1| > B\big) \\ &\leq R\mu([-B,B])/A + R\Pi_H\big(|X_1| > B\big).\end{aligned}$$

Since f_π is a mixture of the densities f_0, f_1, an application of Markov's inequality yields

$$\Pi_H\big(|X_1| > B\big) \leq \max_j P_{X\sim f_j}\big(|X| > B\big) \leq B^{-\nu} \max_j E_{X\sim f_j}|X|^\nu,$$

which is at most a constant times $B^{-\nu}$ by the assumption. Choosing $b > 1/\nu$, we have $R\Pi_H(|X_1| \geq B) \to 0$.

Since $B = R^b \geq 1$ and μ is equal to either to Lebesgue or counting measure, $\mu([-B, B]) \leq 2B + 1 \leq 3B$. Then

$$R\mu([-B, B])/A \leq 3R^{1+b-a},$$

which tends to zero for $a > 1 + b$, so that any $a > 1 + \nu^{-1}$ is permissible. ■

For the following two lemmas recall the definition $\hat{S}_0 = \hat{S}_0(t) = \{i : \hat{\varphi}_i = 1\}$, where $\hat{\varphi}$ is as in Definition 1, so that $\hat{K} = |\hat{S}_0|$ is characterised by

$$\frac{1}{\hat{K}} \sum_{i=1}^{\hat{K}} \hat{\ell}_{(i)} \leq t < \frac{1}{\hat{K}+1} \sum_{i=1}^{\hat{K}+1} \hat{\ell}_{(i)}.$$

where, by convention, the left inequality holds if $\hat{K} = 0$, and $\hat{\ell}_{(N+1)} = \infty$ so that the right inequality holds if $\hat{K} = N$. Recall the definition

$$\hat{t} := \operatorname{postFDR}_{\hat{H}}(\hat{\varphi}) = \frac{1}{\hat{K}} \sum_{i=1}^{\hat{K}} \hat{\ell}_{(i)}.$$

Lemma 13. *In the setting of Theorem 2, $E_H \hat{t} \to \min(t, \pi_0)$.*

Proof Since $0 \leq \hat{t} \leq 1$, it's enough to show that $\hat{t} \to \min(t, \pi_0)$ in probability. By Lemma 15, we have

$$\frac{1}{N} \sum_{i=1}^{N} \hat{\ell}_i(X) \to \pi_0 \quad \text{in probability.} \tag{37}$$

By monotonicity of the average of increasing numbers, we have

$$\hat{t} \leq \frac{1}{N} \sum_{i=1}^{N} \hat{\ell}_{(i)} = \frac{1}{N} \sum_{i=1}^{N} \hat{\ell}_i,$$

and by construction we note also that $\hat{t} \leq t$, hence $\hat{t} \leq \min(t, \pi_0) + o_p(1)$.

If $t = 0$ we trivially have the matching lower bound $\hat{t} \geq \min(t, \pi_0) - o_p(1)$. If $t > 0$, we decompose relative to the event $\mathcal{C} = \{\hat{K} = N\}$. Observe, using (37), that

$$\hat{t}\mathbb{1}_{\mathcal{C}} = \mathbb{1}_{\mathcal{C}} \frac{1}{N} \sum_{i=1}^{N} \hat{\ell}_i \geq \mathbb{1}_{\mathcal{C}} \pi_0 - o_p(1).$$

By definition of $\hat{K}$ we also have

$$t\mathbb{1}_{\mathcal{C}^c} < \frac{1}{\hat{K}+1} \sum_{i=1}^{\hat{K}+1} \hat{\ell}_{(i)} \mathbb{1}_{\mathcal{C}^c} = \frac{\hat{K}}{\hat{K}+1} \hat{t}\mathbb{1}_{\mathcal{C}^c} + \frac{\hat{\ell}_{(\hat{K}+1)}}{\hat{K}+1} \mathbb{1}_{\mathcal{C}^c},$$

hence, since $\hat{\ell}_{(\hat{K}+1)} \leq 1$ on $\mathcal{C}^c$,

$$\hat{t}\mathbb{1}_{\mathcal{C}^c} > \frac{\hat{K}+1}{\hat{K}} t\mathbb{1}_{\mathcal{C}^c} - \frac{\hat{\ell}_{(\hat{K}+1)}}{\hat{K}} \mathbb{1}_{\mathcal{C}^c} \geq t\mathbb{1}_{\mathcal{C}^c} - \frac{1}{\hat{K}}.$$

By Lemma 14, $\hat{K} \to \infty$ in probability for any $t > 0$, so that the above display implies $\hat{t}\mathbb{1}_{\mathcal{C}^c} > t\mathbb{1}_{\mathcal{C}^c} - o_p(1)$ and hence

$$\hat{t} > t\mathbb{1}_{\mathcal{C}^c} + \pi_0 \mathbb{1}_{\mathcal{C}} - o_p(1) \geq \min(t, \pi_0) - o_p(1),$$

proving the lower bound. ■

The next lemma shows that $\hat\varphi$ makes, with probability tending to 1, a number of discoveries of order N. The proof goes via comparing $\hat\ell_i$ to some ℓ_i', which closely approximates ℓ_i and allows for the use of ergodicity arguments. Note that one could alternatively compare to ℓ_i^∞ as is done in the proof of Theorem 3 (see also Appendix A.2); by using ℓ_i' instead we avoid the need for Condition G when proving Theorem 2.

Recall the definition of constants $C = C(\mathcal{I})$ from Section 4.3.

Lemma 14. *In the setting of Theorem 2, for all $t > 0$, there exists $a = a(t, \mathcal{I}) > 0$ such that*

$$\Pi_H(|\hat S_0| > aN) \to 1.$$

Proof The definition of $\hat\lambda$ trivially implies $\hat\lambda \geq t$, so that

$$\{i : \hat\ell_i < t\} \subseteq \{i : \hat\ell_i < \hat\lambda\} \subseteq \hat S_0.$$

For $A \in \mathbb{N}$ write

$$\ell_i'(X) := \Pi_H(\theta_i = 0 \mid X_{i-A}, \dots, X_{i+A}), \quad A < i \leq N - A.$$

By Lemma 16, there exist $A = A(t)$ and events G_N of probability tending to 1 such that

$$\Big\{\#\{i : A < i \leq N - A,\ |\hat\ell_i(X) - \ell_i'(X)| > t/2\} \leq N\delta_N\Big\},$$

for some $\delta_N \to 0$. On G_N, we observe that

$$\#\{i \leq N : \hat\ell_i < t\} \geq \#\{i : A < i \leq N - A,\ \ell_i' < t/2\} - N\delta_N,$$

hence it suffices to show that there exists $c > 0$ such that $\#\{i : A < i \leq N - A : \ell_i' < t/2\} > cN$ with probability tending to 1.

By ergodicity (i.e. applying Lemma 20 with $g(x) = \mathbb{1}\{x < t/2\}$) we have for any $\varepsilon > 0$

$$\Pi_H\Big(\#\big\{i : A < i \leq N - A : \ell_i' < t/2\big\} > \big(N - 2A\big)\big(\Pi_H(\ell_i' < t/2) - \varepsilon\big)\Big) \to 1,$$

hence it suffices to show that $\Pi_H(\ell_i' < t/2) \neq 0$.

Fix i satisfying $A < i \leq N - A$. For $\alpha, \beta \in \{0,1\}^A$ write

$$\eta_{\alpha,\beta} = \pi_{\alpha_1} \prod_{a<A} Q_{\alpha_a,\alpha_{a+1}} Q_{\beta_a,\beta_{a+1}}.$$

Introducing the notation $\theta_a^b = (\theta_a, \theta_{a+1}, \dots, \theta_b) \in \mathbb{R}^{b+1-a}$, we note that

$$\Pi_H(\theta_{i-A}^{i+A} = (\alpha, 0, \beta)) = \eta_{\alpha,\beta} Q_{\alpha_A,0} Q_{0,\beta_1}, \quad \Pi_H(\theta_{i-A}^{i+A} = (\alpha, 1, \beta)) = \eta_{\alpha,\beta} Q_{\alpha_A,1} Q_{1,\beta_1}.$$

Define

$$p_0 = \sum_{\alpha,\beta \in \{0,1\}^A} Q_{\alpha_A,0} f_0(X_i) Q_{0,\beta_1} \eta_{\alpha,\beta} \prod_{a \leq A} f_{\alpha_a}(X_{i-A+a-1}) f_{\beta_a}(X_{i+a})$$
$$p_1 = \sum_{\alpha,\beta \in \{0,1\}^A} Q_{\alpha_A,1} f_1(X_i) Q_{1,\beta_1} \eta_{\alpha,\beta} \prod_{a \leq A} f_{\alpha_a}(X_{i-A+a-1}) f_{\beta_a}(X_{i+a}),$$

and observe that

$$\ell_i'(X) = \Pi_H(\theta_i = 0 \mid X_{i-A}^{i+A}) = \frac{p_0}{p_0 + p_1}.$$

Note that each term in the sum defining p_1 is at least $\delta^2 f_1(X_i)/f_0(X_i)$ times the corresponding term in the sum defining p_0, with $\delta > 0$ as in Assumption C, hence

$$p_1 \geq p_0 \delta^2 \frac{f_1(X_i)}{f_0(X_i)}, \quad \text{so that} \quad \ell_i'(X) \leq \frac{1}{1+\delta^2(f_1(X)/f_0(X))}.$$

In view of Assumption A, assume without loss of generality that there exists $x^* \in \mathbb{R} \cup \{\pm\infty\}$ such that $f_1(x)/f_0(x) \to \infty$ as $x \uparrow x^*$. Then we deduce that for some $u = u(t,\delta) > 0$,

$$\Pi_H(\ell_i' < t/2) \geq \Pi_H\Big(\frac{f_1(X_i)}{f_0(X_i)} > \frac{2-t}{t\delta^2}\Big) \geq \pi_1 P_{X\sim f_1}(x^* - u \leq X \leq x^*) > 0,$$

as required. ■

Lemma 15. *In the setting of Theorem 2,*

$$\frac{1}{N}\sum_{i=1}^N \hat{\ell}_i(X) \to \pi_0$$

in probability as $N \to \infty$.

Proof It is required to prove, for $\varepsilon > 0$ arbitrary, that

$$\Pi_H\Big(\Big|\frac{1}{N}\sum_{i=1}^N \hat{\ell}_i(X) - \pi_0\Big| > \varepsilon\Big) \to 0.$$

By Lemma 16, defining

$$\ell_i'(X) = \Pi_H(\theta_i = 0 \mid X_{i-A}, \dots, X_{i+A}), \quad A < i \leq N - A,$$

there exists $A = A(\varepsilon)$ for which, with probability tending to 1,

$$\#\{i : A < i \leq N-A,\ |\hat{\ell}_i(X) - \ell_i'(X)| > \varepsilon/2\} \leq N\delta_N.$$

On the event on which the last line holds we can decompose:

$$\Big|\frac{1}{N}\sum_{i=1}^N \hat{\ell}_i(X) - \pi_0\Big| \leq \frac{2A}{N} + \varepsilon/2 + \delta_N + \frac{1}{N}\Big|\sum_{i=A+1}^{N-A}(\ell_i'(X) - \pi_0)\Big|.$$

Finally, by ergodicity of $\ell_i'(X)$ (see Lemma 20) we have

$$\Pi_H\Big(\frac{1}{N}\Big|\sum_{i=A+1}^{N-A}(\ell_i'(X)-\pi_0)\Big| > \varepsilon/4\Big) \leq \Pi_H\Big(\frac{1}{N-2A}\Big|\sum_{i=A+1}^{N-A}(\ell_i'(X) - E_H[\ell_i'(X)])\Big| > \varepsilon/4\Big) \to 0,$$

where we have used that

$$E_H[\ell_i'(X)] = E_H\Pi_H(\theta_i = 0 \mid X_{i-A}, \dots, X_{i+A}) = \Pi_H(\theta_i = 0) = \pi_0.$$

The result follows. ■

Lemma 16. *For $A \in \mathbb{N}$, define*

$$\ell_i'(X) = \Pi_H(\theta_i = 0 \mid X_{i-A}, \dots, X_{i+A}), \quad A < i \leq N - A.$$

For any fixed $\varepsilon > 0$, there exists $A = A(\varepsilon)$ and $\delta_N \to 0$ such that

$$\#\{i : A < i \leq N - A,\ |\hat{\ell}_i(X) - \ell_i'(X)| > \varepsilon\} \leq N\delta_N, \quad \textit{with probability tending to 1.}$$

A similar result holds in the limit $A \to \infty$, see Lemma 19 below.

Proof Essentially, this is a consequence of Lemma 9 and exponential mixing – hence forgetfulness – of the Markov chain θ, the former telling us that $\hat{\ell}_i \approx \ell_i$ and the latter that ℓ_i is nearly independent of X_j if $|j - i|$ is large so that $\ell_i' \approx \ell_i$. Precisely, Lemma 9 tells us that there exist events G_N of probability tending to 1 on which

$$\Big\{\#\{i \leq N : |\hat{\ell}_i(X) - \ell_i(X)| > \varepsilon_N'\} \leq N\delta_N\Big\},$$

for some $\varepsilon_N' \to 0$; in particular note $\varepsilon_N' < \varepsilon/2$ for N large. Next, we apply Proposition 4.3.23iii of Cappé et al. (2005). Our Assumption C implies that Assumption 4.3.24 therein holds, so by the consequent Lemma 4.3.25 one sees that the $\rho_0(y)$ in the proposition can be replaced by $\rho = (1-2\delta)/(1-\delta)$. Applying the proposition with $j = k - A$ yields

$$|\Pi_H(\theta_k = 0 \mid X_1, \dots, X_n) - \Pi_H(\theta_k = 0 \mid X_{k-A}, \dots, X_n)| < 2\rho^A, \quad k > A.$$

Any two-state Markov chain is reversible, hence by time-reversal we similarly obtain

$$|\Pi_H(\theta_k = 0 \mid X_{k-A}, \dots, X_n) - \Pi_H(\theta_k = 0 \mid X_{k-A}, \dots, X_{k+A})| < 2\rho^A,$$

and hence

$$|\ell_k(X) - \ell_k'(X)| < 4\rho^A, \quad A < k \leq N - A.$$

Choose $A = A(\varepsilon)$ so that $4\rho^A < \varepsilon/2$; then, on G_N and for N large, an application of the triangle inequality yields

$$\#\{i : A < i \leq N - A,\ |\hat{\ell}_i(X) - \ell_i'(X)| > \varepsilon\} \leq N\delta_N,$$

and the result follows. ■

A.2 Lemmas for Theorem 3

We may concretely define ℓ_i^∞ as the almost sure limit

$$\ell_i^\infty(X) = \lim_{K\to\infty} \Pi_H(\theta_i = 0 \mid X_{-K}, \dots, X_K); \tag{38}$$

this limit is well defined by a standard martingale convergence theorem.

Lemma 17. *In the setting of Theorem 2, assume that the distribution function of the variable $f_1(X_1)/f_0(X_1)$ is continuous and strictly increasing on $(0, \infty)$. Then the distribution function of $\ell_i^\infty(X)$ is continuous and strictly increasing on $[0, 1]$.*

Note that atomicity of $\ell_i(X)$ relates to that of $f_1(X_i)/f_0(X_i)$, rather than that of X_i itself, since for example the distribution of ℓ_1 is atomic when $N = 1$ if $\Pi_H(f_1(X_1)/f_0(X_1) = c) > 0$ for some constant c. It is therefore unsurprising that the key properties of the distribution of $\ell_i^\infty(X)$ depend on the distribution of the ratio $f_1(X_i)/f_0(X_i)$.

Proof Let G_0 denote the distribution function of $(f_1/f_0)(X_1)$ when $X_1 \sim f_0\mu$ and G_1 the distribution function of $(f_1/f_0)(X_1)$ when $X_1 \sim f_1\mu$.
Define the stationary filter sequence $(\Phi_i^\infty(X))_{i\in\mathbb{Z}}$ by

$$\Phi_i^\infty(X) := \Pi_H(\theta_i = 0 \mid (X_n : n \in \mathbb{Z}, n \leq i)). \tag{39}$$

Using the usual forward-backward equations, see Baum et al. (1970), and taking almost-sure limits one obtains the following forward equation: for each i,

$$\Phi_i^\infty(X) = \frac{[(1-p)\Phi_{i-1}^\infty(X) + q(1-\Phi_{i-1}^\infty(X))]f_0(X_i)}{((1-p)f_0(X_i) + pf_1(X_i))\Phi_{i-1}^\infty(X) + (qf_0(X_i) + (1-q)f_1(X_i))(1-\Phi_{i-1}^\infty(X))}$$

where $p = Q_{01}$ and $q = Q_{10}$, leading to

$$\Phi_i^\infty(X) = \frac{(1-p)\Phi_{i-1}^\infty(X) + q(1-\Phi_{i-1}^\infty(X))}{(1-p+p(f_1/f_0)(X_i))\Phi_{i-1}^\infty(X) + (q+(1-q)(f_1/f_0)(X_i))(1-\Phi_{i-1}^\infty(X))}. \tag{40}$$

That is, if we define $A(\Phi) = (1-p)\Phi + q(1-\Phi)$, then

$$\Phi_i^\infty(X) = \frac{A(\Phi_{i-1}^\infty(X))}{A(\Phi_{i-1}^\infty(X)) + (f_1/f_0)(X_i)(1-A(\Phi_{i-1}^\infty(X)))}. \tag{41}$$

Since conditional on $\Phi_{i-1}^\infty(X)$, X_i has distribution $\left[A(\Phi_{i-1}^\infty(X))f_0(x) + (1-A(\Phi_{i-1}^\infty(X))f_1(x)\right]\mu$, we deduce that $(\Phi_i^\infty(X))_{i\in\mathbb{Z}}$ is a stationary Markov chain (with state space $[0,1]$ and) with transition kernel $K(\Phi, d\Phi')$ given by

$$\begin{aligned} K(\Phi, d\Phi') &= \int \delta_{g(\Phi,x)}(d\Phi')\,[(\Phi(1-p)+(1-\Phi)q)f_0(x) + (\Phi p + (1-\Phi)(1-q))f_1(x)]\,d\mu(x) \\ &= \int \delta_{g(\Phi,x)}(d\Phi')\,[A(\Phi)f_0(x) + (1-A(\Phi))f_1(x)]\,d\mu(x), \end{aligned}$$

where

$$g(\Phi, x) = \frac{A(\Phi)}{A(\Phi) + (f_1/f_0)(x)(1-A(\Phi))}.$$

Then, for each $t \in (0,1)$, we have

$$\Pi_H\left(\Phi_i^\infty(X) \leq t | \Phi_{i-1}^\infty(X)\right) = \Pi_H\left((f_1/f_0)(X_i) \geq \frac{A(\Phi_{i-1}^\infty(X))}{1-A(\Phi_{i-1}^\infty(X))}(1/t-1) \mid \Phi_{i-1}^\infty(X)\right).$$

Recall that $\pi_0 G_0 + \pi_1 G_1$ is assumed to be continuous and strictly increasing on $(0,+\infty)$, and that $\pi_0 > 0$ and $\pi_1 > 0$, so that G_0 and G_1 are both continuous, and on the set where G_0 is not strictly increasing, G_1 is strictly increasing and vice versa. We deduce that

$$\begin{aligned} \Pi_H\left(\Phi_i^\infty(X) \leq t \mid \Phi_{i-1}^\infty(X)\right) = A(\Phi_{i-1}^\infty(X))&\left[1 - G_0\left(\frac{A(\Phi_{i-1}^\infty(X))}{1-A(\Phi_{i-1}^\infty(X))}\left(\frac{1}{t}-1\right)\right)\right] \\ &+ \left(1-A(\Phi_{i-1}^\infty(X))\right)\left[1 - G_1\left(\frac{A(\Phi_{i-1}^\infty(X))}{1-A(\Phi_{i-1}^\infty(X))}\left(\frac{1}{t}-1\right)\right)\right]. \end{aligned}$$

Then $\Phi_i^\infty(X)$ has, conditionally on $\Phi_{i-1}^\infty(X)$, a continuous and strictly increasing distribution function on $(0,1)$. The same holds for $\Phi_i^\infty(X)$ since for all t,

$$\Pi_H\left(\Phi_i^\infty(X) \leq t\right) = E_H[\Pi_H\left(\Phi_i^\infty(X) \leq t \mid \Phi_{i-1}^\infty(X)\right)].$$

That is, $\Phi_i^\infty(X)$ has (conditionally on $\Phi_{i-1}^\infty(X)$ and unconditionally) no atoms and support $(0,1)$.

The same ideas used to derive (40) and (41) allow us to show that for all i,

$$\ell_i^\infty(X) = \frac{(1-p)\Phi_i^\infty(X)\ell_{i+1}^\infty(X)}{A(\Phi_i^\infty(X))} + \frac{p\Phi_i^\infty(X)(1-\ell_{i+1}^\infty(X))}{1-A(\Phi_i^\infty(X))}. \tag{42}$$

Let $C(\Phi) = \frac{\Phi(1-\Phi)}{A(\Phi)(1-A(\Phi))}$ and notice that for any $p, q \in (0,1)$ there exists $a = a(p,q) < 1$ such that for all $\Phi \in (0,1)$, $|1-p-q|C(\Phi) \le a$. Then an easy recursion yields

$$\ell_i^\infty(X) = \frac{p\Phi_i^\infty(X)}{1-A(\Phi_i^\infty(X))} + \sum_{k\ge 1}(1-p-q)^k C(\Phi_i^\infty(X))C(\Phi_{i+1}^\infty(X))\cdots C(\Phi_{i+k-1}^\infty(X))\frac{p\Phi_{i+k}^\infty(X)}{1-A(\Phi_{i+k}^\infty(X))}.$$

Indeed, since for any $\Phi \in (0,1)$, $|1-p-q|C(\Phi) \le a(p,q) < 1$, the series converges almost surely. We see that for each i, $\ell_i^\infty(X)$ is a function of $(\Phi_k^\infty(X))_{k\ge i}$, and we have

$$\ell_i^\infty(X) = \frac{p\Phi_i^\infty(X)}{1-A(\Phi_i^\infty(X))} + (1-p-q)C(\Phi_i^\infty(X))\ell_{i+1}^\infty(X).$$

It follows that for all t,

$$\begin{aligned}&\Pi_H(\ell_i^\infty(X) \le t | \Phi_{i-1}^\infty(X))\\ =&E_H\Big[\Pi_H\Big((1-p-q)C(\Phi_i^\infty(X))\ell_{i+1}^\infty(X) \le t - \frac{p\Phi_i^\infty(X)}{1-A(\Phi_i^\infty(X))} \mid \Phi_i^\infty(X)\Big) \mid \Phi_{i-1}^\infty(X)\Big].\end{aligned} \tag{43}$$

Define the function F_ℓ by

$$F_\ell(t;\Phi_{i-1}^\infty(X)) = \Pi_H\left(\ell_i^\infty(X) \le t | \Phi_{i-1}^\infty(X)\right);$$

note that by stationarity F_ℓ does not depend on i. Then by (43), if $(1-p-q) > 0$, we have

$$F_\ell(t;\Phi_{i-1}^\infty(X)) = E_H\Big[F_\ell\Big(\frac{1}{(1-p-q)C(\Phi_i^\infty(X))}\Big(t - \frac{p\Phi_i^\infty(X)}{1-A(\Phi_i^\infty(X))}\Big);\Phi_i^\infty(X)\Big) \mid \Phi_{i-1}^\infty(X)\Big];$$

that is, for any t and any $\Phi \in (0,1)$,

$$F_\ell(t;\Phi) = \int F_\ell\Big(\frac{1}{(1-p-q)C(x)}\left(t - \frac{px}{1-A(x)}\right);x\Big)K(\Phi,dx). \tag{44}$$

Similarly, if $(1-p-q) < 0$, defining the function $\tilde{F}_\ell$ by $\tilde{F}_\ell(t,\Phi) = \lim_{s\to t, s<t} F_\ell(t;\Phi)$,

$$F_\ell(t;\Phi) = \int \left[1 - \tilde{F}_\ell\left(\frac{1}{(1-p-q)C(x)}\left(t - \frac{px}{1-A(x)}\right);x\right)\right]K(\Phi,dx). \tag{45}$$

Note that under Assumption C, $(1-p-q) \neq 0$.

Finally, the fact that $\Phi_i^\infty(X)$ has no atoms and support $(0,1)$ (both conditionally on $\Phi_{i-1}^\infty(X)$ and unconditionally) implies, together with equations (44) and (45), that whatever the sign of $(1-p-q)$, the function $t \mapsto E_H[F_\ell(t;\Phi_{i-1}^\infty(X))]$ is continuous and strictly increasing, which is to say that the distribution function of $\ell_i^\infty(X)$ is continuous and strictly increasing. ■

Lemma 18. *Under the conditions of Theorem 3, writing* $\ell_i^\infty(X) = \Pi_H(\theta_i = 0 \mid (X_n)_{n\in\mathbb{Z}})$, *the function* m *defined by*

$$m(\lambda) = E_H[\ell_i^\infty(X) \mid \ell_i^\infty(X) < \lambda]$$

is continuous and strictly increasing on $(0,1)$, *and* $m(\lambda) < \lambda$ *for all* $\lambda \in (0,1)$.

Proof For any bounded random variable U and any $a < b$ such that $P(U < a) > 0$, we have

$$\begin{aligned} E[U \mid U < b] &= E[U \mid U < a]P(U < a \mid U < b) + E[U \mid a \leq U < b]P(U \geq a \mid U < b) \\ &= E[U \mid U < a](1 - P(U \geq a \mid U < b)) + E[U \mid a \leq U < b]P(U \geq a \mid U < b), \end{aligned}$$

hence

$$E[U \mid U < b] - E[U \mid U < a] = \frac{P(a \leq U < b)}{P(U < b)}\Big(E[U \mid a \leq U < b] - E[U \mid U < a]\Big). \tag{46}$$

Note now that $E[U \mid U < a] < a$: indeed, if

$$V \overset{d}{=} (U - a) \mid \{U < a\},$$

then $V \leq 0$ and V is strictly negative with positive probability, hence $E[V] < 0$. Taking $U = \ell_i^\infty$ yields that $m(\lambda) < \lambda$ as claimed. Continuing with the proof of continuity and monotonicity, we similarly note that $a \leq E[U \mid a \leq U < b] < b$. Using also that U is bounded, so that $E[U \mid U < a] \geq -c$ for some $c < \infty$, we deduce that

$$0 < E[U \mid a \leq U < b] - E[U \mid U < a] < b + c.$$

Returning to (46) we see for a general bounded random variable U that $x \mapsto E[U \mid U < x]$ is strictly increasing on $\{x : P(U < x) > 0\}$ if $P(a \leq U < b) > 0$ for all a, b, and continuous if $P(a \leq U < b) \to 0$ as $b - a \to 0$. Taking $U = \ell_i^\infty$, we conclude by Lemma 17, which tells us that the distribution function of ℓ_i^∞ is continuous and strictly increasing and also implies that $\Pi_H(\ell_i^\infty < \lambda) > 0$ for all $\lambda > 0$. ■

Lemma 19. *Recall the definition*

$$\ell_i^\infty(X) = \Pi_H(\theta_i = 0 \mid (X_n : n \in \mathbb{Z})).$$

There exist $\delta_N, \xi_N, \xi_N' \to 0$ such that with probability tending to 1,

$$\begin{aligned} \#\{i : 1 \leq i \leq N,\ |\ell_i(X) - \ell_i^\infty(X)| > \xi_N\} &\leq N\delta_N \\ \#\{i : 1 \leq i \leq N,\ |\hat{\ell}_i(X) - \ell_i^\infty(X)| > \xi_N'\} &\leq N\delta_N. \end{aligned}$$

Proof Define $\ell_i'(X) = \Pi_H(\theta_i = 0 \mid X_{i-A_N}, \dots, X_{i+A_N})$. As in Lemma 16, we may argue using Proposition 4.3.23iii of Cappé et al. (2005) that for a suitable sequence $A_N \to \infty$ satisfying $A_N/N \to 0$, that

$$\#\{i \leq N : |\ell_i(X) - \ell_i'(X)| > 4\rho^{A_N}\} \leq 2A_N.$$

Recalling from (38) that $\ell_i^\infty(X)$ is formally defined as an almost sure limit of $\ell_i'(X)$ as $A_N \to \infty$, so that $\ell_i' \to \ell_i^\infty$ in probability also, this proves the first bound. The second bound then follows after an appeal to Lemma 9. ■

Lemma 20 (Ergodic theorems). *The sequences ℓ_i' and ℓ_i^∞, defined for $A \in \mathbb{N}$ by*

$$\begin{aligned} \ell_i'(X) &= \Pi_H(\theta_i = 0 \mid X_{i-A}, \dots, X_{i+A}), \quad A < i \leq N - A, \\ \ell_i^\infty(X) &= \Pi_H(\theta_i = 0 \mid (X_n : n \in \mathbb{Z})), \end{aligned}$$

are ergodic, so that for any bounded function g,

$$\frac{1}{N}\sum_{i=1}^N g(\ell_i') \to E_\pi[g(\ell_1')], \quad \textit{a.s. (hence also in probability),}$$

and similarly for ℓ_i^∞.

Proof These are standard ergodicity results for functions of Markov chains (see for example Durrett, 2019, Chapter 6). In the case of ℓ_i' one can also note that $g(\ell_i'(X))$ is a function of the Markov chain $(\theta_{i-A},\dots,\theta_{i+A},X_{i-A},\dots,X_{i+A})$ to reduce to the ergodic theorem for Markov chains themselves. ■

We gather some results for the mFDR, defined as in (16).

Lemma 21. *a. For any multiple testing procedure ψ,*

$$\mathrm{mFDR}_H(\psi) \le a \quad \textit{if and only if} \quad E_H \sum_{i\le N} (\ell_i - a)\psi_i \le 0, \tag{47}$$

with equality in one implying equality in the other.

b. Define the class $(\varphi_{\lambda,H} : \lambda \in [0,1])$ as in (8). For $\lambda \in [0,1]$, we have

$$\mathrm{mFDR}_H(\varphi_{\lambda,H}) \le \lambda, \quad \textit{with equality iff } \lambda = 0. \tag{48}$$

c. The map $\lambda \mapsto \mathrm{mFDR}_H(\varphi_{\lambda,H})$ is non-decreasing on $[0,1]$. In the setting of Theorem 2, for each $\lambda < \lambda'$ there exists $c = c(\lambda,\lambda') > 0$ such that for all N large enough we have

$$\mathrm{mFDR}_H(\varphi_{\lambda',H}) \ge \mathrm{mFDR}_H(\varphi_{\lambda,H}) + c. \tag{49}$$

Proof Recalling the convention $0/0 = 0$ for defining the mFDR, the first part holds trivially when $E_H \sum \psi_i = 0$, and from rearranging the definition when $E_H \sum \psi_i > 0$ (note that $E_H \mathbb{1}\{\theta_i = 0\} = E_H E_H[\mathbb{1}\{\theta_i = 0\} \mid X] = E_H \ell_i(X)$). Part (b) similarly is trivial for λ such that $\varphi_{\lambda,H} = 0$ with probability 1. If instead $\varphi_{\lambda,H}$ is not almost surely the zero vector, there exists k such that with positive probability $\varphi_k = 1$ (and hence $\ell_k < \lambda$); then $U = (\ell_k - \lambda)\varphi_k$ satisfies $U \le 0$ and $\Pi_H(U < 0) > 0$, which together imply that $E_H[U] < 0$, so that

$$E_H \sum_{i\le N} (\ell_i - \lambda)\varphi_i < 0,$$

implying (48) by part (a). For part (c), writing $a = \mathrm{mFDR}_H(\varphi_{\lambda,H}) \le \lambda$ we have, using part (a) to obtain the second line,

$$\begin{aligned} E_H \sum_{i\le N} (\ell_i - a)\mathbb{1}\{\ell_i < \lambda'\} &= E_H[\sum_{i\le N} (\ell_i - a)\mathbb{1}\{\ell_i < \lambda\}] + E_H[\sum_{i\le N} (\ell_i - a)\mathbb{1}\{\lambda \le \ell_i < \lambda'\}]. \\ &= E_H[\sum_{i\le N} (\ell_i - a)\mathbb{1}\{\lambda \le \ell_i < \lambda'\}] \\ &\ge (\lambda - a) E_H[\#\{i \le N : \lambda \le \ell_i < \lambda'\}]. \end{aligned}$$

This last expression is non-negative, and the fact that the map $\lambda \mapsto \mathrm{mFDR}_H(\varphi_{\lambda,H})$ is non-decreasing follows by part (a). For (49), we may assume without loss of generality that $\lambda > 0$, and hence that $\lambda - a > 0$. By Lemma 19 there exists a sequence $\xi_N \to 0$ such that $E_H \#\{i : |\ell_i - \ell_i^\infty| > \xi_N\}/N \to 0$ as $N \to \infty$, and we decompose

$$E_H[\#\{i : \lambda \le \ell_i < \lambda'\}] \ge E_H \#\{i : \lambda + \xi_N \le \ell_i^\infty < \lambda' - \xi_N\} - E_H \#\{i : |\ell_i - \ell_i^\infty| > \xi_N\}.$$

Lemma 17 tells us that under the assumptions of Theorem 3 the distribution function of ℓ_i^∞ is strictly increasing, so that for N large enough that $\lambda + \xi_N < \lambda' - \xi_N$ the first term on the right in the latest display is of order N and the second is of smaller order. We deduce that, for $c = c(\lambda,\lambda') > 0$ small enough, and for all N large enough,

$$E_H[\sum_{i\le N} (\ell_i - (a+c))\mathbb{1}\{\ell_i < \lambda'\}] \ge cN - cE_H[\sum_{i\le N} \mathbb{1}\{\ell_i < \lambda'\}] \ge 0,$$

so that $\mathrm{mFDR}_H(\varphi_{\lambda',H}) \geq a+c$, implying (49). ■

Lemma 22. *In the setting of Theorem 2, define the class* $(\varphi_{\lambda,H} : \lambda \in [0,1])$ *as in* (8)*, and define the mTDR and mFDR as in* (15) *and* (16)*. Then for each* $\lambda \in (0,1)$ *we have*

$$\mathrm{mTDR}_H(\varphi_{\lambda,H}) = \sup\{\mathrm{mTDR}_H(\psi) : \mathrm{mFDR}_H(\psi) \leq \mathrm{mFDR}_H(\varphi_{\lambda,H})\}.$$

Remarks. i. A version of this result in the HMM setting originates in Sun and Cai (2009), but to avoid a monotonicity property needed therein we instead adapt the proof Lemma 9.2 of Rebafka et al. (2019) (see also the proof of Cai et al. 2019, Theorem 1). The proof is valid for ℓ-value procedures in any (correctly specified) model, not just the hidden Markov model (1).

ii. The result need not in general hold for $\lambda = 0$, since $\mathrm{mFDR}_H(\psi) = 0$ whenever $E_H[\ell_i(X)\psi_i(X)] = 0$ for all i, so that if $\Pi_H(\ell_i(X)=0) > 0$, the test ψ defined by $\psi_i(X) = \mathbb{1}\{\ell_i(X) = 0\}$ has positive probability of making at least one true discovery, so that $\mathrm{mTDR}_H(\psi) > 0$, while $\mathrm{mFDR}_H(\psi) = 0$.

iii. In general, $\{\mathrm{mFDR}_H(\varphi_{\lambda,H}) : \lambda \in [0,1]\}$ is a proper subset of $[0,1]$, and consequently the class $\varphi_{\lambda,H}$ need not be optimal for every threshold. In particular, the supremum of the set is generally strictly smaller than one, and – especially in discrete data settings – there may be jump discontinuities in the function $\lambda \mapsto \mathrm{mFDR}_H(\varphi_{\lambda,H})$. The first of these does not cause any issues, since $\mathrm{mTDR}_H(\varphi_{1,H}) = 1 = \sup_\psi \mathrm{mTDR}_H(\psi)$ (provided $\theta_i = 1$ with positive probability, which is true in the current setting by Assumption C), while Lemma 23 overcomes the issues raised in the second case in the setting of Theorem 3.

Proof Fix $\lambda \in (0,1)$ and write φ for $\varphi_{\lambda,H}$. Lemma 21 tells us that $a = \mathrm{mFDR}_H(\varphi)$ satisfies $a < \lambda$, and implies that if $\mathrm{mFDR}_H(\psi) \leq a$ then

$$E_H \sum_{i\leq N} (\ell_i - a)(\varphi - \psi) \geq 0. \tag{50}$$

We show that, for all i,

$$(\ell_i - a)(\varphi_i - \psi_i) \leq \frac{\lambda - a}{1-\lambda}(1-\ell_i)(\varphi_i - \psi_i). \tag{51}$$

Indeed, if $\varphi_i = 1$, then $\ell_i < \lambda$, so that

$$\ell_i - a < \frac{1-\ell_i}{1-\lambda}(\lambda - a),$$

and multiplying by $\varphi_i - \psi_i \geq 0$ yields the inequality, while if $\varphi_i = 0$, then $\ell_i \geq \lambda > a$, so that

$$\ell_i - a \geq \frac{1-\ell_i}{1-\lambda}(\lambda - a),$$

and multiplying by $\varphi_i - \psi_i \leq 0$ yields the inequality.

Now, since $a < \lambda < 1$, so that $(1-\lambda)/(\lambda - a) > 0$, we deduce from (50) and (51) that

$$E_H \sum_{i\leq N} (1-\ell_i)(\varphi_i - \psi_i) \geq 0.$$

Finally, by definition,

$$\mathrm{mTDR}_H(\varphi) = \frac{E_H[\sum_{i\leq N}(1-\ell_i)\varphi_i]}{N\pi_1}, \quad \mathrm{mTDR}_H(\psi) = \frac{E_H[\sum_{i\leq N}(1-\ell_i)\psi_i]}{N\pi_1},$$

hence $\mathrm{mTDR}_H(\varphi) \geq \mathrm{mTDR}_H(\psi)$ as claimed. ■

Lemma 23. *In the setting of Theorem 3, define the map*

$$g : x \mapsto \sup\{\mathrm{mTDR}_H(\psi) : \mathrm{mFDR}_H(\psi) \le x\},$$

where the supremum is defined over multiple testing procedures ψ. Then for sequences x_N, y_N such that $|x_N - y_N| \to 0$, we have

$$|g(x_N) - g(y_N)| \to 0 \quad \text{as } N \to \infty.$$

[Note that g depends implicitly on N, so that this does not simply say that g is continuous.]

Proof Prompted by Lemma 22, we focus on tests ψ of the form $\varphi_{\lambda,H}$, $\lambda \in [0,1]$ and define, for $N \ge 1$,

$$\lambda_N = \sup\{\lambda : \mathrm{mFDR}_H(\varphi_{\lambda,H}) \le x_N\},$$
$$\mu_N = \sup\{\lambda : \mathrm{mFDR}_H(\varphi_{\lambda,H}) \le y_N\}.$$

One has the following dichotomies, as for the postFDR (recall (12)), implied by the fact that the map $\lambda \mapsto \mathrm{mFDR}_H(\varphi_{\lambda,H})$ is non-decreasing (by Lemma 21) and left continuous (by the definition of $\varphi_{\lambda,H}$):

$$\begin{aligned} \mathrm{mFDR}_H(\varphi_{\lambda,H}) \le x_N &\iff \lambda \le \lambda_N, \\ \mathrm{mFDR}_H(\varphi_{\lambda,H}) \le y_N &\iff \lambda \le \mu_N. \end{aligned} \tag{52}$$

Suppose (for a contradiction) that $|\lambda_N - \mu_N| \not\to 0$. Without loss of generality we may assume that for some $\delta > 0$ and some subsequence N_j we have $\lambda_{N_j} - \mu_{N_j} > \delta$ for all $j \in \mathbb{N}$. Then $\lambda_{N_j} - \tilde{\mu}_{N_j} > \delta/2$ for large j, and by restricting to a further subsequence if necessary we may assume that for some $\lambda > \mu$ we have $\lambda_{N_j} \ge \lambda > \mu > \mu_{N_j}$ for all j. Now Lemma 21 tells us that there exists a constant $= c(\lambda, \mu) > 0$ such that for N large enough

$$\mathrm{mFDR}_H(\varphi_{\lambda,H}) - \mathrm{mFDR}_H(\varphi_{\mu,H}) > c.$$

Using (52) we deduce

$$x_{N_j} \ge \mathrm{mFDR}_H(\varphi_{\lambda,H}) > \mathrm{mFDR}_H(\varphi_{\mu,H}) + c > y_{N_j} + c,$$

so that $x_{N_j} - y_{N_j} > c$, contradicting that $|x_N - y_N| \to 0$. We deduce that necessarily $|\lambda_N - \mu_N| \to 0$.

Now set $\tilde{\lambda}_N = \min(\lambda_N + 1/N, 1)$ and $\tilde{\mu}_N = \min(\mu_N + 1/N, 1)$. Then (52), together with Lemma 22 (and Remark iii thereafter for the cases $\tilde{\lambda}_N = 1$, $\tilde{\mu}_N = 1$) implies that

$$\mathrm{mTDR}_H(\varphi_{\lambda_N,H}) \le g(x_N) \le \mathrm{mTDR}_H(\varphi_{\tilde{\lambda}_N,H})$$
$$\mathrm{mTDR}_H(\varphi_{\mu_N,H}) \le g(y_N) \le \mathrm{mTDR}_H(\varphi_{\tilde{\mu}_N,H}).$$

We prove that $|\mathrm{mTDR}_H(\varphi_{\lambda_N,H}) - \mathrm{mTDR}_H(\varphi_{\mu_N,H})| \to 0$ as a consequence of the fact that $|\lambda_N - \mu_N| \to 0$. Since also $|\tilde{\lambda}_N - \lambda_N| \to 0$, $|\tilde{\mu}_N - \mu_N| \to 0$, the same proof will imply that each of $\mathrm{mTDR}_H(\varphi_{\lambda_N,H})$, $\mathrm{mTDR}_H(\varphi_{\tilde{\lambda}_N,H})$, $\mathrm{mTDR}_H(\varphi_{\mu_N,H})$ and $\mathrm{mTDR}_H(\varphi_{\tilde{\mu}_N,H})$ differ by at most $o(1)$, allowing us to conclude.

Assume for notational convenience that $\lambda_N \ge \mu_N$. The denominator in the expressions defining each of the mTDR's is $E_H\#\{i : \theta_i = 1\} = N\pi_1$, and we see that

$$\mathrm{mTDR}_H(\varphi_{\lambda_N,H}) = \mathrm{mTDR}_H(\varphi_{\mu_N,H}) + \frac{E_H\#\{i : \theta_i = 1, \mu_N \le \ell_i < \lambda_N\}}{N\pi_1}.$$

By Lemma 19 there exists a sequence $\xi_N \to 0$ such that $E_H\#\{i : |\ell_i - \ell_i^\infty| > \xi_N\}/N \to 0$ as $N \to \infty$. Lemma 17 tells us that the distribution function of ℓ_i^∞ is continuous – and hence uniformly continuous – and we see that

$$\begin{aligned}&N^{-1}E_H\#\{i : \theta_i = 1, \mu_N \le \ell_i < \lambda_N\}\\ \le&\Pi_H(\mu_N - \xi_N \le \ell_1^\infty < \lambda_N + \xi_N) + N^{-1}E_H\#\{i : |\ell_i - \ell_i^\infty| > \xi_N\} \to 0,\end{aligned}$$

as $N \to \infty$, proving the claim. ■

Appendix B. Auxiliary Results for the upper bounds of Section 3

B.1 Well-definedness of the Estimators

Lemma 24. *In the setting of Theorem 5, there exist $(h_l)_{l\in\mathbb{N}}$ (not depending on H) uniformly supremum-norm bounded such that $O^{L_0} = (E_H[h_l(X_1) \mid \theta_1 = j]_{l\le L_0, j\le J}) \in \mathbb{R}^{L_0\times J}$ satisfies*

$$\sigma_J(O^{L_0}) \ge C,$$

uniformly in $L_0 \ge \underline{L}$, for some $C, \underline{L}$ depending on the parameters $f_j,\ j \le J$.

Proof For $L > L'$, $\sigma_J(O^L) > \sigma_J(O^{L'})$ because $O^{L'}$ is a submatrix of O^L (see for example Stewart and Sun, 1990, Chapter 1, Theorem 4.4). So it suffices to show that $\sigma_J(O^{\underline{L}}) > 0$ for some $\underline{L}$.

Choose a countable family of sets $\mathcal{A} = \{A_1, \dots\}$ generating the Borel σ-algebra on $\mathbb{R}$, for example $\mathcal{A} = \{(-\infty, q) : q \in \mathbb{Q}\}$, and let $h_l = \mathbb{1}_{A_l}$. Suppose for a contradiction that $\sigma_J(O^L) = 0$ for all $L \in \mathbb{N}$, or, put another way, that the J vectors $(\langle h_l, f_j\rangle_{l\le L}) \in \mathbb{R}^L$, $j \le J$ are linearly dependent for all $L \in \mathbb{N}$, so that there exist $a_1^L, \dots, a_J^L \in [-1, 1]$ for which $\sum_j |a_j^L| = 1$ and $\sum_j a_j^L \langle h_l, f_j\rangle = 0$ for all $l \le L$. By Bolzano–Weierstrass, there is a sequence $L_n \to \infty$ such that for each $j \le J$, $a_j^{L_n}$ converges to some a_j^∞, and note that necessarily $(a_j^\infty)_{j\le J}$ is not the zero vector. For each $l \in \mathbb{N}$, we have that

$$\langle h_l, \sum_{j\le J} a_j^\infty f_j\rangle = \sum_{j\le J} a_j^\infty \langle h_l, f_j\rangle = \lim_{n\to\infty} \sum_{j\le J} a_j^{L_n} \langle h_l, f_j\rangle = 0.$$

Since $\operatorname{span}\{h_l : l \in \mathbb{N}\}$ corresponds to the simple functions, which are dense in L^2, and since $\sum_j a_j^\infty f_j$ is a continuous function, the latter is the zero function, contradicting that the functions $f_j,\ j \le J$ are linearly independent. ■

Lemma 25. *Under the assumptions of Theorem 5, define $\hat{P}$ and $(\hat{M}^x, \hat{B}^x,\ x \in \mathbb{R})$ as in Algorithm 1. Then*

a. The map $x \mapsto \hat{M}^x$ is continuous. For any $\kappa > 0$, there exists $c = c(\kappa, \mathcal{H})$ such that the event

$$\mathcal{A} = \{\|\hat{P} - P\| \le cL_0 r_N,\ \sup_{x\in\mathbb{R}} \|\hat{M}^x - M^x\| \le cL_0^2 r_N\}$$

(is measurable and) has probability at least $1 - N^{-\kappa}$ for N large.

b. On $\mathcal{A}$, for N large enough $\hat{P}$ has rank J, and the matrices

$$\tilde{B}^x = (\hat{V}^\intercal P \hat{V})^{-1} \hat{V}^\intercal M^x \hat{V}, \quad x \in \mathbb{R}, \tag{53}$$

are well defined.

c. *On $\mathcal{A}$, for some $C > 0$ depending on both the constant c of $\mathcal{A}$ and on $\mathcal{H}$, we have for N large enough*

$$\sup_{x\in\mathbb{R}} \max(\|\hat{B}^x\|, \|\tilde{B}^x\|) \leq C L_0^{1/2}, \tag{54}$$

$$\sup_{x\in\mathbb{R}} \|\tilde{B}^x - \hat{B}^x\| \leq C L_0^2 r_N. \tag{55}$$

Proof Lemma 29 and Lemma 30 together imply that for suitable $c = c(\kappa, \mathcal{H})$,

$$\Pi_H(\|\hat{P} - P\| \leq c L_0 r_N, \ \sup_{x\in\mathbb{Q}} \|\hat{M}^x - M^x\| \leq c L_0^2 r_N) \geq 1 - N^{-\kappa}.$$

[In fact a union bound yields this with $2N^{-\kappa}$ in place of $N^{-\kappa}$, but the factor 2 can be removed by initially considering some $\kappa' > \kappa$.] We prove the claimed continuity of the map $x \mapsto \hat{M}^x$; it will follow that

$$\{\sup_{x\in\mathbb{Q}} \|\hat{M}^x - M^x\| \leq c L_0^2 r_N\} = \{\sup_{x\in\mathbb{R}} \|\hat{M}^x - M^x\| \leq c L_0^2 r_N\},$$

which implies measurability and the probability bound for $\mathcal{A}$. This continuity results from the assumed Lipschitz continuity of K. Indeed, if Λ is the Lipschitz constant for K, observe that if $|x - y| < \delta$ then for any n

$$|K_L(x, X_{n+1}) - K_L(y, X_{n+1})| \leq \sup_{t\in\mathbb{R}} |K_L(x,t) - K_L(y,t)| \leq \sup_{|u-v|<2^L\delta} 2^L |K(u) - K(v)| \leq 2^{2L}\Lambda\delta,$$

hence, for some $C = C(\mathcal{H})$,

$$\|\hat{M}^x - \hat{M}^y\| \leq \frac{L_0}{N^{1/2}} \max_l \|h_l\|_\infty^2 \max_{n\leq N} |K_L(x, X_{n+1}) - K_L(y, X_{n+1})| \leq C \frac{L_0 2^{2L}}{N^{1/2}} |x - y|.$$

Next, in view of the assumption on O made in the algorithm, Lemma 34 implies that $\sigma_J(P)$ is bounded away from zero for large N and consequently by Lemma 35a, on $\mathcal{A}$ and for N large we have that $\hat{P}$ is of rank J and that $\hat{V}^\intercal P \hat{V}$ is invertible (recall that $L_0^5 r_N \to 0$ by assumption, so that the condition of Lemma 35 – that $\|\hat{P} - P\| < \sigma_J(P)/3$ – holds eventually). Then Lemma 11 tells us that $\tilde{B}^x$ is well defined for each $x \in \mathbb{R}$ and can be expressed as $(QO^\intercal\hat{V})^{-1} D^x QO^\intercal\hat{V}$. It follows, using Lemma 35b and eq. (27), that on $\mathcal{A}$, for a constant $c = c(\mathcal{H})$ and any $x \in \mathbb{R}$ we have

$$\|\tilde{B}^x\| \leq \kappa(QO^\intercal\hat{V}) \max_j |K_L[f_j](x)| \leq c L_0^{1/2}$$

for N large (recall $\kappa(A) := \|A\|\|A^{-1}\|$ is the condition number of a matrix).

Finally, Lemma 35c tells us that on $\mathcal{A}$, for N large enough that $cL_0 r_N < \sigma_J(P)/3$,

$$\|\tilde{B}^x - \hat{B}^x\| \leq 3.2\Big[\frac{\|\hat{M}^x - M^x\|}{\sigma_J(P)} + \frac{\|M^x\|\|\hat{P} - P\|}{\sigma_J(P)^2}\Big], \quad \forall x \in \mathbb{R}.$$

Noting that $\|M^x\| \leq cL_0$ for some $c = c(\mathcal{H})$ by Lemma 34, we deduce (55). The bound for $\|\hat{B}^x\|$ then follows from the bound for $\|\tilde{B}^x\|$ by the triangle inequality. ■

Lemma 26. *Recall* $\operatorname{sep}(B)$ *denotes the eigen-separation of a matrix B, in that if B has eigenvalues $\lambda_1, \ldots, \lambda_J$ then* $\operatorname{sep}(B) = \min_{j\neq j'} |\lambda_j - \lambda_{j'}|$. *On the event $\mathcal{A}$ of Lemma 25, define $B^{a,u} \equiv \tilde{B}^{a,u}$ as in Algorithm 1 for $V = \hat{V}$:*

$$\tilde{B}^{a,u} = \sum a_i \tilde{B}^{u_i}, \qquad \tilde{B}^x = (\hat{V}^\intercal P \hat{V})^{-1} \hat{V}^\intercal M^x \hat{V}.$$

Define

$$\mathcal{D}_N = \Big\{\frac{j}{2^N} : j \in \mathbb{Z}\Big\} \cap [-N, N],$$
$$\mathbb{D}_N = \{(a,u) \in \mathcal{D}_N^{J(J-1)/2} \times \mathcal{D}_N^{J(J-1)/2} : \sum_i |a_i| \le 1\}.$$

Then there exists a constant c depending only on $f_1, \dots, f_J$ and (strictly) positive when they are all distinct such that, on $\mathcal{A}$,

$$\max\{\operatorname{sep}(\tilde{B}^{a,u}) : (a,u) \in \mathbb{D}_N\} \ge c,$$

for all N large.

Remark. Recall, as remarked after Algorithm 1, that proving this result for $V = \hat{V}$ implies it holds for any V such that $B^x = (V^\intercal P V)^{-1}(V^\intercal M^x V)$ is well-defined.

Proof In view of Lemma 11, $\tilde{B}^{a,u}$, being a linear combination of simultaneously diagonalisable matrices, is diagonalisable for any a, u, with eigenvalues

$$(\sum_i a_i K_L[f_j](u_i))_{j \le J}.$$

Recall that $\|K_L[f_j] - f_j\|_\infty \to 0$ as $L = L(N) \to \infty$ by (26). It follows by the triangle inequality that

$$\max_{\mathbb{D}_N} |\sum_i a_i (K_L[f_j](u_i) - f_j(u_i))| \to 0,$$

hence

$$\max_{\mathbb{D}_N} \operatorname{sep}(\tilde{B}^{a,u}) = \max_{\mathbb{D}_N} \min_{j \ne j'} \Big|\sum_i a_i K_L[f_j - f_{j'}](u_i)\Big| > \tfrac{1}{2} \max_{\mathbb{D}_N} \min_{j \ne j'} \Big|\sum_i a_i \big(f_j(u_i) - f_{j'}(u_i)\big)\Big|, \tag{56}$$

for N large, provided this latter quantity is strictly positive.

Next, let U_N denote $[-N,N]^{J(J-1)/2}$. Observe that, since $f_j \in C^s(\mathbb{R})$ for each $j \le J$,

$$(a,u) \mapsto \min_{j \ne j'} |\sum_i a_i (f_j(u_i) - f_{j'}(u_i))|$$

is uniformly continuous on $\mathbb{R}^{J(J-1)/2} \times \mathbb{R}^{J(J-1)/2}$, so that

$$\tfrac{1}{2} \max_{(a,u) \in \mathbb{D}_N} \min_{j \ne j'} \Big|\sum_i a_i \big(f_j(u_i) - f_{j'}(u_i)\big)\Big| > \tfrac{1}{4} \sup_a \sup_{u \in U_N} \min_{j \ne j'} \Big|\sum_i a_i \big(f_j(u_i) - f_{j'}(u_i)\big)\Big| \tag{57}$$

for N large, provided this latter quantity is strictly positive. The supremum on the right can be extended: while at first we must take the supremum over (a such that $\sum |a_i| \le 1$ and) $u \in U_N$, the result remains true taking the supremum instead over all $u \in \mathbb{R}^{J(J-1)/2}$, at least for N large, using that $f_j(u) \to 0$ as $u \to \infty$. [That is, when the right side of (57) is strictly positive, the supremum over $u \in \mathbb{R}^{J(J-1)/2}$ is attained on U_N for N large.] We now prove that

$$\sup_{a,u} \min_{j \ne j'} \Big|\sum_i a_i \big(f_j(u_i) - f_{j'}(u_i)\big)\Big| > 0.$$

Choose for each pair $j \ne j'$ some $x \in \mathbb{R}$ such that $f_j(x) \ne f_{j'}(x)$, and collect these x into the vector u. For each $j \ne j'$, writing i for an index such that $f_j(u_i) \ne f_{j'}(u_i)$, the set $\{v \in \mathbb{R}^{J(J-1)/2} : \langle v, f_j(u_i) - f_{j'}(u_i) \rangle = 0\}$ is a proper subspace of $\mathbb{R}^{J(J-1)/2}$, so the union over these $J(J-1)/2$ spaces is not equal to $\mathbb{R}^{J(J-1)/2}$ (for example it has Lebesgue measure zero) and we may choose a

in the complement of the union. Scale invariance means that moreover we may assume a satisfies $\sum_i |a_i| = 1$. Then $|\sum_i a_i(f_j(u_i) - f_{j'}(u_i))| > 0$ for each $j \neq j'$, as required.

Finally, combining also with (56) and (57) we deduce that

$$\max(\operatorname{sep}(\tilde{B}^{a,u}) : (a,u) \in \mathbb{D}_N) > \tfrac{1}{4} \sup_{a,u} \min_{j \neq j'} \Big| \sum_i a_i(f_j(u_i) - f_{j'}(u_i)) \Big| > 0,$$

concluding the proof. ■

Lemma 27. *In the setting of Theorem 5, let $\mathcal{A}$ be the event of Lemma 25. Define $\hat{B}^x = \hat{B}^{x,L_0,L}$ as in Algorithm 1 and $\tilde{B}^x = \tilde{B}^{x,L_0,L}$ as in* (53). *For $a, u \in \mathbb{R}^{J(J-1)/2}$ define $\hat{B}^{a,u} = \sum a_i \hat{B}^{u_i}$, $\tilde{B}^{a,u} = \sum a_i \tilde{B}^{u_i}$. Then there exists a constant $c = c(\mathcal{H}) > 0$ such that, for $\hat{a}, \hat{u}$ as in Algorithm 1, on the event $\mathcal{A}$ we have*

$$\operatorname{sep}(\hat{B}^{\hat{a},\hat{u}}) > c, \tag{58}$$

$$\operatorname{sep}(\tilde{B}^{\hat{a},\hat{u}}) > c, \tag{59}$$

for N large. Note that (58) *implies in particular that $\hat{B}^{\hat{a},\hat{u}}$ has J distinct eigenvalues and so is diagonalisable.*

Proof By Lemma 25, on $\mathcal{A}$ the matrices $\hat{B}^x, \tilde{B}^x$ are well-defined and satisfy for some $C = C(\mathcal{H})$

$$\sup_x \|\tilde{B}^x - \hat{B}^x\| \leq C L_0^2 r_N, \quad \sup_x \max(\|\tilde{B}^x\|, \|\hat{B}^x\|) \leq C L_0^{1/2}.$$

By the triangle inequality, we deduce that

$$\|\hat{B}^{a,u}\| \leq \sum |a_i| \|\hat{B}^{u_i}\| \leq \sup_x \|\hat{B}^x\| \leq C L_0^{1/2},$$

and similarly $\|\tilde{B}^{a,u}\| \leq C L_0^{1/2}$. Let $(a_N, u_N) \in \operatorname{argmax}_{\mathbb{D}_N}(\operatorname{sep}(\tilde{B}^{a,u}))$ and recall by assumption that

$$\operatorname{sep}(\tilde{B}^{a_N,u_N}) > c \quad \text{uniformly in } N \text{ large enough, for some } c > 0.$$

[As noted in the remark after Lemma 26, choosing $V = \hat{V}$ in Algorithm 1, and hence replacing B^x defined therein with $\tilde{B}^x$, is valid on $\mathcal{A}$.] We apply the Ostrowski–Elsner theorem (Theorem 36) to $A = \hat{B}^{a,u}$, $B = \tilde{B}^{a,u}$ to see for a constant $C = C(\mathcal{H})$ that for any a, u we have

$$\min_\tau \max_j |\lambda_{\tau(j)}(\tilde{B}^{a,u}) - \lambda_j(\hat{B}^{a,u})| \leq C L_0^{(J-1)/(2J)} (L_0^2 r_N)^{1/J},$$

where $\lambda_j, j \leq J$ are maps taking matrices to their eigenvalues. This last expression tends to zero as $N \to \infty$ (since by assumption $L_0^{(J+3)/2} r_N \to 0$) and in particular it is smaller than $\operatorname{sep}(\tilde{B}^{a_N,u_N})/5$ for N large.

By the triangle inequality we deduce that on $\mathcal{A}$,

$$\operatorname{sep}(\hat{B}^{a_N,u_N}) \geq \operatorname{sep}(\tilde{B}^{a_N,u_N}) - 2 \sup_{a,u} \min_\tau \max_j |\lambda_{\tau(j)}(\tilde{B}^{a,u}) - \lambda_j(\hat{B}^{a,u})| \geq (3/5) \operatorname{sep}(\tilde{B}^{a_N,u_N}).$$

It follows by definition of $\hat{a}, \hat{u}$ that

$$\operatorname{sep}(\hat{B}^{\hat{a},\hat{u}}) \geq \operatorname{sep}(\hat{B}^{a_N,u_N}) \geq (3/5) \operatorname{sep}(\tilde{B}^{a_N,u_N}),$$

proving (58). Applying the triangle inequality again we conclude that

$$\operatorname{sep}(\tilde{B}^{\hat{a},\hat{u}}) \geq (1/5) \operatorname{sep}(\tilde{B}^{a_N,u_N}),$$

proving (59). ■

B.2 Concentration of Empirical Estimators

We note the following concentration results for Markov chains, adapted as in Proposition 13 of De Castro et al. (2016) from results of Paulin (2015), which will allow us to control the errors of the empirical estimators $\hat{P}$ and $\hat{M}^x$. The pseudo-spectral gap of a chain is defined in Paulin (2015), wherein it is noted that its reciprocal is equivalent to the mixing time. The bracketing numbers $N_{[]}(\mathcal{T}, \|\cdot\|_{L^2(P)}, \varepsilon)$ are defined as the smallest number of pairs of functions $(\underline{f}, \bar{f})$ such that every $g \in \mathcal{T}$ is bracketed by one of the pairs, where $(\underline{f}, \bar{f})$ brackets g if $\underline{f} \leq g \leq \bar{f}$ pointwise.

Lemma 28. *Let Y be a stationary Markov chain taking values in $\mathcal{Y}$ with pseudo-spectral gap $\gamma_{\mathrm{ps}} > 0$, with law denoted P. Let $\mathcal{T}$ be some countable class of real valued and measurable functions on $\mathcal{Y}$. Assume there exist $\sigma, b > 0$ such that for all $t \in \mathcal{T}$, $\|t\|_{L^2(P)} \leq \sigma$ and $\|t\|_\infty \leq b$. Suppose that the $L^2(P)$ bracketing entropy*

$$H_{[]}(\mathcal{T}, \|\cdot\|_{L^2(P)}, \varepsilon) := \log N_{[]}(\mathcal{T}, \|\cdot\|_{L^2(P)}, \varepsilon),$$

is upper bounded by some $\bar{H}(\varepsilon)$, achievable using brackets of L^∞-diameter at most b. Then for fixed $t \in \mathcal{T}$ we have

$$P(|\sum (h(Y_i) - Eh(Y_1))| \geq x) \leq 2 \exp\Big(-\frac{x^2 \gamma_{\mathrm{ps}}}{8(N + 1/\gamma_{\mathrm{ps}})\sigma^2 + 20bx}\Big), \tag{60}$$

and there exists $C > 0$ depending only on a lower bound for γ_{ps} such that

$$P\Big(\sup_{t \in \mathcal{T}} \sum_{n=1}^N (t(Y_n) - Et) \geq C[A + \sigma\sqrt{Nx} + bx]\Big) \leq \exp(-x), \tag{61}$$

where

$$A = \sqrt{N} \int_0^\sigma \sqrt{\bar{H}(u) \wedge N}\, \mathrm{d}u + (b + \sigma)\bar{H}(\sigma).$$

Proof The first claim is proved by Paulin (2015, Theorem 3.4) (but note there is an updated version of the paper on arXiv). For the second, observe that the proof of the same theorem gives the following bound for the Laplace transform of $S = \sum (t(Y_n) - Et)/b$:

$$E \exp(\lambda S) \leq \exp\Big(\frac{2(N + 1/\gamma_{\mathrm{ps}})(\sigma^2/b^2)}{\gamma_{\mathrm{ps}}} \lambda^2 \Big(1 - \frac{10\lambda}{\gamma_{\mathrm{ps}}}\Big)^{-1}\Big). \tag{62}$$

One now appeals to Theorem 6.8 of Massart (2007) and the consequent Corollary 6.9. While the theorem is stated for independent random variables, the proof uses this condition only when applying Lemma 6.6 of the same reference, a version of which holds also in the current setting thanks to (62). ■

Lemma 29. *In the setting of Theorem 5 and defining $P, \hat{P}$ as in Algorithm 1, for any $\kappa > 0$ there exists $C = C(\kappa, \mathcal{H})$ such that*

$$\Pi_H\Big(\|\hat{P} - P\| > CL_0(N/\log N)^{-1/2}\Big) \leq N^{-\kappa}.$$

Proof Noting that $Y_n = (X_n, X_{n+1}, X_{n+2}, \theta_n, \theta_{n+1}, \theta_{n+2})$ defines a stationary Markov chain and upper bounding the L^2-norm by the supremum norm, we apply (60) to deduce that

$$\begin{aligned}
&\Pi_H\Big(\Big|\frac{1}{N}\sum_{i=1}^N h_{ij}(Y_n) - E_H[h_{ij}]\Big| > C\Big(\frac{\log N}{N}\Big)^{1/2}\Big) \\
\leq &2 \exp\Big(-\frac{C^2 \gamma_{\mathrm{ps}} N \log N}{8(N + 1/\gamma_{\mathrm{ps}})\|h_{ij}\|_\infty^2 + 20C(N \log N)^{1/2}\|h_{ij}\|_\infty}\Big),
\end{aligned}$$

where $h_{ij}(Y_n) = h_i(Y_{n,1})h_j(Y_{n,3})$ and where γ_{ps} is the pseudo-spectral gap of the chain Y_n. We note that $\|h_{ij}\|_\infty^2 \le \|h_i\|_\infty^2 \|h_j\|_\infty^2$ is bounded by assumption. The pseudo spectral gap is also bounded: by Proposition 3.4 of Paulin (2015) its reciprocal is controlled up to a constant by the mixing time of the Markov chain Y_n, which is equal to the mixing time of the chain $(\theta_n, \theta_{n+1}, \theta_{n+2})_n$. This latter quantity is bounded since the assumption that Q is irreducible and aperiodic on a finite state space implies that θ mixes exponentially, at a rate governed (again, in view of Paulin 2015, Proposition 3.4) by the pseudo spectral gap of Q itself and $\min_j \pi_j$.

We deduce that for a constant $c = c(\mathcal{H})$ we have

$$\Pi_H\Big(\Big|\frac{1}{N}\sum h_{ij}(Y_n) - E_H[h_{ij}]\Big| > C\Big(\frac{\log N}{N}\Big)^{1/2}\Big) \le 2\exp(-C^2 c \log(N)).$$

For any $\kappa > 0$, choosing $C = C(\kappa, c)$ large enough, this last probability is smaller than $N^{-\kappa}$ as claimed. ■

Lemma 30. *In the setting of Theorem 5, define $M^x = M^{x,L_0,L}$, $\hat{M}^x = \hat{M}^{x,L_0,L}$ as in Algorithm 1, and recall that we chose L such that $2^L \asymp (N/\log N)^{1/(1+2s)}$ and assumed that $L_0^5 r_N \to 0$. For any $\kappa > 0$ there exists $C = C(\kappa, \mathcal{H})$ such that*

$$\Pi_H\Big(\sup_{x\in\mathbb{Q}}\|\hat{M}^x - M^x\| \ge C L_0^2 (N/\log N)^{-s/(1+2s)}\Big) \le N^{-\kappa}.$$

Proof As in Lemma 29 we note that the pseudo-spectral gap of the chain

$$Y_n = (X_n, X_{n+1}, X_{n+2}, \theta_n, \theta_{n+1}, \theta_{n+2})$$

is bounded away from zero provided the same is true of $\min_j \pi_j$ and the pseudo-spectral gap of Q itself, which holds by Assumption C' (see also Section 4.3). We apply Lemma 28 to the family $\mathcal{T} = \{\pm h_i \otimes K_L(x,\cdot) \otimes h_j : i, j \le L_0, x \in \mathbb{Q}\}$. Recall we assume that $\max(\|h_l\|_\infty : l \le L_0)$ is bounded independently of L_0. Lemma 31 implies, for some $C = C(\mathcal{H})$, the bracketing entropy bound

$$H_{[]}(\mathcal{T}, \|\cdot\|_{L^2(\Pi_H)}, \varepsilon) \le \bar{H}(\varepsilon) = C\log(L_0 2^L \varepsilon^{-1}), \quad \varepsilon \le \sigma,$$

where we may take

$$b = \sigma^2 = C2^L,$$

with the bound on σ^2 following from the calculations

$$\sup_{x\in\mathbb{Q}, i,j\le L_0} \|h_i \otimes K_L(x,\cdot) \otimes h_j\|_{L^2(\Pi_H)}^2 \le \max_{i\le L_0}\|h_i\|_\infty^4 \|f_\pi\|_\infty \sup_{x\in\mathbb{Q}} \int K_L(x,y)^2\,\mathrm{d}y,$$
$$\int K_L(x,y)^2\,\mathrm{d}y = 2^{2L}\int K(2^L(x-y))^2\,\mathrm{d}y = 2^L \int K(z)^2\,\mathrm{d}z \le 2^{L+1}\|K\|_\infty^2.$$

An application of Jensen's inequality yields the standard bound

$$\int_0^x \sqrt{\log(1/u)}\,\mathrm{d}u \le x\sqrt{1+\log(1/x)} \le x\Big(1+\sqrt{\log(1/x)}\Big). \tag{63}$$

Performing suitable substitutions we deduce that

$$\int_0^\sigma \sqrt{\log(L_0^{1/4}2^L/u)}\,\mathrm{d}u = L_0^{1/4}2^L \int_0^{\sigma/(2^L L_0^{1/4})} \sqrt{\log(1/v)}\mathrm{d}v \le \sigma\Big(1+\sqrt{\log(L_0^{1/4}2^L/\sigma)}\Big) \le C\sqrt{L2^L},$$

for some constant C, since by assumption $L_0^5 r_N \to 0$, which implies that $\log(L_0) \le \log N \asymp L$. Noting that $(b+\sigma)\bar{H}(\sigma) \le CL2^L$ for some C, we deduce that

$$\Pi_H\Big(\sup_{t\in\mathcal{T}}\sum_{n=1}^N \big(t(Y_n) - E_H t\big) \ge C[\sqrt{N2^L}(\sqrt{L} + \sqrt{\kappa\log N}) + 2^L(L+\kappa\log N)]\Big) \le \exp(-\kappa\log N).$$

Since $2^L \asymp (N/\log N)^{1/(1+2s)}$ we find, bounding the operator norm by the L_0^2 times the maximum of the entries, that as claimed, for some $C' = C'(\kappa)$ we have

$$\Pi_H\Big(\sup_{x\in\mathbb{Q}}\|\hat{M}^x - M^x\| \ge C' L_0^2 (N/\log N)^{-s/(1+2s)}\Big) \le N^{-\kappa}. \tag{64}$$

■

Recall from Assumption B that f_π has a bounded νth absolute moment. Recall from Section 4.3 the definition of a constant $C = C(\mathcal{H})$.

Lemma 31. *Let K_L be as in* (25)*, let $(h_l : l \le L_0)$ be as in Algorithm 1, and define $\mathcal{T} = \{h_i \otimes K_L(t,\cdot)\otimes h_j : i,j \le L_0, t\in\mathbb{R}\}$. Then there exists a constant $C = C(\mathcal{H}) > 0$ such that, with brackets whose L^∞-diameter is at most $C2^L$, one achieves the following bound for the bracketing numbers for $\varepsilon \le C2^L$:*

$$N_{[]}(\mathcal{T}, \|\cdot\|_{L^2(\Pi_H)}, \varepsilon) \le CL_0^2 \max(2^{2L(1+1/\nu)}\varepsilon^{-(1+2/\nu)}, 1). \tag{65}$$

Proof The kernel K (from which K_L is constructed) is assumed to be bounded, continuous, Lipschitz, and supported in $[-1,1]$, see before (25).

Let $\mathcal{U} = \{K_L(t,\cdot) : t\in\mathbb{R}\}$. Then writing $h = \max_l \|h_l\|_\infty$,

$$N_{[]}(\mathcal{T}, \|\cdot\|_{L^2(\Pi_H)}, 4\varepsilon h^2) \le L_0^2 N_{[]}(\mathcal{U}, \|\cdot\|_{L^2(\Pi_H)}, \varepsilon) \tag{66}$$

since given brackets $[\underline{v}_k, \overline{v}_k], k \le N_{\mathcal{U}}$ of $L^2(\Pi_H)$-diameter at most ε for $\mathcal{U}$, we can define

$$\begin{aligned}\underline{t}_{ikj} &= h_i \otimes \underline{v}_k \otimes h_j \mathbb{1}_{h_i\otimes 1\otimes h_j \ge 0} + h_i \otimes \overline{v}_k \otimes h_j \mathbb{1}_{h_i\otimes 1\otimes h_j < 0},\\ \overline{t}_{ikj} &= h_i \otimes \overline{v}_k \otimes h_j \mathbb{1}_{h_i\otimes 1\otimes h_j \ge 0} + h_i \otimes \underline{v}_k \otimes h_j \mathbb{1}_{h_i\otimes 1\otimes h_j < 0}\end{aligned}$$

to obtain brackets $[\underline{t}_{ikj}, \overline{t}_{ikj}], i,j \le L_0, k \le N_{\mathcal{U}}$ for $\mathcal{T}$ whose $L^2(\Pi_H)$-diameter is at most $4h^2\varepsilon$ and whose L^∞-diameter is at most h^2 times that of the brackets for $\mathcal{U}$.

Under Assumption B, there exists a constant $C = C(\mathcal{H}) > 0$ such that for $T_\varepsilon = C2^{(2L+2)/\nu}\varepsilon^{-2/\nu}$ we have $\Pi_H(|X_1| > T_\varepsilon) \le (\|K\|_\infty 2^{L+1})^{-2}\varepsilon^2$. Observe that for any t such that $|t| > T_\varepsilon + 1$, the support of $K_L(t,\cdot)$ does not intersect $[-T_\varepsilon, T_\varepsilon]$. It follows for any such t that $\underline{v} = -\|K\|_\infty 2^L \mathbb{1}_{[-T_\varepsilon,T_\varepsilon]^c}$ and $\overline{v} = \|K\|_\infty 2^L \mathbb{1}_{[-T_\varepsilon,T_\varepsilon]^c}$ bracket $K_L(t,\cdot)$; the $L^2(\Pi_H)$-diameter of this bracket is at most ε and the L^∞-diameter at most $2^{L+1}\|K\|_\infty$. Writing $\mathcal{U}_\varepsilon = \{K_L(t,\cdot) : |t| \le T_\varepsilon + 1\}$, we deduce that

$$N_{[]}(\mathcal{U}, \|\cdot\|_{L^2(\Pi_H)}, \varepsilon) \le N_{[]}(\mathcal{U}_\varepsilon, \|\cdot\|_{L^2(\Pi_H)}, \varepsilon) + 1. \tag{67}$$

To bound the right side, observe that $|K_L(t,x) - K_L(s,x)| \le 2^{2L}\Lambda|s-t|$ for each $s,t,x\in\mathbb{R}$, where Λ denotes the Lipschitz constant of K. Since the set $[-T_\varepsilon - 1, T_\varepsilon + 1]$ is compact, we deduce that for a constant $C = C(\mathcal{H})$

$$N_{[]}(\mathcal{U}_\varepsilon, \|\cdot\|_{L^2(\Pi_H)}, \varepsilon) \le C\max\big(2^{2L(1+1/\nu)}\varepsilon^{-(1+2/\nu)}, 1\big),$$

see for example Theorem 2.7.11 in van der Vaart and Wellner (1996) (applied with $2^{2L+1}\Lambda\varepsilon$ in place of ε, and in view of the proof of which the brackets can be taken to have L^∞-diameter at most $\varepsilon \le C2^L$). Together with (66) and (67), this yields the result. ■

B.3 Matrix Approximation Theory Arguments

Lemma 32. *Define $\mathcal{A}$ as in Lemma 25. In the setting of Theorem 5, define $\hat{R}$ as in Algorithm 1 for $2^L \asymp (N/\log N)^{1/(1+2s)}$, and define $\tilde{R}$ to have columns equal to the normalised columns of $QO^\intercal \hat{V}$. Then, on $\mathcal{A}$, $\hat{R}$ is well-defined and*

$$\|\hat{R} - \tilde{R}_\tau\| \leq \|\hat{R} - \tilde{R}_\tau\|_F \leq C L_0^{7/2} r_N,$$

for some $C = C(\mathcal{H})$ and some permutation τ, where $\tilde{R}_\tau$ is obtained by permuting the columns of $\tilde{R}$ according to τ.

Remark. Strictly speaking the columns of $\hat{R}$, as eigenvectors of $\hat{B}^{\hat{a},\hat{u}}$, are defined only up to signs, and this result holds only for one set of choices of signs. However, the estimators $\tilde{f}_j^L(x) = (\hat{R}^{-1}\hat{B}^x\hat{R})_{jj}$ are unaffected by the choices of signs, hence we may assume without loss of generality that these signs are chosen appropriately for the lemma to hold.

Proof Lemma 27 tells us on $\mathcal{A}$ that $\hat{B}^{\hat{a},\hat{u}}$ is diagonalisable, so that $\hat{R}$ is well defined, and moreover that

$$\min\bigl(\operatorname{sep}(\hat{B}^{\hat{a},\hat{u}}), \operatorname{sep}(\tilde{B}^{\hat{a},\hat{u}})\bigr) > c,$$

for some constant $c = c(\mathcal{H}) > 0$. Now we apply Lemma C.3 from Anandkumar et al. (2012), which says, as a consequence of the Bauer–Fike theorem, that if

$$\varepsilon = \kappa(\tilde{R}) \operatorname{sep}(\tilde{B}^{\hat{a},\hat{u}})^{-1} \|\hat{B}^{\hat{a},\hat{u}} - \tilde{B}^{\hat{a},\hat{u}}\|$$

is smaller than 1/2, then there exists a permutation τ such that

$$\|\hat{R} - \tilde{R}_\tau\| \leq \|\hat{R} - \tilde{R}_\tau\|_F \leq 4J^{1/2}(J-1)\|\tilde{R}^{-1}\|\varepsilon.$$

By construction $\sum |\hat{a}_i| \leq 1$, hence by the triangle inequality and Lemma 25, on $\mathcal{A}$ we have

$$\|\hat{B}^{\hat{a},\hat{u}} - \tilde{B}^{\hat{a},\hat{u}}\| \leq \sum_i |\hat{a}_i| \|\hat{B}^{\hat{u}_i} - \tilde{B}^{\hat{u}_i}\| \leq \sup_x \|\hat{B}^x - \tilde{B}^x\| \leq C L_0^2 r_N,$$

for some $C = C(\mathcal{H})$. By Lemma 35b, we have $\kappa(\tilde{R}) \leq CL_0$ and $\|\tilde{R}^{-1}\| \leq CL_0^{1/2}$. We deduce that $\varepsilon \to 0$ on $\mathcal{A}$, hence is smaller than 1/2 for large N, and the result follows. ■

One could directly use the Ostrowski–Elsner theorem (Theorem 36) to obtain a version of Theorem 5 with a suboptimal estimation rate. We here go through the slightly circuitous route of using Theorem 36 to prove an eigen-separation condition (i.e. Lemma 27) and deducing Lemma 32 because we may then apply the following lemma, adapted from Lemma C.4 of Anandkumar et al. (2012), to obtain a near-minimax rate instead.

Lemma 33. *Suppose $(A_t : t \in \mathcal{T})$ are $J \times J$ matrices simultaneously diagonalised by a matrix R with unit norm columns:*

$$R^{-1} A_t R = \operatorname{diag}(\lambda_{t,1}, \dots, \lambda_{t,J}), \;\; t \in \mathcal{T}.$$

Let $\hat{R}$ be a matrix such that for some permutation τ of $\{1, \dots, J\}$ we have

$$\|\hat{R} - R_\tau\| := \varepsilon_R \leq (1/2)\|R^{-1}\|^{-1},$$

where R_τ has is obtained by permuting the columns of R according to τ. Assume

$$\lambda_{\max} := \sup_t \max_j |\lambda_{t,j}| < \infty.$$

For matrices $(\hat{A}_t : t \in \mathcal{T})$, *write*

$$\varepsilon_A := \sup_t \|A_t - \hat{A}_t\|,$$

and define

$$\hat{\lambda}_{t,j} = e_j^\intercal \hat{R}^{-1}\hat{A}_t\hat{R}e_j.$$

Then

$$\sup_t \max_j |\hat{\lambda}_{t,j} - \lambda_{t,\tau(j)}| \leq 4\kappa(R)[\varepsilon_A + \lambda_{\max}\|R^{-1}\|\varepsilon_R].$$

Proof Let $\hat{\zeta}_j^\intercal$ be the jth row of $\hat{R}^{-1}$, let $\hat{\xi}_j$ be the jth column of $\hat{R}$, and define ζ_j, ξ_j correspondingly with respect to the matrix R_τ obtained by permuting the columns of R according to τ. Then $\lambda_{t,\tau(j)} = \zeta_j^\intercal A_t \xi_j$, $\hat{\lambda}_{t,j} = \hat{\zeta}_j^\intercal \hat{A}_t \hat{\xi}_j$, and we have

$$\begin{aligned}
|\hat{\lambda}_{t,j} - \lambda_{t,\tau(j)}| &= |\hat{\zeta}_j^\intercal \hat{A}_t \hat{\xi}_j - \zeta_j^\intercal A_t \xi_j| \\
&= |\hat{\zeta}_j^\intercal \hat{A}_t (\hat{\xi}_j - \xi_j) + \hat{\zeta}_j^\intercal (\hat{A}_t - A_t)\xi_j + (\hat{\zeta}_j^\intercal - \zeta_j^\intercal) A_t \xi_j| \\
&\leq \|\hat{\zeta}_j^\intercal\| \|\hat{A}_t\| \|\hat{\xi}_j - \xi_j\| + \|\hat{\zeta}_j^\intercal\| \|\xi_j\| \varepsilon_A + \|A_t \xi_j\| \|\hat{\zeta}_j - \zeta_j\|
\end{aligned}$$

Using Lemma 37, we have that

$$\|\hat{R}^{-1} - R_\tau^{-1}\| \leq \|R^{-1}\|^2 \varepsilon_R / (1 - \|R^{-1}\|\varepsilon_R),$$

and we further note the following:

- $\|\zeta_j^\intercal\| = \|e_{\tau(j)}^\intercal R^{-1}\| \leq \|R^{-1}\|$, and $\|\hat{\zeta}_j^\intercal - \zeta_j^\intercal\| \leq \|\hat{R}^{-1} - R_\tau^{-1}\| \leq \|R^{-1}\|^2\varepsilon_R/(1 - \|R^{-1}\|\varepsilon_R)$, so that also $\|\hat{\zeta}_j^\intercal\| \leq \|\zeta_{\tau(j)}^\intercal\| + \|\hat{\zeta}_j - \zeta_{\tau(j)}\| \leq \|R^{-1}\|/(1 - \|R^{-1}\|\varepsilon_R)$.
- $\|\xi_j\| \leq \|R\|$, and $\|\hat{\xi}_j - \xi_j\| \leq \|\hat{R} - R_\tau\| = \varepsilon_R$.
- $\|A_t\| = \|R \operatorname{diag}(\lambda_{t,\cdot})R^{-1}\| \leq \kappa(R)\lambda_{\max}$, and $\|\hat{A}_t\| \leq \|A_t\| + \varepsilon_A \leq \kappa(R)\lambda_{\max} + \varepsilon_A$.
- $\|A_t\xi_j\| = |\lambda_{t,\tau(j)}| \|\xi_j\| \leq \lambda_{\max}\|R\|$.

Then, continuing the inequalities from the display, we have

$$\begin{aligned}
|\hat{\lambda}_{t,j} - \lambda_{t,\tau(j)}| &\leq \frac{\|R^{-1}\|}{1 - \|R^{-1}\|\varepsilon_R}\Big[(\kappa(R)\lambda_{\max} + \varepsilon_A)\varepsilon_R + \|R\|\varepsilon_A\Big] + \lambda_{\max}\|R\|\|R^{-1}\|^2 \frac{\varepsilon_R}{1 - \|R^{-1}\|\varepsilon_R} \\
&\leq \frac{\kappa(R) + \|R^{-1}\|\varepsilon_R}{1 - \|R^{-1}\|\varepsilon_R}\varepsilon_A + 2\lambda_{\max}\kappa(R)\frac{\|R^{-1}\|\varepsilon_R}{1 - \|R^{-1}\|\varepsilon_R} \\
&\leq (1 + 2\kappa(R))\varepsilon_A + 4\lambda_{\max}\|R^{-1}\|\kappa(R)\varepsilon_R,
\end{aligned}$$

where for the last line we have used that $\|R^{-1}\|\varepsilon_R \leq 1/2$ by assumption. Taking the supremum over $t \in \mathcal{T}$ concludes the result since necessarily $1 + 2\kappa(R) \leq 3\kappa(R) < 4\kappa(R)$. ■

Lemma 34. *Define* $O = O^{L_0}, P = P^{L_0}, (M^x = M^{x,L,L_0} : x \in \mathbb{R})$ *as in Lemma 11 for functions* $(h_l)_{l \leq L_0}$ *satisfying a sup-norm bound uniformly in* L_0 *and assume that* $\sigma_J(O) \geq c > 0$ *uniformly in* $L_0 \geq \underline{L}$ *for some* $\underline{L} = \underline{L}(\mathcal{H})$ *(for example, by choosing* $(h_l : l \leq L_0)$ *as in Lemma 24). Then*

$$\kappa(O) \leq CL_0^{1/2}, \quad \sigma_J(P) \geq c', \quad \textit{and} \quad \|M^x\| \leq C'L_0,$$

for some constants $c', C, C' > 0$*, uniformly in* $L_0 \geq \underline{L}$ *and all* L*.*

Proof Given the assumed bound on $\sigma_J(O)$, to control $\kappa(O)$ it remains to bound $\|O\|$, since one has the standard expression $\kappa(O) := \|O\|\|O^{-1}\| \equiv \|O\|/\sigma_J(O)$. Then it suffices to note, using Cauchy–Schwarz and the fact that $|\langle f_j, h_l\rangle| = |\int h_l(x)f_j(x)\,\mathrm{d}x| \leq \|h_l\|_\infty$, that

$$\|O\|^2 = \sup_{\|v\|=1} \sum_j \Big(\sum_l v_l \langle f_j, h_l\rangle\Big)^2 \leq \max_l \|h_l\|_\infty^2 J L_0. \tag{68}$$

Next, Assumption C' implies $\sigma_J(Q) > 0$ and $\sigma_J(\mathrm{diag}(\pi)) = \min_j \pi_j > 0$. Using submultiplicativity of σ_J (see Lemma 37) and the expression $P = O\,\mathrm{diag}(\pi)Q^2O^\intercal$ (from Lemma 11), we have

$$\sigma_J(P) = \sigma_J(O\,\mathrm{diag}(\pi)Q^2O^\intercal) \geq \sigma_J(O)\sigma_J(\mathrm{diag}(\pi))\sigma_J(Q)^2\sigma_J(O^\intercal) \geq c'(H) > 0.$$

For M^x, the expression $M^x = O\,\mathrm{diag}(\pi)QD^xQO^\intercal$ from Lemma 11 similarly yields

$$\|M^x\| \leq \|O\|^2\|Q\|^2 \max_j |K_L[f_j](x)|.$$

Recalling that $\|K_L[f_j]\|_\infty$ is bounded (see (27)) we deduce the result. ∎

The following collects several useful results from De Castro et al. (2017) and Anandkumar et al. (2012).

Lemma 35. *Define $O, \hat{O}, P, \hat{P}$ and $M^x, \hat{M}^x, \tilde{B}^x, \hat{B}^x, x \in \mathbb{R}$ as in Lemma 11, Algorithm 1, and* (53)*. Assume $\sigma_J(O) \geq c > 0$ uniformly in $L_0 \geq \underline{L}$, so that by Lemma 34 we also have $\sigma_J(P) > 0$ and $\kappa(O) \leq CL_0^{1/2}$ for some C. On the event $\mathcal{B} = \{\|\hat{P} - P\| < \sigma_J(P)/3\}$, for $L_0 \geq \underline{L}$ and N large enough we have the following.*

a. *$\sigma_J(\hat{P}) > c/2$. Writing $\hat{V}$ and V for matrices of orthonormal right singular vectors of $\hat{P}$ and P respectively we have $\sigma_J(\hat{V}^\intercal V)^2 \geq 3/4$, and consequently $\hat{V}^\intercal P\hat{V}$ is invertible.*

b. *$\kappa(QO^\intercal\hat{V}) \leq CL_0^{1/2}$, $\|\tilde{R}^{-1}\| \leq C'L_0^{1/2}$ and $\kappa(\tilde{R}) \leq C''L_0$, where $\tilde{R}$ is the matrix whose columns are those of $QO^\intercal\hat{V}$ but rescaled to have unit norm.*

c. *For any $x \in \mathbb{R}$,*

$$\|\tilde{B}^x - \hat{B}^x\| \leq 3.2\Big[\frac{\|\hat{M}^x - M^x\|}{\sigma_J(P)} + \frac{\|M^x\|\|\hat{P} - P\|}{\sigma_J(P)^2}\Big].$$

Proof We throughout use various basic properties of σ_J, κ, which are summarised in Lemma 37 below.

a. By Lemma 34, $\sigma_J(P) > 0$. The result then follows from standard approximation theory. In particular Lemma C.1 part 2 of Anandkumar et al. (2012) tells us that $\sigma_J(\hat{P}) > \sigma_J(P)/3 > 0$. That $\sigma_J(V^\intercal\hat{V})^2 \geq 3/4$ on $\mathcal{B}$ is given by Lemma C.1 part 3 of the same reference and submultiplicativity of σ_J yields

$$\sigma_J(\hat{V}^\intercal P\hat{V}) = \sigma_J(\hat{V}^\intercal(VV^\intercal)P(VV^\intercal)\hat{V}) \geq \sigma_J(V^\intercal\hat{V})^2\sigma_J(V^\intercal PV) \geq (3/4)\sigma_J(P) > 0,$$

which implies invertibility of $\hat{V}^\intercal P\hat{V}$.

b. Observe that

$$\kappa(QO^\intercal\hat{V}) = \frac{\|QO^\intercal\hat{V}\|}{\sigma_J(QO^\intercal\hat{V})} \leq \frac{\|QO^\intercal\|}{\sigma_J(QO^\intercal V)\sigma_J(V^\intercal\hat{V})}.$$

We have $\sigma_J(QO^\intercal V) = \sigma_J(QO^\intercal)$ and we deduce that $\kappa(QO^\intercal\hat{V}) \leq (4/3)^{1/2}\kappa(QO^\intercal) \leq 2\kappa(Q)\kappa(O)$ by part a. Assumption C' implies $\kappa(Q) < \infty$. For R, see Lemma C.5 of Anandkumar et al. (2012), which tells us that $\|\tilde{R}^{-1}\| \leq \kappa(QO^\intercal\hat{V})$ and $\kappa(\tilde{R}) \leq \kappa(QO^\intercal\hat{V})^2$.

c. One adapts the proof of Lemma F.4 in De Castro et al. (2017), decomposing

$$\|\tilde{B}^x - \hat{B}^x\| \leq \|(\hat{V}^\intercal \hat{P}\hat{V})^{-1}\| \|\hat{V}^\intercal (M^x - \hat{M}^x)\hat{V}\| + \|\hat{V}^\intercal M^x \hat{V}\| \|(\hat{V}^\intercal P\hat{V})^{-1} - (\hat{V}^\intercal \hat{P}\hat{V})^{-1}\|,$$

then using Lemma 37 with $\hat{A} = \hat{V}^\intercal \hat{B}^x \hat{V}$, $A = \hat{V}^\intercal \tilde{B}^x \hat{V}$, noting that in part a we showed $\|(\hat{V}^\intercal P\hat{V})^{-1}\| \equiv \sigma_J(\hat{V}^\intercal P \hat{V})^{-1} \leq (4/3)\sigma_J(P)^{-1}$.

■

Theorem 36 (Ostrowski–Elsner, e.g. Stewart and Sun 1990, Chapter IV, Theorem 1.4)**.** *For a matrix $U \in \mathbb{R}^{J\times J}$, write $(\lambda_i(U) : i \leq J)$ for the eigenvalues of U. Then for matrices $A, B \in \mathbb{R}^{J\times J}$ we have*

$$\min_\tau \max_j |\lambda_{\tau(j)}(A) - \lambda_j(B)| \leq (2J-1)(\|A\| + \|B\|)^{(J-1)/J} \|A - B\|^{1/J}, \tag{69}$$

where the minimum is over permutations τ.

Lemma 37. *Let A and $\hat{A}$ be matrices such that A is invertible and $\|A - \hat{A}\| < \|A^{-1}\|^{-1}$. Then $\hat{A}$ is invertible and*

$$\|\hat{A}^{-1} - A^{-1}\| \leq \frac{\|A^{-1}\|^2 \|A - \hat{A}\|}{1 - \|A^{-1}\|\|A - \hat{A}\|}.$$

We also have the following: $\kappa(A) = \kappa(A^\intercal)$; $\sigma_J(A) = \sigma_J(A^\intercal)$; $\sigma_J(A) = \sigma_J(AW^\intercal)$ for any matrix W whose columns are orthonormal and whose domain is $\mathbb{R}^J$; $\sigma_J(AB) \geq \sigma_J(A)\sigma_J(B)$, and $\kappa(AB) \leq \kappa(A)\kappa(B)$ for matrices A, B.

Proof For the first see Theorem 2.5 in Chapter III of Stewart and Sun (1990) The other results can be found in Chapter I.4 of the same reference. ■

B.4 Sketch Proof of Theorem 4

The arguments used to prove Theorem 5 work also in this discrete setting, given the following observations and slight adaptations. To ease notation we assume that $f_j(x) = 0$ for all $x \leq 0$ and $j \leq J$. We make the following definitions, which correspond to taking $h_l = \mathbb{1}_l$, i.e. $h_l(x) = \mathbb{1}\{x = l\}$, and replacing $K_L(x, y)$ by $\mathbb{1}\{x = y\}$:

$$\begin{aligned} M^x = M^{x,L_0} &= \Pi_H(X_1 = l, X_2 = x, X_3 = m)_{l,m\leq L_0}, \quad x \in \mathbb{N} \\ P = P^{L_0} &= \Pi_H(X_1 = l, X_3 = m)_{l,m\leq L_0}, \\ O = O^{L_0} &= \Pi_H(X_1 = l \mid \theta_1 = j)_{l\leq L_0, j\leq J}, \\ D^x &= (\operatorname{diag} \Pi_H(X_2 = x \mid \theta_2 = j)_{j\leq J}) \equiv \operatorname{diag}((O_{xj})_j). \end{aligned}$$

The proof of Lemma 11 is unchanged with these adjusted definitions, and we adapt the definitions in Algorithm 1 correspondingly:

$$\begin{aligned} \hat{M}^x &= \Big(\frac{1}{N}\sum_{n\leq N} \mathbb{1}_l(X_n)\mathbb{1}_x(X_{n+1})\mathbb{1}_m(X_{n+2})\Big)_{l,m\leq L_0}, \\ \hat{P} &= \Big(\frac{1}{N}\sum_{n\leq N} \mathbb{1}_l(X_n)\mathbb{1}_m(X_{n+2})\Big)_{l,m\leq L_0}, \\ \hat{B}^x &= (\hat{V}^\intercal \hat{P}\hat{V})^{-1}\hat{V}^\intercal \hat{M}^x \hat{V}, \end{aligned}$$

with $\hat{V}$ comprising right singular vectors of $\hat{P}$.

Observe that the proofs of Lemmas 24 and 34 work in the current setting for the current choice of the h_l [indeed, thanks to the disjoint support of h_l, h_m for $l \neq m$ one can improve the bound in eq. (68) to $\|O^{L_0}\| \leq J$], and similarly a version of Lemma 26 holds by choosing $\mathbb{D}_N = \mathcal{A}_N \times \mathcal{U}_N$ for sequences of finite sets $\mathcal{A}_N \subset \mathbb{R}$, $\mathcal{U}_N \subset \mathbb{N}$ such that $\cup_N \mathcal{U}_N = \mathbb{N}$ and $\cup_N \mathcal{A}_N$ is dense in $\{a \in \mathbb{R} : \sum_i |a_i| \leq 1\}$.

Next note that a version of the Glivenko–Cantelli theorem gives control over $\sup_{x\in\mathbb{N}} \|\hat{M}^x - M^x\|$ for our new definitions of $\hat{M}^x, M^x$; we give here a slightly indirect proof of this fact by reusing the machinery of Lemma 30. Indeed, inspecting the proof of Lemma 31, one deduces that for $\mathcal{U} = \{\mathbb{1}_{\{t\}} : t \in \mathbb{R}\}$,

$$N_{[]}(\mathcal{T}, \|\cdot\|_{L^2(\Pi_H)}, \varepsilon) \leq L_0^2 N_{[]}(\mathcal{U}, \|\cdot\|_{L^2(\Pi_H)}, \varepsilon/(4h^2)).$$

Mimicking the proof of Glivenko–Cantelli, to bound the latter quantity one choose M of order ε^{-2} and $-\infty = t_0 < t_1 < \cdots < t_M = +\infty$ such that $\Pi_H(X_1 \in [t_m, t_{m+1}))$ is roughly ε^2. Then the functions $\underline{u}_k = 0$, $\overline{u}_k = \mathbb{1}_{[t_k, t_{k+1})}$ bracket $\mathcal{U}$, with $L^2(\Pi_H)$-diameter of order ε and L^∞-diameter 1, yielding for a constant C that

$$H_{[]}(\mathcal{T}, \|\cdot\|_{L^2(\Pi_H)}, \varepsilon) \leq C \log(L_0 \varepsilon^{-1}).$$

In view of the standard bound (see (63))

$$\int_0^x \sqrt{\log(1/u)}\, \mathrm{d}u \leq x(1 + \sqrt{\log(1/x)}),$$

and recalling as in Lemma 29 that the chain

$$Y_n = (X_n, X_{n+1}, X_{n+2}, \theta_n, \theta_{n+1}, \theta_{n+2})$$

has pseudo-spectral gap bounded away from zero by Assumption C', it follows that Lemma 28, applied with $b = \sigma = 1$, yields

$$\Pi_H\big(\sup_{x\in\mathbb{N}} |\hat{M}^x_{ij} - M^x_{ij}| > C(N^{-1/2} + N^{-1/2}\sqrt{u} + N^{-1}u)\big) \leq \exp(-u).$$

We note that $\|\hat{M}^x - M^x\| \leq L_0 \max_{ij} |\hat{M}^x_{ij} - M^x_{ij}|$. Combining with (the proof of) Lemma 29, for any $c_N \to \infty$, we may choose suitable $L_0 \to \infty$ and $u \to \infty$ to deduce

$$\Pi_H(\|\hat{P} - P\| \leq c_N N^{-1/2},\ \sup_{x\in\mathbb{N}} \|\hat{M}^x - M^x\| \leq c_N N^{-1/2}) \to 1.$$

The rest of the proof exactly mirrors that of Theorem 5 (note that, in only seeking a rate in probability, we avoid the need for a log factor which would appear in this proof if seeking a rate in expectation).

Appendix C. Proof of the Lower Bound

For the lower bound for simplicity we consider in details the (in view of the multiple testing application) most relevant case $J = 2$. The case of $J \geq 3$ is broadly similar in spirit and is briefly discussed at the end of this section.

Let $\mathfrak{S}_2$ denote the set of all permutations of $\{0, 1\}$. Define, for $s, R > 0$ and for the Hölder space C^s defined in Assumption E,

$$\mathcal{C}^s(R) = \Big\{ f \in C^s : f \geq 0,\ \int_{\mathbb{R}} f = 1,\ \|f\|_{C^s} \leq R \Big\}. \tag{70}$$

Parameters. The unknown parameters are $H = (Q, \pi, \mathbf{f})$, where $\mathbf{f} = (f_0, f_1)$ denotes the vector of emission densities. Denoting by P_{f_i} the distribution of density f_i on $\mathbb{R}$, $i = 0, 1$, the distribution of the observations $X = (X_1, \ldots, X_N)$ is

$$\Pi_H = \Pi_H^{(N)} = \sum_{\mathbf{v} \in \{0,1\}^N} w_{\mathbf{v}} \bigotimes_{j=1}^{N} P_{\mathbf{f}_{v_j}},$$

where $w_{\mathbf{v}}$ denotes the probability under the Markov chain to observe the successive sequence of states $(v_1, \ldots, v_N) \in \{0,1\}^N$; that is, $w_{(v_1,\ldots,v_N)} = \pi_{v_1} Q_{v_1,v_2} \cdots Q_{v_{N-1},v_N}$.

Class $\mathcal{H}_{sep}$ of well-separated parameters. Let $\mathcal{F}_{sep}$ be a class of pairs $\mathbf{f} = (f_0, f_1)$ that are well-separated in the following sense, for a (small) $d > 0$ to be chosen:

$$\mathcal{F}_{sep} = \left\{ \mathbf{f} = (f_0, f_1) \in \mathcal{C}^s(R)^2 : \ |(f_1 - f_0)(0)| \geq d, \ |P_{f_1}([-1,1]) - P_{f_0}([-1,1])| \geq d \right\}. \tag{71}$$

We define, for given Q, π,

$$\mathcal{H}_{sep} = \mathcal{H}_{sep}(Q, \pi, R, d, s) = \{ H = (Q, \pi, \mathbf{f}) : \ \mathbf{f} \in \mathcal{F}_{sep} \}. \tag{72}$$

Minimax risk. For $\mathbf{f} = (f_0, f_1)$ and $\mathbf{g} = (g_0, g_1)$ two pairs of real functions, denote

$$\rho(\mathbf{f}, \mathbf{g}) = \min_{\varphi \in \mathfrak{S}_2} \left(\|g_{\varphi(0)} - f_0\|_\infty + \|g_{\varphi(1)} - f_1\|_\infty \right). \tag{73}$$

The loss ρ is a pseudo-metric, verifying the axioms of a distance except that one can have $\rho(\mathbf{f}, \mathbf{g}) = 0$ for $\mathbf{f} \neq \mathbf{g}$. We note that one could also consider the equivalent loss obtained by replacing the sum in (73) with a maximum.

Let us consider the minimax risk

$$R_n = R_n(\mathcal{H}_{sep}) = \inf_{\mathbf{T}=(T_0,T_1)} \sup_{H \in \mathcal{H}_{sep}} E_H \left[\rho\left(\mathbf{T}, \mathbf{f}\right) \right]. \tag{74}$$

Since $E[\min(X, Y)] \leq \min(EX, EY)$, one notes that

$$R_n \leq \inf_{\mathbf{T}=(T_1,T_2)} \sup_{H \in H_{sep}} \left[\min_{\varphi \in \mathfrak{S}_2} \left(E_H \|T_{\varphi(1)} - f_1\|_\infty + E_H \|T_{\varphi(2)} - f_2\|_\infty \right) \right]. \tag{75}$$

In view of Section 4.3 (and constructing $\hat{f}_0, \hat{f}_1$ using $L_0 = 2$, $h_1 = 1$, $h_2 = \mathbb{1}_{[-1,1]}$ in Algorithm 1), Theorem 5 provides a procedure for which the last quantity is bounded from above by a multiple of $r_N = (N/\log N)^{-s/(2s+1)}$. The next result provides the corresponding minimax lower bound. Note that the lower bound in Proposition 38 is pointwise in Q and π, and thus continues to hold if π, Q are allowed to vary in some set.

Proposition 38. *Consider $J = 2$ classes, and fix both $\pi = (\pi_0, \pi_1) \in [0,1]^2$ and Q a 2×2 transition matrix. Given $s, R > 0$, let $\mathcal{H}_{sep}$ be as in* (72) *for a small enough $d = d(s, R)$, and let $R_n = R_n(\mathcal{H}_{sep})$ be as in* (74)*. Then there exists $C = C(s, R) > 0$ such that, for N large enough,*

$$R_n(\mathcal{H}_{sep}) \geq C \left(\frac{\log N}{N} \right)^{\frac{s}{2s+1}}.$$

Proof We reduce the estimation problem to a classification problem in a standard way. Suppose the two sets of densities $\{f_0^{(m)}, \ 0 \leq m \leq M\}$ and $\{f_1^{(m)}, \ 0 \leq m \leq M\}$ are such that for some $0 < s_0, s_1 < C_0$,

$$\min\{\|f_1^{(i)} - f_0^{(j)}\|_\infty, \ 0 \leq i, j \leq M\} \geq C_0, \tag{76}$$

$$\min\{\|f_0^{(i)} - f_0^{(j)}\|_\infty : 0 \leq i, j \leq M, i \neq j\} \geq 2s_0, \tag{77}$$

$$\min\{\|f_1^{(i)} - f_1^{(j)}\|_\infty : 0 \leq i, j \leq M, i \neq j\} \geq 2s_1. \tag{78}$$

It follows that the family of functions $\mathbf{f}^{(m)} = (f_0^{(m)}, f_1^{(m)})$ is $2(s_0+s_1)$-separated in terms of ρ, since for $m \neq m'$,

$$\begin{aligned}\rho(\mathbf{f}^{(m)}, \mathbf{f}^{(m')}) &\geq \min\Big(\|f_0^{(m)} - f_0^{(m')}\|_\infty + \|f_1^{(m)} - f_1^{(m')}\|_\infty, \|f_1^{(m)} - f_0^{(m')}\|_\infty + \|f_0^{(m)} - f_1^{(m')}\|_\infty\Big) \\ &\geq \min(2(s_0+s_1), 2C_0) = 2(s_0+s_1) =: 2S.\end{aligned}$$

For a given estimator $\mathbf{T}$ of $\mathbf{f} \in \{\mathbf{f}^{(0)}, \dots, \mathbf{f}^{(M)}\}$, let $j^*(\mathbf{T})$ be the index j such that $\mathbf{f}^{(j)}$ is the closest to $\mathbf{T}$ in the ρ pseudo-distance. Since the family $(\mathbf{f}^{(m)}, m \in \{0, \dots, M\})$ is $2S$-separated, we have $\rho(\mathbf{T}, \mathbf{f}^{(m)}) \geq S \mathbb{1}\{j^*(\mathbf{T}) \neq m\}$. Writing $H_m = (Q, \pi, \mathbf{f}^{(m)})$, we have

$$\begin{aligned}\sup_{H\in\mathcal{H}_{sep}} E_H\left[\rho\left(\mathbf{T}, \mathbf{f}\right)\right] &\geq \max_{0\leq m\leq M} E_{H_m}\left[\rho\left(\mathbf{T}, \mathbf{f}^{(m)}\right)\right] \\ &\geq S \max_{0\leq m\leq M} \Pi_{H_m}[j^*(\mathbf{T}) \neq m] \geq S p_{e,M}, \end{aligned} \tag{79}$$

where $p_{e,M} = \inf_\psi \max_{0\leq m\leq M} \Pi_{H_m}[\psi \neq m]$, with the infimum being over all classifiers ψ. Taking the infimum with respect to $\mathbf{T}$ in (79), one obtains $R_n(\mathcal{H}_{sep}) \geq S p_{e,M}$.

Lemma 40 shows that in order to bound $p_{e,M}$ from below it suffices to bound $\mathrm{KL}(\Pi_{H_m}, \Pi_{H_0})$ from above, where $\mathrm{KL}(P, Q)$ denotes the Kullback-Leibler divergence between distributions P and Q with densities p, q,

$$\mathrm{KL}(P, Q) = E_P\Big[\log\Big(\frac{p}{q}\Big)\Big]. \tag{80}$$

By convexity of the map $(x, y) \to x \log(x/y)$, writing $\mathbf{v} = (v_j) \in \{0, 1\}^N$, one obtains

$$\mathrm{KL}(\Pi_{H_m}, \Pi_{H_0}) \leq \sum_{\mathbf{v}\in\{0,1\}^N} w_{\mathbf{v}}\, \mathrm{KL}\left(\bigotimes_{j=1}^N P_{f_{v_j}^{(m)}}, \bigotimes_{j=1}^N P_{f_{v_j}^{(0)}}\right).$$

For a given $\mathbf{v} \in \{0, 1\}^N$, let $n_i(\mathbf{v})$, $i = 0, 1$, denote the number of elements of $\mathbf{v}$ equal to i. The tensorisation property of the KL divergence implies

$$\mathrm{KL}\left(\bigotimes_{j=1}^N P_{f_{v_j}^{(m)}}, \bigotimes_{j=1}^N P_{f_{v_j}^{(0)}}\right) = n_0(\mathbf{v})\, \mathrm{KL}(P_{f_0^{(m)}}, P_{f_0^{(0)}}) + n_1(\mathbf{v})\, \mathrm{KL}(P_{f_1^{(m)}}, P_{f_1^{(0)}}),$$

where $n_0(\mathbf{v}), n_1(\mathbf{v})$ are both at most N.

Let us now choose functions $f_0^{(m)}, f_1^{(m)}$, satisfying eqs. (76) to (78) for which we have good control over $\mathrm{KL}(f_j^{(m)}, f_j^{(0)})$, $j = 0, 1$ and $1 \leq m \leq M$. For ϕ the standard normal density and $g_{m,A}$ defined as in Lemma 39 to follow, set

$$\begin{aligned} f_0^{(m)}(x) &= g_{m,A}(x), & m &\geq 1, & f_0^{(0)}(x) &= r\phi(rx), \\ f_1^{(m)}(x) &= g_{m,A}(x - 2/r), & m &\geq 1, & f_1^{(0)}(x) &= r\phi(r(x - 2/r)), \end{aligned}$$

where we choose

$$A = c_0\Big(\frac{\log N}{N}\Big)^{\frac{s}{2s+1}}, \quad M = \Big\lceil\Big(\frac{N}{\log N}\Big)^{\frac{1}{2s+1}}\Big\rceil, \tag{81}$$

with r, c_0 small, but fixed, positive constants. Note firstly that for r, c_0 small enough (and N large enough) each pair $(f_0^{(m)}, f_1^{(m)})$ is in $\mathcal{F}_{sep}$ for given $R > 0$ and a small enough constant $d > 0$. Indeed, examining the definition of $g_{m,A}$ from Lemma 39, we see for all $0 \leq m \leq M$ that we have

$$|f_1^{(m)}(0) - f_0^{(m)}(0)| = r(\phi(0) - \phi(2));$$

that $P_{f_0^{(m)}}[-1,1] = \int_{-r}^{r} \phi$ and $P_{f_1^{(m)}}[-1,1] = \int_{-r-2}^{r-2} \phi$. So, for $r < 1/2$,

$$|P_{f_0^{(m)}}[-1,1] - P_{f_1^{(m)}}[-1,1]| \geq 2r(\phi(1/2) - \phi(3/2)),$$

so that the last two constraints in (71) are fulfilled for small enough $d = d(r)$. We further note by Lemma 39 that for suitably small c_0, r, we have both that (77) and (78) hold for $s_0 = s_1 = A/2$, and

$$\mathrm{KL}\left(P_{f_0^{(m)}}, P_{f_0^{(0)}}\right) \leq C\frac{A^2}{M} \leq C\frac{c_0^2 \log N}{N}, \qquad \mathrm{KL}\left(P_{f_1^{(m)}}, P_{f_1^{(m)}}\right) \leq C\frac{A^2}{M} \leq C\frac{c_0^2 \log N}{N}.$$

Putting the previous bounds together leads to

$$\begin{aligned}\mathrm{KL}(\Pi_{H_m}, \Pi_{H_0}) &\leq N \cdot \mathrm{KL}\left(P_{f_0^{(m)}}, P_{f_0^{(0)}}\right) + N \cdot \mathrm{KL}\left(P_{f_1^{(m)}}, P_{f_1^{(0)}}\right) \\ &\leq Cc_0^2 \log N.\end{aligned}$$

In particular, one can bound from above

$$\frac{1}{M}\sum_{m=1}^{M} \mathrm{KL}(\Pi_{H_m}, \Pi_{H_0}) \leq Cc_0^2 \log N \leq (\log M)/10,$$

provided that c_0 is a small enough constant, and we deduce by Lemma 40 that $p_{e,M} := \inf_\psi \max_{0\leq m\leq M} \Pi_{H_m}[\psi \neq m]$ is greater than a positive constant. Finally, recalling (79), we have

$$R_n(\mathcal{H}_{sep}) \geq Sp_{e,M},$$

with $S = 2(s_0 + s_1) = 2A$. The proposition follows from the choice of A in (81). ■

Recall the definition (70) of $\mathcal{C}^s(R)$.

Lemma 39. *Let ψ be a C^∞ function with support in $(-1/2, 1/2)$ such that $\|\psi\|_\infty = 1$ and $\int_{\mathbb{R}} \psi = 0$. Let $\phi(\cdot)$ denote the standard normal density and for $m \in \{1, \dots, M\}$ and some $A, r > 0$ and integer $M \geq 2$, set $g_0(x) = r\phi(rx)$ and*

$$g_{m,A}(x) = r\phi(rx) + A\psi(Mx - m + 1/2).$$

Then for $s, R > 0$, the functions g_0 and $g_{m,A}$ are densities belonging to $\mathcal{C}^s(R)$ provided $AM^s \leq R/2$ and r, A are small enough,

$$\|g_0 - g_{m,A}\|_\infty = \|g_{m,A} - g_{p,A}\|_\infty = A, \qquad (\text{for all } m \neq p),$$

and, for P_g the distribution with density g on $\mathbb{R}$, and some $C = C(r) > 0$, any $m \in \{1, \dots, M\}$,

$$\mathrm{KL}(P_{g_{m,A}}, P_{g_0}) \leq CA^2/M.$$

Proof For the statement on supremum norms, it suffices to note that the maps $x \to \psi(Mx - m)$ have disjoint support for different m's. For the KL bounds, one expands the logarithm at the order 2 in a neighbourhood of 0. ■

Lemma 40. *For a family of points $(H_m)_{0\leq m\leq M}$ in $\mathcal{H}_{sep}$ with $M \geq 2$, let*

$$p_{e,M} = \inf_\psi \max_{0\leq m\leq M} \Pi_{H_m}[\psi \neq m], \tag{82}$$

where the infimum is over all possible measurable ψ taking values in $\{1,\dots,M\}$. Suppose, for $\alpha < 1/8$,

$$\frac{1}{M}\sum_{m=1}^{M} \mathrm{KL}(\Pi_{H_m},\Pi_{H_0}) \le \alpha \log M.$$

Then

$$p_{e,M} \ge \frac{\sqrt{M}}{1+\sqrt{M}}\left(1-2\alpha-\sqrt{\frac{2\alpha}{\log M}}\right).$$

Proof This follows from combining Proposition 2.3 and (the proof of) Theorem 2.5 in Tsybakov (2009). ■

Lower bound in the case of $J \ge 3$. One first updates the risk by setting, with $\mathfrak{S}_J$ denoting the set of permutations of $\{0,\dots,J-1\}$,

$$\rho(\mathbf{f},\mathbf{g}) = \min_{\varphi\in\mathfrak{S}_J}\left(\|g_{\varphi(1)}-f_1\|_\infty + \cdots + \|g_{\varphi(J)}-f_J\|_\infty\right).$$

The main change concerns the definition of the class of separated functions $\mathcal{F}_{sep}$. When $J > 2$, the spectral argument used in the proof of the upper bound requires a control on a singular value as in Lemma 24. Let us consider, as in (the proof of) Lemma 24, a fixed collection of functions h_l, for instance a given countable collection of functions bounded in supremum norm and generating the Borel σ-algebra (e.g. a countable number of interval indicators). Recall from Lemma 11 the definition of the matrix, for $L \ge 1$ an integer, for given Q and π and for $\boldsymbol{f} = (f_1,\dots,f_J)$ the vector of emission densities,

$$O^L = O^L(\boldsymbol{f}) = (E_{Q,\pi,\boldsymbol{f}}[h_l(X_1) \mid \theta_1 = j])_{l\le L, j\le J}.$$

Let us further write $\sigma_J^L(f_1,\dots,f_J) = \sigma_J(O^L(f_1,\dots,f_J))$ as a shorthand for the Jth largest singular value of the matrix O^L. In particular, note that if the conclusion of Lemma 24 holds for some emission densities $f_1,\dots,f_J$, then $\sigma_J^L(f_1,\dots,f_J)$ is bounded away from 0 for some suitable integer $L \ge 1$.

The class we consider for $J \ge 3$ is then defined as, for given $s, R > 0$,

$$\mathcal{F}_{sep} = \left\{\mathbf{f} = (f_1,\dots,f_J) \in \mathcal{C}^s(R)^J : \min_{i\ne j}|(f_i-f_j)(0)| \ge d,\ \ \sigma_J^L(f_1,\dots,f_J) \ge d\right\}, \qquad (83)$$

for $d > 0$ a small enough constant and L a large enough integer. One further defines, for given Q, π,

$$\mathcal{H}_{sep} = \mathcal{H}_{sep}(Q,\pi,R,s,d,L) = \{H = (Q,\pi,\mathbf{f}) :\ \mathbf{f} \in \mathcal{F}_{sep}\}.$$

One can then state a proposition analogous to that of Proposition 38 with the obvious modifications of the notation to correspond to $J \ge 3$ states and the updated definition of $\mathcal{F}_{sep}$ as in (83). For brevity we omit the statement, just noting that the corresponding uniform upper bound result holds, as noted in Section 4.3. We now give a brief sketch of the proof of the lower bound, the arguments being broadly similar to the case of $J = 2$ states.

One defines perturbations using the same idea as for $J = 2$, just adding further translated functions for new states: for $j = 1,\dots,J$ and $m = 1,\dots,M$,

$$f_j^{(m)}(x) = g_{m,A}(x-2(j-1)/r)\ , \qquad f_j^{(0)}(x) = r\phi(r(x-2(j-1)/r)),$$

where the function $g_{m,A}$ is still defined as in Lemma 39 and the choice of A, M is as in the case $J = 2$. The proof is then nearly identical to the one in the case $J = 2$. One difference is in checking that the functions $f_1^{(m)},\dots,f_J^{(m)}$ for $m = 0,1,\dots,M$ are in $\mathcal{F}_{sep}$. In order to verify that

$\sigma_J^L(f_1^{(m)},\dots,f_J^{(m)}) \geq d$ for a small $d>0$ and large enough $L\geq 1$, it suffices to note that this holds true for $\sigma_J^L(f_1^{(0)},\dots,f_J^{(0)})$, which follows by applying Lemma 24, noticing that the functions $f_1^{(0)},\dots,f_J^{(0)}$ are linearly independent (as Gaussian densities with different tail behaviours). Next it suffices to notice, by using standard matrix perturbation theory, that for any given integer $L\geq 1$, the quantity $|\sigma_J(O^L(f_1^{(m)},\dots,f_J^{(m)}))-\sigma_J(O^L(f_1^{(0)},\dots,f_J^{(0)}))|$ scales with the constant A in the definition of the perturbations; in particular, the difference vanishes as $N\to\infty$, which implies that the last condition in the definition of $\mathcal{F}_{sep}$ is met. The verification of the other conditions in the definition of $\mathcal{F}_{sep}$ and the rest of the proof are nearly identical (using the updated definition of the perturbations in the last display) to the arguments in the case $J=2$ and are omitted.

Appendix D. Notation

We give notation assuming, as in Section 3, that there are a (known) number J of hidden states $\{1,\dots,J\}$ (recall that $J=2$ for Section 2 and the proofs of results therein, with hidden states labelled 0 and 1, and the notation is adapted accordingly).

HMM parameters.

$X=(X_n)_{n\leq N}$ (or $(X_n)_{n\leq N+2}$ for convenience, or $(X_n)_{n\in\mathbb{N}}$ for some of the proofs and lemmas) the data, drawn from the HMM (1).

$\theta=(\theta_n)_{n\leq N}$ the vector of hidden states, taking values in $\{1,\dots,J\}^N$.

Q,π the transition matrix of θ and its stationary (and initial) distribution.

μ a dominating measure on the space $\mathcal{X}=\mathbb{R}$ (equipped with the usual Borel σ-algebra) in which X_1 takes values. Throughout we take μ to equal Lebesgue measure on $\mathbb{R}$ or counting measure on $\mathbb{Z}\subset\mathbb{R}$.

$f_1,\dots,f_J$ the emission densities, i.e. f_j is the density of X_1 conditional on $\theta_1=j$.

f_π the density of X_1; this is only used in the two-state case so $f_\pi=\pi_0f_0+\pi_1f_1$.

H $=(Q,\pi,f_1,\dots,f_J)$, $\hat{H}=(\hat{Q},\hat{\pi},\hat{f}_1,\dots,\hat{f}_J)$.

Π_H,E_H the law of X for parameter H and the associated expecation operator.

$\mathcal{H},\mathcal{I}$**:** see Section 4.3. [Also note that $C=C(\mathcal{H})$ is allowed to depend on the kernel K and the functions $(h_l)_{l\in\mathbb{N}}$ and sets $\mathbb{D}_N$ of Algorithm 1 since these can be chosen universally.]

ν,x^* constants as in Assumptions A and B.

δ a lower bound for $\min_{i,j}Q_{ij}$.

Multiple testing.

FDP, FDR, TDR, postFDR, mFDR, mTDR, see eqs. (2) to (5), (15) and (16) (also (10) for an alternative characterisation of postFDR).

$\ell_i\equiv\ell_i(X)\equiv\ell_{i,H}(X)=\Pi_H(\theta_i=0\mid X)$; $\hat{\ell}_i=\ell_{i,\hat{H}}$; $\ell_i'=\Pi_H(\theta_i=0\mid X_{i-A},\dots,X_{i+A})$ for some A; $\ell_i^\infty=\Pi_H(\theta_i=0\mid(X_n)_{n\in\mathbb{Z}})$.

Φ_i^∞ $=\Pi_H(\theta_i=0\mid(X_n:n\in\mathbb{Z},n\leq i))$.

$\varphi_{\lambda,H}$ $=(\mathbb{1}\{\ell_{i,H}<\lambda\})_{i\leq N}$

$\hat{\lambda}$ $=\sup\{\lambda:\mathrm{postFDR}_{\hat{H}}(\varphi_{\lambda,\hat{H}})\leq t\}$. λ^* the solution to $E[\ell_i^\infty\mid\ell_i^\infty<\lambda^*]=\min(t,\pi_0)$.

$\hat{\varphi}$ $\equiv\hat{\varphi}^{(t)}=\varphi_{\hat{\lambda},\hat{H}}$ when there are no ties in ℓ-values, or given by Definition 1 when there may be ties.

$\hat{S}_0$ $=\{i:\hat{\varphi}_i=1\}$, $\hat{K}=|\hat{S}_0|$.

ε_N some rate of consistency of estimators in (14).

Estimation.

r_N $=(N/\log N)^{-s/(1+2s)}$.

$h_1,\dots,h_{L_0}$, where L_0 is either constant or diverges slowly to infinity; bounded functions such that "witness" the linear independence of $f_1,\dots,f_J$ (see Algorithm 1 and Lemma 24).

K, K_L, a convolution kernel, see (25).

$M^x \equiv M^{x,L_0,L} = (E_H[h_i(X_1)K_L(x,X_2)h_j(X_3)]_{i,j\le L_0}) \in \mathbb{R}^{L_0\times L_0}$.

$P \equiv P^{L_0} = (E_H[h_i(X_1)h_j(X_3)]_{i,j\le L_0}) \in \mathbb{R}^{L_0\times L_0}$

$O = O^{L_0} = (E_H[h_i(X_1) \mid \theta_1 = a]_{i\le L_0, a\le J}) \in \mathbb{R}^{L_0\times J}$

$D = D^x = \operatorname{diag}((K_L[f_j](x))_{j\le J})$, i.e. the diagonal matrix whose diagonal entries are $D_{jj} = K_L[f_j](x)$.

$V = V^{L_0} \in \mathbb{R}^{L_0\times J}$ a matrix such that $V^\intercal PV$ is invertible. Specifically, we either take V to equal a matrix of orthonormal right singular vectors of P (so that $\sigma_J(V^\intercal PV) = \sigma_J(P)$) or, on the event of Lemma 25, to equal $\hat{V}$ (defined in Algorithm 1).

$B^x = B^{x,L_0} = [V^\intercal PV]^{-1}V^\intercal M^xV \equiv [QO^\intercal V]^{-1}D^xQO^\intercal V$.

$\hat{M}^x, \hat{P}, \hat{O}, \hat{V}$ empirical versions of M^x, P, O, V, B^x (see Algorithm 1, p23).

$\hat{B}^x = [\hat{V}^\intercal\hat{P}\hat{V}]^{-1}\hat{V}^\intercal\hat{M}^x\hat{V}$, $\tilde{B}^x = [\hat{V}^\intercal P\hat{V}]^{-1}\hat{V}^\intercal M^x\hat{V}$, $\hat{B}^{a,u} = \sum a_i\hat{B}^{u_i}$ and $\tilde{B}^{a,u} = \sum a_i\tilde{B}^{u_i}$ for $a, u \in \mathbb{R}^{J(J-1)/2}$ such that $\sum|a_i| \le 1$.

$\operatorname{sep}(B) = \min_{i\ne j}|\lambda_i - \lambda_j|$ the "eigen-separation" of a matrix $B \in \mathbb{R}^{J\times J}$, with eigenvalues $\lambda_1, \dots, \lambda_J$.

$\hat{a}, \hat{u}, \mathbb{D}_N$ See Algorithm 1, p23.

$\hat{R}$ a matrix of normalised columns diagonalising $\hat{B}^{\hat{a},\hat{u}}$, $\tilde{R}$ a matrix whose columns are those of $QO^\intercal\hat{V}$ but scaled to have unit Euclidean norm (which therefore diagonalises $\tilde{B}^{a,u}$ for any a, u).

$\mathcal{A} = \{\|\hat{P} - P\| \le cL_0r_N,\ \|\hat{M}^x - M^x\| \le cL_0^2r_N\ \forall x \in \mathbb{R}\}$ the event of Lemma 25.

C^s the usual space of locally Hölder smooth functions, equipped with the usual Hölder norm $\|\cdot\|_{C^s(\mathbb{R})}$ (see Assumption E). Note that since we consider density functions, we could equivalently use the space of globally Hölder smooth functions.

$\mathcal{C}^s(R)$ the subspace of C^s consisting of probability density functions with Hölder norm bounded by R.

$\mathfrak{S}_2$ the set of all permutations on $\{0,1\}$.

$\rho(\mathbf{f},\mathbf{g}) = \min_{\varphi\in\sigma_2}\big(\|g_{\varphi(0)} - f_0\|_\infty + \|g_{\varphi(1)} - f_1\|_\infty\big)$, for $\mathbf{f} = (f_0, f_1)$, $\mathbf{g} = (g_0, g_1)$.

$\mathcal{F}_{sep} = \big\{\mathbf{f} = (f_0, f_1) \in \mathcal{C}^s(R) :\ |(f_1 - f_0)(0)| \ge d,\ |P_{f_1}([-1,1]) - P_{f_0}([-1,1])| \ge d\big\}$.

$\mathcal{H}_{sep} = \{H = (Q, \pi, \mathbf{f}) :\ \mathbf{f} \in \mathcal{F}_{sep}\}$, for some arbitrary (fixed) Q, π. [Taking the union over certain Q, π, this can be viewed as a subset of $\mathcal{H}$ defined in Section 4.3.]

Miscellaneous.

$\|\cdot\|, \|\cdot\|_F, \|\cdot\|_\infty$ the Euclidean norm on vectors or the corresponding operator norm on matrices, the Frobenius norm on matrices, and the L^∞ (supremum) norm on functions taking values in $\mathbb{R}$.

$\sigma_j(A)$ the jth largest singular value of a matrix A.

$\kappa(A) = \sigma_1(A)/\sigma_J(A) = \|A\|\|A^{-1}\|$ for a matrix with smaller dimension J, the condition number of the matrix A.

$o(1), o_p(1)$ The usual little-oh notation: $a_N = o(1)$ if $a_N \to 0$ as $N \to \infty$, $a_N = o_p(1)$ if $a_N \to 0$ in probability as $N \to \infty$.

$N_{[]}, H_{[]}$**:** The bracketing numbers/entropy, wherein $N_{[]}(\mathcal{T}, \|\cdot\|_{L^2(P)}, \varepsilon)$ is the smallest number of pairs of functions $(\underline{f}, \bar{f})$ such that every $g \in \mathcal{T}$ is bracketed by one of the pairs, where $(\underline{f}, \bar{f})$ brackets g if $\underline{f} \le g \le \bar{f}$ pointwise, and $H_{[]}(\mathcal{T}, \|\cdot\|_{L^2(P)}, \varepsilon) := \log N_{[]}(\mathcal{T}, \|\cdot\|_{L^2(P)}, \varepsilon)$.

References

K. Abraham, I. Castillo, and E. Roquain. Sharp multiple testing boundary for sparse sequences. Arxiv eprint 2109.13601, 2021. URL `https://arxiv.org/pdf/2109.13601`.

G. Alexandrovich, H. Holzmann, and A. Leister. Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, 103(2):423–434, 2016.

A. Anandkumar, D. J. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden Markov models. In *25th Annual Conference On Learning Theory*, page 33.1–33.34, 2012.

B. Bárány and I. Kolossváry. On the absolute continuity of the Blackwell measure. *J. Stat. Phys.*, 159(1): 158–171, 2015.

L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, 37:1554–1563, 1966.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, 41:164–171, 1970.

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.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188, 2001.

P. J. Bickel, Y. Ritov, and T. Rydén. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist.*, 26(4):1614–1635, 1998.

D. Blackwell. The entropy of functions of finite-state Markov chains. In *Transactions of the first Prague conference on information theory, Statistical decision functions, random processes held at Liblice near Prague from November 28 to 30, 1956*, pages 13–20. Publishing House of the Czechoslovak Academy of Sciences, Prague, 1957.

T. T. Cai, W. Sun, and W. Wang. Covariate-assisted ranking and screening for large-scale two-sample inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 81(2):187–234, 2019.

E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 80(3):551–577, 2018.

O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York, 2005. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, with Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

I. Castillo and E. Roquain. On spike and slab empirical Bayes multiple testing. *Ann. Statist.*, 48(5): 2548–2574, 2020.

G. Cleanthous, A. G. Georgiadis, G. Kerkyacharian, P. Petrushev, and D. Picard. Kernel and wavelet density estimators on manifolds and more general metric spaces. *Bernoulli*, 26(3):1832–1862, 2020.

Y. De Castro, E. Gassiat, and C. Lacour. Minimax adaptive estimation of nonparametric hidden Markov models. *J. Mach. Learn. Res.*, 17:Paper No. 111, 43, 2016.

Y. De Castro, E. Gassiat, and S. Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *IEEE Trans. Inform. Theory*, 63(8):4758–4777, 2017.

R. Douc and C. Matias. Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, 7(3):381–420, 2001.

R. Durrett. *Probability—theory and examples*, volume 49 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019.

B. Efron. Size, power and false discovery rates. *Ann. Statist.*, 35(4):1351–1377, 2007a.

B. Efron. Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.*, 102(477): 93–103, 2007b.

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.

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

H. Finner, T. Dickhaus, and M. Roters. Dependency and false discovery rate: asymptotics. *Ann. Statist.*, 35(4):1432–1455, 2007.

E. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space non parametric hidden Markov models and applications. *Stat. Comput.*, 26(1-2):61–71, 2016.

S. Ghosal and A. van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2017.

E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York, 2016.

R. Heller and S. Rosset. Optimal control of false discovery criteria in the two-group model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 83(1):133–155, 2021.

L. Lehéricy. Nonasymptotic control of the MLE for misspecified nonparametric hidden Markov models. ArXiv eprint 1807.03997, 2018. URL `https://arxiv.org/pdf/1807.03997`.

L. Lehéricy. State-by-state minimax adaptive estimation for nonparametric hidden Markov models. *J. Mach. Learn. Res.*, 19:Paper No. 39, 46, 2018.

P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

D. Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electron. J. Probab.*, 20:no. 79, 32, 2015.

T. Petrie. Probabilistic functions of finite-state Markov chains. *Proc. Nat. Acad. Sci. U.S.A.*, 57:580–581, 1967.

T. Rebafka, E. Roquain, and F. Villers. Graph inference with clustering and false discovery rate control. ArXiv eprint 1907.10176, 2019. URL `https://arxiv.org/abs/1907.10176`.

E. Roquain and N. Verzelen. On using empirical null distributions in Benjamini–Hochberg procedure. ArXiv eprint 1912.03109, 2020. URL `https://arxiv.org/pdf/1912.03109`.

M. Sesia, C. Sabatti, and E. J. Candès. Gene hunting with hidden Markov model knockoffs. *Biometrika*, 106(1):1–18, 2019.

H. Shu, B. Nan, and R. Koeppe. Multiple testing for neuroimaging via hidden Markov random field. *Biometrics*, 71(3):741–750, 2015.

G. W. Stewart and J. G. Sun. *Matrix perturbation theory*. Computer Science and Scientific Computing. Academic Press, Inc., Boston, MA, 1990.

J. D. Storey. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.*, 31 (6):2013–2035, 2003.

W. Su and X. Wang. Hidden Markov model in multiple testing on dependent count data. *J. Stat. Comput. Simul.*, 90(5):889–906, 2020.

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.

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.

A. Touron. Consistency of the maximum likelihood estimator in seasonal hidden Markov models. *Stat. Comput.*, 29(5):1055–1075, 2019.

A. B. Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

A. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3. doi: 10.1007/978-1-4757-2545-2. With applications to statistics.

X. Wang, A. Shojaie, and J. Zou. Bayesian hidden Markov models for dependent large-scale multiple testing. *Comput. Statist. Data Anal.*, 136:123–136, 2019.

Z. Wei, W. Sun, K. Wang, and H. Hakonarson. Multiple testing in genome-wide association studies via hidden Markov models. *Bioinfo.*, 25(21):2802–2808, 2009.

W. B. Wu. On false discovery control under dependence. *Ann. Statist.*, 36(1):364–380, 2008.

J. Xie, T. T. Cai, J. Maris, and H. Li. Optimal false discovery rate control for dependent data. *Stat. Interface*, 4(4):417–430, 2011.

C. Yau, O. Papaspiliopoulos, G. O. Roberts, and C. Holmes. Bayesian non-parametric hidden Markov models with applications in genomics. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(1):37–57, 2011.

W. Zucchini, I. L. MacDonald, and R. Langrock. *Hidden Markov models for time series*, volume 150 of *Monographs on Statistics and Applied Probability.* CRC Press, Boca Raton, FL, second edition, 2016.