Journal of the
Royal Statistical Society

SERIES B
Statistical
Methodology

J. R. Statist. Soc. B (2009)
71, *Part* 2, *pp.* 393–424

Large-scale multiple testing under dependence

Wenguang Sun and T. Tony Cai

University of Pennsylvania, Philadelphia, USA

[Received April 2008. Revised August 2008]

Summary. The paper considers the problem of multiple testing under dependence in a compound decision theoretic framework. The observed data are assumed to be generated from an underlying two-state hidden Markov model. We propose oracle and asymptotically optimal data-driven procedures that aim to minimize the false non-discovery rate FNR subject to a constraint on the false discovery rate FDR. It is shown that the performance of a multiple-testing procedure can be substantially improved by adaptively exploiting the dependence structure among hypotheses, and hence conventional FDR procedures that ignore this structural information are inefficient. Both theoretical properties and numerical performances of the procedures proposed are investigated. It is shown that the procedures proposed control FDR at the desired level, enjoy certain optimality properties and are especially powerful in identifying clustered non-null cases. The new procedure is applied to an influenza-like illness surveillance study for detecting the timing of epidemic periods.

Keywords: Compound decision problem; False discovery rate; Hidden Markov models; Local significance index; Multiple testing under dependence

1. Introduction

Observations arising from large-scale multiple-comparison problems are often dependent. For example, in microarray experiments, different genes may cluster into groups along biological pathways and exhibit high correlation. In public health surveillance studies, the observed data from different time periods and locations are often serially or spatially correlated. Other examples include the analysis of data from functional magnetic resonance imaging and multistage clinical trials, where the observations are also dependent in some fashion. Multiple-testing procedures, especially the false discovery rate FDR (Benjamini and Hochberg, 1995) analyses, have been widely used to screen over these massive data sets to identify a few interesting cases. However, these procedures rely heavily on the independence assumption, and the correlation between hypotheses, which is often treated as a nuisance parameter, is typically ignored.

The outcomes of a multiple-testing procedure can be summarized as in Table 1. FDR is defined as $E(N_{10}/R|R>0)\,\Pr(R>0)$. To compare the power of different FDR procedures, we define the false non-discovery rate FNR (Genovese and Wasserman, 2002), equal to $E(N_{01}/S|S>0)\,\Pr(S>0)$. We call an FDR procedure *valid* if it controls FDR at a prespecified level α, and *optimal* if it has the smallest FNR among all FDR procedures at level α. The marginal false discovery rate $\text{mFDR}=E(N_{10})/E(R)$ is asymptotically equivalent to the FDR measure in the sense that $\text{mFDR}=\text{FDR}+O(m^{-1/2})$ under weak conditions (Genovese and Wasserman, 2002). Similarly, the marginal false non-discovery rate is given by $\text{mFNR}=E(N_{01})/E(S)$.

Address for correspondence: T. Tony Cai, Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340, USA.
E-mail: tcai@wharton.upenn.edu

© 2009 Royal Statistical Society

Table 1. Classification of tested hypotheses

Hypothesis	*Claimed non-significant*	*Claimed significant*	*Total*
Null	N_{00}	N_{10}	m_0
Non-null	N_{01}	N_{11}	m_1
Total	S	R	m

The effects of correlation on FDR procedures have been discussed by Benjamini and Yekutieli (2001), Finner and Roters (2002), Owen (2005), Sarkar (2006) and Efron (2007), among others. Efron (2007) noted that correlation may result in overly liberal or overly conservative testing procedures and so must be accounted for in deciding which hypotheses should be reported as non-null hypotheses. Qiu *et al.* (2005) showed that the correlation effects can substantially deteriorate the performance of many FDR procedures. However, the works by Benjamini and Yekutieli (2001), Farcomeni (2007) and Wu (2008) show that FDR is controlled at the nominal level by the Benjamini and Hochberg (1995) step-up procedure and the adaptive p-value procedure (Benjamini and Hochberg, 2000; Genovese and Wasserman, 2004) under different dependence assumptions, supporting the universal use of the conventional FDR procedures that were developed for independent tests without any adjustments.

In dealing with the effects of correlation on an FDR procdure, the validity issue has been overemphasized, and the efficiency issue is largely ignored. The FDR procedures that are developed under the independence assumption, even valid, may suffer from substantial loss of efficiency when the dependence structure is highly informative. The works by Yekutieli and Benjamini (1999), Genovese *et al.* (2006) and Benjamini and Heller (2007) showed that incorporating scientific, spatial information into a multiple-testing procedure may greatly improve the efficiency. However, these approaches are either based on resampling the p-values or rely on prior information, such as well-defined clusters or prespecified weights. The correlation structure is not modelled and the optimality essentially remains unknown.

A hidden Markov model (HMM) is an effective tool for modelling the dependence structure and has been widely used in areas such as speech recognition, signal processing and DNA sequence analysis; see Rabiner (1989), Churchill (1992), Krogh *et al.* (1994) and Ephraim and Merhav (2002), among others. In the context of multiple testing, an HMM assumes that the sequence of the unobservable states forms a Markov chain $(\theta_i)_1^m = (\theta_1, \ldots, \theta_m)$, where $\theta_i = 1$ if hypothesis i is non-null, and $\theta_i = 0$ otherwise. The observed data $\mathbf{x} = (x_1, \ldots, x_m)$ are independently generated conditionally on the hidden states $(\theta_i)_1^m$. When positive dependence exists in an HMM, one expects that the non-null hypotheses ($\theta_i = 1$) appear in clusters or clumps. This is a natural feature of many data sets arising from time series analysis and spatial data analysis. For example, in the influenza-like illness (ILI) study that we analyse in Section 5, the epidemic periods (measured weekly) tend to last for weeks and thus to cluster temporally; this dependence structure can be well described by using an HMM.

In this paper, the problem of multiple testing under HMM dependence is studied in a compound decision theoretic framework. We first propose an oracle testing procedure in an ideal setting where the HMM parameters are assumed to be known. Under mild conditions, the oracle procedure is shown to be optimal in the sense that it minimizes mFNR subject to a constraint on mFDR. Our approach is distinguished from the conventional methods in that the procedure proposed is built on a new test statistic (the local index of significance, LIS) instead of the p-values. Unlike p-values, LIS takes into account the observations in adjacent locations by exploiting the local dependence structure in the HMM. The precision of individual tests is hence improved by pooling information from different samples.

As a motivating example, we generate a Markov chain of Bernoulli variables $(\theta_i)_1^m$ and observations $(x_i)_1^m$ according to the mixture model $x_i|\theta_i \sim (1-\theta_i)\, N(0,1) + \theta_i\, N(\mu,1), i=1,\ldots,1000$. A comparison of the Benjamini and Hochberg (1995) step-up procedure BH, adaptive p-value procedure AP (assuming that the proportion of non-null hypotheses is known) and the oracle procedure OR, which is developed in Section 3, is shown in Fig. 1. We can see that all three procedures control FDR at the nominal level, and BH is conservative. In addition, OR has much lower FNR level than BH and AP, and the gain in efficiency is substantial when μ is small to moderate. It is important to note that, comparing with AP, the gain in efficiency of OR is not at the price of a higher FDR. To show how this power is achieved, we present in Table 2 the outcomes of BH, AP and OR in testing two clusters of significant hypotheses in one experiment ($\mu=2$), where '○' denotes a null hypothesis and '•' denotes a non-null hypothesis. It can been seen that BH and AP can only reject individual hypotheses with extremely small

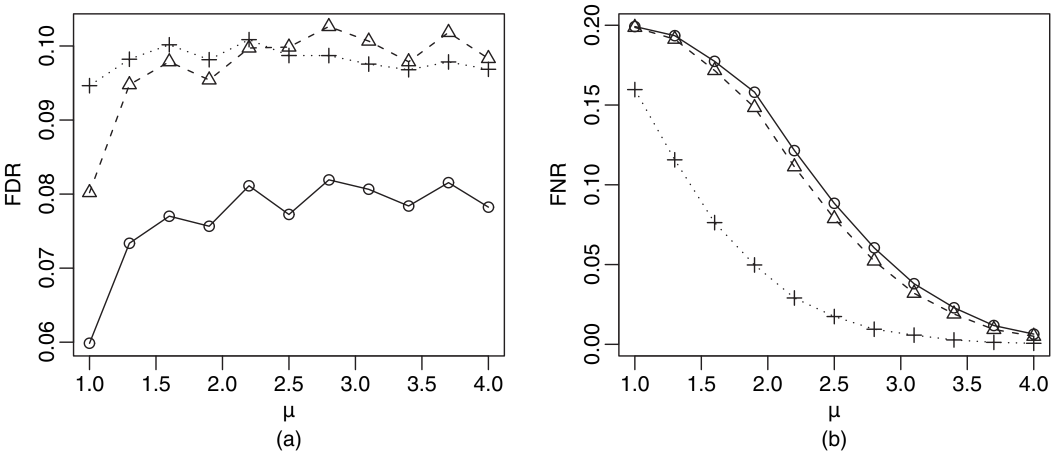

Fig. 1. Comparison of BH (○), AP (△) and OR (+) in an HMM (the FDR level is set at 0.10): (a) FDR *versus* μ; (b) FNR *versus* μ

Table 2. Comparison of procedures BH and OR in an HMM

Sequence	*State*	*p-value*	*BH procedure*	*AP procedure*	*OR procedure*
121	•	0.07	○	○	•
122	•	0.001	•	•	•
123	•	<0.001	•	•	•
124	•	0.02	○	○	•
⋮	⋮	⋮	⋮	⋮	⋮
177	•	0.01	○	○	•
178	•	<0.001	•	•	•
179	•	0.13	○	○	•
180	•	<0.001	•	•	•
181	•	0.004	○	•	•
182	•	0.15	○	○	○

p-values, whereas OR tends to identify the entire cluster of non-null hypotheses. We demonstrate in Section 3 that this is a typical phenomenon that can be expected, because, in deciding the level of significance of a particular hypothesis, OR pools information from adjacent locations by exploiting the local dependence structure in the observed sequence.

We then develop a data-driven procedure that mimics the oracle procedure by plugging in consistent estimates of the unknown HMM parameters. The data-driven procedure is shown to be *asymptotically optimal* in the sense that it attains both the FDR and the FNR levels of the oracle procedure asymptotically. The oracle and data-driven procedures are compared with the conventional p-value-based procedures by using simulation studies in Section 4. The results indicate the favourable performance of the newly proposed procedures. Our findings show that the correlation between hypotheses is highly informative in simultaneous inference and can be exploited to construct more efficient testing procedures. Our procedure is especially powerful in identifying weak signals if they are clustered in large groups. This indicates that dependence can make the testing problem easier and is a blessing if incorporated properly in a testing procedure.

The paper is organized as follows. Section 2 studies the problem of multiple testing under dependence in a compound decision theoretic framework. In Section 3 we propose an oracle procedure and a data-driven procedure for FDR control under dependence and investigate their theoretical properties. Simulation studies are performed in Section 4 to compare the proposed procedures with conventional FDR procedures in various settings. In Section 5, the new procedure is applied to an ILI surveillance study for identifying epidemic periods. The proofs are given in Appendix A and Appendix B.

2. Compound decision problem in a hidden Markov model

In this section, we develop a compound decision theoretic framework for both the weighted classification and the multiple-testing problems in an HMM. It is shown in Section 2.1 that these two problems are essentially equivalent under mild conditions. Therefore we first study the optimal rule in a weighted classification problem in Section 2.2 and then use the results to solve the multiple-testing problem in Section 3.

Let $\mathbf{x} = (x_1, \ldots, x_m)$ be a vector of observed values with associated unknown states $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. Assume that, conditionally on $\boldsymbol{\theta}$, the x_is are independent. Suppose that we are interested in inference of the unknown θ_is based on the observed $\mathbf{x}$. This involves solving m component problems simultaneously and is called a *compound decision problem* (Robbins, 1951). Different state space and correlation structure can be assumed for θ_is. In this paper, we consider an HMM, where θ_is are Bernoulli variables and distributed as a Markov chain. We assume that $(\theta_i)_1^m = (\theta_1, \ldots, \theta_m)$ is stationary, irreducible and aperiodic. Specifically, the transition probabilities are homogeneous and bounded away from 0 and 1, i.e. $a_{jk} = P(\theta_i = k|\theta_{i-1} = j), 0 \leqslant j, k \leqslant 1$, do not depend on i, with the standard stochastic constraints $0 < a_{jk} < 1$, $a_{j0} + a_{j1} = 1$. The convergence theorem of a Markov chain (theorem 5.5.1 in Durrett (2005)) implies that

$$\frac{1}{m} \sum_{i=1}^{m} I(\theta_i = j) \rightarrow \pi_j$$

almost surely as $m \rightarrow \infty$. The Bernoulli variables $\theta_1, \ldots, \theta_m$ are identically distributed (but correlated) with $P(\theta_i = j) = \pi_j$. Let $X_i|\theta_i \sim (1 - \theta_i)F_0 + \theta_i F_1$. Denote by $\mathcal{A} = \{a_{jk}\}$ the transition matrix, $\boldsymbol{\pi} = (\pi_0, \pi_1)$ the stationary distribution, $\mathcal{F} = \{F_0, F_1\}$ the observation distribution and $\vartheta = (\mathcal{A}, \boldsymbol{\pi}, \mathcal{F})$ the collection of all HMM parameters.

In a compound decision problem where the goal is to separate the non-null hypotheses ($\theta_i = 1$) from the null hypotheses ($\theta_i = 0$), the solution can be represented by a general decision rule

$\boldsymbol{\delta}=(\delta_1,\ldots,\delta_m)\in\{0,1\}^m$, with $\delta_i=1$ indicating that hypothesis i is rejected and $\delta_i=0$ otherwise. When the relative cost of a false positive (type I error) to a false negative (type II error) result is specified, we can study a weighted classification problem where the goal is to construct $\boldsymbol{\delta}$ that minimizes the expectation of the loss function

$$L_\lambda(\boldsymbol{\theta},\boldsymbol{\delta})=\frac{1}{m}\sum_i\{\lambda(1-\theta_i)\delta_i+\theta_i(1-\delta_i)\}, \tag{1}$$

with $\lambda>0$ the weight for a false positive result. Alternatively, if the goal is to discover as many true findings as possible while incurring a relatively low proportion of false positive findings, we can study a multiple-testing problem where the goal is to find $\boldsymbol{\delta}$ that has the smallest FNR among all FDR procedures at level α. The multiple-testing and weighted classification problems are closely connected. Specifically, the solutions to both problems can be represented by a binary decision rule $\boldsymbol{\delta}\in\{0,1\}^m$. In addition, if a stricter constraint on the false positive results is desirable, we can either set a smaller FDR level or put a larger penalty λ in loss function (1). The next subsection introduces a so-called monotone ratio condition under which the two problems are 'equivalent'.

2.1. Connection between multiple testing and weighted classification

Consider a stationary, irreducible and aperiodic Markov chain $(\theta_i)_1^m$ and observations $\mathbf{x}=(x_1,\ldots,x_m)$ generated according to the conditional probability model

$$P(\mathbf{x}|\boldsymbol{\theta},\mathcal{F})=\prod_{i=1}^m P(x_i|\theta_i,\mathcal{F}), \tag{2}$$

where $(\theta_i)_1^m\in\{0,1\}^m$, $P(x_i<x|\theta_i=j)=F_j(x)$, $j=0,1$ and $\mathcal{F}=(F_0,F_1)$. Denote by f_0 and f_1 the corresponding densities. Suppose that $\boldsymbol{\delta}$ is a general decision rule which is defined in terms of a statistic $\mathbf{T}(\mathbf{x})=(T_1(\mathbf{x}),\ldots,T_m(\mathbf{x}))$ and vector $\mathbf{c}=c\mathbf{1}$ such that $\boldsymbol{\delta}(\mathbf{T},\mathbf{c})=I(\mathbf{T}<\mathbf{c})=[I(T_i<c):i=1,\ldots,m]$. Let $G_i^j(t)=P(T_i<t|\theta_i=j)$, $j=0,1$, be the conditional cumulative distribution functions (CDFs) of $T_i(\mathbf{x})$. Recall that $P(\theta_i=j)=\pi_j$; the marginal CDF of $T_i(\mathbf{x})$ is then given by $G_i(t)=P(T_i<t)=\pi_0\,G_i^0(t)+\pi_1\,G_i^1(t)$. Define the average conditional CDFs of $\mathbf{T}$,

$$G^j(t)=\frac{1}{m}\sum_{i=1}^m G_i^j(t)$$

and average conditional probability density functions (PDFs) of $\mathbf{T}$, $g^j(t)=(\mathrm{d}/\mathrm{d}t)G^j(t)$, $j=0,1$. For a vector $\mathbf{x}=(x_i)_1^m$ with associated unknown states $(\theta_i)_1^m$, we consider a class of test statistics $\mathcal{T}$ such that, for each $\mathbf{T}(\mathbf{x})\in\mathcal{T}$, $g^0(t)$ and $g^1(t)$, the average conditional PDFs of $\mathbf{T}$, satisfy a monotone ratio condition (MRC):

$$g^1(t)/g^0(t) \text{ is monotonically decreasing in } t. \tag{3}$$

The MRC generalizes the monotone likelihood ratio condition in Sun and Cai (2007) in that it reduces to the monotone likelihood ratio condition when the tests are independent.

The MRC class $\mathcal{T}$ is fairly general. For example, the condition in Genovese and Wasserman (2002, 2004) and Storey (2002) that the non-null CDF of p-value $G^1(t)$ is concave and twice differentiable implies that the p-value vector $\mathbf{P}=(P_1,\ldots,P_m)$ and the weighted p-value vector $\mathbf{P}_w=(P_1/w_1,\ldots,P_m/w_m)$ (Genovese *et al.*, 2006), where $\Sigma_i w_i=m$, belong to the MRC class $\mathcal{T}$. This can be shown by first calculating the average conditional PDFs of the weighted p-value vector, $g^0(t)=1$ (the null distribution is uniform) and $g^1(t)$; then noting that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{g^1(t)}{g^0(t)}\right\}=\sum_{i=1}^{m}\left(\frac{w_i^2}{m}\right)G^{1\prime\prime}(w_it)<0.$$

In addition, test statistics that are defined on the basis of the z-values, such as the local false discovery rate (Efron *et al.*, 2001), and the oracle test statistic in an HMM, which is defined in Section 3, also belong to MRC class $\mathcal{T}$ (see corollary 1). The following theorem shows that the MRC is a desirable condition for inference in an HMM.

Theorem 1. Consider an HMM defined as in model (2). Let $\boldsymbol{\delta}$ be a decision rule such that $\boldsymbol{\delta}(\mathbf{T},c)=I(\mathbf{T}<c\mathbf{1})$ with $\mathbf{T}\in\mathcal{T}$. Then

(a) mFDR of $\boldsymbol{\delta}(\mathbf{T},c)$ is strictly increasing in the threshold c,
(b) mFNR of $\boldsymbol{\delta}(\mathbf{T},c)$ is strictly decreasing in c and
(c) in the weighted classification problem, the optimal cut-off c (for classification statistic $\mathbf{T}$) that minimizes the classification risk is strictly decreasing in λ, the relative weight for a false positive result.

The following theorem makes the connection between the multiple-testing and weighted classification problems.

Theorem 2. Consider an HMM defined as in model (2). Suppose that the classification risk with the loss function

$$L_\lambda(\boldsymbol{\theta},\boldsymbol{\delta})=\frac{1}{m}\sum_{i=1}^{m}\{\lambda(1-\theta_i)\delta_i+\theta_i(1-\delta_i)\}$$

is minimized by $\boldsymbol{\delta}^\lambda\{\boldsymbol{\Lambda},c(\lambda)\}=I\{\boldsymbol{\Lambda}<c(\lambda)\mathbf{1}\}$, so that $\boldsymbol{\Lambda}$ is the optimal statistic in the weighted classification problem. If $\boldsymbol{\Lambda}$ belongs to $\mathcal{T}$, then $\boldsymbol{\Lambda}$ is also the optimal statistic in the multiple-testing problem in the sense that, for each mFDR level α, there is a unique $c(\alpha)$ such that $\boldsymbol{\delta}^\alpha\{\boldsymbol{\Lambda},c(\alpha)\}=I\{\boldsymbol{\Lambda}<c(\alpha)\mathbf{1}\}$ controls mFDR at level α with the smallest mFNR among all testing rules in $\mathcal{D}$, where $\mathcal{D}$ is the collection of all testing rules of the form $\boldsymbol{\delta}=I(\mathbf{T}<c\mathbf{1})$ with $\mathbf{T}\in\mathcal{T}$.

Theorem 2 highlights two fundamental problems in the development of an optimal testing procedure: deriving the optimal test statistic $\boldsymbol{\Lambda}$ and setting the threshold for $\boldsymbol{\Lambda}$. The first problem is studied in Section 2.2 by solving an equivalent weighted classification problem. The second problem, which has been the focus of the FDR literature (mainly in terms of p-values), is discussed in terms of the optimal test statistic $\boldsymbol{\Lambda}$ in Section 3.

2.2. Oracle classification rule

We shall derive an optimal classification rule $\boldsymbol{\delta}^\lambda$ in an HMM as stated in theorem 2 and describe a recursive procedure for its implementation. We begin by considering an ideal set-up in which an oracle knows the HMM parameters $\vartheta=(\mathcal{A},\boldsymbol{\pi},\mathcal{F})$. Then the oracle's response to the weighted classification problem is given by the following theorem.

Theorem 3. Consider an HMM defined as in model (2). Suppose that the true parameter in the HMM, $\vartheta=(\mathcal{A},\boldsymbol{\pi},\mathcal{F})$, is known. Then the classification risk with loss function defined in equation (1) is minimized by the oracle classification rule $\boldsymbol{\delta}(\boldsymbol{\Lambda},1/\lambda)=(\delta_1,\ldots,\delta_m)$, where

$$\Lambda_i(\mathbf{x})=\frac{P_\vartheta(\theta_i=0|\mathbf{x})}{P_\vartheta(\theta_i=1|\mathbf{x})}\tag{4}$$

and $\delta_i=I\{\Lambda_i(\mathbf{x})<1/\lambda\}$ for $i=1,\ldots,m$.

Remark 1. Given ϑ, the oracle classification statistic $\Lambda_i(\mathbf{x})$ can be expressed in terms of the forward and backward density variables, which are defined as $\alpha_i(j) = f_\vartheta\{(x_t)_1^i, \theta_i = j\}$ and $\beta_i(j) = f_\vartheta\{(x_t)_{i+1}^m | \theta_i = j\}$ respectively (note that the dependence of $\alpha_i(j)$ on $(x_t)_1^i$ has been suppressed, and similarly for $\beta_i(j)$). It can be shown that $P_\vartheta(\mathbf{x}, \theta_i = j) = \alpha_i(j)\,\beta_i(j)$ and hence $\Lambda_i(\mathbf{x}) = \alpha_i(0)\beta_i(0)/\alpha_i(1)\beta_i(1)$. The forward variable $\alpha_i(j)$ and backward variable $\beta_i(j)$ can be calculated recursively by using the *forward–backward procedure* (Baum *et al.*, 1970; Rabiner, 1989). Specifically, we initialize $\alpha_1(j) = \pi_j\, f_j(x_1)$ and $\beta_m(j) = 1$; then by induction we have

$$\alpha_{i+1}(j) = \left\{ \sum_{k=0}^{1} \alpha_i(k) a_{kj} \right\} f_j(x_{i+1})$$

and

$$\beta_i(j) = \sum_{k=0}^{1} a_{jk}\, f_k(x_{i+1})\, \beta_{i+1}(k).$$

Corollary 1. Consider the oracle classification statistic $\boldsymbol{\Lambda}$ that is given in theorem 3. Let $G_i^j(t) = P(\Lambda_i < t | \theta_i = j)$,

$$G^j(t) = \frac{1}{m} \sum_{i=1}^{m} G_i^j(t)$$

and $g^j(t) = (\mathrm{d}/\mathrm{d}t) G^j(t)$, $i = 1, \ldots, m$, $j = 0, 1$, be the conditional CDFs, average conditional CDFs and average conditional PDFs of $\boldsymbol{\Lambda}$ respectively. Then $g^1(t)/g^0(t) = (1/t)\pi_0/\pi_1$. In particular, the oracle classification statistic $\boldsymbol{\Lambda}(\mathbf{x})$ belongs to the MRC class $\mathcal{T}$.

3. Oracle and data-driven procedures for multiple testing under dependence

Theorems 2 and 3, together with corollary 1, imply that $\boldsymbol{\Lambda}(\mathbf{x})$ is the optimal statistic for multiple testing. Since $\Lambda_i(\mathbf{x})$ is increasing in $P_\vartheta(\theta_i = 0 | \mathbf{x})$, an optimal multiple-testing rule in an HMM can be written in the form of $\boldsymbol{\delta} = [I\{P_\vartheta(\theta_i = 0 | \mathbf{x}) < t\} : i = 1, \ldots, m]$. Define the *local index of significance*, LIS, for hypothesis i by

$$\mathrm{LIS}_i = P_\vartheta(\theta_i = 0 | \mathbf{x}). \qquad (5)$$

LIS depends only on x_i and reduces to Efron's local false discovery rate Lfdr in the independent case, i.e. $\mathrm{LIS}_i(\mathbf{x})$ simplifies to $\mathrm{Lfdr}(x_i) = (1 - p)\, f_0(x_i)/f(x_i)$, where p is the proportion of non-null hypotheses and f is the marginal PDF.

It is important to note that the traditional framework for multiple testing confines attention to procedures that essentially involve ranking and thresholding p-values, whereas under our framework the optimal statistic is LIS. In this section, we first give some intuition on why LIS is more appropriate for testing correlated hypotheses by comparing the use of the p-value, Lfdr and LIS from a compound decision theoretical view. We then turn to the development of an oracle procedure that is based on LIS and a data-driven procedure that mimics the oracle procedure. Theoretical properties of these procedures are investigated, showing that both procedures enjoy certain optimality properties. Simulation studies are performed in Section 4, demonstrating that the procedures proposed are superior to conventional FDR approaches for testing correlated hypotheses.

3.1. p-value, Lfdr and LIS

Sun and Cai (2007) studied the multiple-testing problem in a compound decision theoretic framework and showed that the FDR procedures that threshold p-values are in general inefficient, and a testing procedure that thresholds Lfdr is optimal for independent tests. The gain in efficiency of the Lfdr approach is due to the fact that it produces more efficient rankings of the hypotheses than traditional p-value-based approaches. When determining the level of significance of a hypothesis, a p-value approach considers each hypothesis separately, whereas an Lfdr approach considers the m hypotheses simultaneously by incorporating the distributional information of the z-values in the Lfdr statistic. However, the validity of the Lfdr procedure is questionable in the dependent case because the marginal distribution of z-values is no longer well defined; see Qiu *et al.* (2005). In addition, both p-value and Lfdr approaches are inefficient when the tests are correlated, as we shall explain shortly.

Let $\boldsymbol{\delta}$ be a general decision rule; then $\boldsymbol{\delta}$ is *symmetric* if $\boldsymbol{\delta}\{\tau(\mathbf{x})\}=\tau\{\boldsymbol{\delta}(\mathbf{x})\}$ for all permutation operators τ (Copas, 1974). In situations where we expect the non-null hypotheses to appear in clusters, it is natural to treat differently a hypothesis surrounded by non-null from one surrounded by null hypotheses. However, these two hypotheses are exchangeable when a symmetric rule is applied. The FDR procedures that threshold p-value or Lfdr are symmetric rules, so they are not desirable when hypotheses are correlated. Storey (2007) considered an optimal discovery procedure, which maximizes the expected number of true positive results subject to a constraint on the expected number of false positive results. The optimal discovery procedure is also a symmetric rule and is only optimal in a subclass of testing rules. Therefore it is inefficient in testing hypotheses arising from an HMM.

By contrast, we consider decision rule $\boldsymbol{\delta}(\text{LIS}, \lambda)=\{I\{\text{LIS}_i(\mathbf{x})<\lambda\}:i=1,\ldots,m\}$. It is easy to see that $\boldsymbol{\delta}(\text{LIS}, \lambda)$ is asymmetric, and the order of the sequence $(x_i)_1^m$ is accounted for in deciding the level of significance of hypothesis i. In particular, the local dependence structure is captured by the HMM, and the operation of the forward–backward procedure implies that a large or small observation will respectively increase or decrease the level of significance of its neighbours. The performance of the testing procedure is hence improved by pooling information from adjacent locations. In addition, the signal-to-noise ratio is increased since the information from the whole sequence is integrated to calculate the LIS value of a single hypothesis. Therefore, LIS is more robust against local disturbance, which further increases the efficiency of our testing procedure.

3.2. Oracle testing procedure

We have shown that the optimal testing procedure is of the form $\boldsymbol{\delta}=[I(\text{LIS}_i<\lambda):i=1,\ldots,m]$. The next step is to derive an appropriate cut-off λ for a given FDR level. We begin by considering an ideal situation in which an oracle knows the HMM parameter ϑ. The MRC, which is defined in expression (3), implies that mFNR is a decreasing function of mFDR; therefore, the oracle's response to this thresholding problem is to choose λ that 'spend' all mFDR so that mFNR is minimized.

Next we derive the cut-off of LIS for a given FDR level. The general idea in the derivation, which has been used in Genovese and Wasserman (2004), Newton *et al.* (2004) and Sun and Cai (2007), is first to estimate FDR for a given cut-off; then to search for the largest cut-off c such that $\widehat{\text{FDR}}(c)\leqslant\alpha$. Let $\text{LIS}_{(1)}(\mathbf{x}),\ldots,\text{LIS}_{(m)}(\mathbf{x})$ be the ranked test statistics and $H_{(1)},\ldots,H_{(m)}$ be corresponding hypotheses. Let $R_\lambda=\Sigma_{i=1}^m I(\text{LIS}_i<\lambda)$, $V_\lambda=\Sigma_{i=1}^m I(\text{LIS}_i<\lambda,\theta_i=0)$ and $Q(\lambda)=E(V_\lambda)/E(R_\lambda)$ be the number of rejections, number of false positive results and mFDR yielded by decision rule $\boldsymbol{\delta}=[I(\text{LIS}_i<\lambda):i=1,\ldots,m]$ respectively. It can be shown by using the double-expectation theorem that

$$E(V_\lambda)=E\left[\sum_{i=1}^m I\{\text{LIS}_i(\mathbf{x})<\lambda\}\text{LIS}_i(\mathbf{x})\right].$$

If k hypotheses are rejected, then the expected number of false positive results can be approximated by $\hat{V}(k) = \Sigma_{i=1}^{k} \mathrm{LIS}_{(i)}(\mathbf{x})$ and mFDR can be approximated by

$$\hat{Q}(k) = \frac{1}{k} \sum_{i=1}^{k} \mathrm{LIS}_{(i)}(\mathbf{x}).$$

Note that $\hat{Q}(k)$ is increasing in k since

$$\hat{Q}(k+1) - \hat{Q}(k) = \frac{1}{(k^2+k)} \sum_{i=1}^{k} \{\mathrm{LIS}_{(k+1)}(\mathbf{x}) - \mathrm{LIS}_{(i)}(\mathbf{x})\} > 0.$$

We shall choose the largest k such that the mFDR level is controlled at level α. Hence we propose the following step-up procedure:

$$\text{let } k = \max\left\{ i : \frac{1}{i} \sum_{j=1}^{i} \mathrm{LIS}_{(j)}(\mathbf{x}) \leqslant \alpha \right\}; \text{ then reject all } H_{(i)},\ i = 1, \ldots, k. \tag{6}$$

The testing procedure that is given in expression (6) is referred to as the *oracle testing procedure* OR. The next theorem shows that OR is valid for FDR control under dependence.

Theorem 4. Consider an HMM defined as in model (2). Define test statistic $\mathrm{LIS}_i(\mathbf{x}) = P_\vartheta(\theta_i = 0|\mathbf{x}), i = 1, \ldots, m$. Let $\mathrm{LIS}_{(1)}, \ldots, \mathrm{LIS}_{(m)}$ be the ranked LIS values and $H_{(1)}, \ldots, H_{(m)}$ the corresponding hypotheses. Then the oracle testing procedure (6) controls FDR at α.

Each multiple-testing procedure involves two steps: ranking the hypotheses and then choosing a cut-off along the rankings. The LIS and p-value usually produce different rankings of the hypotheses. To illustrate this, we revisit the example that was presented in Table 2 of Section 1 and contrast the levels of significance of each hypothesis given by the p-value and LIS in Table 3. It is interesting to note that BH ranks hypothesis 177 higher than hypothesis 179, whereas OR ranks hypothesis 179 higher than hypothesis 177 (note that hypothesis 179 has a smaller LIS value because it is surrounded by two very significant observations). We set the FDR level at 0.10; then the cut-off for the p-value given by BH is 0.003, and the cut-off for LIS given by OR is 0.334. We can see that, among the six non-null hypotheses in the second cluster (hypotheses 177–182), two are identified by BH, and five are identified by OR. This illustrates the benefit of taking into account the local dependence structure when ranking the hypotheses. The gain in

Table 3. Levels of significance suggested by *p*-value and LIS

Sequence	*State*	*p-value*	*LIS*	*BH procedure*	*OR procedure*
121	●	0.07	0.296	○	●
122	●	0.001	0.011	●	●
123	●	<0.001	<0.001	●	●
124	●	0.02	0.159	○	●
⋮	⋮	⋮	⋮	⋮	⋮
177	●	0.01	0.092	○	●
178	●	<0.001	<0.001	●	●
179	●	0.13	0.017	○	●
180	●	<0.001	<0.001	●	●
181	●	0.004	0.046	○	●
182	●	0.15	0.473	○	○

efficiency of OR is substantial when there are many structured weak signals, as we shall see in the simulation studies that are performed in Section 4.

3.3. A data-driven testing procedure

In practice, the HMM parameters ϑ are unknown. We can first estimate the unknown quantities by $\hat{\vartheta}$, and then plug-in $\hat{\vartheta}$ to obtain $\hat{\mathrm{LIS}}_i$. The maximum likelihood estimate (MLE) is commonly used and is strongly consistent and asymptotically normal under certain regularity conditions (Baum and Petrie, 1966; Leroux, 1992; Bickel *et al.*, 1998). The MLE can be computed by using the EM algorithm or other standard numerical optimization schemes, such as the gradient search, or downhill simplex algorithm. These methods were reviewed by Ephraim and Merhav (2002).

Let $\hat{\vartheta}$ be an estimate of the HMM parameter ϑ. Define the plug-in test statistic $\hat{\mathrm{LIS}}_i(\mathbf{x}) = P_{\hat{\vartheta}}(\theta_i = 0|\mathbf{x})$. For given $\hat{\vartheta}$, $\hat{\mathrm{LIS}}_i$ can be computed by using the forward–backward procedure. Denote by $\hat{\mathrm{LIS}}_{(1)}(\mathbf{x}), \ldots, \hat{\mathrm{LIS}}_{(m)}(\mathbf{x})$ the ranked plug-in test statistics and $H_{(1)}, \ldots, H_{(m)}$ the corresponding hypotheses. In light of the oracle procedure (6), we propose the following data-driven testing procedure:

$$\text{let } k = \max\left\{ i : \frac{1}{i} \sum_{j=1}^{i} \hat{\mathrm{LIS}}_{(j)}(\mathbf{x}) \leqslant \alpha \right\}; \text{ then reject all } H_{(i)},\ i = 1, \ldots, k. \tag{7}$$

The testing procedure that is given in expression (7) is referred to as the *local index of significance testing procedure* LIS. We shall show that the performance of OR is asymptotically attained by LIS under the following standard assumptions on the HMM. These assumptions guarantee that good estimates of the model parameters can be constructed on the basis of the observed data. In particular, assumptions 1–3 were assumed by Bickel *et al.* (1998) to show that the MLE for HMM parameters is asymptotically normal. Leroux (1992) showed that assumption 4 is satisfied by the MLE in an HMM under some additional regularity conditions. Assumption 5 is satisfied when, for example, a Gaussian mixture model is assumed.

Assumption 1. $\{\theta_i\}_1^m$ is an irreducible, aperiodic and stationary Markov chain that is characterized by $\vartheta_0 = (\mathcal{A}_0, \boldsymbol{\pi}_0, \mathcal{F}_0)$. ϑ_0 is an interior point of the parameter space Θ.

Assumption 2. Denote by $\mathcal{A}_\vartheta = (a_{ij}(\vartheta))$ the transition matrix and $\boldsymbol{\pi}_\vartheta = (\pi_0(\vartheta), \pi_1(\vartheta))$ the stationary distribution when the underlying HMM parameters are ϑ. There are $\gamma > 0$ and $\varepsilon_0 > 0$ such that, for all $|\vartheta - \vartheta_0| < \gamma$ and all $i, j = 0, 1$, $a_{ij}(\vartheta) \geqslant \varepsilon_0 > 0$ and $\pi_i(\vartheta) \geqslant \varepsilon_0 > 0$.

Assumption 3. Let f_0 and f_1 be the conditional PDFs of x_i. There is a $\gamma > 0$ such that $P_{\vartheta_0}\{\rho_0(X_1) = \infty | \theta_1 = i\} < 1$ for all i, where

$$\rho_0(x) = \sup_{|\vartheta - \vartheta_0| < \gamma} \max_{0 \leqslant i, j \leqslant 1} \{f_i(x)/f_j(x)\}.$$

Assumption 4. $\hat{\vartheta}$ is a consistent estimate of ϑ_0.

Assumption 5. f_0 and f_1 are continuous and positive over the real line, and $\inf_x\{f_0(x)/f_1(x)\} = 0$ for all $|\vartheta - \vartheta_0| < \gamma$.

Without loss of generality, we assume that the γs in assumptions 2, 3 and 5 agree.

We now turn to the asymptotic properties of the plug-in procedure. Theorem 5 shows that the rejection sets that are yielded by OR and LIS are asymptotically equivalent in the sense that the ratio of the number of rejections and the ratio of the number of true positive results yielded by the two procedures approach 1 as $m \to \infty$.

Theorem 5. Consider an HMM defined as in model (2). Let R and $\hat{R}$, and V and $\hat{V}$ be the number of rejections and number of false positive results that are yielded by the OR and LIS procedures respectively. If assumptions 1–5 hold, then $\hat{R}/R \to^{\mathrm{p}} 1$ and $\hat{V}/V \to^{\mathrm{p}} 1$.

Theorem 6 below, together with theorem 4, implies that FDR is controlled at level $\alpha + o(1)$ by LIS, so the LIS procedure is asymptotically valid. Theorem 6 also shows that the performance of OR is attained by the LIS procedure asymptotically in the sense that the FNR level that is yielded by LIS approaches that of OR as $m \to \infty$; therefore the LIS procedure is asymptotically efficient.

Theorem 6. Consider an HMM defined as in model (2). Let $\mathrm{FDR}_{\mathrm{OR}}$ and $\mathrm{FDR}_{\mathrm{LIS}}$, and $\mathrm{FNR}_{\mathrm{OR}}$ and $\mathrm{FNR}_{\mathrm{LIS}}$ be the FDR levels and FNR levels that are yielded by OR and LIS respectively. If assumptions 1–5 hold, then $\mathrm{FDR}_{\mathrm{OR}} - \mathrm{FDR}_{\mathrm{LIS}} \to 0$. In addition, if at least a fixed proportion of hypotheses are not rejected, then $\mathrm{FNR}_{\mathrm{OR}} - \mathrm{FNR}_{\mathrm{LIS}} \to 0$ as $m \to \infty$.

4. Simulation studies

In this section, we investigate the numerical performance of the OR and LIS procedures and compare them with traditional FDR procedures that have been developed for independence tests, including the Benjamini and Hochberg (1995) step-up procedure BH and the adaptive p-value procedure AP. BH and AP are considered for comparison because they are known to control FDR when the hypotheses are generated from an underlying HMM (Wu, 2008). The validity of AP requires a conservative or consistent estimate of p_0, the proportion of true null hypotheses. For an HMM defined as in model (2) with stationary distribution $\boldsymbol{\pi} = (\pi_0, \pi_1)$, we have $p_0 \to^{\mathrm{p}} P(\theta_i = 0) = \pi_0$ and $\boldsymbol{\pi}\mathcal{A} = \boldsymbol{\pi}$. The transition matrix $\mathcal{A} = \{a_{ij}\}$ can be consistently estimated by its MLE $\hat{\mathcal{A}} = \hat{a}_{ij}$; therefore, a consistent estimate of p_0 is given by $\hat{p}_0 = \hat{a}_{10}/(\hat{a}_{01} + \hat{a}_{10})$. This estimate is used for AP in our simulation study.

In our simulation, we assume that $X_i|\theta_i = 0 \sim N(0, 1)$ and $X_i|\theta_i = 1 \sim F_1$, where F_1 is a normal mixture. The normal mixture model can be used to approximate a large collection of distributions and is used in a wide range of applications; see Magder and Zeger (1996) and Efron (2004). The MLE of the HMM parameters in a normal mixture model can be obtained by using the EM algorithm. General methods for estimating HMM parameters in other mixture models such as exponential and Poisson mixtures were discussed in Ephraim and Merhav (2002). One difficulty is that these algorithms assume that L, the number of components in the non-null mixture, is known, but in many applications this is not so. Consistent estimates of L can be obtained by using the method that was proposed by Kiefer (1993) and Liu and Narayan (1994), among others. Alternatively, we can use likelihood-based criteria, such as the Akaike or Bayesian information criterion BIC to select the number of components in the normal mixture.

In this section, we first introduce the EM algorithm for estimating the HMM parameters in a normal mixture model; then we perform simulations to investigate the numerical performance of OR, LIS, BH and AP for testing correlated hypotheses that are generated from an HMM. Finally, we investigate the robustness of LIS under model misspecification and give some practical recommendations for the choice of L when it is unknown.

4.1. EM algorithm in a hidden Markov model for a normal mixture model

The likelihood function for complete data $[(x_i)_1^m, (\theta_i)_1^m]$ in an HMM is

$$P\{\vartheta|(x_i)_1^m, (\theta_i)_1^m\} = \pi_{\theta_1} \prod_{i=2}^{m} a_{\theta_{i-1}\theta_i} \prod_{i=1}^{m} f_{\theta_i}(x_i).$$

Table 4. EM algorithm for normal mixtures in an HMM

```
1. Take initial guesses for model parameters: $\pi_i^{(0)}, a_{ij}^{(0)}, \mu_0^{(0)}, \sigma_0^{2(0)}, c_l^{(0)}, \mu_l^{(0)}, \sigma_l^{2(0)}$
2 (E-step). Compute the following quantities:
    (a) the forward variable $\alpha_i(j) = P_\vartheta\{(x_k)_1^i, \theta_i = j\}$;
    (b) the backward variable $\beta_i(j) = P_\vartheta\{(x_k)_{i+1}^m | \theta_i = j\}$;
    (c) the LIS variable $\gamma_i(j) = \alpha_i(j)\,\beta_i(j)/\{\alpha_i(0)\,\beta_i(0) + \alpha_i(1)\,\beta_i(1)\}$;
    (d) the transition variable $\xi_i(j,k) = P_\vartheta(\theta_i = j, \theta_{i+1} = k | x_1^m) = \gamma_i(j) a_{jk}\, f_k(x_{i+1})\,\beta_{i+1}(k)/\beta_i(j)$;
    (e) the weight variable $w_i(l) = P_\vartheta(x_i \sim f_{1l} | x_1^m) = \gamma_i(1) c_l\, f_{1l}(x_i)/f_1(x_i)$
3 (M-step). Update the model parameters:
    (a) $\pi_j^{(t)} = \gamma_1^{(t-1)}(j)$;
    (b) $a_{jk}^{(t)} = \{\Sigma_{i=1}^{m-1} \xi_i^{(t-1)}(jk)\}/\Sigma_{i=1}^{m-1} \gamma_i^{(t-1)}(j)$;
    (c) $\mu_0^{(t)} = \{\Sigma_{k=1}^m \gamma_k^{(t-1)}(0) x_k\}/\Sigma_{k=1}^m \gamma_k^{(t-1)}(0)$;
    (d) $\sigma_0^{2(t)} = \{\Sigma_{k=1}^m \gamma_k^{(t-1)}(0)(x_k - \mu_0^{(t)})^2\} \Sigma_{k=1}^m \gamma_k^{(t-1)}(0)$;
    (e) $c_l^{(t)} = \{\Sigma_{i=1}^m \omega_i^{(t-1)}(l)\}/\Sigma_{i=1}^m \gamma_i^{(t-1)}(1)$;
    (f) $\mu_l^{(t)} = \{\Sigma_{i=1}^m \omega_i^{(t-1)}(l) x_i\}/\Sigma_{i=1}^m \omega_i^{(t-1)}(1)$;
    (g) $\sigma_l^{2(t)} = \{\Sigma_{i=1}^m \omega_i^{(t-1)}(l)(x_i - \mu_l^{(t)})^2\}/\Sigma_{i=1}^m \omega_i^{(t-1)}(1)$.
4. Iterate the E-step and M-step until convergence
```

Let f_0 be $N(0,1)$ and $f_1(x_i) = \Sigma_{l=1}^L c_l N(x_i|\mu_l, \sigma_l^2)$, where $\Sigma c_l = 1$. Note that, although many software packages are available for estimating normal mixtures in an HMM, the settings are somewhat different from ours because the L-components in the alternative are usually treated as different states, whereas we consider the normal mixture f_1 as one single state (non-null). The EM algorithm for estimating the HMM parameters in our setting is summarized in Table 4. In a normal mixture model, the likelihood function is unbounded when a parameter approaches a boundary point (Kiefer and Wolfowitz, 1956), which may result in non-convergence of the EM algorithm. A restrained parameter space or penalized method should be used when this happens (Hathaway, 1985; Ciuperca *et al.*, 2003).

4.2. Comparison in a normal mixture model

We first assume that L, the number of components in a non-null mixture, is known or estimated correctly from the data. The situation where L is misspecified is considered in Section 4.3. In all simulations, we choose the number of hypotheses $m = 3000$ and the number of replications $N = 500$. The software for implementing the EM algorithm and OR and LIS procedures is available at `http://stat.wharton.upenn.edu/~tcai/paper/html/FDR-HMM.html`.

4.2.1. Simulation study 1: $L = 1$

The Markov chain $(\theta_i)_1^m$ is generated with the initial state distribution $\boldsymbol{\pi}^0 = (\pi_0, \pi_1) = (1, 0)$ and transition matrix

$$\mathcal{A} = \begin{pmatrix} 0.95 & 0.05 \\ 1 - a_{11} & a_{11} \end{pmatrix}.$$

The observations $(x_i)_1^m$ are generated conditionally on $(\theta_i)_1^m$: $x_i|\theta_i = 0 \sim N(0,1)$; $x_i|\theta_i = 1 \sim N(\mu, 1)$. Fig. 2 compares the performance of BH, AP, OR and LIS. In Figs 2(a)–2(c) we choose $\mu = 2$ and plot FDR, FNR and the average number of true positives ATP yielded by BH, AP, OR and LIS as functions of a_{11}. In Figs 2(d)–2(f) we choose $a_{11} = 0.8$ and plot FDR, FNR and ATP as functions of μ. The nominal FDR in all simulations is set at level 0.10.

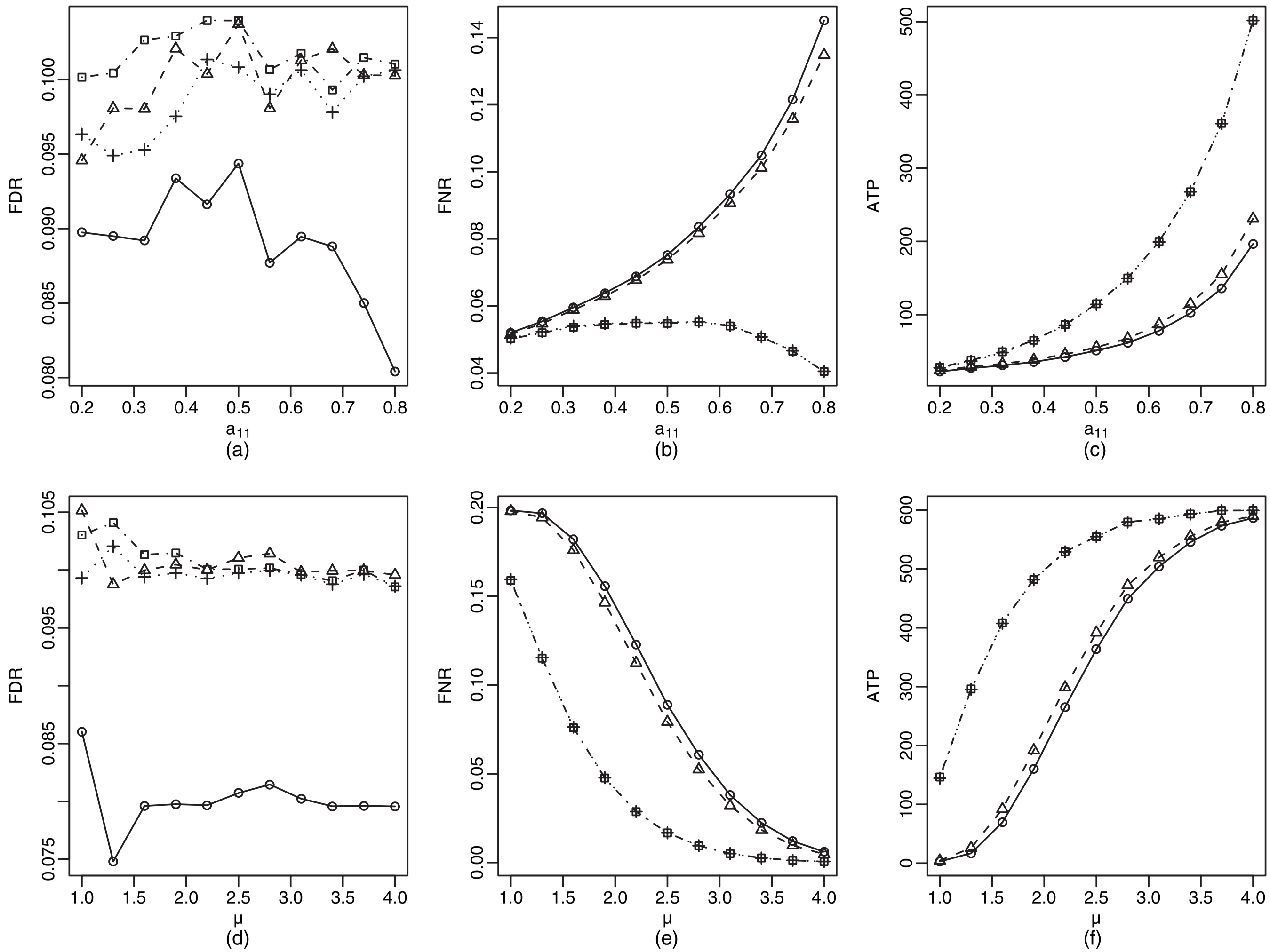

Fig. 2. Comparison of BH (○), AP (△), OR (+) and LIS (□) in an HMM with simple alternative (the FDR level is set at 0.10): (a) FDR *versus* a_{11}; (b) FNR *versus* a_{11}; (c) ATP *versus* a_{11}; (d) FDR *versus* μ; (e) FNR *versus* μ; (f) ATP *versus* μ

From Fig. 2(a), we can see that the FDR levels of all four procedures are controlled at 0.10 asymptotically, and procedure BH is conservative. From Figs 2(b) and 2(c), we can see that the two lines of the oracle procedure and LIS procedure are almost overlapped, indicating that the performance of the oracle procedure is attained by the LIS procedure asymptotically. In addition, the two p-value-based procedures are dominated by the LIS procedure and the difference in FNR and ATP levels becomes larger as a_{11} increases. a_{11} is the transition probability from a non-null case to a non-null case; therefore it controls how likely the non-null cases cluster together. It is interesting to observe that the p-value procedures have higher FNR levels as the non-null cases cluster in larger groups. In contrast, the FNR levels of the LIS procedure decrease as a_{11} increases. This observation shows that, if modelled appropriately, the positive dependence is a blessing (FNR decreases in a_{11}); but, if it is ignored, the positive dependence may become a disadvantage. In situations where the non-null cases are prevented from forming into clusters ($a_{11} < 0.5$), the LIS procedure is still more efficient than BH and AP, although the gain in efficiency is not as much as the situation where $a_{11} > 0.5$.

Fig. 2(d) similarly shows that all procedures are valid and BH is conservative. In Figs 2(e) and 2(f), we plot the FNR and ATP levels as functions of the non-null mean μ. We can see that BH and AP are dominated by LIS, and the difference is large when μ is small to moderate. This is because the LIS procedure can integrate information from adjacent locations, so it is still very efficient even when the signals are weak.

4.2.2. *Simulation study 2: $L \geqslant 2$*

The initial state distribution is $\boldsymbol{\pi}^0 = (1, 0)$, and the transition matrix is

$$\mathcal{A} = \begin{pmatrix} 0.95 & 0.05 \\ 0.2 & 0.8 \end{pmatrix}.$$

The results for comparison are displayed in Fig. 3, where we plot the FDR, FNR and ATP that are yielded by procedures BH, AP, OR and LIS as functions of μ. In Figs 3(a)–3(c), the non-null distribution is a two-component normal mixture $0.5\,N(\mu, 1) + 0.5\,N(2, 1)$. In Figs 3(d)–3(f), the non-null distribution is a three-component normal mixture $0.4\,N(\mu, 1) + 0.3\,N(1, 1) + 0.3\,N(3, 1)$. The nominal FDR in all simulations is set at level 0.10.

We can similarly make the following observations.

(a) All procedures (BH, AP, OR and LIS) control FDR at the nominal level asymptotically, and BH is conservative.
(b) Both BH and AP are dominated by OR and LIS, and the gain in efficiency of OR and LIS is especially large when the signal is weak (small μ) or the average cluster size is large (large a_{11}).
(c) The performances of OR and LIS are similar, as suggested by theorem 6.

The results from both simulation studies show that the dependence actually makes the testing problem 'easier' in the sense that a testing procedure becomes more precise as the dependence increases. So it is desirable to estimate the correlation structure and to incorporate it into a multiple-testing procedure.

4.3. *Model misspecification and practical recommendations*

In many practical applications, the number of components in the non-null mixture L is unknown, yet the information is needed by the algorithms that are used to maximize the likelihood function. We recommend the methods that were mentioned earlier, such as Kiefer's method or BIC, to choose appropriate L. Meanwhile, we perform the following simulation study to investigate

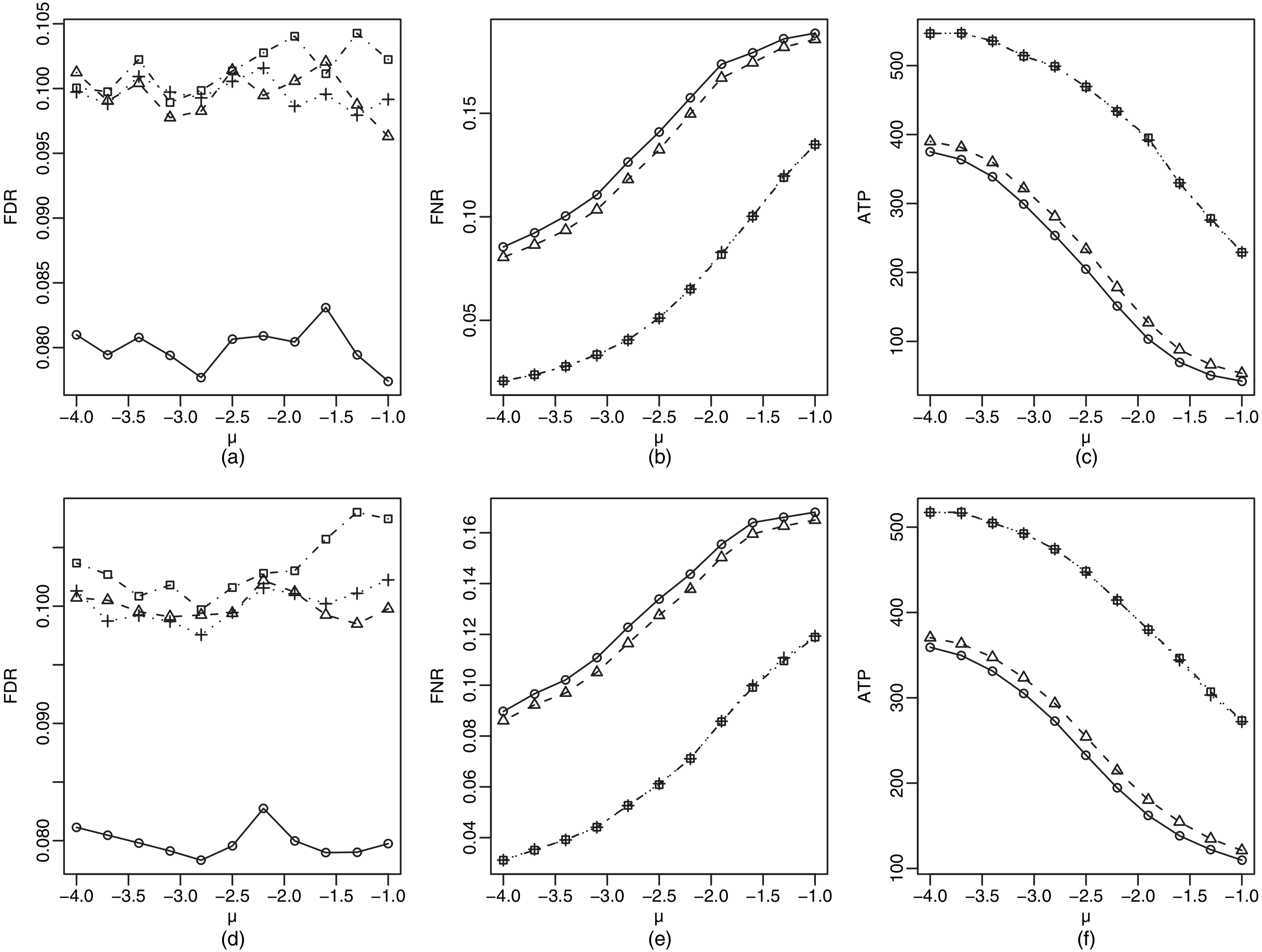

Fig. 3. Comparison of BH (○), AP (Δ), OR (+) and LIS (□) in an HMM (the FDR level is set at 0.10): (a)–(c) two-component alternative; (d)–(f) three-component alternative

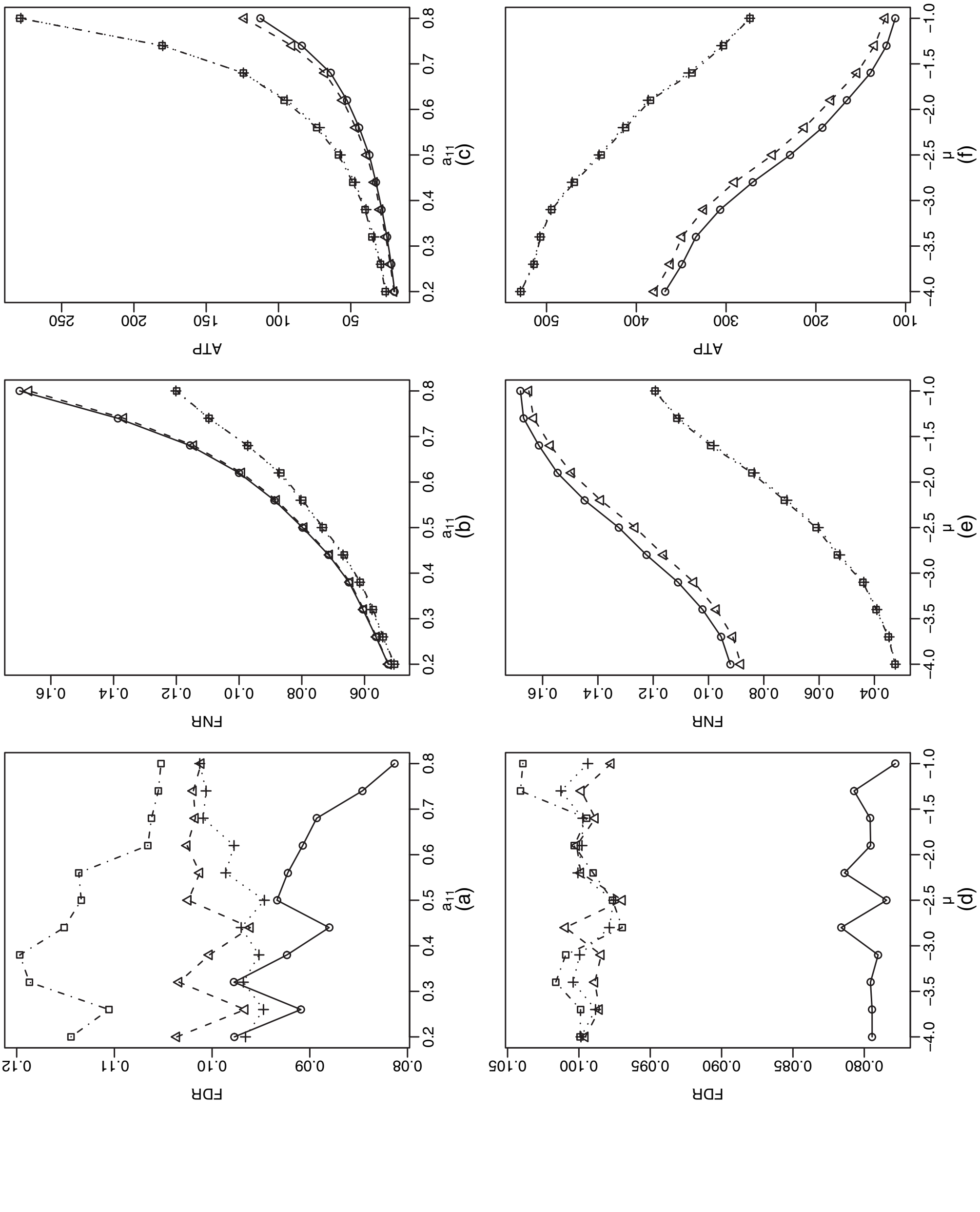

Fig. 4. Comparison of BH (○), AP (Δ), OR (+) and LIS (□) in an HMM when the alternative is misspecified (the FDR level is set at 0.10): (a) FDR *versus* a_{11}; (b) FNR *versus* a_{11}; (c) ATP *versus* a_{11}; (d) FDR *versus* μ; (e) FNR *versus* μ; (f) ATP *versus* μ

the robustness of our testing procedure when L is misspecified. Specifically, we consider the performance of a two-component normal mixture model when the true number of components is greater than 2.

4.3.1. Simulation study 3: misspecified alternative

π^0 and $\mathcal{A}$ are the same as in simulation study 1. Suppose that the non-null distribution is a three-component normal mixture $0.4\,N(\mu,1)+0.3\,N(1,1)+0.3\,N(3,1)$, but we misspecify it as a two-component normal mixture: $p_1\,N(\mu_1,\sigma_1^2)+p_2\,N(\mu_2,\sigma_2^2)$. The comparison results are displayed in Fig. 4.

In Figs 4(a)–4(c) we choose $\mu=-2$ and plot the FDR, FNR and ATP that are yielded by procedures BH, AP, OR and LIS as functions of a_{11}. In Figs 4(d)–4(f), we choose $a_{11}=0.8$ and plot the FDR, FNR and ATP that are yielded by procedures BH, AP, OR and LIS as functions of μ. The nominal FDR in all simulations is set at level 0.10.

We can see that FDR is still controlled at the nominal level by LIS, except for a few cases where a_{11} is small to moderate. Even in these unfavourable cases, LIS does not break down and the actual FDR levels are acceptable. In addition, the gain in efficiency of LIS over BH and AP is significant. Additional simulation results for the cases of $L=4$ and $L=5$ show that the LIS procedure (with the two-component working model) is robust against model misspecification.

5. Application to epidemiologic surveillance data

The analysis and interpretation of the massive databases that are collected routinely from public health surveillance systems are important for prevention and control of epidemic diseases. The increased complexity and scale of these databases present new challenges in epidemiologic research. The surveillance data are typically collected at regular time intervals in the form of epidemiologic indicators, such as incidence rates for a given period of time and in a given population. As a motivating example, we describe the ILI data that were collected from the Sentinelles Network, a national computerized surveillance system in France (`http://websenti.b3e.jussieu.fr/sentiweb`). An ILI is defined as the combination of a sudden fever of at least 39 °C with respiratory signs and myalgia. Weekly ILI incidence rates are standardized according to the sizes of the underlying population as well as the representativeness of the participating physicians. The data for incidence rates between January 1985 and February 2008, which contain 1216 time points, are shown in Fig. 5(a).

The report of a health event based on past data can be classified into one of the two categories—aberration or usual. However, conventional methods in time series analysis, such as auto-regressive integrated moving average models, assume a single underlying probability distribution and stationarity of the underlying sequence. This assumption may not hold in many public health surveillance studies. For example, ILI epidemic data often present irregularly abrupt changes over time. Strat and Carrat (1999) demonstrated that ILI data can be better described by using a two-state HMM, with two states respectively representing a low level dynamic (usual) that may vary according to a seasonal pattern and a high level dynamic (aberration) in which the incidence rate increases sharply at irregular intervals.

Aberrations in the usual incidence rates may provide a signal of an epidemic or clues to possible causes. The timely and accurate detection of these aberrations is important to curtail an outbreak or to identify important risk factors of the disease of interest. The identification of the timing of ILI epidemics involves the simultaneous testing of a large number of hypotheses that correspond to different time periods. Good sensitivity and low false detection rate are among

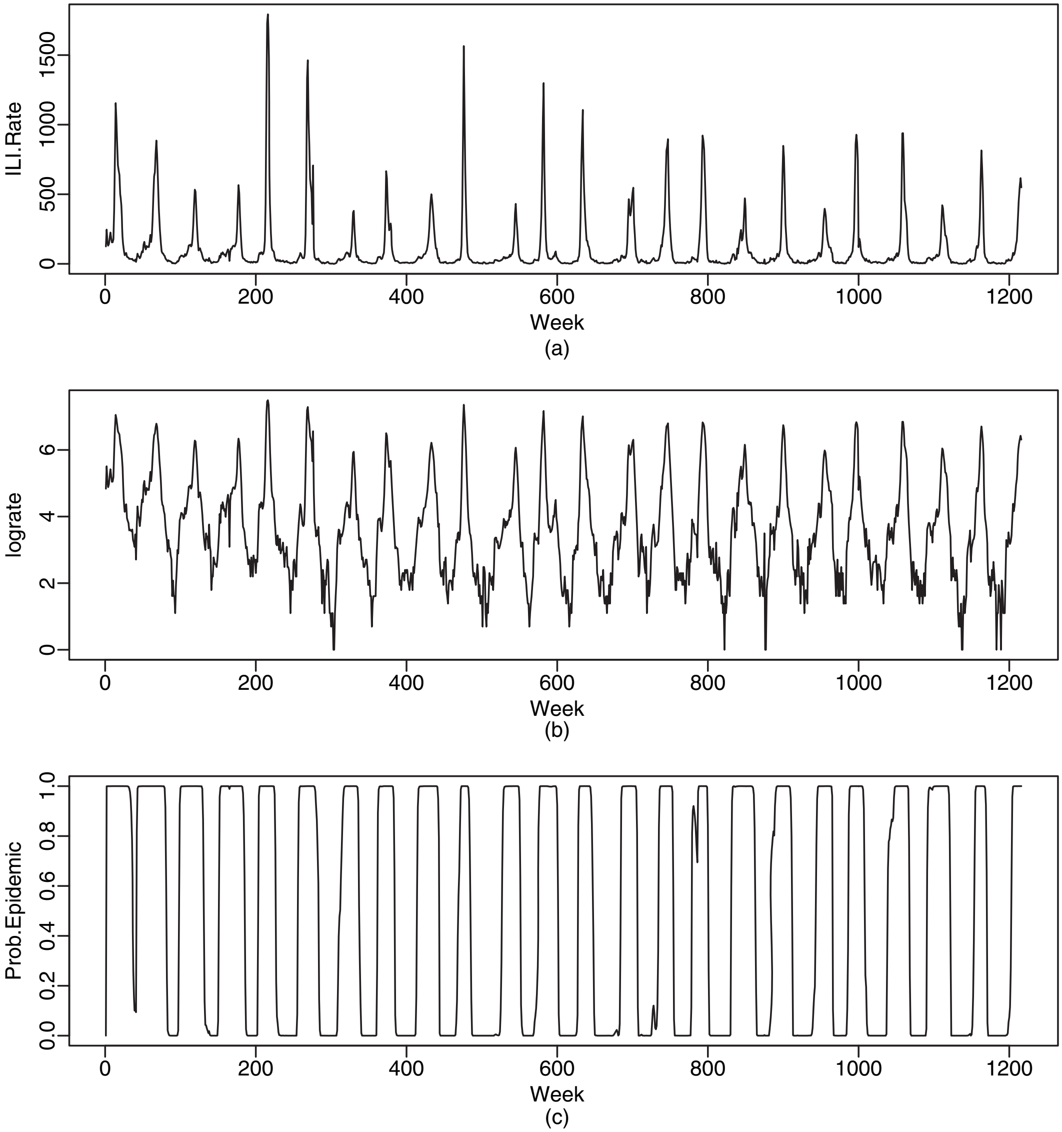

Fig. 5. Weekly ILI (France, 1984–2008): (a) weekly ILI rate; (b) log-transformed weekly ILI rate; (c) estimated probability of ILI epidemic over time

the top concerns when a statistical cut point is defined for the state of ‘aberration’. Specifically, we wish to maintain good power to detect true aberrations so that interventions or investigations can be effectively put into place. At the same time, we wish to guard against too many false positive results to avoid the waste of a large amount of human and financial resources.

The FDR approach is particularly useful as an exploratory tool to compromise these two goals and has recently received much attention in the practice of epidemic disease surveillance and control. For example, Castro and Singer (2006) demonstrated in a scenario of geographical disease surveillance analysis that FDR is a better and efficient alternative to familywise error rate and unadjusted per comparison testing procedures. However, conventional p-value-based

FDR approaches that were originally developed for independent tests are inappropriate for analysis of ILI surveillance data since they ignore the serial correlation between observations. Another difficulty is that p-values are difficult to obtain for each time period because the 'theoretical null distribution' essentially remains unknown. In particular, the null distribution, which corresponds to the state of low level dynamic, needs be estimated on the basis of the past data. In this section, we describe how to use our LIS procedure for detecting aberrations in disease incidence rates over time.

The observed data are assumed to be a mixture of two hidden dynamics—one for the low level incidence rates, and another for high level incidence rates—and the hidden states are distributed as a Markov chain. The goal is to identify time periods that correspond to the state of high level dynamic. The original ILI data are highly skewed. Standard methods for transformation of scale can be used to make the data closer to a normal sample. Both the Box–Cox procedure and Atkinson's score method (Box and Cox, 1964; Atkinson, 1973; Weisberg, 1985) suggest a logarithm transformation of the data. The log-transformed data are shown in Fig. 5(b).

For the log-transformed data, we assume that the null distribution is normal $N(\mu_0, \sigma_0^2)$ and the non-null distribution is a normal mixture $\Sigma_{l=1}^{L} c_l\, N(\mu_l, \sigma_l^2)$, where $\mu_l > \mu_0$, $l = 1, \ldots, L$. The number of components L in the non-null distribution can be determined by using BIC. Specifically, let L be the number of components in the mixture, ϑ_L be the HMM parameters, $P(\vartheta_L|\mathbf{x})$ be the likelihood function and $\hat{\vartheta}_L$ the corresponding MLE. BIC is then defined as $\log\{P(\hat{\vartheta}_L|\mathbf{x})\} - \{f(L)/2\}\log(m)$, where $f(L)$ is the number of parameters that are needed to be estimated in the HMM. We vary L and compare different mixture models; the results are summarized in Table 5. It can be seen that BIC is in favour of a two-component normal mixture model for the non-null distribution.

The LIS procedure is then applied to the ILI data at FDR level 0.001 by assuming a two-component normal mixture for the alternative. The estimated probabilities of ILI epidemic over time are shown in Fig. 5(c). A total of 512 time periods are identified as the high level dynamic, and all these time periods appear in clusters. For the years from 2000 to 2007, the epidemic periods identified are (the durations of these epidemics are given in the parentheses) as follows:

week 47, 1999–week 7, 2000 (13 weeks); week 41, 2000–week 16, 2001 (28 weeks);
week 47, 2001–week 13, 2002 (19 weeks); week 51, 2002–week 16, 2003 (18 weeks);
week 40, 2003–week 6, 2004 (19 weeks); week 48, 2004–week 13, 2005 (18 weeks);
week 40, 2005–week 14, 2006 (28 weeks); week 51, 2006–week 11, 2007 (13 weeks).

Table 5. Comparison of Gaussian mixture models fitted to the ILI data

L	*Null* f_0	*Alternative* f_1	*Transition matrix* $\mathcal{A}$	*log(P)*	*f(L)*	*BIC*
1	$N(2.50, 0.81^2)$	$N(4.93, 1.00^2)$	$\begin{pmatrix} 0.04 & 0.96 \\ 0.96 & 0.04 \end{pmatrix}$	−1717.49	7	−1742.36
2	$N(2.37, 0.76^2)$	$0.50\,N(3.97, 0.42^2) + 0.50\,N(5.57, 0.85^2)$	$\begin{pmatrix} 0.04 & 0.96 \\ 0.96 & 0.04 \end{pmatrix}$	−1655.92	10	−1691.43
3	$N(2.21, 0.70^2)$	$0.32\,N(3.58, 0.32^2) + 0.42\,N(4.44, 0.56^2) +$ $0.26\,N(6.10, 0.59^2)$	$\begin{pmatrix} 0.05 & 0.95 \\ 0.97 & 0.03 \end{pmatrix}$	−1646.41	13	−1692.58
4	$N(2.37, 0.76^2)$	$0.20\,N(3.70, 0.25^2) + 0.31\,N(4.18, 0.43^2) +$ $0.27\,N(5.05, 0.57^2) + 0.22\,N(6.30, 0.50^2)$	$\begin{pmatrix} 0.04 & 0.96 \\ 0.96 & 0.04 \end{pmatrix}$	−1645.41	16	−1702.24

The epidemics usually start from the end of a year and end in the beginning of the following year. However, the patterns of these time periods are quite irregular and unpredictable. For example, the epidemics could start as early as week 41 and as late as week 51, and the durations of these epidemics range from 13 weeks to 28 weeks. Comparing with testing procedures that were proposed by Genovese *et al.* (2006), Benjamini and Heller (2007) and Wei and Li (2008), our testing procedure does not rely on prior information such as prespecified weights or well-defined clusters for different time periods. This property is attractive because the information is usually unavailable in disease surveillance data. In addition, by contrast with the conventional p-value-based procedures, the serial correlation between consecutive observations is exploited by the LIS statistic and hence the significant time periods can be identified by groups. This advantage is of great scientific interest, since a group of significant time periods, rather than individual periods, is more informative in characterizing an outbreak of disease. Furthermore, the LIS statistic at the current time point can be updated on the basis of upcoming observations and can serve as an early warning statistic, which is very useful in the practice of epidemic disease surveillance and decision making.

6. Discussion

In the present paper we have focused on large-scale multiple testing under a special form of dependence—an HMM for the hypotheses. Although the HMM dependence is very useful in many applications, extensions of the LIS procedure to more general forms of dependence would be of great interest from both the theoretical and the practical perspectives. To generalize the optimality result of the LIS oracle procedure to other dependence structures, it is in general required that the corresponding Bernoulli variables, which represent the unknown null and non-null states, have a stationary distribution. Otherwise the null and non-null distributions F_0 and F_1 are no longer well defined.

The problem is more challenging when the goal is to develop a 'good' data-driven procedure since the optimality of the oracle procedure does not guarantee good performance of the corresponding data-driven procedure. The asymptotic optimality of data-driven procedures requires that the estimates of the unknown model parameters are consistent. However, to the best of our knowledge, such theoretical results (consistency of the estimates) for other dependence structures, such as for a higher dimensional random field, have not been developed in the literature. Hence the optimality of the LIS procedure may be lost in the estimation step. In addition, the implementation of the data-driven procedure for other correlation structures may be very complicated. The forward–backward procedure and EM algorithm for an HMM are known to be efficient and relatively easy to programme. However, such efficient algorithms may not exist, for other dependence structures. Much research is needed for developing optimal multiple-testing procedures under general dependence structure.

Acknowledgements

We thank the Associate Editor and the referee for detailed and constructive comments which have helped to improve the presentation of the paper.

Appendix A: Proofs of main results

We shall prove theorems 3–6 and corollary 1. The proofs of theorem 1 and 2 follow similar lines to those in Sun and Cai (2007). For brevity, we omit these proofs here. Some of the technical lemmas are proved in Appendix B.

A.1. Proof of theorem 3

The posterior distribution of $\boldsymbol{\theta}$ is $P_{\boldsymbol{\theta}}|\mathbf{x}(\boldsymbol{\theta}|\mathbf{x}) \propto \pi(\boldsymbol{\theta})\ P(\mathbf{x}|\boldsymbol{\theta})$. The posterior risk is

$$\begin{aligned} E_{\boldsymbol{\theta}|\mathbf{x}}\{L(\boldsymbol{\theta},\boldsymbol{\delta})\} &= \frac{1}{m}\sum_{k=1}^{m} E_{\boldsymbol{\theta}|\mathbf{x}}\{\lambda(1-\theta_k)\delta_k + \theta_k(1-\delta_k)\} \\ &= \frac{1}{m}\sum_{k=1}^{m}\{\lambda\delta_k\ P(\theta_k=0|\mathbf{x}) + (1-\delta_k)P(\theta_k=1|\mathbf{x})\}. \end{aligned}$$

Given that $\vartheta=(\mathcal{A},\boldsymbol{\pi},\mathcal{F})$ is known, $P_{\vartheta}(\theta_k=0|\mathbf{x})$ and $P_{\vartheta}(\theta_k=1|\mathbf{x})$ can be obtained by using the forward–backward procedure (remark 1). The oracle classification rule is therefore given by $\delta_k = I\{\Lambda_k(\mathbf{x}) = P_{\vartheta}(\theta_k=0|\mathbf{x})/P_{\vartheta}(\theta_k=1|\mathbf{x}) < 1/\lambda\}$.

A.2. Proof of corollary 1

Consider a weighted classification problem with loss function

$$L_{\lambda}(\boldsymbol{\theta},\boldsymbol{\delta}) = \frac{1}{m}\sum_{i=1}^{m}\{\lambda\delta_i(1-\theta_i) + (1-\delta_i)\theta_i\}.$$

Take $\lambda=1/t$. Let $t^*>0$. Suppose that $\boldsymbol{\delta}(\boldsymbol{\Lambda},t^*)=(\boldsymbol{\Lambda}<t^*\mathbf{1})$ is used for classification. Note that $E\{(1-\theta_i)\delta_i\} = P(\theta_i=0, T_i<t^*) = \pi_0 G_i^0(t^*)$ and $E\{\theta_i(1-\delta_i)\} = P(\theta_i=1, T_i>t^*) = \pi_1\ \tilde{G}_i^1(t^*)$; the risk is

$$R = \frac{1}{m}\sum_{i=1}^{m}\left\{\frac{1}{t}\pi_0\ G_i^0(t^*) + \pi_1\ \tilde{G}_i^1(t^*)\right\} = \frac{1}{t}\pi_0\ G^0(t^*) + \pi_1 - \pi_1\ G^1(t^*).$$

The optimal cut-off t^* that minimizes this risk satisfies $g^1(t^*)/g^0(t^*) = (1/t)\pi_0/\pi_1$. Meanwhile, from theorem 3 we conclude that the optimal cut-off t^* is given by $t^*=1/\lambda=t$. Hence $g^1(t)/g^0(t)=(1/t)\pi_0/\pi_1$.

A.3. Proof of theorem 4

Denote by $T_i(\mathbf{x}) = \mathrm{LIS}_i(\mathbf{x}) = P_{\vartheta}(\theta_i=0|\mathbf{x})$. Let R be the number of rejections and V the number of false positive results that are yielded by the oracle procedure (6). It is easy to see that the threshold λ satisfies $T_{(R)}(\mathbf{x}) < \lambda \leqslant T_{(R+1)}(\mathbf{x})$. We let $\lambda = T_{(R+1)}(\mathbf{x})$. Define $V/R=0$ if $R=0$. Note that R is known given $\mathbf{x}$; we have

$$\mathrm{FDR} = E\left(\frac{V}{R}\right) = E\left\{E\left(\frac{V}{R}\middle|\mathbf{x}\right)\right\} = E\left\{\frac{1}{R}E(V|\mathbf{x})\right\}$$

and

$$E(V|\mathbf{x}) = E\left[\sum_{i=1}^{m} I\{T_i(\mathbf{x})<\lambda, \theta_k=0\}|\mathbf{x}\right] = \sum_{i=1}^{R} T_{(i)}(\mathbf{x}).$$

Therefore, $\mathrm{FDR} = E\{(1/R)\,\Sigma_{i=1}^R\ T_{(i)}(\mathbf{x})\}$. Since the oracle procedure (6) guarantees that $(1/R)\,\Sigma_{i=1}^R\ T_{(i)}(\mathbf{x}) \leqslant \alpha$ for all realizations of $\mathbf{x}$, FDR is controlled at level α.

A.4. Proof of theorem 5

Consider an infinite dimensional HMM: $(\{\theta_i\}_{-\infty}^{+\infty}, \{x_i\}_{-\infty}^{+\infty})$. Let $T_i = P_{\vartheta}(\theta_i=0|\{x_i\}_1^m)$, $\hat{T}_i = P_{\hat{\vartheta}}(\theta_i=0|\{x_i\}_1^m)$, $T_i^{\infty} = P_{\vartheta}(\theta_i=0|\{x_i\}_{-\infty}^{\infty})$ and $\hat{T}_i^{\infty} = P_{\hat{\vartheta}}(\theta_i=0|\{x\}_{-\infty}^{\infty})$, $i=1,\ldots,m$. Let ξ_i be the test statistic for hypothesis i. We consider the following testing procedure: let

$$k = \max\left\{i : \frac{1}{i}\sum_{j=1}^{i}\xi_{(j)}(\mathbf{x}) \leqslant \alpha\right\};$$

then reject all $H_{(i)}$, $i=1,\ldots,k$, where $\xi_{(i)}$s are the ranked values of ξ_is. When ξ is replaced with T, $\hat{T}$, $T\infty$ and $\hat{T}^{\infty}$, the corresponding procedures are respectively denoted by δ_{OR}, δ_{PI}, $\delta_{\mathrm{OR}}^{\infty}$ and $\delta_{\mathrm{PI}}^{\infty}$. Let the number of rejections, number of false positive and number of false negative results yielded by δ_{OR}, δ_{PI}, $\delta_{\mathrm{OR}}^{\infty}$ and $\delta_{\mathrm{PI}}^{\infty}$ be (R, V, S), ($\hat{R}$, $\hat{V}$, $\hat{S}$), (R^{∞}, V^{∞}, S^{∞}) and ($\hat{R}^{\infty}$, $\hat{V}^{\infty}$, $\hat{S}^{\infty}$) respectively. Note that $\{\theta_i\}_{-\infty}^{\infty}$ is stationary, irreducible and aperiodic; then, according to Leroux (1992), $\{x_i\}_{-\infty}^{\infty}$ is ergodic. The generalization (to a two-sided sequence) of theorem 6.1.3 in Durrett (2005) implies that $\{T_i^{\infty}\}$ is also ergodic.

The proof is outlined as follows. First we establish the asymptotic equivalence between $\delta_{\mathrm{PI}}^{\infty}$ and $\delta_{\mathrm{OR}}^{\infty}$ (lemma 5). It is indicated by lemma 1 that most T_is only differ from T_i^{∞}s by an exponentially small quantity; the asymptotic equivalence of δ_{OR} and $\delta_{\mathrm{OR}}^{\infty}$ as well as the asymptotic equivalence of δ_{PI} and $\delta_{\mathrm{PI}}^{\infty}$ is further established (lemmas 7–10). Then the asymptotic equivalence between δ_{OR} and δ_{PI} follows.

By definition, T_i^{∞}, $i=1,\ldots,m$, are identically distributed. Let $(T_i^{\infty}|\theta_i=j)\sim G_j^{\infty}$, $j=1,2$; then $T_i^{\infty}\sim G^{\infty}=\pi_0 G_0^{\infty}+\pi_1 G_1^{\infty}$. Define $R_{\lambda}^{\infty}=\Sigma_{i=1}^{m} I(T_i^{\infty}<\lambda)$ and $V_{\lambda}^{\infty}=\Sigma_{i=1}^{m} I(T_i^{\infty}<\lambda)$; then $E(R_{\lambda}^{\infty})=m\,G^{\infty}(\lambda)$ and $E(V_{\lambda}^{\infty})=\pi_0 m\,G_0^{\infty}(\lambda)$. Therefore, the mFDR that is yielded by $\delta(\mathbf{T}^{\infty},\lambda)=[I(T_i^{\infty}<\lambda):i=1,\ldots,m]$ is $Q_{\mathrm{OR}}^{\infty}(\lambda)=\pi_0\,G_0^{\infty}(\lambda)/G^{\infty}(\lambda)$. We assume that $T_i^{\infty}\in\mathcal{T}$; then theorem 1 implies that $Q_{\mathrm{OR}}^{\infty}(\lambda)$ is increasing in λ. Let $\lambda_{\mathrm{OR}}^{\infty}=\sup\{\lambda:Q_{\mathrm{OR}}^{\infty}(\lambda)\leqslant\alpha\}$. The proofs of lemmas 1–7 are given in Appendix B.

Lemma 1. Let $\tau_0(x)=\{1+\varepsilon_0^{-2}\rho_0(x)\}^{-1}$ and $L<[m/2]$. For a given k, let $L_1=1\vee(k-L)$ and $L_2=m\wedge(k+L)$. If assumptions 1–3 hold, then, for any ϑ such that $|\vartheta-\vartheta_0|\leqslant\gamma$, $E_{\vartheta_0}|P_{\vartheta}(\theta_k=0|x_{L_1}^{L_2})-P_{\vartheta}(\theta_k=0|x_1^m)|<C_0\beta_0^L$ for some $\beta_0<1$.

Lemma 2. If assumptions 1–3 hold, then the CDF $G^{\infty}(t)=P(T_i^{\infty}<t)$ is continuous. In addition $G^{\infty}(t)>0$ for any given $t>0$.

Lemma 3. Let R and $\hat{R}$ be the number of rejections that are yielded by δ_{OR} and δ_{PI} respectively. If assumptions 1–3 hold, then $R\rightarrow\infty$, $\hat{R}\rightarrow\infty$ almost surely as $m\rightarrow\infty$.

Lemma 4. Let $Q_{\mathrm{OR}}^{\infty}(\lambda)$ be the mFDR that is yielded by $\boldsymbol{\delta}=[I(T_i^{\infty}<\lambda):i=1,\ldots,m]$, and $\lambda_{\mathrm{OR}}^{\infty}=\sup\{\lambda:Q_{\mathrm{OR}}^{\infty}(\lambda)\leqslant\alpha\}$. If assumptions 1–5 hold, then there is an ε_0 such that $\lambda_{\mathrm{OR}}^{\infty}\geqslant\alpha+\varepsilon_0$.

Lemma 5. Let R^{∞} and $\hat{R}^{\infty}$ be the number of rejections that are yielded by $\delta_{\mathrm{OR}}^{\infty}$ and $\delta_{\mathrm{PI}}^{\infty}$ for a given FDR level α. If assumptions 1–3 hold, then $\hat{R}^{\infty}/R^{\infty}\rightarrow^{\mathrm{p}}1$. Similarly, let V^{∞} and $\hat{V}^{\infty}$ be the number of true positive results that are yielded by $\delta_{\mathrm{OR}}^{\infty}$ and $\delta_{\mathrm{PI}}^{\infty}$; then $\hat{V}^{\infty}/V^{\infty}\rightarrow^{\mathrm{p}}1$.

Lemma 6. Let $S_k=\{i:T_i\leqslant T_{(k)}\}$, $\hat{S}_k=\{i:\hat{T}_i\leqslant\hat{T}_{(k)}\}$, $S_k^{\infty}=\{i:T_i^{\infty}\leqslant T_{(k)}^{\infty}\}$ and $\hat{S}_k^{\infty}=\{i:\hat{T}_i^{\infty}\leqslant\hat{T}_{(k)}^{\infty}\}$. If assumptions 1–3 hold, then

$$E\left|\frac{1}{k}\sum_{i\in S_k}T_i-\frac{1}{k}\sum_{i\in S_k^{\infty}}T_i^{\infty}\right|\rightarrow 0$$

as $k\rightarrow\infty$. Similarly,

$$E\left|\frac{1}{k}\sum_{i\in\hat{S}_k}\hat{T}_i-\frac{1}{k}\sum_{i\in\hat{S}_k^{\infty}}\hat{T}_{(i)}^{\infty}\right|\rightarrow 0$$

as $k\rightarrow\infty$.

The next several lemmas establish the asymptotic equivalence (in terms of the number of rejections and number of false positive results) between δ_{OR} and $\delta_{\mathrm{OR}}^{\infty}$ as well as δ_{PI} and $\delta_{\mathrm{PI}}^{\infty}$. Let $\alpha_*=\inf\{0\leqslant t\leqslant 1:G^{\infty}(t)=1\}$. We shall assume that $G^{\infty}(t)$ is continuous and strictly increasing in $(0,\alpha_*)$. Consider two situations:

(a) $\lambda_{\mathrm{OR}}^{\infty}\geqslant\alpha_*$, and
(b) $\lambda_{\mathrm{OR}}^{\infty}<\alpha_*$.

Lemma 7. Let R^{∞}, R, V^{∞} and $\hat{V}$ be defined as in theorem 5. If $\lambda_{\mathrm{OR}}^{\infty}\geqslant\alpha_*$ and assumptions 1–5 hold, then $R^{\infty}/R\rightarrow^{\mathrm{p}}1$ and $V^{\infty}/V\rightarrow^{\mathrm{p}}1$.

Lemma 8. Assume that $0<\lambda_{\mathrm{OR}}^{\infty}<\alpha_*$. Let R, R^{∞}, $\hat{R}$ and $\hat{R}^{\infty}$ be the number of rejections that are yielded by δ_{OR}, $\delta_{\mathrm{OR}}^{\infty}$, δ_{PI} and $\delta_{\mathrm{PI}}^{\infty}$ respectively. If conditions 1–5 hold, then $E|R^{\infty}/R-1|\rightarrow 0$, and $E|\hat{R}^{\infty}/\hat{R}-1|\rightarrow 0$ as $m\rightarrow\infty$.

Proof. We show the results for only δ_{PI} and $\delta_{\mathrm{PI}}^{\infty}$. The results for δ_{OR} and $\delta_{\mathrm{OR}}^{\infty}$ follow similar arguments. Note that, as $\hat{R}\rightarrow\infty$ almost surely, the definition of δ_{PI} implies that $(1/\hat{R})\,\Sigma_{j=1}^{\hat{R}}\hat{T}_{(j)}\leqslant\alpha<\{1/(\hat{R}+1)\}\,\Sigma_{j=1}^{\hat{R}+1}\hat{T}_{(j)}$. Also note that

$$E\left|\frac{1}{\hat{R}}\sum_{j=1}^{\hat{R}}\hat{T}_{(j)}-\frac{1}{\hat{R}+1}\sum_{j=1}^{\hat{R}+1}\hat{T}_{(j)}\right|=\left|\frac{\sum_{j=1}^{\hat{R}}(\hat{T}_{(j)}-\hat{T}_{(\hat{R}+1)})}{\hat{R}(\hat{R}+1)}\right|\leqslant E\left|\frac{1}{\hat{R}+1}\right|\rightarrow 0;$$

we have $E|(1/\hat{R})\,\Sigma_{j=1}^{\hat{R}}\hat{T}_{(j)}-\alpha|\rightarrow 0$. Similarly, $E|(1/\hat{R}^{\infty})\,\Sigma_{j=1}^{\hat{R}^{\infty}}\hat{T}_{(j)}^{\infty}-\alpha|\rightarrow 0$. Therefore,

$$E\left|\frac{1}{\hat{R}}\sum_{j=1}^{\hat{R}}\hat{T}_{(j)}-\frac{1}{\hat{R}^{\infty}}\sum_{j=1}^{\hat{R}^{\infty}}\hat{T}^{\infty}_{(j)}\right|\rightarrow 0. \tag{8}$$

Assume that $E|\hat{R}^{\infty}/\hat{R}-1|\rightarrow 0$ is not true; then there is an $\varepsilon_0>0$ such that, for any $M>0$, $E|\hat{R}^{\infty}/\hat{R}-1|>\varepsilon_0$ holds for some $m\geqslant M$. If $\hat{R}>\hat{R}^{\infty}$, some algebra gives

$$\begin{aligned}\left|\frac{1}{\hat{R}}\sum_{j=1}^{\hat{R}}\hat{T}_{(j)}-\frac{1}{\hat{R}^{\infty}}\sum_{j=1}^{\hat{R}^{\infty}}\hat{T}^{\infty}_{(j)}\right|&=\left|\frac{1}{\hat{R}}\sum_{j=1}^{\hat{R}}\hat{T}_{(j)}-\frac{1}{\hat{R}}\sum_{j=1}^{\hat{R}}\hat{T}^{\infty}_{(j)}+\frac{1}{\hat{R}}\sum_{j=1}^{\hat{R}}\hat{T}^{\infty}_{(j)}-\frac{1}{\hat{R}^{\infty}}\sum_{j=1}^{\hat{R}^{\infty}}\hat{T}^{\infty}_{(j)}\right|\\&\geqslant\left|1-\frac{\hat{R}^{\infty}}{\hat{R}}\right|\left|\hat{T}^{\infty}_{(\hat{R}^{\infty}+1)}-\frac{1}{\hat{R}^{\infty}}\sum_{j=1}^{\hat{R}^{\infty}}\hat{T}^{\infty}_{(j)}\right|-\left|\frac{1}{\hat{R}}\sum_{j=1}^{\hat{R}}\hat{T}_{(j)}-\frac{1}{\hat{R}}\sum_{j=1}^{\hat{R}}\hat{T}^{\infty}_{(j)}\right|.\end{aligned}$$

In Appendix B we show that $\hat{T}^{\infty}_{(\hat{R}^{\infty}+1)}\geqslant\lambda^{\infty}_{\text{OR}}+o_p(1)>\alpha+o_p(1)$. Also note that $E|(1/\hat{R})\,\Sigma_{j=1}^{\hat{R}}\hat{T}_{(j)}-(1/\hat{R})\,\Sigma_{j=1}^{\hat{R}}\hat{T}^{\infty}_{(j)}|\rightarrow 0$ and $E|(1/\hat{R}^{\infty})\,\Sigma_{j=1}^{\hat{R}^{\infty}}\hat{T}^{\infty}_{(j)}-\alpha|\rightarrow 0$; then, for any $M>0$,

$$E\left|\frac{1}{\hat{R}}\sum_{j=1}^{\hat{R}}\hat{T}_{(j)}-\frac{1}{\hat{R}^{\infty}}\sum_{j=1}^{\hat{R}^{\infty}}\hat{T}^{\infty}_{(j)}\right|>\varepsilon_0|\lambda^{\infty}_{\text{OR}}-\alpha|+o(1)$$

holds for some $m\geqslant M$. This is a contradiction to result (8).

Now consider the case $\hat{R}<\hat{R}^{\infty}$. Note that $E|(1/\hat{R})\,\Sigma_{i=1}^{\hat{R}}\,\hat{T}_{(i)}-(1/\hat{R})\,\Sigma_{i=1}^{\hat{R}}\,\hat{T}^{\infty}_{(i)}|\rightarrow 0$; we have $(1/\hat{R})\,\Sigma_{i=1}^{\hat{R}}\,\hat{T}^{\infty}_{(i)}=\alpha+o_p(1)$. So $\hat{T}^{\infty}_{(\hat{R}+1)}\geqslant\alpha+o_p(1)$. Then

$$\left|\frac{1}{\hat{R}^{\infty}}\sum_{j=1}^{\hat{R}^{\infty}}\hat{T}^{\infty}_{(j)}-\frac{1}{\hat{R}}\sum_{j=1}^{\hat{R}}\hat{T}_{(j)}\right|\geqslant\left|1-\frac{\hat{R}}{\hat{R}^{\infty}}\right|\left|\frac{1}{\hat{R}^{\infty}-\hat{R}}\sum_{j=\hat{R}+1}^{\hat{R}^{\infty}}\hat{T}^{\infty}_{(j)}-\frac{1}{\hat{R}}\sum_{j=1}^{\hat{R}}\hat{T}^{\infty}_{(j)}\right|-\left|\frac{1}{\hat{R}}\sum_{j=1}^{\hat{R}}\hat{T}_{(j)}-\frac{1}{\hat{R}}\sum_{j=1}^{\hat{R}}\hat{T}^{\infty}_{(j)}\right|.$$

Let $\alpha<\eta<\min(\alpha_*,\lambda^{\infty}_{\text{OR}})$, $S_1=\{i:\hat{T}^{\infty}_{(R)}<\hat{T}^{\infty}_i<\eta\}$ and $S_2=\{i:\eta<\hat{T}^{\infty}_i<T^{\infty}_{(\hat{R}^{\infty})}\}$. Note that, as $\hat{T}^{\infty}_{(\hat{R}+1)}\geqslant\alpha+o_p(1)$, we have

$$\frac{1}{\hat{R}^{\infty}-\hat{R}}\sum_{j=\hat{R}+1}^{\hat{R}^{\infty}}\hat{T}^{\infty}_{(j)}\geqslant\frac{1}{\hat{R}^{\infty}-\hat{R}}\sum_{j\in S_1}\hat{T}^{\infty}_{(j)}+\frac{1}{\hat{R}^{\infty}-\hat{R}}\sum_{j\in S_2}\hat{T}^{\infty}_{(j)}\geqslant\alpha+\frac{|S_2|}{\hat{R}^{\infty}-\hat{R}}(\eta-\alpha)+o_p(1),$$

where $|S_2|=\Sigma_{i=1}^{m}I(i\in S_2)$. We apply the ergodic theorem to obtain $(1/m)|S_2|=G^{\infty}(\lambda^{\infty}_{\text{OR}})-G^{\infty}(\eta)+o_p(1)$ and $(1/m)(\hat{R}^{\infty}-\hat{R})\leqslant G(\lambda^{\infty}_{\text{OR}})-G(\alpha)+o_p(1)$. Hence

$$\frac{1}{\hat{R}^{\infty}-\hat{R}}\sum_{j=\hat{R}+1}^{\hat{R}^{\infty}}\hat{T}^{\infty}_{(j)}\geqslant\alpha+\frac{G^{\infty}(\lambda^{\infty}_{\text{OR}})-G^{\infty}(\eta)}{G(\lambda^{\infty}_{\text{OR}})-G(\alpha)}(\eta-\alpha)+o_p(1).$$

Note that $G^{\infty}(t)$, the CDF of T^{∞}_i, is strictly increasing in t over the interval $(0,\alpha_*)$, implying that $\nu_0=[\{G^{\infty}(\lambda^{\infty}_{\text{OR}})-G^{\infty}(\eta)\}/\{G(\lambda^{\infty}_{\text{OR}})-G(\alpha)\}](\eta-\alpha)>0$. Hence

$$\left|\frac{1}{\hat{R}^{\infty}}\sum_{j=1}^{\hat{R}^{\infty}}\hat{T}^{\infty}_{(j)}-\frac{1}{\hat{R}}\sum_{j=1}^{\hat{R}}\hat{T}_{(j)}\right|\geqslant\left|1-\frac{\hat{R}}{\hat{R}^{\infty}}\right|\nu_0+o_p(1).$$

We take expectations on both sides, and note that, as both $(1/\hat{R}^{\infty})\,\Sigma_{j=1}^{\hat{R}^{\infty}}\,\hat{T}^{\infty}_{(j)}$ and $(1/\hat{R})\,\Sigma_{j=1}^{\hat{R}}\,\hat{T}_{(j)}$ are bounded, we must have $E|\hat{R}^{\infty}/\hat{R}-1|\rightarrow 0$.

Lemma 9. Assume that $0<\lambda^{\infty}_{\text{OR}}<\alpha_*$. Let $\hat{\lambda}_{\text{OR}}$, $\hat{\lambda}^{\infty}_{\text{OR}}$, $\hat{\lambda}_{\text{PI}}$ and $\hat{\lambda}^{\infty}_{\text{PI}}$ be the threshold that is yielded by $\delta_{\text{OR}},\delta^{\infty}_{\text{OR}},\delta_{\text{PI}}$ and $\delta^{\infty}_{\text{PI}}$ respectively. If assumptions 1–5 hold, then $\hat{\lambda}_{\text{OR}}-\hat{\lambda}^{\infty}_{\text{OR}}\rightarrow^{\text{p}}0$, and $\hat{\lambda}_{\text{PI}}-\hat{\lambda}^{\infty}_{\text{PI}}\rightarrow^{\text{p}}0$.

Proof. We shall show only the first part of the lemma; the second part follows similar arguments. We claim that $\hat{\lambda}_{\mathrm{OR}}-\hat{\lambda}^{\infty}_{\mathrm{OR}}\rightarrow^{\mathrm{p}}0$ is true for, if not, there are ε_0 and δ_0 such that, for any $M>0$, $P(|\hat{\lambda}^{\infty}_{\mathrm{OR}}-\hat{\lambda}_{\mathrm{OR}}|\geqslant\varepsilon_0)>\delta_0$ holds for some $m\geqslant M$, $m\in\mathbb{Z}^{+}$. We shall consider two cases.

(a) For $\hat{\lambda}_{\mathrm{OR}}\geqslant\hat{\lambda}^{\infty}_{\mathrm{OR}}+\varepsilon_0$, let $L=m^{\kappa}$, where $0<\kappa<1$. Define set $S=\{i:i<L+1 \text{ or } i>m-L-1\}$ and its complement $S^{\mathrm{c}}=\{i:L+1\leqslant i\leqslant m-L-1\}$. In Appendix B we show that, for $i\in S$, $|T_i-T_i^{\infty}|<s_i$, where

$$s_i=\prod_{j=i-L+1}^{i-1}\exp\{-2\,\tau_0(x_j)\}+\prod_{j=i+1}^{i+L-1}\exp\{-2\,\tau_0(x_j)\}.$$

Denote by $|S|$ the cardinality of S. Note that, when $\hat{\lambda}_{\mathrm{OR}}\geqslant\hat{\lambda}^{\infty}_{\mathrm{OR}}+\varepsilon_0$, we have $I(T_i<\hat{\lambda}_{\mathrm{OR}})+I(|T_i-T_i^{\infty}|>\varepsilon_0/2)\geqslant I(T_i^{\infty}<\hat{\lambda}^{\infty}_{\mathrm{OR}}+\varepsilon)$. It follows that

$$\frac{1}{m}\left\{\sum_{i=1}^{m}I(T_i<\hat{\lambda}_{\mathrm{OR}})\right\}\geqslant\frac{1}{m}\left\{\sum_{i=1}^{m}I\left(T_i^{\infty}<\hat{\lambda}^{\infty}_{\mathrm{OR}}+\frac{\varepsilon_0}{2}\right)\right\}-\frac{1}{m}\left\{\sum_{i\in S}I\left(s_i>\frac{\varepsilon_0}{2}\right)\right\}-\frac{1}{m}|S^{\mathrm{c}}|.$$

Recall that the sequence $\{s_i\}_{i\in S}$ is ergodic and that $E(s_k)<C_0\beta_0^{L}$ for some $0<\beta_0<1$; it follows that $(1/m)\,\Sigma_{i\in S}\,s_k=o_p(1)$. Therefore,

$$\begin{aligned}\frac{1}{m}R=\frac{1}{m}\left\{\sum_{i=1}^{m}I(T_i<\hat{\lambda}_{\mathrm{OR}})\right\}&\geqslant\frac{1}{m}\left\{\sum_{i=1}^{m}I\left(T_i^{\infty}<\hat{\lambda}^{\infty}_{\mathrm{OR}}+\frac{\varepsilon_0}{2}\right)\right\}+o_p(1)\\&=G^{\infty}\left(\lambda^{\infty}_{\mathrm{OR}}+\frac{\varepsilon}{2}\right)+o_p(1).\end{aligned}$$

However, we have that $(1/m)R^{\infty}=G^{\infty}(\lambda^{\infty}_{\mathrm{OR}})$, so $R/R^{\infty}=G^{\infty}(\lambda^{\infty}_{\mathrm{OR}}+\varepsilon_0/2)/G^{\infty}(\lambda^{\infty}_{\mathrm{OR}})+o_p(1)$. Note that, as we assume $\lambda^{\infty}_{\mathrm{OR}}<\alpha_*$, we can choose ε_0 such that $\lambda^{\infty}_{\mathrm{OR}}+\varepsilon_0/2<\alpha_*$. Recall that $G^{\infty}(t)$ is strictly increasing in t over $(0,\alpha_*)$; we conclude that $R/R^{\infty}\geqslant 1+\nu_0+o_p(1)$ for some $\nu_0>0$, which contradicts lemma 8.

(b) For $\hat{\lambda}_{\mathrm{OR}}\geqslant\hat{\lambda}^{\infty}_{\mathrm{OR}}+\varepsilon_0$, we can similarly show that $R/R^{\infty}=G^{\infty}\{\lambda^{\infty}_{\mathrm{OR}}-(\varepsilon_0/2)\}/G^{\infty}(\lambda^{\infty}_{\mathrm{OR}})+o_p(1)\leqslant 1-\nu_0'+o_p(1)$, for some $\nu_0>0$, which contradicts lemma 8.

Lemma 10. Assume that $0<\lambda^{\infty}_{\mathrm{OR}}<\alpha_*$. Let V, V^{∞}, $\hat{V}$ and $\hat{V}^{\infty}$ be the number of false positive results that are yielded by δ_{OR}, $\delta^{\infty}_{\mathrm{OR}}$, δ_{PI} and $\delta^{\infty}_{\mathrm{PI}}$ respectively. If conditions 1–4 hold, then $V/V^{\infty}\rightarrow^{\mathrm{p}}1$ and $\hat{V}/\hat{V}^{\infty}\rightarrow^{\mathrm{p}}1$.

Proof. Let $\hat{\lambda}_{\mathrm{OR}}$, $\hat{\lambda}^{\infty}_{\mathrm{OR}}$, $\hat{\lambda}_{\mathrm{PI}}$ and $\hat{\lambda}^{\infty}_{\mathrm{PI}}$ be the threshold that is yielded by δ_{OR}, $\delta^{\infty}_{\mathrm{OR}}$, δ_{PI} and $\delta^{\infty}_{\mathrm{PI}}$ respectively. Let $S_1=\{i:T_i^{\infty}<\hat{\lambda}^{\infty}_{\mathrm{OR}}\}$, $S_2=\{i:T_i<\hat{\lambda}_{\mathrm{OR}}\}$, $S_3=\{i:\hat{T}_i^{\infty}<\hat{\lambda}^{\infty}_{\mathrm{PI}}\}$ and $S_4=\{i:\hat{T}_i<\hat{\lambda}_{\mathrm{OR}}\}$. Let $S_A=S_1\cap S_2$ and $S_B=S_3\cap S_4$. Denote by $|S|$ the cardinality of set S. We let $S^*=(S_1\backslash S_2)\cup(S_2\backslash S_1)=S_1^*\cup S_2^*$, where $S_1^*=\{i:T_i^{\infty}<\hat{\lambda}^{\infty}_{\mathrm{OR}}$ and $T_i\geqslant\hat{\lambda}_{\mathrm{OR}}\}$ and $S_2^*=\{i:T_i^{\infty}\geqslant\hat{\lambda}^{\infty}_{\mathrm{OR}}$ and $T_i<\hat{\lambda}_{\mathrm{OR}}\}$. It is shown in Appendix B that $\hat{\lambda}^{\infty}_{\mathrm{OR}}\rightarrow^{\mathrm{p}}\lambda^{\infty}_{\mathrm{OR}}$. The ergodic theorem implies that $(1/m)R^{\infty}=G^{\infty}(\lambda^{\infty}_{\mathrm{OR}})+o_p(1)$. First we show that $|S_A|/|S_1|\rightarrow^{\mathrm{p}}1$, which is equivalent to showing that $(1/m)|S_1^*|\rightarrow^{\mathrm{p}}0$, and $(1/m)|S_2^*|\rightarrow^{\mathrm{p}}0$. For $K>0$, let $S_{1K}^*=\{i:T_i^{\infty}<\hat{\lambda}^{\infty}_{\mathrm{OR}}-1/K$ and $T_i>\hat{\lambda}_{\mathrm{OR}}\}$. Hence

$$\frac{1}{m}|S_1^*|\leqslant\frac{1}{m}|S_{1K}^*|+\frac{1}{m}\sum_{i=1}^{m}I\left(\hat{\lambda}^{\infty}_{\mathrm{OR}}-\frac{1}{K}\leqslant T_i^{\infty}<\hat{\lambda}^{\infty}_{\mathrm{OR}}\right).\tag{9}$$

Now we show that $(1/m)|S_{1K}^*|\rightarrow^{\mathrm{p}}0$. Note that

$$I\left(\left\{T_i^{\infty}<\hat{\lambda}^{\infty}_{\mathrm{OR}}-\frac{1}{K}\right\}\cap\{T_i>\hat{\lambda}_{\mathrm{OR}}\}\right)\leqslant I\left(|\hat{\lambda}^{\infty}_{\mathrm{OR}}-\hat{\lambda}_{\mathrm{OR}}|>\frac{1}{2K}\right)+I\left(|T_i^{\infty}-T_i|>\frac{1}{2K}\right);$$

we have

$$\frac{1}{m}|S_{1K}^*|\leqslant I\left(|\hat{\lambda}^{\infty}_{\mathrm{OR}}-\hat{\lambda}_{\mathrm{OR}}|>\frac{1}{2K}\right)+\frac{1}{m}\sum_{i=1}^{m}I\left(s_i>\frac{1}{2K}\right).$$

The ergodic theorem implies that $(1/m)\,\Sigma_{i=1}^{m}\,I(s_i>1/2K)\rightarrow^{\mathrm{p}}0$. Together with lemma 9 we conclude that $(1/m)|S_{1K}^*|\rightarrow^{\mathrm{p}}0$ for any $K>0$. However, we can choose K sufficiently large that the second term in inequality (9) is small. Thus we have $(1/m)|S_1^*|\rightarrow^{\mathrm{p}}0$. Similarly we can show that $(1/m)|S_2^*|\rightarrow^{\mathrm{p}}0$. Therefore, $|S_A|/|S_1|\rightarrow^{\mathrm{p}}1$ and $V/V^{\infty}\rightarrow^{\mathrm{p}}1$. The second part of the lemma follows similar arguments.

A.5. Proof of theorem 5 (continued)

Note that $\hat{R}/R = (\hat{R}/\hat{R}^{\infty})(\hat{R}^{\infty}/R^{\infty})R^{\infty}/R$. It follows from lemmas 5, 7 and 8 that $\hat{R}/R \xrightarrow{\text{p}} 1$. The second part of the lemma can be proved similarly.

A.6. Proof of theorem 6

By convention (e.g. Genovese and Wasserman (2002)), we define the false discovery proportion

$$\text{FDP} = \begin{cases} N_{10}/R & \text{if } R > 0, \\ 0 & \text{otherwise} \end{cases}$$

and the false non-discovery proportion

$$\text{FNP} = \begin{cases} N_{01}/S & \text{if } S > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Observe that both V/R and $\hat{V}/\hat{R}$ are bounded above by 1, and that

$$\frac{V}{R} - \frac{\hat{V}}{\hat{R}} = \frac{V}{R}\left(1 - \frac{R}{\hat{R}}\right) + \frac{\hat{V}}{\hat{R}}\left(\frac{V}{\hat{V}} - 1\right);$$

together with theorem 5, we conclude that $E(\text{FDP}_{\text{OR}}) - E(\text{FDR}_{\text{LIS}}) \rightarrow 0$. Let U and $\hat{U}$ be the number of false negative results. To show the second part of the theorem, observe that

$$\frac{U}{S} - \frac{\hat{U}}{\hat{S}} = \frac{U}{S}\left(1 - \frac{S}{\hat{S}}\right) + \frac{\hat{U}}{\hat{S}}\left(\frac{U}{\hat{U}} - 1\right).$$

If at least a fixed proportion of hypotheses are not rejected, then it is easy to show that $S/\hat{S} \xrightarrow{\text{p}} 1$ and $U/\hat{U} \xrightarrow{\text{p}} 1$. Also note that, as both U/S and $\hat{U}/\hat{S}$ are bounded above by 1, we have $E(\text{FNP}_{\text{OR}}) - E(\text{FNP}_{\text{LIS}}) \rightarrow 0$.

Appendix B: Proofs of other results

We first present four additional lemmas before giving the proof of lemma 1. Lemmas 11 and 12, which essentially extend the results in Baum and Petrie (1966), were stated in Bickel and Ritov (1996) without proofs. For completeness, we provide the proofs here.

Lemma 11. Let $\tau_0(x) = \{1 + \varepsilon_0^{-2}\rho_0(x)\}^{-1}$ and assume that conditions 1–3 hold; then, for all ϑ such that $|\vartheta - \vartheta_0| < \gamma$, $P_{\vartheta}(\theta_{k+1} = j|\theta_k = i, x_1^m) \geqslant \tau_0(x_{k+1})$.

Proof. Let j and j' be two states. Then

$$\begin{aligned}
\frac{P_{\vartheta}(\theta_{k+1} = j|\theta_k = i, x_1^m)}{P_{\vartheta}(\theta_{k+1} = j'|\theta_k = i, x_1^m)} &= \frac{P_{\vartheta}(\theta_{k+1} = j, \theta_k = i, x_1^m)}{P_{\vartheta}(\theta_{k+1} = j', \theta_k = i, x_1^m)} \\
&= \frac{P_{\vartheta}(\theta_{k+1} = j, x_1^m|\theta_k = i)}{P_{\vartheta}(\theta_{k+1} = j', x_1^m|\theta_k = i)} \\
&= \frac{\sum_{j_0=0}^{1} P_{\vartheta}(\theta_{k+1} = j, \theta_{k+2} = j_0, x_{k+1}^m|\theta_k = i)}{\sum_{j_0=1}^{1} P_{\vartheta}(\theta_{k+1} = j', \theta_{k+2} = j_0, x_{k+1}^m|\theta_k = i)} \\
&= \frac{\sum_{j_0=0}^{1} A_{ij}A_{jj_0}\, f_j(x_{k+1})\, P_{\vartheta}(x_{k+2}^m|\theta_{k+2} = j_0)}{\sum_{j_0=0}^{1} A_{ij'}A_{j'j_0}\, f_{j'}(x_{k+1})\, P_{\vartheta}(x_{k+2}^m|\theta_{k+2} = j_0)}.
\end{aligned}$$

Note that $A_{ij} \geqslant \varepsilon_0$ for all i and j and $P_{\vartheta}(x_{k+1}|j)/P_{\vartheta}(x_{k+1}|j') \leqslant \rho_0(x_{k+1})$; we have

$$P_{\vartheta}(\theta_{k+1} = j|\theta_k = i, x_1^m) < \varepsilon_0^{-2}\,\rho_0(x_{k+1})\, P_{\vartheta}(\theta_{k+1} = j'|\theta_k = i, x_1^m).$$

Since $\Sigma_{j=0}^{1} P_\vartheta(\theta_{k+1}=j|\theta_k=i, x_1^m)=1$, we conclude that for all i and j

$$P_\vartheta(\theta_{k+1}=j|\theta_k=i, x_1^m) \geqslant \{1+\varepsilon_0^{-2}\,\rho_0(x_{k+1})\}^{-1} \equiv \tau_0(x_{k+1}).$$

Lemma 12. Let $M_d^+(k,\vartheta)=\max_i\{P_\vartheta(\theta_k=j|x_1^m, \theta_{k-d}=i)\}$ and define $M_d^-(k,\vartheta)$ as the corresponding minimum. Assume that conditions 1–3 hold; then, for all ϑ such that $|\vartheta-\vartheta_0|\leqslant\gamma$,

$$|M_d^+(k,\vartheta)-M_d^-(k,\vartheta)| \leqslant \prod_{j=k-d+1}^{k-1} \{1-2\,\tau_0(x_j)\}, \tag{10}$$

and the inequality holds independently of j.

Proof.

$$\begin{aligned}P_\vartheta(\theta_k=j|x_1^m, \theta_{k-d}=i) &= \sum_{j_0=0}^{1} P_\vartheta(\theta_k=j, \theta_{k-d+1}=j_0|x_1^m, \theta_{k-d}=i)\\ &= \sum_{j_0=0}^{1} P_\vartheta(\theta_k=j|x_1^m, \theta_{k-d+1}=j_0)\, P_\vartheta(\theta_{k-d+1}=j_0|x_1^m, \theta_{k-d}=i).\end{aligned}$$

Then, according to lemma 11,

$$M_d^+(k,\vartheta) \leqslant \{1-\tau_0(x_{k-d+1})\}M_{d-1}^+(k,\vartheta)+\tau_0(x_{k-d+1})M_{d-1}^-(k,\vartheta).$$

Similarly we have

$$M_d^-(k,\vartheta) \geqslant \{1-\tau_0(x_{k-d+1})\}M_{d-1}^-(k,\vartheta)+\tau_0(x_{k-d+1})M_{d-1}^+(k,\vartheta);$$

thus

$$M_d^+(k,\vartheta)-M_d^-(k,\vartheta) \leqslant \{1-2\,\tau_0(x_{k-d+1})\}\{M_{d-1}^+(k,\vartheta)-M_{d-1}^-(k,\vartheta)\}.$$

Note that, as $\tau_0(x_j)\leqslant 1/5$ (since we always have $\varepsilon_0\leqslant\frac{1}{2}$ and $\rho_0(x)\geqslant 1$), $M_1^+(k,\vartheta)=1$ and $M_1^-(k,\vartheta)=0$, the proof is complete by induction. □

Complementary to lemmas 11 and 12 are results concerning the time-reversed HMM $\{(\theta_k', X_k')\}$, which are summarized in the following lemma. The proof is similar to those of lemmas 11 and 12 and hence is omitted.

Lemma 13. Let $\tau_0(x)=\{1+\varepsilon_0^{-2}\,\rho_0(x)\}^{-1}$. Then $P_\vartheta(\theta_k=j|\theta_{k+1}=i, x_m^1)\geqslant\tau_0(x_k)$ for all ϑ such that $|\vartheta-\vartheta_0|<\gamma$. In addition, let $M_d'^+(k,\vartheta)=\max_i\{P_\vartheta(\theta_k=j|x_m^1, \theta_{k+d}=i)\}$ and define $M_d'^-(k,\vartheta)$ as the corresponding minimum. If assumptions 1–3 hold, then

$$|M_d'^+(k,\vartheta)-M_d'^-(k,\vartheta)| \leqslant \prod_{j=k+d-1}^{k+1} \{1-2\,\tau_0(x_j)\} \tag{11}$$

for all ϑ such that $|\vartheta-\vartheta_0|\leqslant\gamma$, and the inequality holds independently of j. $\tau_0(x)$ can be chosen to be the same for the original and the time-reversed HMMs.

Lemma 14. Let $\tau_0(x)=\{1+\varepsilon_0^{-2}\rho_0(x)\}^{-1}$ and $L<[m/2]$. For a given k, let $L_1=1\vee(k-L)$ and $L_2=m\wedge(k+L)$. Assume that conditions 1–3 hold; then, for any ϑ such that $|\vartheta-\vartheta_0|\leqslant\gamma$,

$$P_\vartheta(\theta_k=0|x_{L_1}^{L_2})-P_\vartheta(\theta_k=0|x_1^m) < \prod_{i=k+1}^{L_2-1} \exp\{-2\,\tau_0(x_i)\} \qquad \text{if } L_1=1,$$

and

$$P_\vartheta(\theta_k=0|x_{L_1}^{L_2})-P_\vartheta(\theta_k=0|x_1^m) < \prod_{i=L_1+1}^{k-1} \exp\{-2\,\tau_0(x_i)\} \qquad \text{if } L_2=m.$$

For $L_1>1$ and $L_2<m$, we have

$$P_\vartheta(\theta_k=0|x_{L_1}^{L_2})-P_\vartheta(\theta_k=0|x_1^m) < \prod_{i=L_1+1}^{k-1} \exp\{-2\,\tau_0(x_i)\}+\prod_{i=k+1}^{L_2-1} \exp\{-2\,\tau_0(x_i)\} \tag{12}$$

Proof. Note that

$$|P_\vartheta(\theta_k=0|x_{L_1}^{L_2})-P_\vartheta(\theta_k=0|x_1^m)|\leqslant|P_\vartheta(\theta_k=0|x_{L_1}^{L_2})-P_\vartheta(\theta_k=0|x_1^{L_2})|+|P_\vartheta(\theta_k=0|x_1^{L_2})-P_\vartheta(\theta_k=0|x_1^m)|;$$

we only need to show that

$$|P_\vartheta(\theta_k=0|x_{L_1}^{L_2})-P_\vartheta(\theta_k=0|x_1^{L_2})|\leqslant\prod_{i=L_1+1}^{k-1}\exp\{-2\,\tau_0(x_i)\}$$

and

$$|P_\vartheta(\theta_k=0|x_1^{L_2})-P_\vartheta(\theta_k=0|x_1^m)|\leqslant\prod_{i=k+1}^{L_2-1}\exp\{-2\,\tau_0(x_i)\}.$$

The case of $L_1=1$ is trivial, so we assume that $L_1>1$. Then

$$\begin{aligned}|P_\vartheta(\theta_k=0|x_{L_1}^{L_2})-P_\vartheta(\theta_k=0|x_1^{L_2})|&=\Bigg|\sum_{j=0}^{1}P_\vartheta(\theta_k=0|\theta_{k-L}=j,x_{L_1}^{L_2})\,P_\vartheta(\theta_{k-L}=j|x_{L_1}^{L_2})\\&\qquad-\sum_{j'=0}^{1}P_\vartheta(\theta_k=0|\theta_{k-L}=j',x_{L_1}^{L_2})\,P_\vartheta(\theta_{k-L}=j'|x_1^{L_2})\Bigg|\\&\leqslant\max_{j,j'}|P_\vartheta(\theta_k=0|\theta_{k-L}=j,x_{L_1}^{L_2})-P_\vartheta(\theta_k=0|\theta_{k-L}=j',x_{L_1}^{L_2})|\\&\leqslant\prod_{i=L_1+1}^{k-1}\{1-2\,\tau_0(x_j)\}\\&\leqslant\prod_{i=L_1+1}^{k-1}\exp\{-2\,\tau_0(x_i)\}.\end{aligned}$$

Similarly we can show that $|P_\vartheta(\theta_k=0|x_1^{L_2})-P_\vartheta(\theta_k=0|x_1^m)|\leqslant\Pi_{i=k+1}^{L_2-1}\exp\{-2\,\tau_0(x_i)\}$.

B.1. Proof of lemma 1

We consider the case such that $L_1>1$ and $L_2<m$; other cases are simpler and can be proved in a similar manner. The proof essentially exploits the conditional independence of $\{x_i\}$ given $\{\theta_i\}$. According to lemma 14, we have

$$\begin{aligned}E_0|P_\vartheta(\theta_k=0|x_{L_1}^{L_2})-P_\vartheta(\theta_k=0|x_1^m)|&\leqslant E_0\left[\prod_{i=L_1+1}^{k-1}\exp\{-2\,\tau_0(x_i)\}+\prod_{i=k+1}^{L_2-1}\exp\{-2\,\tau_0(x_i)\}\right]\\&=E_0\,E_0\left[\prod_{i=L_1+1}^{k-1}\exp\{-2\,\tau_0(x_i)\}+\prod_{i=k+1}^{L_2-1}\exp\{-2\,\tau_0(x_i)\}|\theta_1^m\right]\\&=E_0\left(\prod_{i=L_1+1}^{k-1}E_0[\exp\{-2\,\tau_0(x_i)\}|\theta_i]+\prod_{i=k+1}^{L_2-1}E_0[\exp\{-2\,\tau_0(x_i)\}|\theta_i]\right).\end{aligned}$$

The definition of $\tau_0(x)$ and assumption 3 imply that $P_{\vartheta_0}\{\tau_0(X_k)>0|\theta_k=i\}>0$ for all i. Let $\beta_0=\max_i(E_{\vartheta_0}[\exp\{-\tau_0(X_1)\}|\theta_1=i])$. Then $\beta_0<1$ and hence $E_0|P_\vartheta(\theta_k=0|x_{L_1}^{L_2})-P_\vartheta(\theta_k=0|x_1^m)|\leqslant C_0\beta_0^L$.

B.2. Proof of lemma 2

The joint PDF of $\{x_i\}_{-N}^{N}$ is continuous; then the PDF and hence CDF of $P(\theta_0=0|\{x_i\}_{-N}^{N})$ are continuous. According to the martingale convergence theorem, we have $P(\theta_0=0|\{x_i\}_{-N}^{N})\to T_0^\infty$ almost surely. Therefore the CDF of T_0^∞ is also continuous. To show the second part of the lemma, observe that, for $t\in(0,1)$,

$$\begin{aligned}P(0<T_0^\infty<t)&=P\left[0<\frac{P\{\theta_0=0|(x_i)_{-\infty}^{\infty}\}}{P\{\theta_0=1|(x_i)_{-\infty}^{\infty}\}}<\frac{t}{1-t}\right]\\&=P\left[0<\frac{\pi_0\,P\{(x_i)_{-\infty}^{\infty}|\theta_0=0\}}{\pi_1\,P\{(x_i)_{-\infty}^{\infty}|\theta_0=1\}}<\frac{t}{1-t}\right]\\&=P\left[0<\frac{\pi_0\,f_0(x_0)}{\pi_1\,f_1(x_0)}\,\frac{\sum\limits_{j_1j_2}P\{(x_i)_{-\infty}^{1}|\theta_{-1}=j_1\}\,a_{j_10}a_{0j_2}\,P\{(x_i)_1^{\infty}|\theta_1=j_2\}}{\sum\limits_{j_1j_2}P\{(x_i)_{-\infty}^{1}|\theta_{-1}=j_1\}\,a_{j_11}a_{1j_2}\,P\{(x_i)_1^{\infty}|\theta_1=j_2\}}<\frac{t}{1-t}\right].\end{aligned}$$

Condition 2 implies that

$$\varepsilon_0^2\leqslant\frac{\sum\limits_{j_1,j_2}P\{(x_i)_{-\infty}^{1}|\theta_{-1}=j_1\}\,a_{j_10}a_{0j_2}\,P\{(x_i)_1^{\infty}|\theta_1=j_2\}}{\sum\limits_{j_1,j_2}P\{(x_i)_{-\infty}^{1}|\theta_{-1}=j_1\}\,a_{j_11}a_{1j_2}\,P\{(x_i)_1^{\infty}|\theta_1=j_2\}}\leqslant\varepsilon_0^{-2},$$

Also observe that $\min_x\{f_0(x)/f_1(x)\}=0$; we conclude that there are some x^* and ε such that $P(0<T_0^\infty<b)$ is satisfied by $S_\varepsilon=\{\{x_i\}_{-\infty}^{\infty}:x_0\in(x^*-\varepsilon,x^*+\varepsilon)\}$. Note that $f(\mathbf{x})$ is continuous and positive over the sample space, so we have $P(0<T_0^\infty<b)\geqslant P(S_\varepsilon)>0$.

B.3. Proof of lemma 3

We consider only the case that there are some hypotheses that are not rejected. Note that, as the threshold is always greater than α, it would be sufficient to prove that $\Sigma_{k=1}^{m}I(T_k<\alpha)\to\infty$ almost surely. Take $L=m^\kappa$, where $0<\kappa<1$. For $L+1<k<m-L-1$, we have

$$|T_k-T_k^\infty|<\prod_{i=k-L+1}^{k-1}\exp\{-2\,\tau_0(x_i)\}+\prod_{i=k+1}^{k+L-1}\exp\{-2\,\tau_0(x_i)\}.$$

Denote the last quantity by $s_k(\{x_i\}_1^m)$; then $s_k(\{x\}_1^m)$ is ergodic. We apply the ergodic theorem to $s_k(\{x_i\}_1^m)$ to obtain $(1/m)\,\Sigma_{i=L+1}^{m-L-1}\,I(s_k>\alpha/2)-P(s_k>\alpha/2)\to^{\mathrm{p}}0$. We have shown in lemma 1 that $E(s_k)<C_0\beta_0^L$; hence $P(s_k>\alpha/2)\to0$ as $L=m^\kappa\to\infty$. Therefore,

$$\frac{1}{m}\sum_{i=1}^{m}I\Big(|T_k-T_k^\infty|>\frac{\alpha}{2}\Big)\leqslant\frac{2L}{m}+\frac{1}{m}\sum_{i=L+1}^{m-L-1}I\Big(s_k>\frac{\alpha}{2}\Big)\overset{\mathrm{p}}{\to}0.$$

Note that $I(T_k<\alpha)+I(|T_k-T_k^\infty|>\alpha/2)\geqslant I(T_k^\infty<\alpha/2)$. We have

$$\frac{1}{m}\sum_{k=1}^{m}I(T_k<\alpha)+\frac{1}{m}\sum_{k=1}^{m}I(|T_k-T_k^\infty|<\alpha)>\frac{1}{m}\sum_{k=1}^{m}I\Big(T_k^\infty<\frac{\alpha}{2}\Big)\to G^\infty\Big(\frac{\alpha}{2}\Big)\qquad\text{almost surely.}$$

Therefore it holds almost surely that $(1/m)\,\Sigma_{k=1}^{m}\,I(T_k<\alpha)\geqslant G^\infty(\alpha/2)$. In lemma 2 we have shown that $G^\infty(\alpha/2)>0$; hence the result follows. Similarly we can prove the second part of the lemma.

B.4. Proof of lemma 4

Let $0<\gamma<1$. Note that $\hat\lambda_{\mathrm{OR}}>\alpha$, so we have

$$\begin{aligned}\hat Q_{\mathrm{OR}}^\infty(\hat\lambda_{\mathrm{OR}}^\infty)&=\frac{1}{R^\infty}\sum_{i=1}^{m}T_i^\infty\,I(T_i^\infty\leqslant\gamma\alpha)+\frac{1}{R^\infty}\sum_{i=1}^{m}T_i^\infty\,I(\gamma\alpha<T_i^\infty<\hat\lambda_{\mathrm{OR}}^\infty)\\&\leqslant\gamma\alpha\frac{\sum\limits_{i=1}^{m}I(T_i^\infty\leqslant\gamma\alpha)}{\sum\limits_{i=1}^{m}I(T_i^\infty<\hat\lambda_{\mathrm{OR}}^\infty)}+\hat\lambda_{\mathrm{OR}}^\infty\frac{\sum\limits_{i=1}^{m}I(\gamma\alpha<T_i^\infty<\hat\lambda_{\mathrm{OR}}^\infty)}{\sum\limits_{i=1}^{m}I(T_i^\infty<\hat\lambda_{\mathrm{OR}}^\infty)}.\end{aligned}$$

In lemma 16 we show that $\hat{\lambda}_{\mathrm{OR}}^{\infty} \rightarrow^{\mathrm{p}} \lambda_{\mathrm{OR}}^{\infty}$; hence $P(T_i^{\infty} < \hat{\lambda}_{\mathrm{OR}}^{\infty}) \rightarrow P(T_i^{\infty} < \lambda_{\mathrm{OR}}^{\infty})$. The ergodic theorem implies that $(1/m)\,\Sigma_{i=1}^{m} I(T_i^{\infty} < \hat{\lambda}_{\mathrm{OR}}^{\infty}) \rightarrow G^{\infty}(\lambda_{\mathrm{OR}}^{\infty})$ almost surely. Therefore,

$$\hat{Q}_{\mathrm{OR}}^{\infty}(\hat{\lambda}_{\mathrm{OR}}^{\infty}) \leqslant \gamma\alpha \frac{G^{\infty}(\gamma\alpha)}{G^{\infty}(\lambda_{\mathrm{OR}}^{\infty})} + \lambda_{\mathrm{OR}}^{\infty} \frac{G^{\infty}(\lambda_{\mathrm{OR}}^{\infty}) - G^{\infty}(\gamma\alpha)}{G^{\infty}(\lambda_{\mathrm{OR}}^{\infty})} + o_p(1).$$

Recall that $\hat{Q}_{\mathrm{OR}}^{\infty}(\hat{\lambda}_{\mathrm{OR}}^{\infty}) = (1/R^{\infty})\,\Sigma_{i=1}^{R^{\infty}} T_{(i)}^{\infty} = \alpha + o_p(1)$; we have

$$\lambda_{\mathrm{OR}}^{\infty} \geqslant \sup_{\gamma \in (0,1)} \left\{ \alpha + \frac{(1-\gamma)\,G^{\infty}(\gamma\alpha)}{G^{\infty}(\lambda_{\mathrm{OR}}^{\infty}) - G^{\infty}(\gamma\alpha)} \right\}.$$

Note that the threshold $\hat{\lambda}_{\mathrm{OR}}^{\infty}$ is always greater than α, and $\hat{\lambda}_{\mathrm{OR}}^{\infty} \rightarrow^{\mathrm{p}} \lambda_{\mathrm{OR}}^{\infty}$, so we have $\lambda_{\mathrm{OR}}^{\infty} \geqslant \alpha$. From lemma 2 we have $G^{\infty}(\gamma\alpha) > 0$, and $G^{\infty}(\lambda_{\mathrm{OR}}^{\infty}) - G^{\infty}(\gamma\alpha) > 0$. The proof is complete by choosing $0 < \gamma < \alpha_*$.

B.5. Proof of lemma 5

We first present several lemmas. Lemma 15 shows that the estimated mFDRs that are yielded by $\delta_{\mathrm{OR}}^{\infty}$ and $\delta_{\mathrm{PI}}^{\infty}$ converge to $Q_{\mathrm{OR}}^{\infty}(\lambda)$ almost surely. Lemma 16 shows that the threshold that is yielded by $\delta_{\mathrm{OR}}^{\infty}$ and $\delta_{\mathrm{PI}}^{\infty}$ converges to $\lambda_{\mathrm{OR}}^{\infty}$ in probability.

Lemma 15. Let R_{λ}^{∞} and $\hat{R}_{\lambda}^{\infty}$ be the number of rejections that are yielded by $\boldsymbol{\delta}(\mathbf{T}^{\infty}, \lambda) = [I(T_i^{\infty} < \lambda) : i = 1, \ldots, m]$ and $\boldsymbol{\delta}(\hat{\mathbf{T}}^{\infty}, \lambda) = [I(\hat{T}_i^{\infty} < \lambda) : i = 1, \ldots, m]$ respectively. Define the corresponding estimated false discovery proportion $\hat{Q}_{\mathrm{OR}}^{\infty}(\lambda) = (1/R_{\lambda}^{\infty})\,\Sigma_{i=1}^{R_{\lambda}^{\infty}} T_{(i)}^{\infty}$ and $\hat{Q}_{\mathrm{PI}}^{\infty}(\lambda) = (1/\hat{R}_{\lambda}^{\infty})\,\Sigma_{i=1}^{\hat{R}_{\lambda}^{\infty}} \hat{T}_{(i)}^{\infty}$. If assumptions 1–5 hold, then $\hat{Q}_{\mathrm{OR}}^{\infty}(\lambda) \rightarrow Q_{\mathrm{OR}}^{\infty}(\lambda)$ almost surely and $\hat{Q}_{\mathrm{PI}}^{\infty}(\lambda) \rightarrow Q_{\mathrm{OR}}^{\infty}(\lambda)$ almost surely.

Proof. We show the second part of the theorem; the first part follows simpler arguments. If follows from the continuous mapping theorem that $\hat{T}_1^{\infty} \rightarrow^{\mathrm{p}} T_1^{\infty}$. Therefore, $E\{I(\hat{T}_1^{\infty} < \lambda)\} \rightarrow E\{I(T_1^{\infty} < \lambda)\}$. Observe that, for given $\varepsilon > 0$, $I(\hat{T}_1^{\infty} < \lambda) = I(\hat{T}_1^{\infty} < \lambda)$ holds on the event $A = \{|\hat{T}_1^{\infty} - T_1^{\infty}| \leqslant \varepsilon\}$ unless $\lambda - \varepsilon \leqslant T_1^{\infty} \leqslant \lambda + \varepsilon$. We have that $E\{\hat{T}_1^{\infty} I(\hat{T}_1^{\infty} < \lambda) - T_1^{\infty} I(T_1^{\infty} < \lambda)\} \leqslant P(|\hat{T}_1^{\infty} - T_1^{\infty}| \geqslant \varepsilon) + E(\hat{T}_1^{\infty} - T_1^{\infty}) + P(\lambda - \varepsilon \leqslant T_1^{\infty} \leqslant \lambda + \varepsilon)$. The first term goes to 0 since $\hat{T}_1^{\infty} \rightarrow^{\mathrm{p}} T_1^{\infty}$. The second term goes to 0 by lemma 2.2 of van der Vaart (1998). The third term goes to 0 by the continuity of $G^{\infty}(t)$. Therefore, we have $E\{\hat{T}_1^{\infty} I(\hat{T}_1^{\infty} < \lambda)\} \rightarrow E\{T_1^{\infty} I(T_1^{\infty} < \lambda)\}$. The ergodic theorem implies that $(1/m)\{\Sigma_{i=1}^{m} I(\hat{T}_i^{\infty} < \lambda)\} - E\{I(\hat{T}_1^{\infty} < \lambda)\} \rightarrow 0$ almost surely; hence $(1/m)\{\Sigma_{i=1}^{m} I(\hat{T}_i^{\infty} < \lambda)\} \rightarrow G^{\infty}(\lambda)$ almost surely. Similarly we have $(1/m)\{\Sigma_{i=1}^{m} \hat{T}_i^{\infty} I(\hat{T}_i^{\infty} < \lambda)\} \rightarrow E\{T_1^{\infty} I(T_1^{\infty} < \lambda)\}$ almost surely. Note that $\hat{Q}_{\mathrm{PI}}^{\infty}(\lambda)$ can be written as

$$\hat{Q}_{\mathrm{PI}}^{\infty}(\lambda) = \left\{ \sum_{i=1}^{m} \hat{T}_i^{\infty} I(\hat{T}_i^{\infty} < \lambda) \right\} \Big/ \sum_{i=1}^{m} I(\hat{T}_i^{\infty} < \lambda)$$

and that

$$E\{T_i^{\infty} I(T_i^{\infty} < \lambda)\} = E[I(T_i^{\infty} < \lambda)\, E\{I(\theta_i = 0 | \{x_i\}_{-\infty}^{\infty})\}] = P(T_i^{\infty} < \lambda, \theta_i = 0) = \pi_0\, G_0^{\infty}(\lambda);$$

the result follows from the definition of Q_{OR}^{∞}.

Lemma 16. Denote by $\hat{\lambda}_{\mathrm{OR}}^{\infty}$ and $\hat{\lambda}_{\mathrm{PI}}^{\infty}$ the thresholds that are yielded by $\delta_{\mathrm{OR}}^{\infty}$ and $\delta_{\mathrm{PI}}^{\infty}$ respectively. Assume that conditions 1–4 hold; then $\hat{\lambda}_{\mathrm{OR}}^{\infty} \rightarrow^{\mathrm{p}} \lambda_{\mathrm{OR}}^{\infty}$ and $\hat{\lambda}_{\mathrm{PI}}^{\infty} \rightarrow^{\mathrm{p}} \lambda_{\mathrm{OR}}^{\infty}$.

Proof. Let

$$\hat{Q}_{\mathrm{OR}}^{\infty}(t) = \left\{ \sum_{i=1}^{m} I(T_i^{\infty} \leqslant t) T_i^{\infty} \right\} \Big/ \sum_{i=1}^{m} I(T_i^{\infty} \leqslant t).$$

Note that $\hat{Q}_{\mathrm{OR}}^{\infty}(t)$ is a step function with jump at $T_{(i)}$; for $T_{(k)} < t < T_{(k+1)}$, we construct an envelope for $\hat{Q}_{\mathrm{OR}}^{\infty}(t)$ by using two continuous functions $\underline{\hat{Q}}_{\mathrm{OR}}^{\infty}(t)$ and $\bar{\hat{Q}}_{\mathrm{OR}}^{\infty}(t)$:

$$\underline{\hat{Q}}_{\mathrm{OR}}^{\infty}(t) = \frac{T_{(k+1)}^{\infty} - t}{T_{(k+1)}^{\infty} - T_{(k)}^{\infty}} \hat{Q}_{\mathrm{OR}}^{\infty}(T_{(k-1)}^{\infty}) + \frac{t - T_{(k)}^{\infty}}{T_{(k+1)}^{\infty} - T_{(k)}^{\infty}} \hat{Q}_{\mathrm{OR}}^{\infty}(T_{(k)}^{\infty});$$

$$\bar{\hat{Q}}_{\mathrm{OR}}^{\infty}(t) = \frac{T_{(k+1)}^{\infty} - t}{T_{(k+1)}^{\infty} - T_{(k)}^{\infty}} \hat{Q}_{\mathrm{OR}}^{\infty}(T_{(k)}^{\infty}) + \frac{t - T_{(k)}^{\infty}}{T_{(k+1)}^{\infty} - T_{(k)}^{\infty}} \hat{Q}_{\mathrm{OR}}^{\infty}(T_{(k+1)}^{\infty}).$$

Note that

$$\hat{Q}^{\infty}_{\mathrm{OR}}(T_{(k+1)}) - \hat{Q}^{\infty}_{\mathrm{OR}}(T_{(k)}) = \left(kT_{(k+1)} - \sum_{i=1}^{k} T_{(i)} \right) \Big/ k(k+1) > 0,$$

so we have $\bar{\hat{Q}}^{\infty}_{\mathrm{OR}}(t) \geqslant \hat{Q}^{\infty}_{\mathrm{OR}}(t) \geqslant \bar{\hat{Q}}^{\infty}_{\mathrm{OR}}(t)$, and both $\bar{\hat{Q}}^{\infty}_{\mathrm{OR}}(t)$ and $\underline{\hat{Q}}^{\infty}_{\mathrm{OR}}(t)$ are strictly increasing in t. Also note that $|\bar{\hat{Q}}^{\infty}_{\mathrm{OR}}(t) - \underline{\hat{Q}}^{\infty}_{\mathrm{OR}}(t)| \leqslant 1/R^{\infty}(t)$, where $R^{\infty}(t)$ is the number of rejections that are yielded by $\delta(T^{\infty}, t) = \{I(T_i^{\infty} < t), i = 1, \ldots, m\}$. It is easy to see that $R^{\infty}(t) \rightarrow \infty$ in probability. Therefore $\bar{\hat{Q}}^{\infty}_{\mathrm{OR}}(t) - \underline{\hat{Q}}^{\infty}_{\mathrm{OR}}(t) \rightarrow^{\mathrm{p}} 0$. Recall that $\hat{Q}^{\infty}_{\mathrm{OR}} \rightarrow^{\mathrm{p}} Q^{\infty}_{\mathrm{OR}}$; we have $\underline{\hat{Q}}^{\infty}_{\mathrm{OR}}(t) \rightarrow^{\mathrm{p}} Q^{\infty}_{\mathrm{OR}}$, and $\bar{\hat{Q}}^{\infty}_{\mathrm{OR}}(t) \rightarrow^{\mathrm{p}} Q^{\infty}_{\mathrm{OR}}$. Let $\underline{\hat{\lambda}}^{\infty}_{\mathrm{OR}} = \sup\{t \in (0, 1) : \underline{\hat{Q}}^{\infty}_{\mathrm{OR}}(t) \leqslant \alpha\}$ and $\bar{\hat{\lambda}}^{\infty}_{\mathrm{OR}} = \sup\{t \in (0, 1) : \bar{\hat{Q}}^{\infty}_{\mathrm{OR}}(t) \leqslant \alpha\}$; then $\bar{\hat{\lambda}}^{\infty}_{\mathrm{OR}} \leqslant \hat{\lambda}^{\infty}_{\mathrm{OR}} \leqslant \underline{\hat{\lambda}}^{\infty}_{\mathrm{OR}}$.

We claim that $\underline{\hat{\lambda}}^{\infty}_{\mathrm{OR}} \rightarrow^{\mathrm{p}} \lambda^{\infty}_{\mathrm{OR}}$. If not, there are ε_0 and η_0 such that, for any $M > 0$, $P(|\underline{\hat{\lambda}}^{\infty}_{\mathrm{OR}} - \lambda^{\infty}_{\mathrm{OR}}| > \varepsilon_0) \geqslant 4\eta_0$ holds for some $m \geqslant M$, $m \in \mathbb{Z}^{+}$. Suppose that

$$P(K_m^1) = P(\underline{\hat{\lambda}}^{\infty}_{\mathrm{OR}} > \lambda^{\infty}_{\mathrm{OR}} + \varepsilon_0) \geqslant 2\eta_0. \tag{13}$$

Let $2\delta_0 = Q^{\infty}_{\mathrm{OR}}(\lambda^{\infty}_{\mathrm{OR}} + \varepsilon_0) - \alpha > 0$. Recall that $\underline{\hat{Q}}^{\infty}_{\mathrm{OR}}(t) \rightarrow^{\mathrm{p}} Q^{\infty}_{\mathrm{OR}}(t)$; there is an $M \in \mathbb{Z}^{+}$ such that

$$P(K_m^2) = P\{|\underline{\hat{Q}}^{\infty}_{\mathrm{OR}}(\lambda^{\infty}_{\mathrm{OR}} + \varepsilon_0) - Q^{\infty}_{\mathrm{OR}}(\lambda^{\infty}_{\mathrm{OR}} + \varepsilon_0)| < \delta_0\} \geqslant 1 - \eta_0 \tag{14}$$

holds for all $m \geqslant M$. Let $K_m = K_m^1 \cap K_m^2$; then equations (13) and (14) imply that there is an $m \in \mathbb{Z}^{+}$ such that $P(K_m) \geqslant \eta_0$. However, note that $\underline{\hat{Q}}^{\infty}_{\mathrm{OR}}(t)$ is strictly increasing in t; on K_m we must have $\alpha = \underline{\hat{Q}}^{\infty}_{\mathrm{OR}}(\underline{\hat{\lambda}}^{\infty}_{\mathrm{OR}}) > \underline{\hat{Q}}^{\infty}_{\mathrm{OR}}(\lambda^{\infty}_{\mathrm{OR}} + \varepsilon_0) > Q^{\infty}_{\mathrm{OR}}(\lambda^{\infty}_{\mathrm{OR}} + \varepsilon_0) - \delta_0 = \alpha + \delta_0$. Hence K_m cannot have positive measure. This is a contradiction. Therefore, we must have $\underline{\hat{\lambda}}^{\infty}_{\mathrm{OR}} \rightarrow^{\mathrm{p}} \lambda^{\infty}_{\mathrm{OR}}$. Similarly, we can show that $\bar{\hat{\lambda}}^{\infty}_{\mathrm{OR}} \rightarrow^{\mathrm{p}} \lambda^{\infty}_{\mathrm{OR}}$. Note that $\bar{\hat{\lambda}}^{\infty}_{\mathrm{OR}} \leqslant \hat{\lambda}^{\infty}_{\mathrm{OR}} \leqslant \underline{\hat{\lambda}}^{\infty}_{\mathrm{OR}}$, so we have $\hat{\lambda}^{\infty}_{\mathrm{OR}} \rightarrow^{\mathrm{p}} \lambda^{\infty}_{\mathrm{OR}}$. Similarly we can prove that $\hat{\lambda}^{\infty}_{\mathrm{PI}} \rightarrow^{\mathrm{p}} \lambda^{\infty}_{\mathrm{OR}}$.

B.6. Proof of lemma 5 (continued)

Note that $(1/m)\hat{R}^{\infty} = (1/m)\,\Sigma_{k=1}^{m} I(\hat{T}_k^{\infty} < \hat{\lambda}^{\infty}_{\mathrm{PI}})$; the ergodic theorem implies that $(1/m)\hat{R}^{\infty} - E(\hat{T}_1^{\infty} \leqslant \hat{\lambda}^{\infty}_{\mathrm{PI}}) \rightarrow^{\mathrm{p}} 0$. Also, observe that $\hat{T}_1^{\infty} \rightarrow^{\mathrm{p}} T_1^{\infty}$ and $\hat{\lambda}^{\infty}_{\mathrm{OR}} \rightarrow^{\mathrm{p}} \lambda^{\infty}_{\mathrm{OR}}$; we have $E(\hat{T}_1^{\infty} \leqslant \hat{\lambda}^{\infty}_{\mathrm{PI}}) \rightarrow E(T_1^{\infty} < \lambda^{\infty}_{\mathrm{OR}}) = G^{\infty}_{\mathrm{OR}}(\lambda^{\infty}_{\mathrm{OR}})$. Therefore, $(1/m)\hat{R}^{\infty} \rightarrow^{\mathrm{p}} G^{\infty}_{\mathrm{OR}}(\lambda^{\infty}_{\mathrm{OR}})$. It is easy to show that $(1/m)R^{\infty} \rightarrow^{\mathrm{p}} G^{\infty}_{\mathrm{OR}}(\lambda^{\infty}_{\mathrm{OR}})$. Therefore, $\hat{R}^{\infty}/R^{\infty} \rightarrow^{\mathrm{p}} 1$. The second part of the proof can be shown similarly, by noting that $(1/m)\hat{V}^{\infty} - E(\hat{T}_1^{\infty} < \hat{\lambda}^{\infty}_{\mathrm{PI}}, \theta_1 = 0) \rightarrow 0$ almost surely, $E(\hat{T}_1^{\infty} < \lambda^{\infty}_{\mathrm{OR}}, \theta_1 = 0) \rightarrow G_0^{\infty}(\lambda^{\infty}_{\mathrm{OR}})$ and $(1/m)V^{\infty} \rightarrow^{\mathrm{p}} G_0^{\infty}(\lambda^{\infty}_{\mathrm{OR}})$.

B.7. Proof of lemma 6

We take $L = k^{\kappa}$, where $0 < \kappa < 1$. Let $\tilde{S}_k = \{i : i \in S_k \text{ and } L + 1 < i < m - L - 1\}$, $\tilde{S}_k^{\infty} = \{i : i \in S_k^{\infty} \text{ and } L + 1 < i < m - L - 1\}$ and

$$s_k = \prod_{i=k-L+1}^{k-1} \exp\{-2\,\tau_0(x_i)\} + \prod_{i=k+1}^{k+L-1} \exp\{-2\,\tau_0(x_i)\}.$$

The definitions of S_k and S_k^{∞} imply that

$$\sum_{i \in S_k} T_i - \sum_{i \in S_k} T_i^{\infty} \leqslant \sum_{i \in S_k} T_i - \sum_{i \in S_k^{\infty}} T_i^{\infty} \leqslant \sum_{i \in S_k^{\infty}} T_i - \sum_{i \in S_k^{\infty}} T_i^{\infty}.$$

Therefore,

$$\begin{aligned} E\left|\frac{1}{k}\sum_{i \in S_k} T_i - \frac{1}{k}\sum_{i \in S_k^{\infty}} T_{(i)}^{\infty}\right| &\leqslant E\left|\frac{1}{k}\sum_{i \in S_k} T_i - \frac{1}{k}\sum_{i \in S_k} T_i^{\infty}\right| + E\left|\frac{1}{k}\sum_{i \in S_k^{\infty}} T_i - \frac{1}{k}\sum_{i \in S_k^{\infty}} T_i^{\infty}\right| \\ &\leqslant \frac{2L}{k} + E\left|\frac{1}{k}\sum_{i \in \tilde{S}_k} s_i\right| + E\left|\frac{1}{k}\sum_{i \in \tilde{S}_k^{\infty}} s_i\right| \\ &\leqslant 2\left(\frac{L}{k} + C_0 \beta_0^L\right) \rightarrow 0, \end{aligned}$$

and the first part of the lemma is shown. Define $A_{\delta} = \{\vartheta : |\vartheta - \vartheta_0| \leqslant \delta\}$. Denote by A_{δ}^{c} the complement of A_{δ}. To show the second part of the lemma, we note that $|\hat{T}_i - \hat{T}_i^{\infty}| < s_k$ holds for all $\hat{\vartheta} \in A_{\delta}$, and the consistency of $\hat{\vartheta}$ implies that $P(\hat{\vartheta} \in A_{\delta}^{\mathrm{c}}) \rightarrow 0$. We follow a similar argument to that for the first part to obtain

$$E\left|\frac{1}{k}\sum_{i\in\hat{S}_k}\hat{T}_i-\frac{1}{k}\sum_{i\in\hat{S}_k^{\infty}}\hat{T}_{(i)}^{\infty}\right|\leqslant E\left\{\left|\frac{1}{k}\sum_{i\in\hat{S}_k}\hat{T}_i-\frac{1}{k}\sum_{i\in\hat{S}_k^{\infty}}\hat{T}_{(i)}^{\infty}\right|I(\hat{\vartheta}\in A_\delta)\right\}+P(\hat{\vartheta}\in A_\delta^{\mathrm{c}})$$
$$\leqslant 2\left(\frac{L}{k}+C_0\beta_0^L\right)+o(1)\rightarrow 0,$$

and the second part of the lemma is shown.

B.8. Proof of lemma 7

It follows from the ergodic theorem that $(1/m)R^{\infty}-G^{\infty}(\lambda_{\mathrm{OR}}^{\infty})\rightarrow^{\mathrm{p}}0$. The definition of α_* and the assumption that $\lambda_{\mathrm{OR}}^{\infty}\geqslant\alpha_*$ imply that $G^{\infty}(\lambda_{\mathrm{OR}}^{\infty})=1$. Therefore, $(1/m)R^{\infty}\rightarrow^{0}1$. Now we assume that $R/m\rightarrow^{\mathrm{p}}1$ is not true. Then there are an $\varepsilon_0>0$ and a δ_0 such that, for any $M>0$, $P(R/m\leqslant 1-\varepsilon)\geqslant\delta_0$ holds for some $m\geqslant M$. It follows from lemma 6 that $(1/m)\,\Sigma_{i=1}^{m}T_i=(1/m)\,\Sigma_{i=1}^{m}T_i^{\infty}+o_p(1)\leqslant\alpha+o_p(1)$. Let S_1 be the rejection set that is yielded by δ_{OR} and S_2 be its complement. Then by lemma 4 we have

$$\frac{1}{m}\sum_{i=1}^{m}T_i=\frac{1}{m}\sum_{i\in S_1}T_i+\frac{1}{m}\sum_{i\in S_2}T_i\geqslant\frac{|S_1|}{m}\alpha+\frac{|S_2|}{m}(\alpha+\nu_0)+o_p(1).$$

Note that $S_2/m\geqslant\varepsilon_0$ with positive probability, so, for any $M>0$, $(1/m)\,\Sigma_{i=1}^{m}T_i\geqslant\alpha+\varsigma_0$ with some positive probability for some $m\geqslant M$. This is a contradiction to the fact that $(1/m)\,\Sigma_{i=1}^{m}T_i=\alpha+o_p(1)$. Therefore we must have $R/m\rightarrow^{\mathrm{p}}1$ and the first part of the lemma is shown. The second part of the lemma can be easily shown by noting that the difference between the rejections sets of δ_{OR} and $\delta_{\mathrm{OR}}^{\infty}$ is of a smaller order of m, whereas the true number of rejections by $\delta_{\mathrm{OR}}^{\infty}$ is proportional to m.

References

Atkinson, A. C. (1973) Testing transformations to normality. *J. R. Statist. Soc.* B, **35**, 473–479.

Baum, L. and Petrie, T. (1966) Statistical inference for probabilistic functions of finite Markov Chains. *Ann. Math. Statist.*, **37**, 1554–1563.

Baum, L., Petrie, T., Soules, G. and Weiss, N. (1970) A maximization technique occuring in the statistical analysis of probabilistic functions of Markov Chains. *Ann. Math. Statist.*, **41**, 164–171.

Benjamini, Y. and Heller, R. (2007) False discovery rate for spatial signals. *J. Am. Statist. Ass.*, **102**, 1272–1281.

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

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

Benjamini, Y. and Yekutieli, D. (2001) The control of false discovery rate in multiple testing under dependence. *Ann. Statist.*, **29**, 1165–1188.

Bickel, P. and Ritov, Y. (1996) Inference in Hidden Markov Models I: local asymptotic normality in the stationary case. *Bernoulli*, **2**, 199–228.

Bickel, P., Ritov, Y. and Rydén, T. (1998) Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.*, **26**, 1614–1635.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211–252.

Castro, M. and Singer, B. (2006) A new approach to account for multiple and dependent tests in local statistics of spatial association: controlling the false discovery rate. *Geogr. Anal.*, **38**, 180–208.

Churchill, G. (1992) Hidden Markov chains and the analysis of genome structure. *Comput. Chem.*, **16**, 107–115.

Ciuperca, G., Ridolfi, A. and Idier, J. (2003) Penalized maximum likelihood estimator for normal mixtures. *Scand. J. Statist.*, **30**, 45–59.

Copas, J. (1974) On symmetric compound decision rules for dichotomies. *Ann. Statist.*, **2**, 199–204.

Durrett, R. (2005) *Probability: Theory and Examples*, 3rd edn. Belmont: Duxbury.

Efron, B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Statist. Ass.*, **99**, 96–104.

Efron, B. (2007) Correlation and large-scale simultaneous significance testing. *J. Am. Statist. Ass.*, **102**, 93–103.

Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Ass.*, **96**, 1151–1160.

Ephraim, Y. and Merhav, N. (2002) Hidden Markov processes. *IEEE Trans. Inform. Theory*, **48**, 1518–1569.

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

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

Genovese, C., Roeder, K. and Wasserman, L. (2006) False discovery control with *p* value weighting. *Biometrika*, **93**, 509–524.

Genovese, C. and Wasserman, L. (2002) Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc.* B, **64**, 499–517.

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

Hathaway, R. (1985) A constrained formulation of maximum-likelihood estimation for Normal mixture distributions. *Ann. Statist.*, **13**, 795–800.

Kiefer, J. (1993) Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Trans. Inform. Theory*, **39**, 893–902.

Kiefer, J. and Wolfowitz, J. (1956) Consistency of the maximum-likelihood estimator in the presence of infinitely many identical parameters. *Ann. Math. Statist.*, **27**, 888–906.

Krogh, A., Brown, M., Mian, I., Sjölander, K. and Haussler, D. (1994) Hidden Markov models in computational biology applications to protein modeling. *J. Molec. Biol.*, **235**, 1501–1531.

Leroux, B. (1992) Maximum-likelihood estimation for hidden Markov models. *Stochast. Processes Appl.*, **40**, 127–143.

Liu, C. and Narayan, P. (1994) Order estimation and sequential universal data compression of a hidden Markov source by the method of mixtures. *IEEE Trans. Inform. Theory*, **40**, 1167–1180.

Magder, L. and Zeger, S. (1996) A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *J. Am. Statist. Ass.*, **91**, 1141–1151.

Newton, M., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.

Owen, A. B. (2005) Variance of the number of false discoveries. *J. R. Statist. Soc.* B, **67**, 411–426.

Qiu, X., Klebanov, L. and Yakovlev, A. (2005) Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Statist. Appl. Genet. Molec. Biol.*, **4**, article 34.

Rabiner, L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Robbins, H. (1951) Asymptotically subminimax solutions of compound statistical decision problems. In *Proc. 2nd Berkeley Symp. Mathematical Statistics and Probability*. Berkeley: University of California Press.

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

Storey, J. D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc.* B, **64**, 479–498.

Storey, J. D. (2007) The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Statist. Soc.* B, **69**, 347–368.

Strat, Y. and Carrat, F. (1999) Monitoring epidemiologic surveillance data using hidden Markov models. *Statist. Med.*, **18**, 3463–3478.

Sun, W. and Cai, T. (2007) Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Statist. Ass.*, **102**, 901–912.

van der Vaart, A. (1998) *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Wei, Z. and Li, H. (2008) A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data. *Ann. Appl. Statist.*, **2**, 408–429.

Weisberg, S. (1985) *Applied Linear Regression*. New York: Wiley.

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

Yekutieli, D. and Benjamini, Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Statist. Planng Inf.*, **82**, 171–196.