Bayesian Methods for Two-Sample Comparison

by

Jacopo Soriano

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Li Ma, Advisor

_____
Jim Berger

_____
Mike West

_____
Cliburn Chan

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2015

Abstract

Bayesian Methods for Two-Sample Comparison

by

Jacopo Soriano

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Li Ma, Advisor

_____
Jim Berger

_____
Mike West

_____
Cliburn Chan

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2015

Copyright © 2015 by Jacopo Soriano
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

Two-sample comparison is a fundamental problem in statistics. Given two samples of data, the interest lies in understanding whether the two samples were generated by the same distribution or not. Traditional two-sample comparison methods are not suitable for modern data where the underlying distributions are multivariate and highly multi-modal, and the differences across the distributions are often locally concentrated. The focus of this thesis is to develop novel statistical methodology for two-sample comparison which is effective in such scenarios. Tools from the non-parametric Bayesian literature are used to flexibly describe the distributions. Additionally, the two-sample comparison problem is decomposed into a collection of local tests on individual parameters describing the distributions. This strategy not only yields high statistical power, but also allows one to identify the nature of the distributional difference. In many real-world applications, detecting the nature of the difference is as important as the existence of the difference itself. Generalizations to multi-sample comparison and more complex statistical problems, such as multi-way analysis of variance, are also discussed.

To my family

Contents

List of Figures

List of Abbreviations and Symbols

AMSE	Average mean square error
ANOVA	Analysis of variance
ARM-tree	Accept reject Markov tree
DWT	Discrete wavelet transform
DP	Dirichlet process
DPT	Dependent Pólya tree
HMT	Hidden Markov tree
MCMC	Markov chain Monte Carlo
NIG-HMT	Normal inverse gamma hidden Markov tree
PT	Pólya tree
RSNR	Root signal ratio to noise

Acknowledgements

I would like to express my deepest appreciation to my advisor, Li Ma, for his guidance, enthusiasm and optimism over the past few years. Working with Li has been a great pleasure for me. I am greatful to Mike West for his dedication to research and students. His energy and determination set a great example. I am appreciative of Jim Berger for tolerating all my questions and for many thoughtful discussions. I would like to thank Cliburn Chan for his patience and support when discussing the scientific details of flow cytometry.

In addition, I am thankful to everyone in the Department of Statisical Science for creating such a nurturing and enjoyable environment. In particular, I wish to thank all those people who would gather around noon in front of Old Chem to go for lunch.

Finally, I would like to acknowledge that my research was partially supported by NSF grant DMS-1309057.

1

Introduction

Two-sample comparison is one of the milestones of every introductory statistics class. Given two samples of data, the purpose of a two-sample problem is to understand whether the two processes generating the data are identical or not.

Two-sample comparison is largely used in epidemiology where cases are compared to controls, but it is a fundamental problem of experimental sciences at large. Two samples of data are obtained under different experimental conditions, and the experimenter is interested in learning if the experimental conditions are associated with the outcome. In the context of epidemiology, for instance, healthy patients are compared to patients with a specific disease, or drug-takers are compared with placebo-takers.

With the advancement of technology, the quantity and complexity of data that one can collect has exploded in recent years. Along with the rich information such massive amounts of data contain, there comes a seemingly daunting challenge of recovering the useful information from a sea of complex noisy data. Classical statistical methods for two-sample comparison developed in a world of "small data" are often inadequate for analyzing modern data-sets. Most classical methods are constructed

for univariate data or rely on strong parametric assumptions. The Student's t-test (Student, 1908), for example, compares the mean of the two underlying distributions under the assumption that the data are Gaussian distributed. In contrast, modern data-sets are often multivariate, and cannot be adequately described by simple parametric families.

Since Ferguson (1973)'s seminal work on the Dirichlet process, there has been increasing attention on Bayesian nonparametric models, which provide a principled and flexible framework to avoid restrictive parametric assumptions. Bayesian nonparametric methods have been used in a variety of statistical problems (density estimation, regression, hierarchical modeling, model validation, clustering, etc.) and applications (biostatistics, text modeling, image segmentation, etc.). I refer the reader to Müller and Quintana (2004), Dunson (2010) and Teh and Jordan (2010) for reviews of recent developments and applications of nonparametric Bayesian methods. Despite the wide success of nonparametric Bayes, little work has been done in the context hypothesis testing.

The research presented in this thesis focuses on borrowing concepts and tools from the nonparametric Bayesian literature to develop new two-sample comparison methodology. In particular, I use Pólya trees and Dirichlet process mixtures for multivariate data and wavelet decomposition for univariate functional data. These methods adapt to the features of the data, yielding high statistical power. Furthermore, they allow one to pin down the nature of the differences across the underlying distributions. Detecting the type of distributional differences can be very important in real-world applications. In flow cytometry, for instance, differences in the distribution of blood cells across samples is informative on the immune response of the patient and thus on the evolution of the disease. However, in most cases only a small fraction of the blood cells are involved in the difference, and so there is a need for methods with high statistical power when the differences are locally concentrated.

This thesis is organized as follows. The remainder of this chapter introduces two-sample comparison in a Bayesian setting and the driving philosophy behind the different methods presented in this thesis. Pólya trees, the Dirichlet process, Dirichlet process mixtures of normals, and wavelet decomposition are also introduced. Chapter 2 develops two-sample comparison in the context of Pólya tree priors. Chapter 3 focuses on multi-sample decomposition of variability across distributions in the context of Dirichlet process mixtures of normals. Chapter 4 presents multi-sample comparison for functional data using wavelet decomposition. Finally, Chapter 5 summarizes the main concepts of this thesis and outlines some areas for future research.

1.1 Bayesian Two-Sample Comparison

Assume the data are generated from the following model:

$$X_{1,1}, X_{1,2}, \ldots | \theta_1 \overset{\text{iid}}{\sim} P_{\theta_1}$$

$$X_{2,1}, X_{2,2}, \ldots | \theta_2 \overset{\text{iid}}{\sim} P_{\theta_2},$$

where P_θ is the distribution of the data for $\theta_k = \theta$ and $k = 1, 2$, and the parameters θ_1 and θ_2 have values in some parameter space Θ. The model is called parametric, if Θ has finite dimension; while the model is called nonparametric, if the parameter space has infinite dimension (Bernardo and Smith, 2009).

Two-sample comparison consists of measuring the evidence for the following competing hypotheses:

$$H_0 : P_{\theta_1} = P_{\theta_2} \quad \text{vs} \quad H_1 : P_{\theta_1} \neq P_{\theta_2}. \tag{1.1}$$

Under a Bayesian formulation one places a prior distribution Π on the two hypotheses and the parameters θ_1 and θ_2 reflecting the degree of uncertainty about them. Since there are two competing hypotheses, the prior Π is a mixture of two distributions Π_0 and Π_1:

$$\Pi \overset{d}{=} \Pi_0 \Pr(H_0) + \Pi_1(1 - \Pr(H_0)),$$

where $\Pr(H_0)$ is the marginal prior probability of H_0, and Π_0 and Π_1 are the prior distributions for the pair of parameters θ_1 and θ_2 under the hypotheses H_0 and H_1, respectively. After observing two samples of data $\boldsymbol{x}_1 = (X_{1,1} = x_{1,1}, \ldots, X_{1,n_1} = x_{1,n_1})$ and $\boldsymbol{x}_2 = (X_{2,1} = x_{2,1}, \ldots, X_{2,n_2} = x_{2,n_2})$, one updates the marginal prior probability $\Pr(H_0)$ to the marginal posterior probability through Bayes' theorem:

$$\Pr(H_0|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|H_0)\Pr(H_0)}{p(\boldsymbol{x}|H_0)\Pr(H_0) + p(\boldsymbol{x}|H_1)(1-\Pr(H_0))},$$

where $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, and $p(\boldsymbol{x}|H_j)$ denotes the marginal likelihood of the data under the hypothesis H_j:

$$p(\boldsymbol{x}|H_j) = \int \prod_{k=1}^{2} \prod_{i=1}^{n_k} p_{\theta_k}(x_{k,i}) \Pi_j(d\theta_1, d\theta_2), \quad j = 0, 1.$$

The marginal posterior probability can be very sensitive to the choice of $\Pr(H_0)$, so often the Bayes Factor is used instead:

$$BF = \frac{p(\boldsymbol{x}|H_0)}{p(\boldsymbol{x}|H_1)}.$$

The most common approach is to elicit a single prior distribution Q on Θ and use it as a building-block for both Π_0 and Π_1. Under the prior Π_0, θ_1 and θ_2 are replaced by a common parameter θ_0 distributed according to Q. Under the prior Π_1, instead, the two parameters θ_1 and θ_2 are independent and identically distributed according to Q. This implies that under H_0 the data are generated as follows:

$$\begin{aligned} X_{1,1}, X_{1,2}, \ldots, X_{2,1}, X_{2,2}, \ldots | \theta_0 &\overset{\text{iid}}{\sim} P_{\theta_0} \\ \theta_0 &\sim Q, \end{aligned}$$

and the marginal likelihood of the data is equal to:

$$p(\boldsymbol{x}|H_0) = \int \prod_{k=1}^{2} \prod_{i=1}^{n_k} p_{\theta}(x_{k,i}) Q(d\theta).$$

Under H_1, one instead has:

$$X_{1,1}, X_{1,2}, \ldots | \theta_1 \overset{\text{iid}}{\sim} P_{\theta_1}$$

$$X_{2,1}, X_{2,2}, \ldots | \theta_2 \overset{\text{iid}}{\sim} P_{\theta_2}$$

$$\theta_k \overset{\text{iid}}{\sim} Q, \quad k = 1, 2,$$

and the marginal likelihood of the data is equal to:

$$p(\boldsymbol{x}|H_1) = \prod_{k=1}^{2} \int \prod_{i=1}^{n_k} p_{\theta_k}(x_{k,i}) Q(d\theta_k).$$

This approach is attractive because of its simplicity. The prior Π depends only on two quantities: the marginal prior null $\Pr(H_0)$ and the distribution Q. However, assuming that the two unknown distributions are independent under the alternative can be unrealistic and overly restrictive. Under this assumption there is no borrowing of information across the two samples, while in most applications one might expect some degree of similarity between the two distributions. When this similarity exists, the distribution under the alternative hypothesis is overly parametrized and the test favors the null hypothesis, thus reducing the power of the hypothesis test to detect differences across the two distributions.

In many cases the prior Q is decomposable into multiple mutually independent components:

$$Q(d\theta) = \prod_{j} Q_j(d\theta_j), \quad j = 1, 2, \ldots$$

$$\theta_k = (\theta_{k,1}, \theta_{k,2}, \ldots), \quad k = 1, 2,$$

where Q_j is the prior distribution on $\theta_{k,j}$ for $k = 1, 2$ and $j = 1, 2, \ldots$. Then, one can also decompose the hypothesis testing into the following multiple local hypothesis tests:

$$H_{0,j} : \theta_{1,j} = \theta_{2,1} \quad \text{vs} \quad H_{1,j} : \theta_{1,j} \neq \theta_{2,1} \quad j = 1, 2, \ldots \tag{1.2}$$

To reflect the uncertainty between the two local hypotheses, a natural prior $\tilde{\Pi}_j$ on the pair of parameters $(\theta_{1,j}, \theta_{2,j})$ is the following:

$$\tilde{\Pi}_j \stackrel{d}{=} \Pi_{0,j} \Pr(H_{0,j}) + \Pi_{1,j}(1 - \Pr(H_{0,j})),$$

where $\Pr(H_{0,j})$ is the marginal prior probability of $H_{0,j}$, and $\Pi_{0,j}$ and $\Pi_{1,j}$ are the prior distributions for the pair of parameters $\theta_{1,j}$ and $\theta_{2,j}$ under the hypotheses $H_{0,j}$ and $H_{1,j}$, respectively. Under the prior $\Pi_{0,j}$ the two parameters $\theta_{1,j}$ and $\theta_{2,j}$ are replaced by a common parameter $\theta_{0,j}$ distributed according to some distribution Q_j. Under the prior $\Pi_{1,j}$ the two parameters are independent and identically distributed according to Q_j.

Now the *global hypothesis* (1.1) can be written as a function of the *local hypotheses* (1.2):

$$H_0 : \cap_j H_{0,j} \quad \text{vs.} \quad H_1 : \cup_j H_{1,j}.$$

If one assumes independent priors for each parameter, then the marginal prior null is equal to:

$$\Pr(H_0) = \Pr(\cap_j H_{0,j}) = \prod_j \Pr(H_{0,j}).$$

When only a small subset of the parameters are involved in the difference across the two distributions, one achieves higher statistical power through this model with respect to the standard approach described at the beginning of this chapter. Additionally, one is able to identify the nature of these differences. In fact, one can identify the subset of local null hypotheses that are rejected with high probability, i.e., all the indices j such that $\Pr(H_{0,j}|\boldsymbol{x})$ is small. In many applications identifying the nature of the difference between the two distributions is as important as the existence of the difference itself. This aspect will be illustrated in multiple examples throughout the thesis.

Furthermore, one can introduce dependency across the local hypotheses, thus obtaining borrowing of information. In the context of Pólya trees and wavelets I obtain spatial dependency across the local hypotheses through a hidden Markov tree model (Crouse et al., 1998). If a local hypothesis is rejected, then the nearby local hypotheses are more likely to be rejected. In the context of a hierarchical mixture of normals, instead, the local null probability $\Pr(H_{0,j})$ is shrunk towards the global null probability $\Pr(H_0)$ through a hyperprior.

1.2 Pólya Trees

Pólya trees (PTs) (Ferguson, 1973; Lavine, 1992) are random probability measures on some measurable space Ω. In this section I provide an introduction to PTs and their properties.

Here I assume $\Omega = [0, 1)$, while later in the thesis I will also consider the multivariate case where $\Omega = [0, 1)^p$. The proposed framework is general since an unbounded rectangle can be transformed into a bounded rectangle by applying, for example, a cumulative distribution function transformation to each dimension.

Consider a recursive partition of Ω generated by dyadic cuts. For each set B arising during the partitioning, define B_0 and B_1 to be the two child sets obtained by dividing B with a dyadic cut (see Figure 1.1). Denote with $\mathcal{B}^l$ the collection of sets obtained at the lth level of the partition, with $\mathcal{B}^{(l)} = \cup_{m=0}^{l} \mathcal{B}^m$ the partition tree up to level l, and with $\mathcal{B}^{(\infty)} = \lim_{l\to\infty} \cup_{m=0}^{l} \mathcal{B}^m$ the infinite partition tree.

As shown in Figure 1.1, for a probability measure Q on Ω and $B \in \mathcal{B}^{(\infty)}$, I define the *probability assignment* $\theta(B)$ as the proportion of mass $Q(B)$ assigns to the child set B_0:

$$\theta(B) = Q(B_0)/Q(B) = Q(B_0|B).$$

Since $\mathcal{B}^{(\infty)}$ generates the Borel σ-algebra, then any probability measure Q over Ω

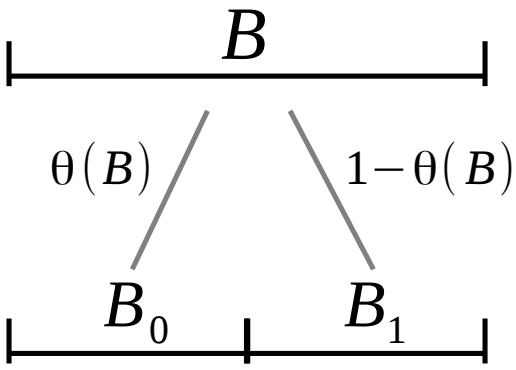

FIGURE 1.1: The probability assignment $\theta(B)$ represents the proportion of mass $Q(B)$ assigned to the child set B_0.

can be mapped injectively to a collection of probability assignments:

$$Q \mapsto \boldsymbol{\theta} = \{\theta(B) : B \in \mathcal{B}^{(\infty)}\}. \tag{1.3}$$

If Q is absolutely continuous, the likelihood for $x_1, \ldots, x_n | Q \overset{\text{iid}}{\sim} Q$ can be written as a function of the collection of probability assignments $\boldsymbol{\theta}$ as follows

$$\Pr(x_1, \ldots, x_n | \boldsymbol{\theta}) = \prod_{B \in \mathcal{B}^{(\infty)}} 2^{n(B)} \Pr(\boldsymbol{n}(B) | \theta(B))$$

$$\Pr(\boldsymbol{n}(B) | \theta(B)) = \theta(B)^{n(B_0)} (1 - \theta(B))^{n(B_1)},$$

where $\boldsymbol{n}(B) = (n(B_0), n(B_1))$, and $n(B_0)$ and $n(B_1)$ represent the number of observations falling in B_0 and B_1.

A prior on Q can be defined through assigning a prior on the probability assignments $\boldsymbol{\theta}$. Assume $\theta(B) \overset{\text{ind}}{\sim} \text{Beta}(\boldsymbol{\alpha}(B))$ where $\boldsymbol{\alpha}(B) = (\alpha(B, 0), \alpha_1(B, 1))$ are fixed and $B \in \mathcal{B}^{(\infty)}$. If the pseudo-counts $\boldsymbol{\alpha}(B)$ satisify the conditions in Theorem 3.3.2 of Ghosh and Ramamoorthi (2003), then the random probability measure Q is said to have a Pólya tree distribution, denoted $Q \sim PT(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}(B) \text{ for } B \in \mathcal{B}^{(\infty)}\}$.

The conditions in Theorem 3.3.2 of Ghosh and Ramamoorthi (2003), which guarantee that the cumulative distribution function is right-continuous, are satisfied, for instance, if $\alpha(B,0) = \alpha_1(B,1) = \alpha(l)$ for all $B \in \mathcal{B}^l$, and $\prod_{l>l^*} \alpha(l) = 0$ for any $l^* > 0$. In addition, if $\sum \alpha(l)^{-1} < \infty$, then the draws from Q are absolutely continuous.

The Pólya tree provides a conjugate family of priors over distributions. Assume that

$$x_1, \ldots, x_n | Q \overset{\text{iid}}{\sim} Q$$
$$Q \sim PT(\boldsymbol{\alpha}),$$

then

$$Q | x_1, \ldots, x_n \sim PT(\alpha | \boldsymbol{x}),$$

where $\alpha|\boldsymbol{x} = \{\boldsymbol{\alpha}(B|\boldsymbol{x}) \text{ for } B \in \mathcal{B}^{(\infty)}\}$, $\boldsymbol{\alpha}(B|\boldsymbol{x}) = (\alpha(B,0) + n(B_0), \alpha_1(B,1) + n(B_1))$ and $n(B) = |x_i \in B : i = 1, \ldots, n|$ for any $B \in \mathcal{B}^{(\infty)}$. The posterior pseudo-counts for the node B are equal to the prior pseudo-counts plus the total number of observations falling in each of the two child sets of B.

Finally, if the conditions for having absolutely continuous draws are satisfied, then the marginal likelihood can be computed analytically:

$$\begin{aligned}\Pr(x_1, \ldots, x_n) &= \int \Pr(x_1, \ldots, x_n | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int \prod_{B \in \mathcal{B}^{(\infty)}} 2^{n(B)} \Pr(\boldsymbol{n}(B) | \theta(B)) \pi(\theta(B)) d\theta(B)\end{aligned}$$

where $D(\boldsymbol{w}) = \Gamma(w_1)\Gamma(w_2)/\Gamma(w_1 + w_2)$ for $\boldsymbol{w} = (w_1, w_2)$. If $\alpha(B,0) = \alpha(B,1)$ and $n(B) < 2$, then $2^{n(B)} D(\boldsymbol{\alpha}(B) + \boldsymbol{n}(B))/D(\boldsymbol{\alpha}(B)) = 1$.

1.3 Dirichlet Process Mixtures

The Dirichlet process (DP) is a stochastic process whose draws are probability distributions (Ferguson, 1973). In this section I provide a definition of the DP, present

some of its properties, and show how it can be used in the context of mixture modeling.

Let H be a distribution over some measurable space Θ and α be a positive real number. Then, Q is DP distributed with baseline distribution H and mass parameter α if

$$(Q(B_1), \ldots, Q(B_r)) \sim \text{Dirichlet}(\alpha H(B_1), \ldots, \alpha H(B_r)),$$

for any finite measurable partition $B_1, \ldots, B_r$ of Θ. I say $Q \sim DP(\alpha, H)$.

The DP is centered around H in the sense that $E(Q(B)) = H(B)$ for any Borel set B. The mass parameter α controls how concentrated Q is around the baseline distribution: $\text{Var}(Q(B)) = H(B)(1 - H(B))/(\alpha + 1)$, for any Borel set B.

The DP is conjugate and the posterior parameters can be easily computed. These features make the DP an attractive tool for Bayesian nonparametric modeling. In particular, if one has observations from a DP:

$$\theta_1, \theta_2, \ldots, \theta_n | Q \overset{\text{iid}}{\sim} Q$$
$$Q \sim DP(\alpha, H),$$

then distribution of Q given the observations has the following distribution:

$$Q|\theta_1, \theta_2, \ldots, \theta_n \sim DP\left(\alpha + n, \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \sum_{i=1}^{n} \frac{1}{n} \delta_{\theta_i}\right).$$

The posterior baseline distribution is a weighted average between the prior baseline distribution H and the empirical distribution. The posterior mass parameter is the sum of the prior mass parameter and the sample size. As more observations are collected, the posterior process concentrates around the posterior baseline distribution.

The DP has the following series-representation:

$$Q(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\lambda_k}(\cdot),$$

where

$$(\pi_1, \pi_2, \ldots) \sim \text{Poisson-Dirichlet}(\alpha)$$

$$\lambda_k \overset{\text{iid}}{\sim} H, \quad k = 1, 2, \ldots$$

The Poisson-Dirichlet distribution was defined by Kingman (1975) as the limiting distribution of a Dirichlet$(\alpha/K, \ldots, \alpha/K)$ as K goes to infinity. The stick-breaking weights of the more common Sethuraman (1994)'s representation of the DP can be obtained by size-biased permutations of the weights of the Poisson-Dirichlet distribution (Pitman and Yor, 1997).

From the series-representation, one can see that the draws from a DP are discrete almost surely. In many applications one is interested in modeling the density from which observations are drawn. In that case one can convolve Q with a kernel, thereby obtaining a random probability distribution which has a density. Specifically, assume data are generated from the so-called mixture model:

$$y_i \overset{\text{iid}}{\sim} F, \quad i = 1, \ldots, n,$$

$$f(\cdot) = \sum_{k \in \mathcal{K}} \pi_k g(\cdot|\lambda_k)$$

where f denotes the density of F, $g(\cdot|\lambda)$ is the mixing kernel with location-scale parameters λ, π_k is the associated mixture weight, and $\mathcal{K}$ is the countable (possibly infinite) index set of the mixture components. The mixture weights live on the $|\mathcal{K}|$-simplex, i.e., $\pi_k \geq 0$ and $\sum_k \pi_k = 1$.

In a Bayesian formulation one places a prior on the mixture weights and the location-scale parameters. If the number of mixture components is finite, i.e., $|\mathcal{K}| < \infty$, then one has a finite mixture model. If the number of mixture components is infinite, then one has a nonparametric model. Within this second class, if Q is used as mixing measure, then one obtains the so-called DP mixture (Escobar and West,

1995):

$$y_i|Q \overset{\text{iid}}{\sim} \int g(y_i|\lambda)Q(d\lambda)$$

$$Q \sim DP(\alpha, H).$$

The framework does not depend on a specific choice of the kernel, however in this thesis I focus on the normal kernel, i.e., $g(\cdot|\theta_k) = N(\cdot|\mu_k, \Sigma_k)$ due to its simplicity and versatility.

1.4 Wavelets

Wavelets are families of orthonormal basis functions over functional spaces. The basis functions are generated by dilations and translations of a function ψ, called the *mother wavelet*:

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), \quad j, k \in \mathbb{Z}.$$

The regularity properties of the functional space depend on the choice of the mother wavelet (Daubechies et al., 1992). The wavelet series representation of a function $y \in L^2(\mathbb{R})$ is:

$$y(t) = \sum_{j,k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t), \tag{1.4}$$

where the wavelet coefficients are defined as follows:

$$d_{j,k} = \int y(t)\psi_{j,k}(t)dt, \quad j, k \in \mathbb{Z}.$$

One of the advantages of the wavelet transform over the Fourier transform is that wavelets decompose the function both in location and scale. The coefficient $d_{j,k}$ describes the behavior of the function around location $2^{-j}k$ at frequencies near 2^j.

In practice, model (1.4) is generally observed only at discrete time points. In particular, if the function y is observed at $T = 2^J$ equally spaced points $\boldsymbol{y} =$

$(y_0, \ldots, y_{T-1})$ for some integer J, then the discrete wavelet transform (DWT) decomposes $\boldsymbol{y}$ into discrete wavelet coefficients $d_{j,k}$ for $j = 0, \ldots, J-1$ and $k = 0, \ldots, 2^j - 1$, and one sampling scale coefficient $c_{0,0}$. The DWT can also be represented as an orthogonal linear transformation $\boldsymbol{d} = \boldsymbol{y}W'$, where the coefficients are organized in lexicographical order, i.e., $\boldsymbol{d} = (d_{0,0}, d_{1,0}, \ldots, d_{J-1,2^{J-1}-1})$.

A second appealing feature of the wavelet transform is that white noise is spread equally among all the wavelet coefficients, while the signal is concentrated only on a small subset of the coefficients. When a noisy function is observed most of the coefficients are very small and represent noise, while the few coefficients which are large in magnitude contain the signal. Many methods have been proposed to separate the functional signal from the noise. In the frequentist literature the most common strategy is to use a thresholding approach, where all the coefficients $d_{j,k}$ smaller than a data-driven threshold are set equal to zero. Some examples are Donoho and Johnstone (1994), Donoho and Johnstone (1995) and Johnstone and Silverman (1997). In the Bayesian literature, instead, the coefficients are modeled through spike and slab priors which shrink the small coefficients towards zero. Some examples are Chipman et al. (1997), Clyde et al. (1998) and Abramovich et al. (1998).

2

Accept-Reject Markov Trees

2.1 Introduction

Comparing two data samples to identify the differences in the underlying distributions is a fundamental inference objective that lies at the heart of many scientific investigations. In numerous modern applications, the data are multi-dimensional, multi-modal and skewed, and the cross-sample differences can be of a variety of shapes and forms. Classic methods such as the k-nearest neighbors test (Schilling, 1986; Henze, 1988) and the Cramér test (Baringhaus and Franz, 2004) cannot be used to identify the corresponding differences because they focus exclusively on testing the global null hypothesis that the two distributions are equal. This entails a need for flexible, nonparametric multivariate two-sample comparison methods that can effectively detect *and* characterize a variety of underlying differences.

Two-sample comparison methods based on Pólya trees (Ferguson, 1973; Lavine, 1992) have emerged recently in the Bayesian literature (Holmes et al., 2009; Chen and Hanson, 2014; Ma and Wong, 2011). Pólya tree priors provide a flexible framework to adaptively approximate arbitrary distributions as well as simple closed-form

expressions for the posterior distribution and the marginal likelihood. Under a Pólya tree prior a probability distribution is decomposed into a collection of local probability assignment coefficients, organized on a multi-resolution dyadic partition tree of the sample space. Each probability assignment coefficient specifies how probability mass is split between the two child sets for each node in the partition tree. Under the null hypothesis the two samples are generated by a single Pólya tree, while under the alternative hypothesis the two samples are generated by two Pólya trees. In Holmes et al. (2009) and Chen and Hanson (2014)'s framework the two Pólya trees are independent under the alternative hypothesis, while Ma and Wong (2011) consider dependent Pólya trees.

I propose a multi-resolution method where two-sample comparison is achieved through carrying out a collection of *local* two-sample tests, each comparing the corresponding pair of probability assignment coefficients on a node in the partition tree and thus corresponding to the two-sample difference at a particular location and scale. The statistical evidence from these local tests is then combined for reporting the overall two-sample difference. This method reframes the two-sample comparison problem under a multiple testing perspective, encompassing the models proposed by Holmes et al. (2009), Chen and Hanson (2014) and Ma and Wong (2011) as special cases.

Existing methods impose stringent constraints among the local tests. In Holmes et al. (2009) and Chen and Hanson (2014), the local tests are either all accepted or all rejected, while in Ma and Wong (2011) if a local null is accepted at a particular location-scale node, then all the nulls defined on the descendants of that node—corresponding to two-sample differences of finer scales at the same location—must also be accepted. These assumptions result in overly conservative multiple testing control and considerable loss of power, especially when the differential structure involves only a small portion of the data. In this work I will relax these constraints

by adaptively defining the total number of local tests, and allowing each local test to have its own rejection/acceptance state. More importantly, I introduce spatial dependency into the local tests to obtain borrowing of information across "neighboring" tests. When two probability distributions differ at one location in the sample space, they tend to also differ in the neighboring parts of the space. The spatial clustering of differences implies that the rejection/acceptance states of the local tests are correlated, and the strength of the correlation depends on the distance between the corresponding nodes in the multi-resolution partition tree.

I introduce a new multi-resolution two-sample comparison method that effectively accounts for this spatial dependency using graphical modeling techniques. In particular, I model the dependency among the rejection/acceptance of the local tests through a Markov tree (Crouse et al., 1998). The Markov tree is a Markov process indexed by nodes in the multi-resolution partition tree, and it transitions between a "reject" and a "accept" state on the nodes, corresponding respectively to whether the local null hypotheses are rejected or accepted. By appropriately specifying the Markov transition probabilities, which determine the conditional rejection probability on a node given the rejection/acceptance on the parent, different levels of dependency can be incorporated. Furthermore, while the standard Markov tree is defined on a *fixed* multi-resolution partition tree, I generalize the model by allowing the multi-resolution partition tree itself to be adaptive to the topological structure of the underlying distributions. This is achieved in a principled manner by embedding a random recursive partition mechanism—a prior on the multi-resolution tree—into the construction of the Markov tree. The marriage between the Markov tree and randomized partitioning results in a fully specified joint generative process for the two probability distributions of interest, which I call the accept-reject Markov tree (ARM-tree).

I provide a mathematically rigorous investigation into the various properties of the

ARM-tree model and construct an efficient inferential recipe for two-sample comparison based on this model. Its Markov nature allows Bayesian inference on two-sample comparison to be carried out analytically and efficiently using "forward-summation-backward-sampling" type of information propagation algorithms (Liu, 2008, Sec. 2). Finally, the corresponding two-sample comparison under the ARM-tree process is asymptotically consistent. An R package implementing the method is available at https://github.com/jacsor/ARMtree.

The chapter is organized as follows. In Section 2.2 I introduce the ARM-tree model, provide guidelines for prior specification that effectively incorporates spatial clustering of differences, and describe the inferential recipe for two-sample comparison. In Section 2.3 I evaluate the performance of the ARM-tree. I first carry out a simulation study to compare its performance to other nonparametric two-sample tests. I then illustrate how to pin-point and visualize where and what the difference is using the posterior rejection/acceptance probabilities for the local tests along with the inferred multi-resolution partition tree. Finally, I apply ARM-tree to a 7-dimensional flow cytometry dataset, for which the proposed method successfully identifies an experimentally validated differential hotspot involving just 0.2% of the cells.

2.2 Model

2.2.1 Dependent Pólya Tree Priors and Multi-Scale Two-Sample Comparison

Given two samples from the following model:

$$x_{1,1}, \ldots, x_{n_1,1} | Q_1 \overset{\text{iid}}{\sim} Q_1$$

$$x_{1,2}, \ldots, x_{n_2,2} | Q_2 \overset{\text{iid}}{\sim} Q_2,$$

the representation (1.3) can be used to describe Q_1 and Q_2 and to decompose the *global hypothesis test*:

$$H_0 : Q_1 = Q_2 \quad \text{vs} \quad H_1 : Q_1 \neq Q_2$$

into a collection of *local tests* on the probability assignments:

$$H_0(B) : \theta_1(B) = \theta_2(B) \quad \text{vs} \quad H_1(B) : \theta_1(B) \neq \theta_2(B),$$

where the subscripts refer to the two distributions. In particular, I assume that under $H_0(B)$:

$$\theta_1(B) = \theta_2(B) = \theta_0(B) \sim \text{Beta}(\boldsymbol{\alpha}_0(B)),$$

while, under $H_1(B)$ the two assignments are drawn independently from Beta priors:

$$\theta_k(B) \sim \text{Beta}(\boldsymbol{\alpha}_k(B)) \quad \text{for } k = 1, 2.$$

Now, the *global hypothesis test* can be formulated in terms of the *local hypotheses*:

$$\begin{aligned} H_0 : &\quad \cap_{B \in \mathcal{B}^{(\infty)}} H_0(B) \\ H_1 : &\quad \cup_{B \in \mathcal{B}^{(\infty)}} H_1(B). \end{aligned}$$

The *global null hypothesis* is rejected if and only if there is at least one *local null hypothesis* rejected. I introduce a state variable $R(B) = 0$ or 1 to indicate whether the hypothesis $H_0(B)$ is *accepted* or *rejected*, and indicate with $\boldsymbol{R} = \{R(B) \text{ for } B \in \mathcal{B}^{(\infty)}\}$ the collection of the state variables. Conditionally on the partition tree and state variables, I define this model as a distribution on a pair of probability measures.

Definition 1. *Given a collection of states $\boldsymbol{R}$, the pair of random probability measures (Q_1, Q_2) is said to have a dependent Pólya tree (DPT) distribution with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{R}$. I write $[Q_1, Q_2 | \boldsymbol{R}] \sim DPT(\boldsymbol{\alpha}, \boldsymbol{R})$.*

The two distributions are dependent in the sense that they share the same conditional structure for some regions. Specifically, $Q_1(B_i|B) = Q_2(B_i|B)$ for $i = 0, 1$ and $B \in \mathcal{B}^{(\infty)}$ such that $R(B) = 0$.

To fully specify a model for (Q_1, Q_2) I need to introduce a prior on the states $\boldsymbol{R} = \{R(B) \text{ for } B \in \mathcal{B}^{(\infty)}\}$. Before describing the proposed prior, I show how this framework encompasses previously proposed methods. The model introduced by Holmes et al. (2009), for instance, can be interpreted as a special case of this framework. They assume that $R(B) = 0$ for all $B \in \mathcal{B}^{(\infty)}$ with probability ρ, otherwise $R(B) = 1$ for all $B \in \mathcal{B}^{(\infty)}$. This implies that $\Pr(H_0) = \rho$, while the $\Pr(H_0|\boldsymbol{x})$ can be expressed as a function of the Bayes factor (BF):

$$\Pr(H_0|\boldsymbol{x}) = \left[BF\frac{\rho}{1-\rho} + 1\right]^{-1} \quad BF = \frac{\Pr(\boldsymbol{x}|H_0)}{\Pr(\boldsymbol{x}|H_1)} = \prod_{B\in\mathcal{B}^{(\infty)}} \frac{\Pr(\boldsymbol{n}(B)|H_0(B))}{\Pr(\boldsymbol{n}(B)|H_1(B))},$$

where under the null

$$\Pr(\boldsymbol{n}(B)|H_0(B)) = \frac{D(\boldsymbol{\alpha}_0(B) + \boldsymbol{n}_1(B) + \boldsymbol{n}_2(B))}{D(\boldsymbol{\alpha}_0(B))}$$

and under the alternative

$$\Pr(\boldsymbol{n}(B)|H_1(B)) = \prod_{k=1,2} \frac{D(\boldsymbol{\alpha}_k(B) + \boldsymbol{n}_k(B))}{D(\boldsymbol{\alpha}_k(B))}.$$

It is often unrealistic that the *local null hypotheses* are either all accepted or all rejected. In particular, when the differences across the two distributions are concentrated only on a small subset of the sample space, local hypotheses that should be rejected are sparse. Thus, the alternative hypothesis under this model is overly parameterized, yielding limited power. Additionally, this model does not allow the identification of the location and scale of the differences, since the *local null hypotheses* are jointly all rejected or all accepted.

Instead of Holmes et al. (2009)'s model where the rejection states $R(B)$ are all identical, one could take an opposite modeling strategy assuming that all the $R(B)$ are mutually independent. Since the number of regions grows geometrically with the level l of the partition tree, one must impose a multiplicity adjustment to control the false rejections. If I assume that $\Pr(R(B) = 0)$ is constant for all regions in a level, then a priori I have:

$$\Pr(H_0) = \prod_{B \in \mathcal{B}^{(\infty)}} \Pr(H_0(B)) = \prod_{l \geq 0} (\rho_l)^{2^l},$$

where $(\rho_l)^{2^l}$ is the probability of being in the null state for all the 2^l probability assignments at level l. For instance, if $\rho_l = \exp\{6\log(\rho)/[(l+1)\pi]^2\}^{2^{-l}}$, we obtain

$$\log(\Pr(H_0)) = \sum_{l \geq 0} 6c/[(l+1)\pi]^2 = \log(\rho).$$

Unfortunately if the prior ρ_l approaches to one at such a rapid rate, the power to detect local differences is very limited. A partial remedy to this problem is to adaptively reduce the total number of hypothesis tests by allowing *early stopping* in the partition tree. Instead of considering the infinite partition tree, some of the branches are pruned and no local tests are carried out within those branches. Furthermore, the assumption that all the states are mutually independent is overly restrictive. In fact, when two probability distributions differ at one location in the sample space, they tend to also differ in the neighboring parts of the space. I can achieve spatial dependency in the local hypothesis tests through the introduction of *Markov tree dependency* into these latent state variables.

Following the strategy used by Wong and Ma (2010) in the context of density estimation, I replace the fixed infinite partition tree $\mathcal{B}^{(\infty)}$ with a random partition tree $\mathcal{T}$ where the branches are randomly pruned. As shown in Figure 2.1, for every node $B \in \mathcal{B}^{(\infty)}$ I introduce stopping variables $S(B) \in \{0, 1\}$, indicating whether

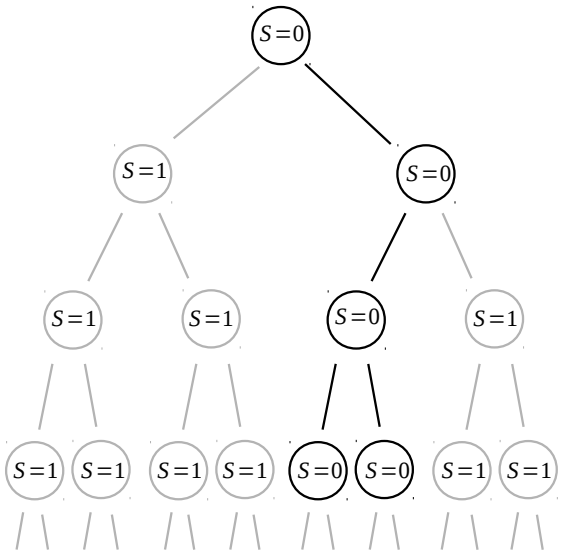

FIGURE 2.1: Representation of the partitioning tree $\mathcal{B}^{(\infty)}$, where for each node $B \in \mathcal{B}^{(\infty)}$ the state variable $S(B)$ is plotted. I highlight in black the unstopped nodes associated to pruned tree $\mathcal{T}$.

the partition tree is stopped at the node B and no more local hypotheses will be carried out within that branch. For ease of notation, I indicate with $\mathcal{B}^{(\infty)}(B)$ the infinite sub-partition tree where the root node is B. The generative procedure for $\mathcal{T}$ can be described recursively as follows. The procedure starts from the root of the tree $B = \Omega$ with $\mathcal{T} = \emptyset$. I draw a stopping variable $S(B) \sim \text{Bernoulli}(\eta(B))$. If $S(B) = 1$, then the partition process stops, and I assign $S(C) = 1$ and $R(C) = 0$ for $C \in \mathcal{B}^{(\infty)}(B)$. Otherwise, if $S(B) = 0$, the set B is split in two child sets B_0 and B_1 with a dyadic cut, and B_0 and B_1 are added to $\mathcal{T}$. For each of the child sets I draw an independent stopping variable and iterate the process until all branches are stopped. I indicate with $\boldsymbol{S} = \{S(B) \text{ for } B \in \mathcal{B}^{(\infty)}\}$ the collection of stopping variables, with $\boldsymbol{\eta} = \{\eta(B) \text{ for } B \in \mathcal{B}^{(\infty)}\}$ and with $\mathcal{T} \sim \text{Tree}(\boldsymbol{\eta})$ the distribution of the pruned tree. I assume that $\eta(B) = \eta$ for all $B \in \mathcal{B}^{(\infty)}$, but one could also let $\eta(B)$ vary with the

level of B. Now, the prior global null can be defined hierarchically as follows:

$$\Pr(H_0|\mathcal{T}) = \prod_{B\in\mathcal{T}} \Pr(H_0(B))$$

$$\mathcal{T} \sim \text{Tree}(\boldsymbol{\eta}).$$

To introduce local dependency across the state variables $R(B)$ I adopt a Markov tree model (Crouse et al., 1998) on $\mathcal{T}$, so that the value of $R(B)$ depends on the state on B's parent, denoted by parent(B), in the partition tree:

$$\Pr\left[R(B) = r'|R(\text{parent}(B)) = r, B \in \mathcal{T}\right] = \rho_{r,r'}(B), \tag{2.1}$$

where $\rho_{r,r'}(B) \geq 0$ for any $r, r' \in \{0,1\}$, $\rho_{r,0}(B) + \rho_{r,1}(B) = 1$ for any $r \in \{0,1\}$, and $B \in \mathcal{B}^{(\infty)} \setminus \Omega$. I call $\rho_{r,r'}(B)$ the transition probabilities, and organize them in a transition probability matrix $\boldsymbol{\rho}(B)$. The root of the tree Ω does not have a parent, thus I introduce an initial state probability $\boldsymbol{\pi} = (\pi_0, \pi_1)$:

$$\Pr\left[R(\Omega) = r|\Omega \in \mathcal{T}\right] = \pi_r, \quad r = 0, 1.$$

A simple and reasonable choice for the prior transition matrix $\boldsymbol{\rho}(B)$ is to make it dependent on the level of B, i.e., $\boldsymbol{\rho}(B) = \boldsymbol{\rho}(l)$. Additionally, I adopt a parsimonious and yet flexible functional form to further reduce the elicitation of the prior to a small number of parameters. More specifically, I define

$$\boldsymbol{\rho}(l) = \begin{bmatrix} 1-\gamma 2^{-l} & \gamma 2^{-l} \\ 1-\gamma\beta^{-l} & \gamma\beta^{-l} \end{bmatrix},$$

for some $\beta \geq 1$ and $\gamma \in [0,1]$. The parameter γ controls the transition from the accept state to the reject state, and the 2^{-l} factor is included to provide adequate control for multiple testing since the expected number of non-stopped regions at level l is equal to $2^l(1-\eta)^{l+1}$ if $\eta(B) = \eta$ for all $B \in \mathcal{B}^{(\infty)}$. Finally, the parameter β determines the spatial clustering of the differential structure. If the process is in the

reject state on a node, then a smaller β indicates that the process is more likely to be in the reject state on its parent, children, and neighbors. If $\beta = 2$, then there is no spatial clustering, i.e., the probability of rejection at B does not depend on the accept/reject state of B's parent.

The distribution of the collection of pairs of latent variables $Z(B) = (S(B), R(B)) \in \mathcal{Z} = \{(1,0),(0,0),(0,1)\}$ can be described jointly:

$$\begin{aligned}\hat{\rho}_{z,z'}(B) &:= \Pr(Z(B) = z' | Z(\text{parent}(B)) = z) \\ &= \begin{cases} \eta(B) & \text{if } z' = (1,0) \text{ and } z = (0,r') \\ (1-\eta(B))\rho_{r,r'}(B) & \text{if } z' = (0,r') \text{ and } z = (0,r) \\ 1 & \text{if } z' = z = (1,0) \\ 0 & \text{if } z' = (0,r') \text{ and } z = (1,0), \end{cases}\end{aligned}$$

for $r, r' = 0, 1$, and $B \in \mathcal{B}^{(\infty)} \setminus \Omega$. For the root of the tree Ω I have:

$$\hat{\pi}_z := \Pr(Z(\Omega) = z) = \begin{cases} \eta(B) & \text{if } z = (1,0) \\ (1-\eta(B))\pi_s(B) & \text{if } z = (0,r), \end{cases}$$

for $r = 0, 1$. Now I can define the distribution of the collection of state variables $\boldsymbol{Z} = \{Z(B) : B \in \mathcal{B}^{(\infty)}\}$.

Definition 2. *The collection of latent variables $\boldsymbol{Z}$ is said to have a hidden Markov tree (HMT) distribution with parameters $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\rho}}$. I write $\boldsymbol{Z} \sim HMT(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\rho}})$.*

Now that I have described the different components of the model, I can write it hierarchically in two stages:

$$\begin{aligned}[Q_1, Q_2 | \boldsymbol{Z}] &\sim DPT(\boldsymbol{\alpha}, \boldsymbol{Z}) \\ \boldsymbol{Z} &\sim HMT(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\rho}}).\end{aligned}$$

Using a similar argument as in Theorem 1 in Wong and Ma (2010), one can show that with probability 1 this procedure will produce a pair of well-defined probability measures (Q_1, Q_2) that are both absolutely continuous with respect to the

Lebesgue measure if $\eta(B)$ is bounded away from zero. Thus, I can formally define this generative model as a distribution on a pair of probability measures.

Definition 3. *The pair of random probability measures (Q_1, Q_2) is said to have a accept-reject Markov tree (ARM-tree) distribution with parameters $\boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{\rho}$ and $\boldsymbol{\alpha}$. I write $(Q_1, Q_2) \sim$ ARM-tree$(\boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$.*

Due to the Markov structure, one can evaluate $\Pr(H_0)$ analytically through "forward-summation" type recursions. To this end, define the following mapping:

$$\Psi(B) := \begin{cases} \Pr\left(\cap_{C \in \mathcal{B}^{(\infty)}(\Omega)} R(C) = 0\right) & \text{if } B = \Omega \\ \Pr\left(\cap_{C \in \mathcal{B}^{(\infty)}(B)} R(C) = 0 | Z(\text{parent}(B)) = (0,0)\right) & \text{if } B \in \mathcal{B}^{(\infty)} \setminus \Omega \end{cases}$$

This quantity represents the probability of accepting all the local hypotheses within a branch starting from the node B, given that the local hypothesis on the parent set of B is neither rejected nor stopped. The mapping $\Psi(B)$ has the following recursive representation:

$$\begin{aligned} \Psi(B) &= \Pr(Z(B) = (1,0) | Z(\text{parent}(B)) = (0,0)) \\ &\quad + \Pr(Z(B) = (0,0) | Z(\text{parent}(B)) = (0,0)) \prod_{i=0,1} \Psi(B_i) \\ &= \eta(B) + (1 - \eta(B)) \rho_{0,0}(B) \prod_{i=0,1} \Psi(B_i), \end{aligned} \tag{2.2}$$

for $B \in \mathcal{B}^{(\infty)} \setminus \Omega$, and $\Psi(B) = \eta(B) + (1 - \eta(B)) \pi_0 \prod_{i=0,1} \Psi(B_i)$, for $B = \Omega$. All local null hypotheses within the branch B are accepted, if the tree is pruned on B or the local null hypothesis on B is accepted and all the local null hypotheses within the branches of both child sets B_0 and B_1 are accepted. Since $\Psi(\cdot) \in [0,1]$, for all regions this quantity can be bounded by terminating the recursion at a deeper finite level. Now I can formally describe how the prior probability that the two distributions are identical can be computed.

Proposition 4. *Let (Q_1, Q_2) have an ARM-tree$(\boldsymbol{\rho}, \boldsymbol{\alpha}, \eta)$ prior. Then the prior probability that the two distributions are identical is given by:*

$$\Pr(H_0) = \Psi(\Omega).$$

2.2.2 Posterior Inference and Posterior Consistency

In this section I find the corresponding posterior of an ARM-tree. The main result (Theorem 7) provides a hierarchical representation of the posterior distribution of the ARM-tree process. Conditionally on the data, the collection of pairs of latent variables $Z(B) = (S(B), R(B))$ are still distributed according to a HMT. Given the $Z(B)$, (Q_1, Q_2) is still distributed according to a DPT. Corollary 8 shows that the recursive representations (2.2) can be used to compute the corresponding posterior quantities by simply replacing the prior parameters with the corresponding posterior parameters.

Moreover, the posterior parameters can be computed recursively. These properties follow from the Markov nature of the ARM-tree process, and the results in this section correspond to a "forward-summation-backward-sampling" information propagation algorithm for computing posterior Markov models (Liu, 2008). More specifically, (2.3) corresponds to the "forward-summation" step for the ARM-tree model while Theorem 7 the "backward-sampling" step. Lemma 6 and the following paragraph show that information propagation can be carried out analytically even when the partition sequence is infinite. Readers less interested in the technical details may directly jump to those results.

Define $\boldsymbol{x}_k(B)$ to be the subset of observations from sample k falling into $B \in \mathcal{B}^{(\infty)}$, i.e., $\boldsymbol{x}_k(B) = \{x_{k,i} : x_{k,i} \in B \text{ for } i = 1, \ldots, n_k\}$, and $\boldsymbol{x}(B) = \{\boldsymbol{x}_1(B), \boldsymbol{x}_2(B)\}$. Similarly, define $\boldsymbol{\theta}_k(B) = \{\theta_k(C) : \text{ for } C \in \mathcal{B}^{(\infty)}(B)\}$ for $k = 1, 2$. If the sample

space is restricted to B, the likelihood of the data is the following:

$$\Pr(\boldsymbol{x}(B)|\boldsymbol{\theta}_1(B),\boldsymbol{\theta}_2(B),\mathcal{T}(B)) = \mu(B)^{-\sum_k n_k(B)} \prod_{C\in\mathcal{T}(B)}\prod_{k=1,2} 2^{n_k(C)} \Pr(\boldsymbol{n}_k(C)|\theta_k(C)),$$

where $\mu(\cdot)$ is the Lebesgue measure, and $\mathcal{T}(B) = \mathcal{T} \cap \mathcal{B}^{(\infty)}(B)$.

Then, the probability assignments and the latent variables can be integrated out:

$$\begin{aligned}\Phi_z(\boldsymbol{x}, B) :&= \Pr(\boldsymbol{x}(B)|Z(\text{parent}(B)) = z) \\ &= \int \Pr(\boldsymbol{x}(B)|\boldsymbol{\theta}_1(B),\boldsymbol{\theta}_2(B),\mathcal{T}(B))\Pi_z(d\boldsymbol{\theta}_1(B), d\boldsymbol{\theta}_2(B), d\boldsymbol{Z}(B)),\end{aligned} \tag{2.3}$$

where $B \in \mathcal{B}^{(\infty)}\backslash\Omega$, and $\Pi_z(\boldsymbol{\theta}_1(B),\boldsymbol{\theta}_2(B),\boldsymbol{Z}(B))$ indicates the joint prior on $\boldsymbol{\theta}_1(B),\boldsymbol{\theta}_2(B)$ and $\boldsymbol{Z}(B)$ given that $Z(\text{parent}(B)) = z$ for $z \in \mathcal{Z}$. For the root Ω there is no parent set, and so $\Phi_z(\boldsymbol{x},\Omega)$ is the marginal likelihood, i.e., $\Phi_z(\boldsymbol{x},\Omega) = \Pr(\boldsymbol{x})$. Given the self-similar nature of ARM-tree, (2.3) can be represented recursively.

Lemma 5. *For every $B \in \mathcal{B}^{(\infty)}$ and $z \in \mathcal{Z}$, $\Phi_z(\boldsymbol{x}, B)$ has the following recursive representation:*

$$\Phi_z(\boldsymbol{x}, B) = \begin{cases} \sum_{z'\in\mathcal{Z}} \hat{\rho}_{z,z'}(B)\Lambda_{z'}(\boldsymbol{x}, B) & \text{if } B \in \mathcal{B}^{(\infty)} \setminus \Omega \\ \sum_{z'\in\mathcal{Z}} \hat{\pi}_{z'}\Lambda_{z'}(\boldsymbol{x},\Omega) & \text{if } B = \Omega, \end{cases}$$

where

$$\begin{aligned}\Lambda_z(\boldsymbol{x}, B) :&= \Pr(\boldsymbol{x}(B)|Z(B) = z) \\ &= \begin{cases} \mu(B)^{-\sum_k n_k(B)} & \text{if } z = (1,0) \\ \Pr(\boldsymbol{n}(B)|H_0(B)) \prod_{i=0,1} \Phi_z(\boldsymbol{x}, B_i) & \text{if } z = (0,0) \\ \Pr(\boldsymbol{n}(B)|H_1(B)) \prod_{i=0,1} \Phi_z(\boldsymbol{x}, B_i) & \text{if } z = (0,1). \end{cases}\end{aligned}$$

Proof. See Appendix A.1. □

This representation of $\Phi_z(\boldsymbol{x}, B)$ is recursive in the sense that one can compute it based on $\Phi_{z'}(\boldsymbol{x}, B_0)$ and $\Phi_{z'}(\boldsymbol{x}, B_1)$ for $z, z' \in \mathcal{Z}$. This recursive representation becomes operational if (2.3) can eventually be expressed in closed form. Lemma 6 provides analytic expressions for some specific regions.

Lemma 6. *For two types of $B \in \mathcal{B}^{(\infty)}$, $\Phi_z(\boldsymbol{x}, B)$ is known analytically:*

1. *Empty regions, i.e. $B : \boldsymbol{x}(B) = \{\emptyset, \emptyset\}$, then $\Phi_z(\boldsymbol{x}, B) = 1$;*

2. *Regions with a single observation, i.e. $B : \boldsymbol{x}(B) = \{x, \emptyset\}$ or $\boldsymbol{x}(B) = \{\emptyset, x\}$, then $\Phi_z(\boldsymbol{x}, B) = 1/\mu(B)$ under the condition that $\alpha_k(B, 0)/\alpha_k(B, 1) = 1$ for all $k = 0, 1, 2$ and $B \in \mathcal{B}^{(\infty)}$.*

Proof. See Appendix A.1. □

For every finite sample size $n_1 + n_2$, there is a finite partitioning level l such that all the nodes of level l in the partitioning tree belong to one of these two types of "terminal" nodes. Thus, (2.3) can be computed recursively from these nodes of the tree up to the root. Finally, Theorem 7 establishes the conjugacy of ARM-tree and provides expression for the posterior parameters.

Theorem 7. *Suppose I observe two groups of i.i.d. samples $\boldsymbol{x}_1 = (x_{1,1}, \ldots, x_{1,n_1})$ and $\boldsymbol{x}_2 = (x_{2,1}, \ldots, x_{2,n_2})$ from two distributions Q_1 and Q_2. If (Q_1, Q_2) have an ARM-tree$(\boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ prior, then*

$$[Q_1, Q_2 | \boldsymbol{Z}, \boldsymbol{x}] \sim DPT(\boldsymbol{\alpha} | \boldsymbol{x}, \boldsymbol{Z})$$

$$\boldsymbol{Z} | \boldsymbol{x} \sim HMT(\hat{\boldsymbol{\pi}} | \boldsymbol{x}, \hat{\boldsymbol{\rho}} | \boldsymbol{x}),$$

where

1. *Initial probabilities:*

$$\hat{\pi}_{z'} | \boldsymbol{x} = \Pr(Z(\Omega) = z' | \boldsymbol{x}) = \hat{\pi}_{z'} \frac{\Lambda_{z'}(\boldsymbol{x}, \Omega)}{\Phi_z(\boldsymbol{x}, \Omega)}$$

   where $z, z' \in \mathcal{Z}$.

2. *Transition probabilities:*

$$\hat{\rho}_{z,z'}(B | \boldsymbol{x}) = \Pr(Z(B) = z' | Z(\textit{parent}(B)) = z, \boldsymbol{x}) = \hat{\rho}_{z,z'}(B) \frac{\Lambda_{z'}(\boldsymbol{x}, B)}{\Phi_z(\boldsymbol{x}, B)}$$

where $B \in \mathcal{B}^{(\infty)} \setminus \Omega$ *and* $z, z' \in \mathcal{Z}$.

3. *Pseudo-counts:*

$$\boldsymbol{\alpha}_0(B|\boldsymbol{x}) = \boldsymbol{\alpha}_0(B) + \boldsymbol{n}_1(B) + \boldsymbol{n}_2(B)$$
$$\boldsymbol{\alpha}_k(B|\boldsymbol{x}) = \boldsymbol{\alpha}_k(B) + \boldsymbol{n}_k(B)$$

for all $B \in \mathcal{B}^{(\infty)}$ *and* $k = 1, 2$.

Proof. See Appendix A.1. □

Corollary 8. *The marginal posterior null probability is* $\Pr(H_0|\boldsymbol{x}) = \Psi(\Omega|\boldsymbol{x})$*, where* $\Psi(\cdot|\boldsymbol{x})$ *has the following recursive representation:*

$$\Psi(B|\boldsymbol{x}) = \Pr(Z(B) = (1,0)|Z(\textit{parent}(B)) = (0,0), \boldsymbol{x}) + \Pr(Z(B) = (0,0)|Z(\textit{parent}(B)) = (0,0), \boldsymbol{x}) \prod_{i=0,1} \Psi(B_i|\boldsymbol{x}),$$

for $B \in \mathcal{B}^{(\infty)} \setminus \Omega$*. For* $B = \Omega$ *I have:*

$$\Psi(B|\boldsymbol{x}) = \Pr(Z(B) = (1,0)|\boldsymbol{x}) + \Pr(Z(B) = (0,0)|\boldsymbol{x}) \prod_{i=0,1} \Psi(B_i|\boldsymbol{x}).$$

The next two theorems establish the consistency for the two-sample test using ARM-tree.

Theorem 9. (Consistency under the alternative) *I observe two independent groups of i.i.d. samples* $\boldsymbol{x}_1 = (x_{1,1}, \ldots, x_{1,n_1})$ *and* $\boldsymbol{x}_2 = (x_{2,1}, \ldots, x_{2,n_2})$ *from two distributions* Q_1 *and* Q_2*, where* $n = n_1 + n_2 \to \infty$*, and* $n_1/n \to \xi$ *for some* $\xi \in (0,1)$*. Let* (Q_1, Q_2) *have an ARM-tree*$(\boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ *prior where* $\eta(B) \geq \eta$*,* $\eta, \pi_0 \in (0,1)$ *and* $\alpha_k(B,0)/\alpha_k(B,1) = 1$ *for all* $B \in \mathcal{B}^{(\infty)}$ *and* $k = 0, 1, 2$*. In addition, if* $\eta(B), \rho_{r,r'}(B) \in (0,1)$ *for all* $r, r' = 0, 1$ *and* $B \in \mathcal{B}^{(l)}$ *for some large enough* l*, then,*

$$\Pr(H_0|\boldsymbol{x}_1, \boldsymbol{x}_2) \xrightarrow{p} 0 \textit{ under } P_1^{(\infty)} \times P_2^{(\infty)},$$

for absolutely continuous P_1 and P_2 provided that there exists $B \in \mathcal{B}^{(\infty)}$ such that $P_1(B_i|B) \neq P_2(B_i|B)$ and $\xi P_1(B_i|B) + (1-\xi)P_2(B_i|B) \neq 1/2$ for $i = 0, 1$.

Proof. See Appendix A.1. □

Theorem 10. (Consistency under the null) *I observe two independent groups of i.i.d. samples $\boldsymbol{x}_1 = (x_{1,1}, \ldots, x_{1,n_1})$ and $\boldsymbol{x}_2 = (x_{2,1}, \ldots, x_{2,n_2})$ from two distributions Q_1 and Q_2, where $n = n_1 + n_2 \to \infty$, and $n_1/n \to \xi$ for some $\xi \in (0, 1)$. Let (Q_1, Q_2) have an ARM-tree$(\boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ prior where $\eta(B) \geq \eta$, $\eta, \pi_0 \in (0, 1)$. In addition, if for some $l \in \mathbb{N}$*

1. *$\rho_{r,r'}(B) \in (0, 1)$ for all $r, r' = 0, 1$ and $B \in \mathcal{B}^{(l)}$.*
2. *$\eta(B) = 1$ for all $s \in \mathcal{S}$, $B \in \mathcal{B}^m$ and for all $m > l$,*

then,

$$\Pr(H_0|\boldsymbol{x}) \xrightarrow{p} 1 \text{ under } P_0^{(\infty)} \times P_0^{(\infty)} \text{ for any absolutely continuous } P_0.$$

Proof. See Appendix A.1. □

2.2.3 Representative Tree

In many applications it is important not only to test for two-sample difference, but also to identify the differential structure. The multi-resolution approach provides a natural means to identifying differences through specifying the location, scale, and effect size of the difference. All such information is encapsulated in the full ARM-tree posterior and can be extracted through posterior Monte Carlo sampling or constructing summaries of the posterior. Here I design a strategy for effectively summarizing such posterior information.

The strategy is to identify a representative partition containing all regions that are *a posteriori* very likely to be in the reject state. A representative tree $\mathcal{T}^*$ can

be constructed using the following tree-growth procedure. Starting from the root $B = \Omega$, add B to $\mathcal{T}^*$, and if $\Pr(\cap_{C \in \mathcal{B}^{(\infty)}(B)} R(C) = 0 | \boldsymbol{x}, B \in \mathcal{T}^*) > 1 - \delta^*$ for some threshold $\delta^* \in (0, 1)$, then I stop the tree growth on B and deem the subtree beneath B as "uninteresting"; otherwise I divide B in B_0 and B_1. For each B_i I repeat this procedure until all branches are stopped.

Once the representative tree is constructed, for each $B \in \mathcal{T}^*$ we report $\Pr(H_1(B)|\boldsymbol{x})$, i.e., the marginal posterior probability of rejecting the local null hypothesis. Additionally, I define a notion of *effect size* to summarize the extent to which the two distributions are different on each node based on the posterior log-odds of the probability assignments of the two samples. The *effect size* on B is given by

$$\begin{aligned} \mathrm{Eff}(B|\boldsymbol{x}) &= \left| \log \frac{E\big(\theta_1(B)|R(B)=1,\boldsymbol{x}\big)/E\big(1-\theta_1(B)|R(B)=1,\boldsymbol{x}\big)}{E\big(\theta_2(B)|R(B)=1,\boldsymbol{x}\big)/E\big(1-\theta_2(B)|R(B)=1,\boldsymbol{x}\big)} \right| \\ &= \left| \log \frac{\alpha_1(B,0|\boldsymbol{x})/\alpha_1(B,1|\boldsymbol{x})}{\alpha_2(B,0|\boldsymbol{x})/\alpha_2(B,1|\boldsymbol{x})} \right|. \end{aligned} \tag{2.4}$$

This quantity is large when the proportion of probability mass in the two children regions B_0, B_1 is very different between the two groups.

2.2.4 Multivariate and Multi-Sample Generalization

In multi-dimensional settings, there are multiple possible *dimensionwise dyadic partitions* on B, some of which are more effective in capturing the topological features in the underlying distributions than others. Such knowledge is typically unavailable *a priori* and it is desirable to allow the partition tree to adapt to the data structures *a posteriori*. I can achieve this in a principled hierarchical Bayesian manner by placing a prior on the cutting direction (Wong and Ma, 2010).

Before discussing the model, I introduce some notations. In a multivariate setting a *dimensionwise dyadic partition* of $B = [a_1, b_1) \times [a_2, b_2) \times \cdots \times [a_p, b_p) \subset \Omega$ refers to a division of B into two halves by splitting in the middle of the support of one of

the p dimensions. For each B, I use $\{B_0^j, B_1^j\}$ to denote the pair of children nodes of B obtained by cutting B along the jth direction. That is, B_0^j is the half with the jth dimension supported on $[a_j, (a_j + b_j)/2)$, and B_1^j is the half with the jth dimension supported on $[(a_j + b_j)/2, b_j)$.

I consider recursive partition sequences of Ω generated by dimensionwise dyadic partitions. Each such partition sequence can be represented by a bifurcating tree. A node in the tree represents a subset of the sample space Ω, and is obtained from a dyadic partition of its parent node. At the top level of the tree, the root (or level-0) node of the tree is the whole space Ω, while at the first level there are two (level-1) nodes obtained by partitioning Ω along one of the p dimensions. Each level-1 node can be partitioned again, defining the next level of the tree and so on.

Let $\mathcal{B}_p^k$ denote the collection of all level-k nodes under all possible recursive partition sequences of Ω. In other words, it is the set of all possible subsets obtainable by sequentially partitioning the sample space k times. Also, let $\mathcal{B}_p^{(k)} = \cup_{i=0}^k \mathcal{B}_p^i$, the collection of all possible nodes up to level k, $\mathcal{B}_p^{(\infty)} = \cup_{i=0}^\infty \mathcal{B}_p^i$, the collection of all possible nodes, and $\mathcal{B}_p^{(\infty)}(B) = \{C \in \mathcal{B}_p^{(\infty)} \ : \ C \subseteq B\}$, the collection of all possible nodes within B.

Now I can describe the generative model for the pruned tree $\mathcal{T}$. Assume the generative procedure reaches a node $B \in \mathcal{B}_p^{(\infty)}$. If the generative procedure does not stop on B, then I draw a partition direction $J(B)$ according to

$$\Pr\big(J(B) = j | S(B) = 0\big) = \lambda_j,$$

where $\lambda_j \geq 0$ for $j = 1, \ldots, p$ and $\sum_j \lambda_j = 1$. If the procedure stops on B, then I define $J(C) = 0$ for all $C \in \mathcal{B}_p^{(\infty)}(B)$. As shown in Figure 2.2, if $S(B) = 0$ and $J(B) = j$, then I make a dimensionwise dyadic partition in dimension j and obtain two child B_0^j and B_1^j which are added to $\mathcal{T}$, and the procedure proceeds as described earlier. In this case the distribution of the tree depends on the collection

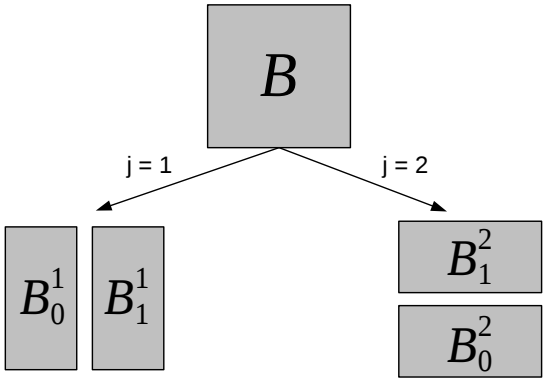

FIGURE 2.2: Recursive partitioning when $\Omega = [0,1)^p$ for $p = 2$. Each region B can be split along p dimensions. The superscripts in the child sets indicate along which direction the parent set was partitioned.

of stopping parameters $\boldsymbol{\eta}$ and the direction probabilities $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$. I say $\mathcal{T} \sim$ Multi-Tree$(\boldsymbol{\eta}, \boldsymbol{\lambda})$. Without prior knowledge about what dimensions are more likely to be involved in characterizing the two distributions, a natural noninformative choice for the direction probabilities is $\lambda_j = 1/p$ for $j = 1, \ldots, p$. Given the pruned tree $\mathcal{T}$, the rest of the model is identical.

The inferential recipe is similar to the univariate case. Lemma 11 provides a recursive representation of the marginal likelihood, while Theorem Theorem 12 establishes the posterior distribution of the ARM-tree.

Lemma 11. *For every $B \in \mathcal{B}_p^{(\infty)}$ and $z \in \mathcal{Z}$, $\Phi_z(\boldsymbol{x}, B)$ has the following recursive representation:*

$$\Phi_z(\boldsymbol{x}, B) = \begin{cases} \sum_{z' \in \mathcal{Z}} \hat{\rho}_{z,z'}(B) \Lambda_{z'}(\boldsymbol{x}, B) & \text{if } B \in \mathcal{B}^{(\infty)} \setminus \Omega \\ \sum_{z' \in \mathcal{Z}} \hat{\pi}_{z'} \Lambda_{z'}(\boldsymbol{x}, \Omega) & \text{if } B = \Omega, \end{cases}$$

where

$$\begin{aligned} \Lambda_z(\boldsymbol{x}, B) &:= \Pr(\boldsymbol{x}(B) | Z(B) = z) \\ &= \begin{cases} \mu(B)^{-\sum_k n_k(B)} & \text{if } z = (1,0) \\ \sum_j \lambda_j \Lambda_{z,j}(\boldsymbol{x}, B) & \text{if } z \neq (1,0). \end{cases} \end{aligned}$$

and

$$\Lambda_{z,j}(\boldsymbol{x}, B) := \Pr(\boldsymbol{x}(B)|Z(B) = z, J(B) = j)$$
$$= \begin{cases} \Pr(\boldsymbol{n}^j(B)|H_0^j(B)) \prod_{i=0,1} \Phi_z(\boldsymbol{x}, B_i^j) & \text{if } z = (0,0) \\ \Pr(\boldsymbol{n}^j(B)|H_1^j(B)) \prod_{i=0,1} \Phi_z(\boldsymbol{x}, B_i^j) & \text{if } z = (0,1), \end{cases}$$

for $\boldsymbol{n}^j(B) = (n_1(B_0^j), n_1(B_1^j), n_2(B_0^j), n_2(B_1^j))$.

Proof. See Appendix A.1. □

Theorem 12. *Suppose I observe two groups of i.i.d. samples* $\boldsymbol{x}_1 = (x_{1,1}, \ldots, x_{1,n_1})$ *and* $\boldsymbol{x}_2 = (x_{2,1}, \ldots, x_{2,n_2})$ *from two distributions* Q_1 *and* Q_2*. If* (Q_1, Q_2) *have an ARM-tree*$(\boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\lambda}, \boldsymbol{\alpha})$ *prior, then the posterior has the following structure:*

1. *Initial probabilities:*

$$\hat{\pi}_{z'}|\boldsymbol{x} = \Pr(Z(\Omega) = z'|\boldsymbol{x}) = \hat{\pi}_{z'} \frac{\Lambda_{z'}(\boldsymbol{x}, \Omega)}{\Phi_z(\boldsymbol{x}, \Omega)}$$

   where $z, z' \in \mathcal{Z}$.

2. *Transition probabilities:*

$$\hat{\rho}_{z,z'}(B|\boldsymbol{x}) = \Pr(Z(B) = z'|Z(\textit{parent}(B)) = z, \boldsymbol{x}) = \hat{\rho}_{z,z'}(B) \frac{\Lambda_{z'}(\boldsymbol{x}, B)}{\Phi_z(\boldsymbol{x}, B)}$$

   where $B \in \mathcal{B}_p^{(\infty)} \setminus \Omega$ *and* $z, z' \in \mathcal{Z}$.

3. *Direction probabilities:*

$$\lambda_j(B, z|\boldsymbol{x}) = \frac{\Lambda_{z,j}(\boldsymbol{x}, B)}{\Lambda_z(\boldsymbol{x}, B)},$$

   where $B \in \mathcal{B}_p^{(\infty)}$, $z' \in \mathcal{Z}$ *and* $j = 1, \ldots, p$.

4. Pseudo-counts:

$$\boldsymbol{\alpha}_0(B|\boldsymbol{x}) = \boldsymbol{\alpha}_0(B) + \boldsymbol{n}_1(B) + \boldsymbol{n}_2(B)$$
$$\boldsymbol{\alpha}_k(B|\boldsymbol{x}) = \boldsymbol{\alpha}_k(B) + \boldsymbol{n}_k(B)$$

for all $B \in \mathcal{B}_p^{(\infty)}$ *and* $k = 1, 2$.

Proof. See Appendix A.1. □

The design of ARM-tree allows comparison across K samples ($K > 2$)—simply let the generative process of ARM-tree generate K distributions $(Q_1, \ldots, Q_K)$ simultaneously by drawing K assignment variables $\theta_1(B), \ldots, \theta_K(B)$ in the probability assignment step. In particular, under $H_0(B)$ the probability assignments are shared across samples:

$$\theta_1(B) = \ldots = \theta_K(B) = \theta_0(B) \sim \text{Beta}(\boldsymbol{\alpha}_0(B)),$$

while under $H_1(B)$ the probability assignments are generated independently:

$$\theta_k(B) \overset{\text{iid}}{\sim} \text{Beta}(\boldsymbol{\alpha}_k(B)), \quad k = 1, \ldots, K.$$

The inferential recipe, computational algorithm, and theoretical results all remain valid.

2.3 Numerical Examples

In this section I provide three numerical examples. The first two are simulated and the last is a real flow cytometry dataset. I compare the performance of ARM-tree for two-sample testing—using $\Pr(H_0|\boldsymbol{x})$ as a test statistic—to that of several other state-of-the-art methods. I then illustrate how to identify and summarize differential structures using the strategy given in Section 2.2.3. The dimensionalities of the three examples are one, two and seven. For all these examples, I set $\pi_0 = 0.7$, $\eta(k) = 0.3$

for $k < 12$ and $\eta(k) = 1$ for $k = 12$, $\beta = 1$, $\gamma = 0.3$ and $\alpha_k(B, 0) = \alpha_k(B, 1) = 0.5$ for all $B \in \mathcal{B}^{(\infty)}$ and $k = 0, 1, 2$. To identify the differential structure I consider the representative partition tree with $\delta^* = 0.8$. I use the range of the data points in each dimension to define the hyper-rectangle Ω.

2.3.1 Example 1

In Example 1 I consider the following two-sample problems in $\mathbb{R}$. For each problem I generate 1,000 datasets, and for each dataset I construct a corresponding "null" dataset by randomly permuting the labels of the two groups.

1. Local shift difference ($n_1 = n_2 = 200$):
   $X_1 \sim 0.9\mathcal{N}(0.2, 0.05^2) + 0.1\mathcal{N}(0.9, 0.01^2)$,
   $X_2 \sim 0.9\mathcal{N}(0.2, 0.05^2) + 0.1\mathcal{N}(0.88, 0.01^2)$.

2. Local dispersion difference ($n_1 = n_2 = 200$):
   $X_1 \sim 0.9\mathcal{N}(0.2, 0.05^2) + 0.1\mathcal{N}(0.8, 0.01^2)$,
   $X_2 \sim 0.9\mathcal{N}(0.2, 0.05^2) + 0.1\mathcal{N}(0.8, 0.04^2)$.

3. Global shift difference ($n_1 = n_2 = 100$):
   $X_1 \sim \mathcal{N}(-0.5, 2^2), \quad X_2 \sim \mathcal{N}(0.5, 2^2)$.

4. Global dispersion difference ($n_1 = n_2 = 50$):
   $X_1 \sim \mathcal{N}(0, 1^2), \quad X_2 \sim \mathcal{N}(0, 2^2)$.

In the first row of Figure 2.3 I plot the pair of density functions for each scenario. In the first two scenarios the difference is located in a small region of the sample, while in the last two the difference is global. In the second row of Figure 2.3 I compare the ROC curves of five different statistics for testing the hypothesis that the two distributions are identical. The other five test statistics are the k-nearest neighbors test (KNN) (Schilling (1986), and Henze (1988)), the Cramér test (Baringhaus and

Franz, 2004), co-OPT (Ma and Wong, 2011), Holmes et al. (2009)'s PT Bayes factor and Chen and Hanson (2014)'s empirical Bayes PT Bayes factor (CH). I use the R package `MTSKNN` (Chen et al., 2010) for the KNN test and the R package `cramer` (Franz, 2006) for the Cramér test.

The ARM-tree substantially outperforms the other methods in the local difference scenarios and behaves robustly in the global difference scenarios. KNN behaves well in the local difference scenarios, but is not efficient in the global scenarios. Cramér, instead, is good in testing large-scale differences, but performs extremely poorly for local differences. I note again that KNN and Cramér tests focus only on the testing of the global null without providing means to summarizing the detected differences.

In Figure 2.4 I illustrate how to identify differences at multiple resolution levels using the posterior representative partition tree $\mathcal{T}^*$ for the local shift difference scenario. On the left plot I use a blue-scale to represent the marginal posterior probability of rejecting the local null hypothesis $\Pr(H_1(B)|\boldsymbol{x})$ on each node of the tree; while on the right plot, I use a red-scale to visualize the effect size. The two distributions share the same normal component on the left, while the difference is located on the right, which is effectively captured in the posterior summary. The spatial clustering of the difference is also reflected in the patch of dark blue and red in the lower right corner of the two plots.

2.3.2 Example 2

I consider the following two-sample problems in $\mathbb{R}^2$. For each problem I again generate 1,000 datasets, and for each dataset I construct a corresponding "null" dataset by randomly permuting the labels of the two groups.

1. Local shift difference ($n_1 = n_2 = 400$):
   $X_1 \sim p_1\mathcal{N}_2(\mu_1, \Sigma_1) + \sum_{k=2}^{5} p_k\mathcal{N}_2(\mu_k, \Sigma_k)$,
   $X_2 \sim p_1\mathcal{N}_2(\mu_1 + \delta, \Sigma_1) + \sum_{k=2}^{5} p_k\mathcal{N}_2(\mu_k, \Sigma_k)$,

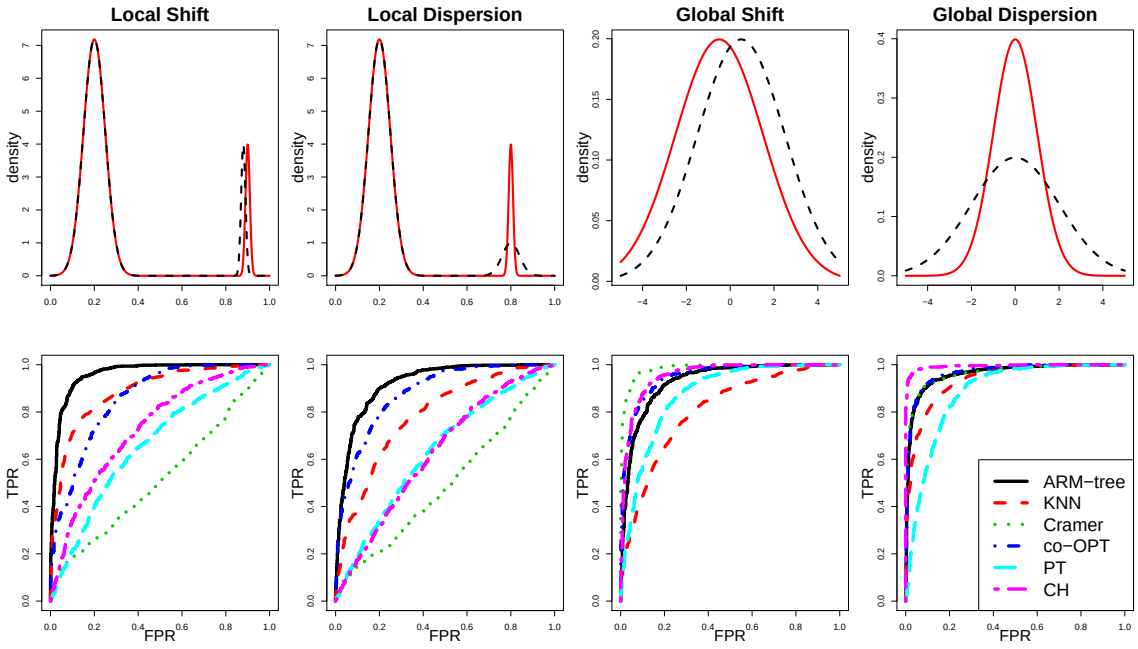

FIGURE 2.3: Two-sample problems in $\mathbb{R}$. First row: densities of the two distributions under the different scenarios - Sample 1 red solid; Sample 2 black dashed. Second row: the ROC curves for each of the testing method considered - ARM-tree black solid; KNN red dash; Cramér green dotted; co-OPT blue dotted dash; PT pale blue long dash; CH pink short-long dash.

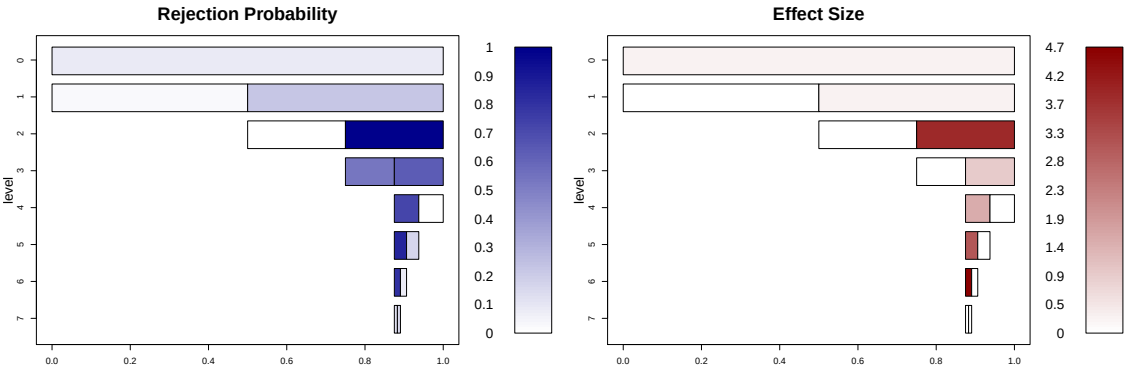

FIGURE 2.4: Nested sequence of partitions for the local shift difference scenario in $\mathbb{R}^1$. On the left, for each region the dark/light blue represents high/low posterior probability of rejecting the local null hypothesis. On the right, for each region the dark/light red represents high/low effect size.

where $\delta = (1, 1)$, while p_k, μ_k and Σ_k are provided in Appendix A.2.

2. Local dispersion difference ($n_1 = n_2 = 400$):
   $X_1 \sim p_1\mathcal{N}_2(\mu_1, \Sigma_1) + \sum_{k=2}^{5} p_k\mathcal{N}_2(\mu_k, \Sigma_k)$,
   $X_2 \sim p_1\mathcal{N}_2(\mu_1, 5\Sigma_1) + \sum_{k=2}^{5} p_k\mathcal{N}_2(\mu_k, \Sigma_k)$,
   where p_k, μ_k and Σ_k are provided in Appendix A.2.

3. Global shift difference ($n_1 = n_2 = 100$): $X_1 \sim \mathcal{N}_2(0, \Sigma)$, $\quad X_2 \sim \mathcal{N}_2(\delta, \Sigma)$, where $\delta = (1, 0)$, while Σ is provided in Appendix A.2.

4. Global dispersion difference ($n_1 = n_2 = 50$): $X_1 \sim \mathcal{N}_2(0, I_2)$, $\quad X_2 \sim \mathcal{N}_2(0, 3I_2)$.

In Figure 2.5 I plot the ROC curve for ARM-tree and the other methods. The results are similar to what I have obtained in the 1D example. The performance of ARM-tree dominates three of the competing methods—KNN, PT and CH—in the sense that ARM-tree is at least as good in all four scenarios. The performance gain over the co-OPT is largely due to incorporation of spatial clustering through Markov dependence, while the gain over the PT is due to both the Markov dependence and the data-adaptive partition sequence. On the other hand, ARM-tree does substantially better than Cramér for local differences, while Cramér is more powerful for global differences.

For a typical dataset in the local shift difference scenario, I illustrate how to identify the differential structure. In Figure 2.6 I plot the regions identified by the representative partition tree $\mathcal{T}^*$. For each region in the partition tree, I show the marginal posterior probability of rejecting the local null hypothesis $\Pr(H_1(B)|\boldsymbol{x})$. Note how the partition sequence adapts to the structure of the data. These regions correctly identify the local differences between the two distributions.

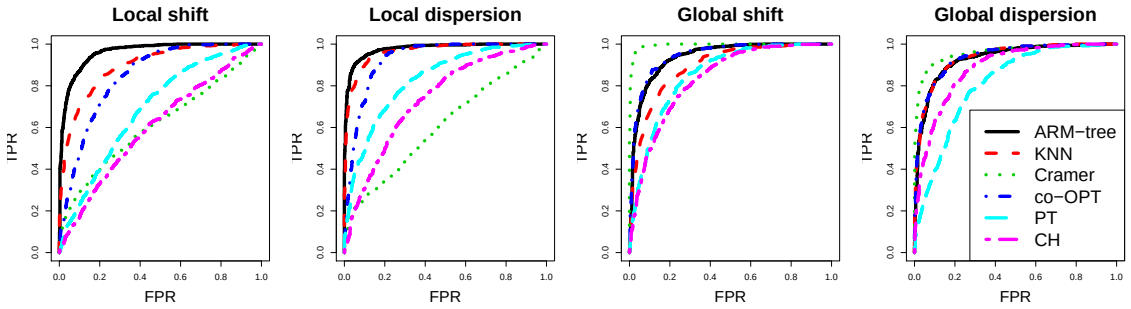

FIGURE 2.5: Two-sample problems in $\mathbb{R}^2$. ROC curves for each of the testing method considered - ARM-tree black solid; KNN red dash; Cramér green dotted; co-OPT blue dotted dash; PT pale blue long dash; CH pink short-long dash.

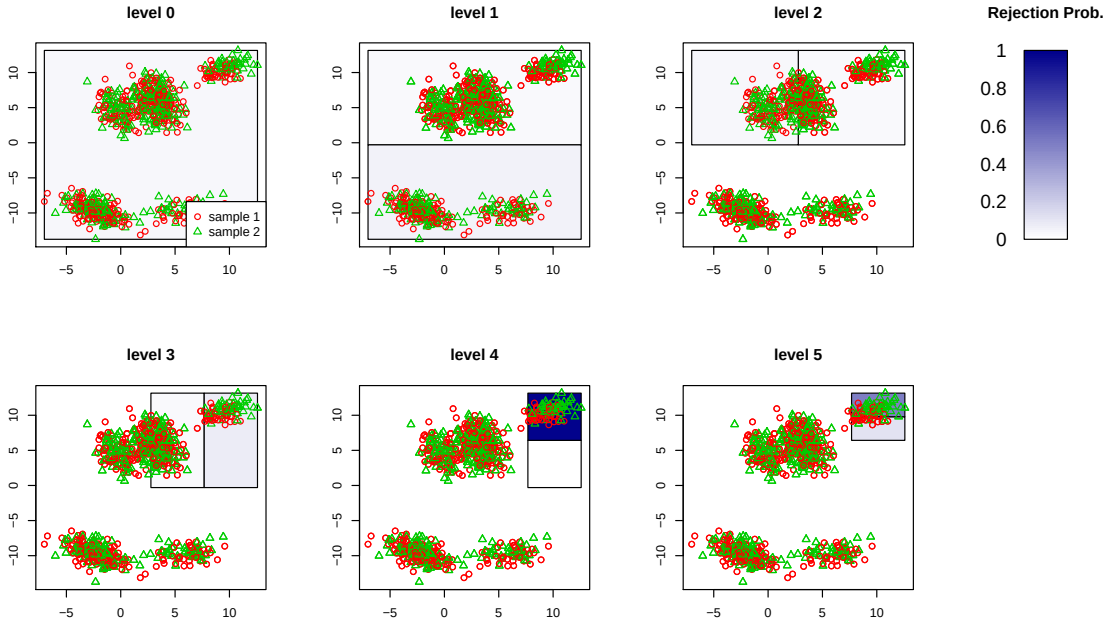

FIGURE 2.6: The representative partition tree for a draw from the local shift difference scenario in $\mathbb{R}^2$. Sample 1 red dots; Sample 2 green triangles. For each region the dark/light blue represents high/low posterior probability of rejecting the local null hypothesis.

2.3.3 A 7-Dimensional Flow Cytometry Dataset

Flow cytometry is a popular laser technology for measuring the protein levels of single cells on thousands of cells simultaneously. In this example I have two blood samples from the same patient, where each sample contains over $300,000$ cells, and for each cell the following 7 markers are measured: FSC-A, FSC-H, SSC-A, Dext, CD4, CD8 and Aqua. In one sample some cells have been transfected via electroporation with a T cell receptor gene specific for Tyrosinase (see Singh et al. (2013) for further details). The dexter conjugated with a fluorescent dye detects Tyrosinase-spefic CD8 T cell receptors, and a higher concentration of transfected cells is expected in the population of cells that are both CD8 and Dext high.

The primary goal of this analysis is to identify cell subpopulations that differentiate the two cell samples. For this reason the existing two-sample tests such as KNN and Cramér are not useful here as one cannot pinpoint the characteristics of the difference using these tests.

I apply the ARM-tree model. The posterior probability of the two samples being equal is virtually 0. Using the representative partition tree (see Figure 2.7) I identify several differential regions with significant differences across the samples. In the representative partition tree the size of each node and the intensity of the color are proportional to the effect size of the associated region. The yellow rectangle highlights the presence of "spatial clustering", a nested sequence of nodes with very large effect sizes. In Figure 2.8 we plot five 2-dimensional projections of the data, highlighting in yellow the differential region with the largest effect size among the regions with $\Pr(H_1(\cdot)|\boldsymbol{x}) > \delta^*$. In particular, $\Pr(H_1(\cdot)|\boldsymbol{x}) = 0.996$ and the effect size is 5.93. In the figures I "smeared" out the cells that do not fall into the detected region. The volume of the region is $1/1024$ of the entire sample space and contains respectively 0.003% and 0.213% of the two cell populations (indicated with red dots

and green triangles). This reported difference indeed involves cells with high CD8 and Dext levels, as is expected based on the scientific knowledge. In Figure 2.8, I mark a probable cluster of transfected cells (green triangles) in the rightmost plot where the data are projected along the CD8 and Dext markers.

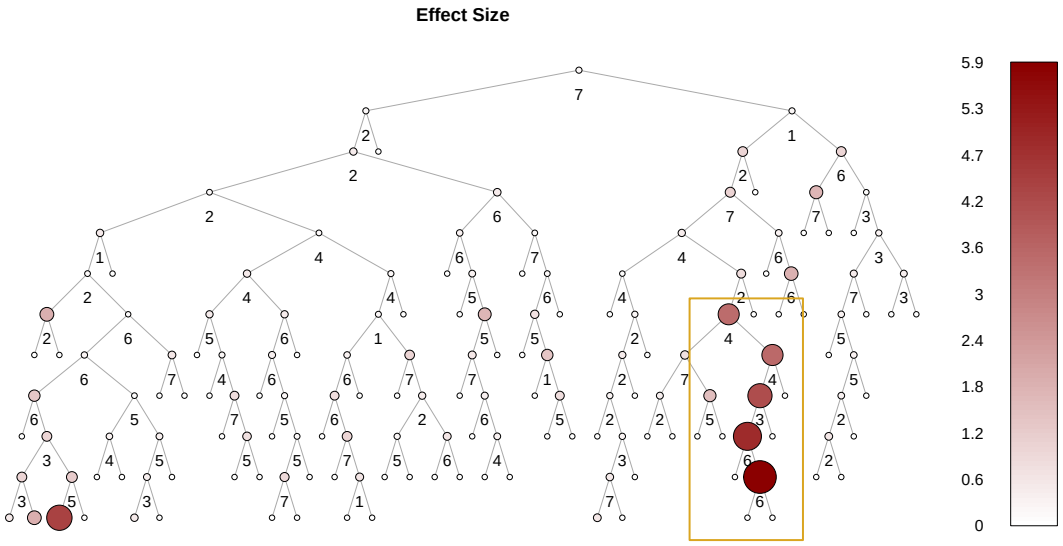

FIGURE 2.7: The representative partition tree for the flow cytometry dataset. The yellow rectangle highlights the presence of "spatial clustering", a nested sequence of nodes with large effect size.

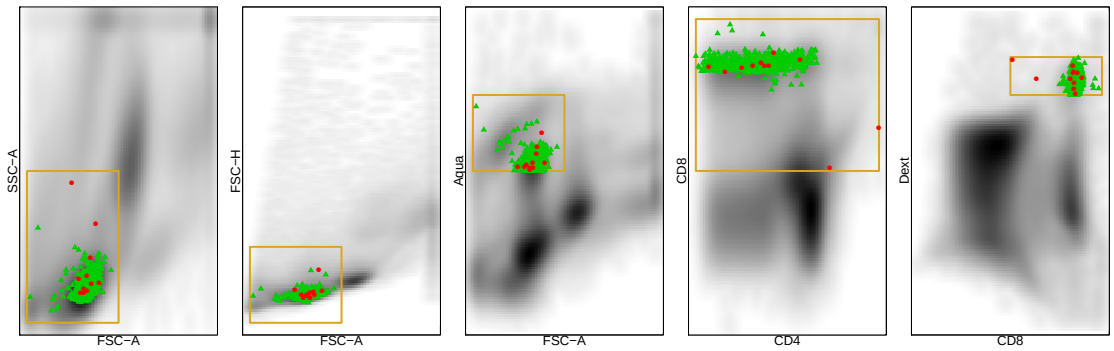

FIGURE 2.8: Five projections of the flow cytometry dataset. For each projection the yellow rectangle highlights the differential region with the largest effect size among the regions with $\Pr(H_1(\cdot)|\boldsymbol{x}) > \delta^*$. The red dots and the green triangles represent respectively the normal and transfected cells within the differential region. In the plot on the far right I observe a cluster of transfected cells (green triangles).

3

Comparison across Mixture Distributions

3.1 Introduction

In recent years there has been growing interest in jointly modeling multiple related distributions. The Bayesian paradigm provides a natural framework to introduce dependency across related distributions through hierarchically linking the priors on the parameters of each distribution, and arbitrary distributions can be effectively approximated by mixing over many simple kernels. A combination of these two strategies has been used by many authors. Lopes et al. (2003) introduced a hierarchical framework to model multiple finite mixture models. Müller et al. (2004) proposed a nonparametric extension of Lopes et al. (2003)'s model replacing finite mixtures with DP mixtures. Cron et al. (2013) proposed the hierarchical DP mixture model, where the mixing measures of the models follow a hierarchical DP (Teh et al., 2006). Griffin et al. (2013), instead, used mixtures of dependent normalized random measures with independent increments, where the dependence is obtained through linear combinations of multiple random measures. The main purpose of all these models is to borrow information across multiple samples to improve inference rather

than to identify differences across distributions. The latter problem is fundamental in many applications. In this work I propose a similar modeling framework where the focus is learning differences across multiple related distributions.

A mixture model is characterized by the number of mixture components, the kernel family and two collections of parameters: the kernel location-scale parameters and the mixture weights associated to each component. I assume that each mixture model shares the same kernel family and the same number of mixture components, while differences across the mixture models can occur in either the location-scale parameters or the mixture weights. The ability to decouple these two sources of variability across the models is of interest in a variety of applied problems.

One example comes from the analysis of flow cytometry experiments in immunology. Flow cytometry is a laser-based technology that measures $p \sim 10$ biomarkers on a large number of cells typically from a blood sample, so each cell is an observation from a distribution in $\mathbb{R}^p$. While mixture modeling has been successfully used to describe single blood samples (Boedigheimer and Ferbas (2008), Chan et al. (2008) and Lo et al. (2008)), there are many open questions in the context of comparison across multiple samples. Identifying variations in the abundance of rare blood cells across multiple samples of the same patient, for instance, can be informative about the evolution of the immune response of the patient. The experimental settings also introduce variability across samples. Small variations in the chemical concentrations across experiments can cause misalignment in the location of the blood cells. The variations in the abundance of rare blood cell populations can be captured by variability in the mixture weights across mixture models. Misalignment in the location of the blood cells can be captured by variations in the location-scale parameters across models. Without separating these two sources of variability, it would be impossible to correctly estimate the variability in the mixture weights.

Current methods impose strong dependency assumptions across the mixture mod-

els which do not allow one to decompose the two sources of variability. In Lopes et al. (2003) and Müller et al. (2004) the location-scale parameters and the mixture weights are either shared across models or completely independent. In Cron et al. (2013) the differences across the models arise only in the mixture weights, while the location-scale parameters are always shared.

In this work I propose a nonparametric hierarchical framework where the modeling of the mixture weights and the location-scale parameters are completely decoupled. Under this framework a subset of the mixture weights are shared across models while the rest are independent. In addition, for each kernel a potential local perturbation in the location parameter allows us to capture misalignment across the samples.

I show that decoupling the mixture weights from the location parameters results in substantially higher power to detect differences across samples in a variety of simulated examples compared to the current methods. I also apply the proposed modeling framework to a flow cytometry dataset in which I am able to detect rare cell subtypes which are differential across samples, after correcting for misalignment across the distributions.

The chapter is organized as follows. In Section 3.2 I introduce the model and provide a recipe for posterior inference through Markov chain Monte Carlo (MCMC). In Section 3.3 I compare my model to current methods and analyze two flow cytometry datasets.

3.2 Method

3.2.1 Comparison across Mixture Models

Assume observations $\boldsymbol{y}_j = (y_{1,j}, \ldots, y_{n_j,j})$ have been collected for $j = 1, \ldots, J$ from J related groups or studies, and the observations of each group are modeled through

a mixture model:

$$y_{i,j} \overset{\text{ind}}{\sim} F_j, \quad i = 1, \dots n_j \quad \text{and} \quad j = 1, \dots, J$$

$$f_j(\cdot) = \sum_{k \in \mathcal{K}} \pi_{j,k} N(\cdot | \mu_{j,k}, \Sigma_{j,k}), \quad j = 1, \dots, J,$$

where f_j denotes the probability density function of F_j. When dealing with observations from related groups, one might expect that the associated distributions would share some common features. To capture these features I link the J models through a hierarchical prior on $\boldsymbol{\pi}_j = (\pi_{j,k} : k \in \mathcal{K})$ and $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})_j = ((\mu_{j,k}, \Sigma_{j,k}) : k \in \mathcal{K})$ for $j = 1, \dots, J$, allowing for ties in the parameters across groups. Unlike previous works, I assume that the ties in the mixture weights and ties in the location-scale parameters are mutually independent. This way one can separate different sources of variability across the models.

To this end I split the index set $\mathcal{K}$ in two disjoint subsets $\mathcal{W}_0$ and $\mathcal{W}_1$. The mixture components in $\mathcal{W}_0$ have the same mixture weights across groups, i.e., $\pi_{j,k} = \pi_{j',k}$ for $j, j' = 1, \dots, J$ and $k \in \mathcal{W}_0$, while the mixture components in $\mathcal{W}_1$ have different mixture weights across groups, i.e., $\pi_{j,k} \neq \pi_{j',k}$ for $j \neq j'$ and $k \in \mathcal{W}_1$. Similarly, I divide the index set $\mathcal{K}$ in two disjoint subsets $\mathcal{P}_0$ and $\mathcal{P}_1$. The subset $\mathcal{P}_0$ contains the indices of the mixture components having the same location-scale parameters across groups, i.e., $(\mu_{j,k}, \Sigma_{j,k}) = (\mu_{j',k}, \Sigma_{j',k})$ for $j, j' = 1, \dots, J$ and $k \in \mathcal{P}_0$, while $\mathcal{P}_1$ is the index set of the mixture component having different location-scale parameters across groups, i.e., $(\mu_{j,k}, \Sigma_{j,k}) \neq (\mu_{j',k}, \Sigma_{j',k})$ for $j \neq j'$ and $k \in \mathcal{P}_1$. As shown in Figure 3.1, I will refer to $\mathcal{W}_1$ as the *weight variation* index set, $\mathcal{P}_1$ as the *local perturbation* index set, and $\mathcal{W}_0 \cap \mathcal{P}_0$ as the *no difference* index set.

Given the defined partitions of the mixture components, I now describe the prior on the mixture weights $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_J)$ and the prior on the location-scale parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_J)$. As far as the mixture weights are concerned, I consider the

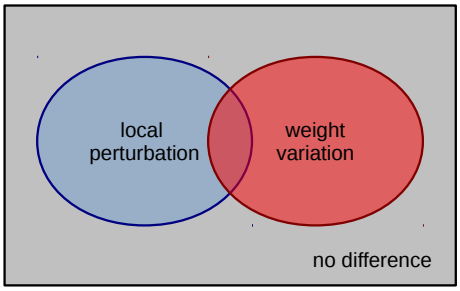

FIGURE 3.1: Venn diagram representing the different types of mixture components.

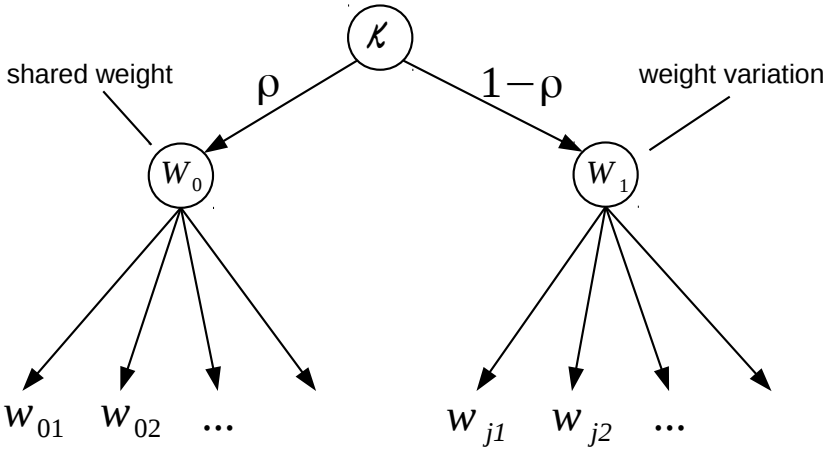

FIGURE 3.2: Hierarchical prior on the mixture weights.

following hierarchical prior:

$$\pi_{j,k} = \begin{cases} \rho w_{0,k} & j = 1, \ldots, J \text{ and } k \in \mathcal{W}_0 \\ (1-\rho) w_{j,k} & j = 1, \ldots, J \text{ and } k \in \mathcal{W}_1, \end{cases}$$

$$\rho \sim \text{Beta}(a_\rho, b_\rho)$$

$$(w_{0,k} : k \in \mathcal{W}_0) \sim \text{Poisson-Dirichlet}(\alpha_0)$$

$$(w_{j,k} : k \in \mathcal{W}_1) \overset{\text{iid}}{\sim} \text{Poisson-Dirichlet}(\alpha_j), \quad j = 1, \ldots, J,$$

where $\mathcal{W}_0$ and $\mathcal{W}_1$ are fixed countably infinite sets. See Figure 3.2 for a visualization of the hierarchical prior on the mixture weights.

For the location-scale parameters, instead, I consider a prior where the index sets $\mathcal{P}_0$ and $\mathcal{P}_1$ are random. As shown in Figure 3.3, for each mixture component the covariance $\Sigma_{j,k} \equiv \Sigma_k$ is shared across the groups, while the mean $\mu_{j,k}$ is either

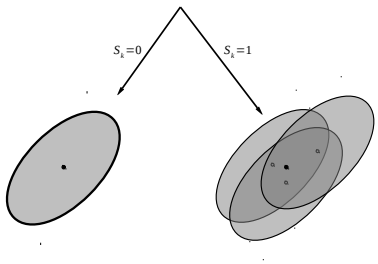

FIGURE 3.3: For each mixture component the covariance is shared across the groups, while the mean is either shared across the groups or centered around a common grand mean.

shared across the groups or centered around a common grand mean $\mu_{0,k}$ based on the partition set of the mixture component:

$$\Sigma_{j,k}^{-1} = \Sigma_k^{-1} \overset{\text{iid}}{\sim} \text{Wishart}(\Psi_1, \nu_1)$$

$$[\mu_{j,k}|\mu_{0,k}, \Sigma_k, \mathcal{P}_0, \mathcal{P}_1] \overset{\text{ind}}{\sim} \delta_{\mu_{0,k}} 1_{\{k\in\mathcal{P}_0\}} + \text{Normal}(\mu_{0,k}, \epsilon\Sigma_k) 1_{\{k\in\mathcal{P}_1\}}$$

$$[\mu_{0,k}|\Sigma_k] \overset{\text{ind}}{\sim} \text{Normal}(m_1, \Sigma_k/k_0)$$

$$\Pr(k \in \mathcal{P}_0|\varphi) = \varphi$$

$$\varphi \sim \text{Beta}(a_\varphi, b_\varphi).$$

The parameter φ controls the assignment of each cluster to $\mathcal{P}_0$ and $\mathcal{P}_1$, while ϵ controls the total amount of local variations between the mean of each group $\mu_{j,k}$ and the grand mean $\mu_{0,k}$ for $k \in \mathcal{P}_1$.

I introduce the following hyperpriors on the hyperparameters of the model:
$\epsilon \sim \text{Uniform}(a_\epsilon, b_\epsilon)$, $m_1 \sim \text{Normal}(m_2, S_2)$, $\Psi_1 \sim \text{Inverse-Wishart}(\Psi_2, \nu_2)$,
$k_0 \sim \text{Gamma}(\tau_1/2, \tau_2/2)$ and $\alpha_j \overset{\text{iid}}{\sim} \text{Gamma}(\tau_{\alpha,1}, \tau_{\alpha,2})$ for $j = 0, \ldots, J$.

3.2.2 Predictive Inference

If $Z_{i,j}$ indicates the latent variable assigning the data point $y_{i,j}$ to a mixture component $k \in \mathcal{K}$, then the likelihood can be written in two stages as follows:

$$[y_{i,j}|Z_{i,j} = k, \mu_{j,k}, \Sigma_{j,k}] \overset{\text{ind}}{\sim} N(y_{i,j}|\mu_{j,k}, \Sigma_{j,k})$$
$$\Pr(Z_{i,j} = k) = \pi_{j,k}.$$

Mixture models have the undesirable property that the likelihood of the model is invariant under permutation of labels of the latent variables. This property makes the mixture weights and location-scale parameters of a mixture models not identifiable, so one cannot make inference on the difference across the distributions directly in the parameter space. However, one can make inference in the space of the observations, where quantities are well identified. In particular, for each observation $y_{i,j}$ one can evaluate the probability of that observation being assigned to a given index set or intersection of index sets. As an example, the prior probability of the data point $y_{i,j}$ belonging to the *weight variation* set, and not to the *location perturbation* set is $E[\mathbb{1}_{\{(Z_{i,j} \in \mathcal{W}_1 \cap \mathcal{P}_0\}}] = E(1-\rho)E(\varphi) = b_\rho/(a_\rho + b_\rho)a_\varphi/(a_\varphi + b_\varphi)$. Given posterior draws of the latent variables $Z_{i,j}^{(1)}, \ldots, Z_{i,j}^{(B)}$, I can easily evaluate the associated posterior probability as follows:

$$E[\mathbb{1}_{\{(Z_{i,j} \in \mathcal{W}_1 \cap \mathcal{P}_0\}}|\boldsymbol{y}] \cong \frac{1}{B}\sum_{b=1}^{B} 1_{\{Z_{i,j}^{(b)} \in \mathcal{W}_1 \cap \mathcal{P}_0^{(b)}\}}, \quad \text{for some large } B.$$

In the flow cytometry application, instead, the interest lies in variations in the abundance of subtypes of blood cells, after controlling for small misalignment in their location across samples. I can identify those cells by looking at the data points $y_{i,j}$ associated to a *weight variation* mixture component independently of their *location perturbation* status, i.e., all the $y_{i,j}$ such that $E[\mathbb{1}_{\{Z_{i,j} \in \mathcal{W}_1\}}|\boldsymbol{y}] > 1 - \epsilon$ for some small $\epsilon > 0$.

3.2.3 Posterior Simulation

Posterior inference is based on MCMC. Exact posterior simulation can be obtained by adapting Müller et al. (2004)'s standard Pólya urn scheme. This scheme is attractive because the random weights are integrated out. However it can be computationally inefficient for large datasets. Alternatively, one can approximate the nonparametric model with a finite model and use a blocked Gibbs sampler (Ishwaran and James, 2001), which is more efficient in term of mixing and computation. In particular, I replace the Poisson-Dirichlet distributions with finite-dimensional Dirichlet distributions:

$$(w_{0,k} : k \in \mathcal{W}_0) \sim \text{Dirichlet}(\alpha_0/K_0)$$

$$(w_{j,k} : k \in \mathcal{W}_1) \overset{\text{iid}}{\sim} \text{Dirichlet}(\alpha_j/K_1), \quad j = 1, \ldots, J,$$

where $K_0 = |\mathcal{W}_0| < \infty$ and $K_1 = |\mathcal{W}_1| < \infty$. This approximation has been studied and used by many authors in a variety of context. See, among others, Neal (2000), Green and Richardson (2001) and Ishwaran and Zarepour (2002). An attractive feature of this approximation over the truncated stick-breaking representation (Ishwaran and James, 2001) is the exchangeability of the random weights. As we will see later in this section, I can leverage this property to improve the mixing of the chain.

Before describing the main steps of the sampling scheme, I introduce the latent perturbation state variable S_k, where $S_k = 0$ if $k \in \mathcal{P}_0$ and $S_k = 1$ if $k \in \mathcal{P}_1$ for $k \in \mathcal{K}$. All of the updates of the hyperparameters (the details are omitted) are conjugate except for $p(\epsilon|\ldots)$ and $p(\alpha_j|\ldots)$, where Metropolis steps are used.

Below I outline the full-conditionals of the sampling scheme:

1. Latent assignments for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$:

$$\Pr(Z_{i,j} = k|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}, y_{i,j}) \propto \pi_{j,k}\text{Normal}(y_{i,j}|\mu_{j,k}, \Sigma_k), \quad k \in \mathcal{K}.$$

2. Mixture weights:

$$[w_{0,1},\ldots,w_{0,K_0}|\boldsymbol{Z},\alpha_0] \sim \text{Dirichlet}(n_{0,1}+\alpha_0/K_0,\ldots,n_{0,K_0}+\alpha_0/K_0)$$

$$[w_{j,1},\ldots,w_{j,K_1}|\boldsymbol{Z},\alpha_j] \overset{\text{ind}}{\sim} \text{Dirichlet}(n_{j,1}+\alpha_j/K_1,\ldots,n_{j,K_1}+\alpha_j/K_1),$$

where $n_{0,k} = |Z_{i,j} = k \ : \ i = 1,\ldots,n_j \text{ and } j = 1,\ldots,J|$ for $k \in \mathcal{W}_0$, and $n_{j,k} = |Z_{i,j} = k \ : \ i = 1,\ldots,n_j|$ for $j = 1,\ldots,J$ and $k \in \mathcal{W}_1$.

3. Latent perturbation state variables for $k \in \mathcal{K}$:

$$\Pr(S_k = 1|\boldsymbol{y},\boldsymbol{Z},\mu_k) = \left(1 + \frac{1-\varphi}{\varphi}\cdot BF_k\right)^{-1},$$

where

$$BF_k = \left(\frac{|\Psi_{1,k}^{(0)}|}{|\Psi_{1,k}^{(1)}|}\right)^{(\nu_1+\sum_j n_{j,k})/2} \prod_j (\epsilon n_{j,k}+1)^{p/2}$$

$$\Psi_{1,k}^{(1)} = \left\{\Psi_1^{-1} + \sum_j \left[SS_{j,k} + \big(\epsilon + \frac{1}{n_{j,k}}\big)^{-1}(\bar{Y}_{j,k}-\mu_k)(\bar{Y}_{j,k}-\mu_k)'\right]\right\}^{-1}$$

$$\Psi_{1,k}^{(0)} = [\Psi_1^{-1} + SS_k + \sum_j n_{j,k}(\bar{Y}_k-\mu_k)(\bar{Y}_k-\mu_k)']^{-1},$$

for $\bar{Y}_{j,k} = \sum_{i:Z_{i,j}=k} Y_{i,j}/n_{j,k}$, $\bar{Y}_k = (\sum_{i,j:Z_{i,j}=k} Y_{i,j})/(\sum_j n_{j,k})$, $SS_{j,k} = \sum_{\{i:Z_{i,j}=k\}}(Y_{i,j}-\bar{Y}_{j,k})(Y_{i,j}-\bar{Y}_{j,k})'$ and $SS_k = \sum_{\{i,j:Z_{i,j}=k\}}(Y_{i,j}-\bar{Y}_k)(Y_{i,j}-\bar{Y}_k)'$.

4. Precision matrices for $k \in \mathcal{K}$:

$$[\Sigma_k^{-1}|\boldsymbol{y},\boldsymbol{Z},S_k,\mu_k] \sim \text{Wishart}\big(\Psi_{1,k}^{(S_k)}, \nu_1 + \sum_j n_{j,k}\big)$$

5. Grand means for $k \in \mathcal{K}$:

$$[\mu_k|\boldsymbol{y},\boldsymbol{Z},S_k,\Sigma_k] \sim \text{Normal}\left(m_{1,k}^{(S_k)}, \Sigma_k/(\sum_j(\epsilon S_k + 1/n_{j,k})^{-1} + k_0)\right),$$

6. Group means for $j = 1, \ldots, J$ and $k \in \mathcal{K}$:

$$[\mu_{j,k}|\mu_k, S_k = 0] \sim \delta_{\mu_k}$$

$$[\mu_{j,k}|\boldsymbol{y}, \boldsymbol{Z}, \mu_k, \Sigma_k, S_k = 1] \sim \text{Normal}\left(\frac{n_{j,k}\bar{Y}_{j,k} + \mu_k/\epsilon}{n_{j,k} + 1/\epsilon}, \Sigma_k/(n_{j,k} + 1/\epsilon)\right).$$

The sampler described above can easily get trapped in some local mode. In particular, a mixture component identical across the J groups can be equally described by a single component with shared mixture weights (1 parameter) or a single component with independent mixture weights (J parameters). Assuming an equal fit to the data, the posterior will favor the simpler model (Jefferys and Berger, 1992), but such a posterior will be highly multi-modal. I introduce a Metropolis step to explore different modes of the posterior distribution by swapping an atom from $\mathcal{W}_0$ with an atom from $\mathcal{W}_1$. Similar strategies to improve the exploration of the sample space have been proposed by Porteous et al. (2012) and Papaspiliopoulos and Roberts (2008).

The proposal distribution is defined as follows. An initial index k' is drawn proportionally to $\sqrt{n_{j,k}}$ for $k \in \mathcal{K}$, where $n_{j,k} = |(i,j) : Z_{i,j} = k|$, and a second index k'' is drawn uniformly from $\mathcal{W}_0$ if $k' \in \mathcal{W}_1$ and uniformly from $\mathcal{W}_1$ if $k' \in \mathcal{W}_0$. Since the proposal is symmetric, the swap is accepted with probability:

$$\min\left(\frac{E_{w,\rho}(\prod_{j,k} \pi_{j,k}^{n_{j,k}}|\boldsymbol{Z}_{\text{new}})}{E_{w,\rho}(\prod_{j,k} \pi_{j,k}^{n_{j,k}}|\boldsymbol{Z})}, 1\right),$$

where $\boldsymbol{Z}_{\text{new}}$ represents the vector of the latent assignments after the swap. Since the mixture components are exchangeable within $\mathcal{W}_0$ and $\mathcal{W}_1$, the acceptance probability depends only on the swapped indices. If I instead use the truncation of the stick breaking process, then the acceptance probability would depend not only on the swapped indices, but also on the indices order within $\mathcal{W}_0$ and $\mathcal{W}_1$ since the mixture

weights would be stochastically ordered. Thus, the choice of type of truncation of the DP is motivated by the improved mixing of the chain.

3.3 Numerical Examples

In this section I provide three numerical examples. In the first example data are simulated under different scenarios, and I compare the performance of my model to other competing methods in term of multi-sample comparison. In the second example I illustrate through a simulated dataset how to identify the different sources of variation across multiple distributions. In the third example I analyze two real flow cytometry datasets.

3.3.1 Example 1

In the first example I consider the following multi-sample problems in $\mathbb{R}^4$. For each problem I generate 100 datasets, and for each of them I generate a corresponding "null" distribution by a random permutation of the group labels. Each problem has three groups and the sample size for each group is 100.

Below I outline the four different problems. For brevity, some of the parameters are omitted here and are instead provided in the Appendix.

1. Single location shift:

$$y_{i,j}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi} \sim \pi_1 N(y_{i,j}|\mu_1 + \delta_j, \Sigma_1) + \sum_{k=2}^{4} \pi_k N(y_{i,j}|\mu_k, \Sigma_k),$$

   where $\delta_j = (j/2, 0, 0, 0)$ for $j = 1, \ldots, 3$, and $\mu_k \sim U(0, 10)$ for $k = 1, \ldots, 4$.

2. Multiple location shifts:

$$y_{i,j}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi} \sim \sum_{k=1}^{4} \pi_k N(y_{i,j}|\mu_k + \frac{j}{10}\mathbb{1}_4, \Sigma_k),$$

where $\mu_k \sim U(0, 10)$ for $k = 1, \ldots, 4$.

3. Local weight difference:

$$y_{i,j}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi} \sim (\pi_1 - 0.04(j-1))N(y_{i,j}|\mu_1, \Sigma_1) + (\pi_2 + 0.04(j-1))N(y_{i,j}|\mu_2, \Sigma_2) + \sum_{k=3}^{4} \pi_k N(y_{i,j}|\mu_k, \Sigma_k), \quad (3.1)$$

where $\boldsymbol{\pi} = (0.09, 0.01, 0.8, 0.1)$ and $\mu_k \sim U(0, 10)$ for $k = 1, \ldots, 4$.

4. Global weight differences:

$$y_{i,j}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi} \sim \sum_{k=1}^{8} \pi_{j,k} N(y_{i,j}|\mu_k, \Sigma_k)$$

$$\pi_j \propto \exp(m_j)$$

$$m_j \sim N(0, S),$$

where $\mu_k \sim U(0, 10)$ for $k = 1, \ldots, 8$.

In Figure 3.4 I compare the ROC curves of three different statistics for testing the hypothesis that the three distributions are identical. For my model I use $E(\rho\varphi|\boldsymbol{y})$ as the test statistic. This quantity represents the proportion of mixture weights associated to mixture components where both mixture weights and location parameters are shared across groups. The other methods are Müller et al. (2004) and Cron et al. (2013). In Müller et al. (2004)'s model each F_j is defined as a mixture of two components: $F_j = \epsilon H_0 + (1 - \epsilon)H_j$ for $j = 1, \ldots, J$. The distribution H_0 represents the common part, and H_j represents the idiosyncratic part. The hyperparameter ϵ controlling the "degree of similarity" across the F_j's has a beta hyperprior. I use $E(\epsilon|\boldsymbol{y})$ as the test statistic. In Cron et al. (2013)'s model each f_j is defined as follows: $f_j(\cdot) = \int g(\cdot|\theta)Q_j(d\theta)$ for $j = 1, \ldots, J$. The mixing measures have the following distributions $Q_j|Q_0 \overset{\text{iid}}{\sim} DP(\alpha_0, Q_0)$ and $Q_0 \sim DP(\alpha, H)$ for some distribution H on the

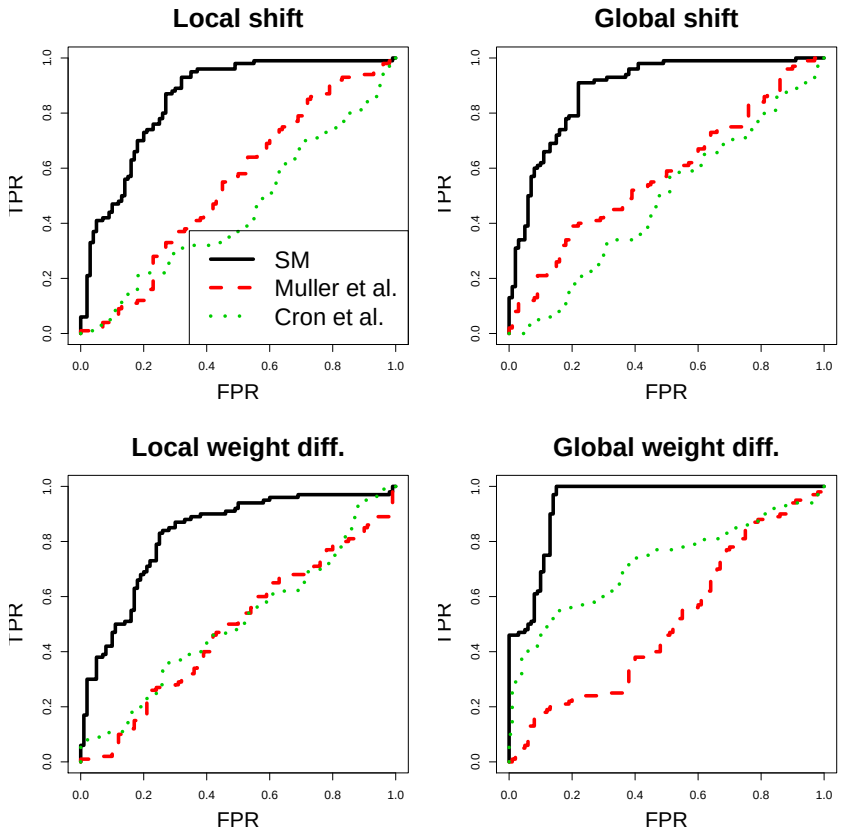

FIGURE 3.4: Multi-sample problems from Example 1. The ROC curve for each of the testing method considered - my method in black solid; Cron et al. (2013)'s method in green dotted; Müller et al. (2004)'s method in red dash.

parameter space. The hyperparameter α_0 controlling the degree of variability across the mixing measure Q_j's has a gamma hyperprior. I use $E(\alpha_0|\boldsymbol{y})$ as the test statistic. I use the R package `DPpackage` (Jara et al., 2011) for fitting Müller et al. (2004)'s model, and the Python package `dpmix` (Cron and Frelinger, 2013) for fitting Cron et al. (2013)'s model. I run each model for 1000 iterations after a 5000 iteration burn-in period. I consider 10 mixture components for Cron et al. (2013)'s model, and $K_0 = K_1 = 10$ for my model. My method outperforms the other methods in all four scenarios. Even though the global weight differences problem should represent the best case scenario for Cron et al. (2013)'s model, the model has limited discriminatory power compared to mine.

3.3.2 Example 2

I consider a numerical example based on mixtures of normals in $\mathbb{R}^4$ to illustrate how one can classify each observation based on the four types of variations across the related distributions (see Figure 3.1). The data are generated as follows:

$$y_{i,1} \sim 0.08N(\mu_{1,1}, I) + 0.9N(\mu_2, 2I) + 0.01N(\mu_3, 0.2I) + 0.01N(\mu_{1,4}, 0.1I)$$

$$y_{i,2} \sim 0.06N(\mu_{2,1}, I) + 0.9N(\mu_2, 2I) + 0.03N(\mu_3, 0.2I) + 0.01N(\mu_{2,4}, 0.1I)$$

$$y_{i,3} \sim 0.04N(\mu_{3,1}, I) + 0.9N(\mu_2, 2I) + 0.05N(\mu_3, 0.2I) + 0.01N(\mu_{3,4}, 0.1I),$$

where $i = 1, \ldots, 1000$, $\mu_{j,1} = (1, 9-j, 1, 8)$, $\mu_2 = (8, 8, 8, 8)$, $\mu_3 = (1, 1, 1, 1)$ and $\mu_{j,4} = (8, j, 8, 1)$. The three plots in the first row of Figure 3.5 show the data projected along the first two dimensions for each of the three distributions. In the second row the three plots show the data points classified based on the most likely type of variation a posteriori. The model correctly identifies the source of variation in each of the four mixture component of the model.

3.3.3 Flow Cytometry

In flow cytometry experiments biomarkers are measured on a large number of blood cells. Different cell subtypes, i.e., groups of cells sharing similar biomarker levels, have distinct functions in the human immune system. Identifying variations in the abundance of subtypes across multiple samples is an important immunological question. The differences across the samples can be minimal, because only a small fraction of the cells ($\lesssim 0.1\%$) is generally involved in the difference. Additionally, the location of a given subtype across samples can slightly change due to experimental variability.

I analyze two datasets where each one contains three samples of 5000 blood cells, and for each cell 6 biomarkers are measured.

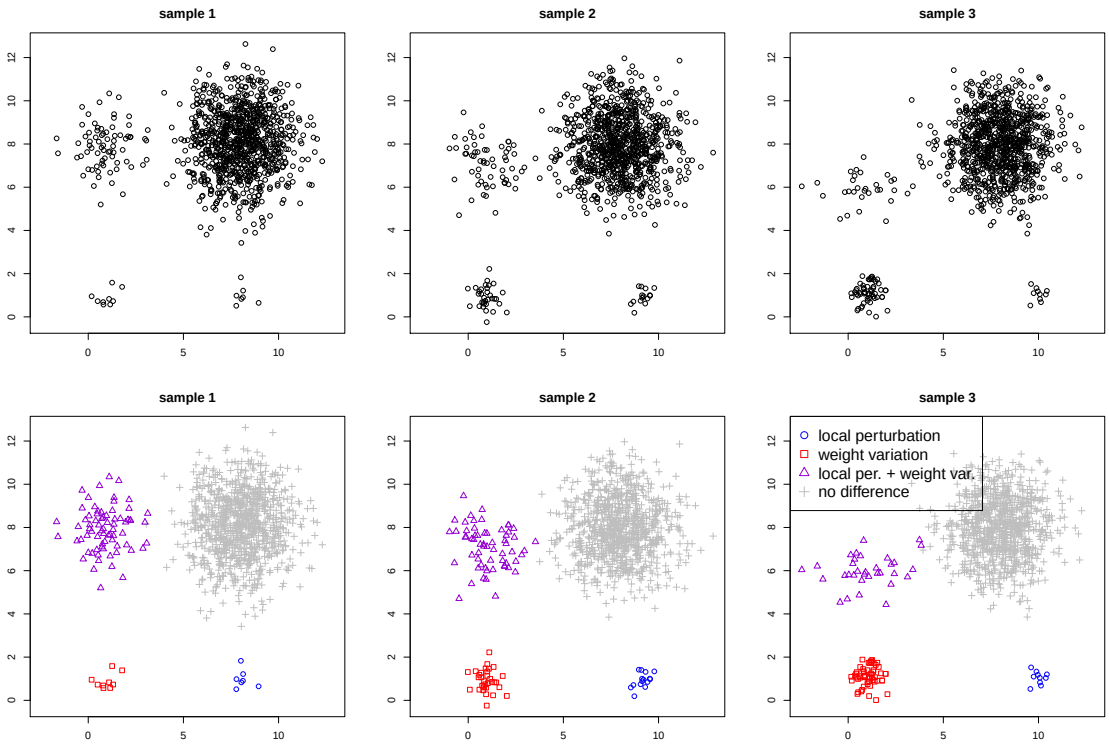

FIGURE 3.5: The three plots in the first row show the data from Example 2 projected along the first two dimensions for each of the three distributions. In the second row the three plots show the data points classified based on the most likely type of variation a posteriori.

Control Study

The blood from a given patient was split in three samples, and each sample was analyzed separately. Since the three samples came from the same patient one would expect to observe experimental variability in the locations of the subtypes, but no variations in the abundance of the different subtypes.

In Figure 3.6 I plot the posterior distributions of ρ, φ and ϵ. I recall that the parameter ρ controls the total mass assigned to mixture components where the mixture weights are shared across the groups. In this dataset the parameter concentrates around one. Thus, there is no evidence of a difference in the mixture weights across the three replicates. The parameter φ controls the fraction of mixture components

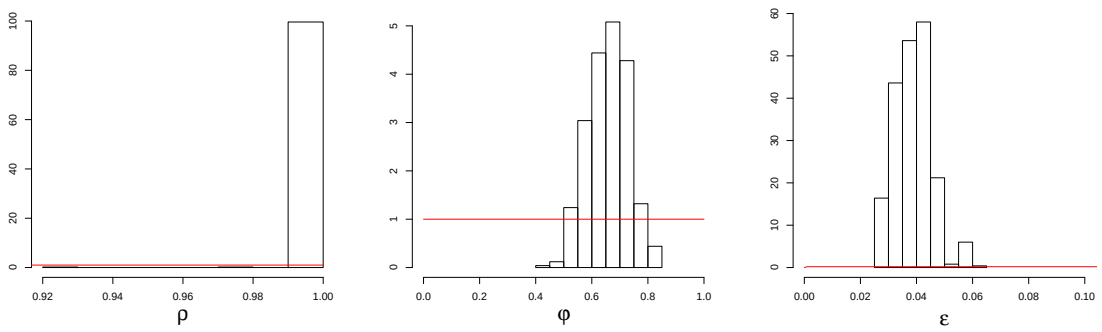

FIGURE 3.6: Histograms of the posterior of ρ, φ and ϵ for the flow cytometry control study. The red lines represent the prior distributions.

where the location of the mixture components is identical across replicates. This parameter does not concentrate around one, implying that the model is detecting across-replicate misalignment in some of the mixture components. Finally, the parameter ϵ concentrates around a very small value, indicating that the misalignment is minimal. This control study shows the importance of allowing for small variations in the mean parameter across the replicates. Additionally, decoupling the two sources of variations allows one to correctly estimate the absence of variations in the mixture weights across the distributions of the three samples.

Different Stimulation Conditions

The blood from a given patient was split in three samples. One sample was left unstimulated, while the two remaining samples were stimulated with CEF and CMV pp65, respectively. Under these stimulation conditions I expect a higher concentration of cells with high values of both the CD4 and the CD8 biomarkers.

I compare two different models. The first one is the model described in Section 3.2. The second one is the same model but without local perturbation in the location of the means across groups, i.e., φ is not random, but fixed equal to 1.

In Figure 3.7 I plot the posterior distributions of ρ, φ and ϵ under the first model.

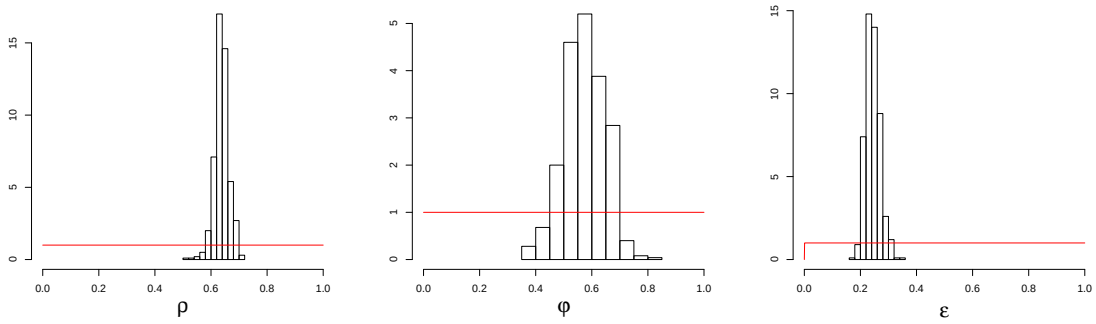

FIGURE 3.7: Histograms of the posterior distributions of ρ, φ and ϵ. The red lines represent the prior distributions.

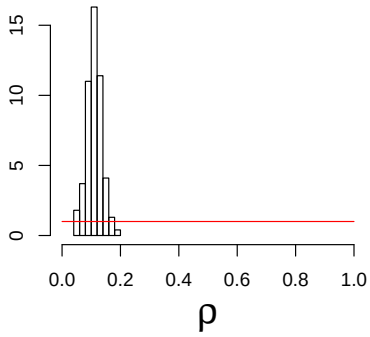

FIGURE 3.8: Histogram of the posterior distribution of ρ when φ is fixed equal to one. The red line represents the prior distribution.

The parameter ρ concentrates around 0.6, indicating that there are differences in the mixture weights across the three samples. Similar to the previous dataset, the parameter φ concentrates around 0.6, indicating the presence of misalignment across at least one of the mixture components. From the last plot one can see that the degree of misalignment is larger in this dataset than in the previous one.

In Figure 3.8 I plot a histogram of posterior draws of ρ from the second model where $\varphi = 1$, i.e. the model that does not allow for small variations in the location of the means across groups. Under this model the estimate of ρ is much smaller. Subtypes of cells misaligned across the samples are described by independent mixture

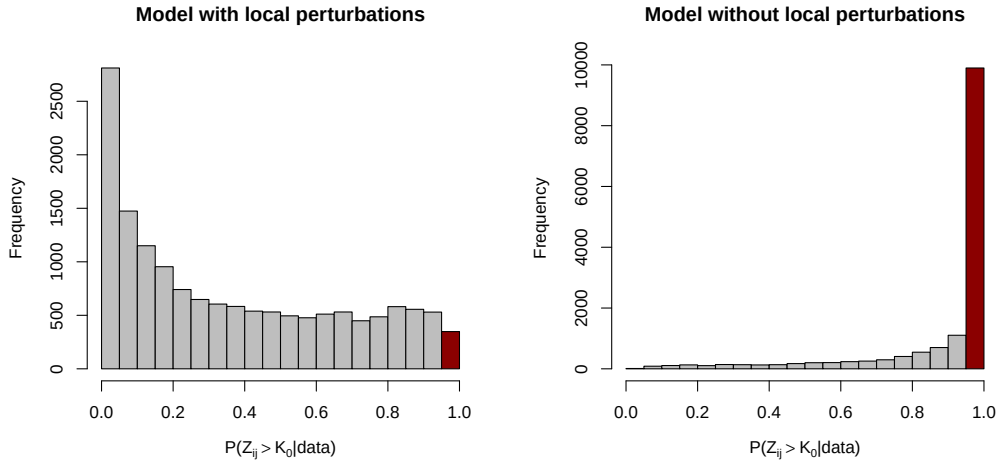

FIGURE 3.9: Histograms of the marginal posterior probabilities of each observations belonging to the *weight variation* index set. On the left when the model allows for local shift, and on the right when the model does not include local shift. In yellow I highlight the points with probability higher than 0.95.

components instead of a shared kernel with perturbation in the location parameter. Thus, the mixture weights of the independent mixture components are very different across groups.

For each observation $y_{i,j}$ one can estimate $E(\mathbb{1}_{\{Z_{i,j}\in\mathcal{W}_1\}}|\boldsymbol{y})$, i.e., the marginal posterior probability of belonging to a mixture component with different mixture weights across the three groups. In Figure 3.9 I plot the histograms of these marginal posterior probabilities. Under the model allowing for local shifts most of the observations are associated to mixture components that have identical mixture weights across groups, while in the model without local shifts most of the observations fall in the opposite category. In Figure 3.10 I show scatter plots of the data for each group, and highlight in color the data points with probability of belonging to the *weight variation* index set higher than 0.95.

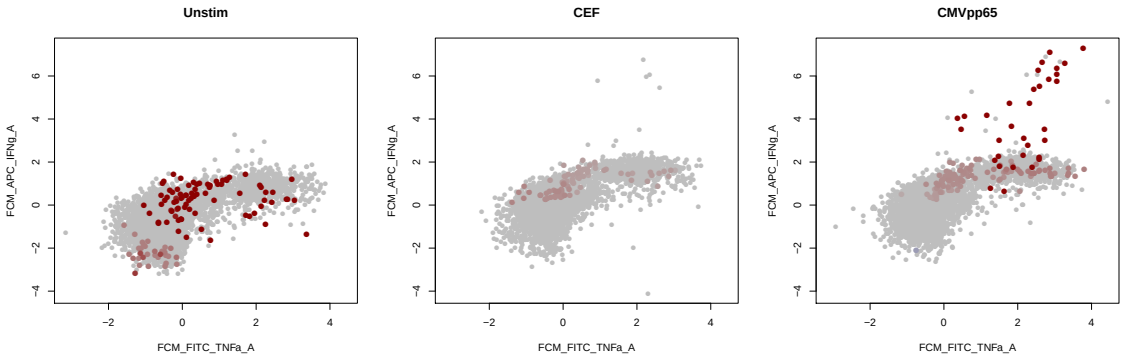

FIGURE 3.10: Scatter plots of the data for each group. I highlight the data points with marginal posterior probabilities of belonging to the *weight variation* index set higher than 0.95.

4

Functional Comparison Using Wavelets

4.1 Introduction

With modern recording equipments patients or "units" can be monitored over time at very high frequency. Although observations are collected only at a fine grid, one can consider the ideal observations to be curves. Then, one can leverage the smoothness properties of the functional space the curve lives in to borrow information between the measurements within each curve. In many applications one is interested in comparing the distribution of such observations across multiple groups of patients. This problem is generally called functional ANOVA since the ideal observations are functions.

A simple approach to this problem is to carry out an ANOVA test – e.g. an F-test – for each time point. This approach divides the global ANOVA problem into testing a collection of local hypotheses, one for each time point. Multiple testing correction must be applied when setting the threshold for determining statistical significance of the differences between the groups. This approach does not take advantage of the time-dependency between the measurements within curves. Thus, when the number

of time points is very large, the multiplicity correction generally washes out the signal.

An alternative approach, to be adopted in this work, is first to apply a basis transformation and then carry out ANOVA in the new basis. In particular, I consider ANOVA comparison under wavelet bases. This approach divides the global ANOVA problem into multiple tests, but now each local test is on the wavelet coefficients associated to a particular time and frequency combination in the wavelet domain. This alternative decomposition is desirable for several reasons. One benefit of the wavelet approach is that it has a "whitening effect" that reduces the correlation in the errors, making the ANOVA model with independent errors more reasonable. A second benefit is that it can substantially improve the efficiency of the individual local tests by concentrating the cross group differences into a small number of location-scale nodes.

Multiple authors have used wavelet decomposition for functional ANOVA problems. Rosner and Vidakovic (2000), Abramovich et al. (2004) and Abramovich and Angelini (2006) proposed frequentist procedures for testing in functional ANOVA. These procedures rely on strong assumptions on the error structure and do not provide quantification of uncertainty in estimation. Morris and Carroll (2006) proposed a fully Bayesian framework for functional mixed-effects models overcoming those limitations. Their model accounts for heterogeneous errors, and uncertainty quantification can be naturally achieved through draws from the posterior distribution. However, it is computationally very intensive and does not include testing.

I propose a Bayesian framework for both testing and uncertainty quantification, while being computationally efficient. In this framework most of the parameters are analytically integrated out, and a few hyperparameters are chosen through an empirical Bayes procedure. Additionally, the proposed model accounts for heterogeneity in the error structure similarly to Morris and Carroll (2006)'s model.

This method uses spike-and-slab priors on the wavelet coefficients associated to a particular time and frequency combination in the wavelet domain to locally test for cross group differences. Borrowing of information across local tests is particularly useful in this framework since the signal concentrates into a small number of nearby location-scale nodes. Thus, I introduce dependency across the local tests through a HMT model (Crouse et al., 1998).

The chapter is organized as follows: In Section 4.2.1 I describe the model in the case of a single functional observation, and in Section 4.2.2 I provide a recipe for posterior inference. In Section 4.2.3 and Section 4.2.5 I extend the model to one-way and multi-way functional ANOVA, respectively. Finally, in Section 4.3 I provide a few simulated numerical examples and the analysis of a dataset arising from physiology.

4.2 Method

4.2.1 The NIG-HMT Model

Suppose one has a single functional observation whose values are attained at T equidistant time points $\boldsymbol{y} = (y_1, \ldots, y_T)$, and

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{f} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(\mathbf{0}, \Sigma_\epsilon), \end{aligned} \tag{4.1}$$

where $\Sigma_\epsilon = \text{diag}(\sigma_1^2, \ldots, \sigma_T^2)$. I wish to recover the unknown function $\boldsymbol{f}$ from the noisy observation $\boldsymbol{y}$ without making parametric assumptions on the structure of $\boldsymbol{f}$. After applying the discrete wavelet transform (DWT) to $\boldsymbol{y}$ I obtain:

$$\boldsymbol{d} = \boldsymbol{z} + \boldsymbol{u},$$

where $\boldsymbol{d} = \boldsymbol{y}W'$, $\boldsymbol{z} = \boldsymbol{f}W'$ and $\boldsymbol{u} = \boldsymbol{\epsilon}W'$ with W being the orthonormal matrix corresponding to the chosen wavelet basis. Due to properties of multivariate Gaussians, $\boldsymbol{u}$ is also a multivariate Gaussian with a diagonal covariance matrix.

The $\boldsymbol{d}$s are referred to as the wavelet coefficients for the observations $\boldsymbol{y}$, and $\boldsymbol{z}$s are the wavelet coefficients of its mean function $\boldsymbol{f}$. The multi-resolution nature of wavelet series allows the components of the corresponding vectors $\boldsymbol{d}, \boldsymbol{z}$ and $\boldsymbol{u}$ to be organized in a bifurcating tree structure, with each element in the corresponding vector associated to a node in the tree. I use the pair of indices (j, k), where $j = 0, \ldots, J = \log_2 T - 1$ and $k = 0, \ldots, 2^j - 1$, to represent the kth node in the jth level of this tree. The two child nodes of node (j, k) are indexed by $(j + 1, 2k)$ and $(j + 1, 2k + 1)$. Correspondingly, for $j > 1$ the parent of (j, k) is indexed by $(j - 1, \lfloor k/2 \rfloor)$. From now on $\mathcal{T}$ is used to denote the collection of indices (j, k) corresponding to all nodes in the bifurcating tree.

The model can be written in a node-specific manner:

$$d_{j,k} = z_{j,k} + u_{j,k} \quad \text{where} \quad u_{j,k} \sim N(0, \sigma^2_{j,k}).$$

In Bayesian wavelet shrinkage, one often adopts a spike-and-slap prior on $z_{j,k}$:

$$z_{j,k} \sim (1 - \pi_{j,k}) \cdot \delta(0) + \pi_{j,k} \cdot N(0, \tau_j \sigma^2_{j,k}),$$

where the hyperparameter τ_j specifies the amount of shrinkage at level j. Specifically, I consider the following parametric structure:

$$\tau_j = 2^{-\alpha j} \tau,$$

for some $\alpha, \tau > 0$. This implies that the wavelet coefficients tend to be smaller as the level j increases. The parameter α controls the rate of shrinkage and thereby controls the smoothness of the functional observations. These hyperparameters can be fixed a priori or chosen through an empirical Bayes approach.

The mixture structure allows the shrinkage to be node-specific and data-adaptive. The same prior can be written hierarchically with the introduction of a hidden shrink-

age state $S_{j,k} \in \mathcal{S} = \{0, 1\}$:

$$z_{j,k}|S_{j,k} \sim \begin{cases} \delta(0) & \text{if } S_{j,k} = 0 \\ N(0, \tau_j \sigma_{j,k}^2) & \text{if } S_{j,k} = 1. \end{cases}$$
$$S_{j,k} \sim \text{Bernoulli}(\pi_{j,k}). \tag{4.2}$$

The error variance is typically unknown, and it can be inferred from the data. In this work I introduce a prior on $\sigma_{j,k}^2$:

$$\sigma_{j,k}^2 \sim \text{Inv-Gamma}(\nu + 1, \nu\sigma_0^2), \tag{4.3}$$

where the hyperparameters ν and σ_0^2 can be either set a priori or determined by an empirical Bayes approach. The inverse-Gamma prior maintains conjugacy to the Gaussian model, and consequently the marginal likelihood can be evaluated analytically.

Note that as $\nu \to \infty$, $\sigma_{j,k}^2 \xrightarrow{p} \sigma_0^2$. This limit corresponds to the case where all the error variances are set to be equal to σ_0^2. In certain applications where homogeneous error variances are reasonable, this will be an appropriate model specification. In other applications where the error variance might be heterogeneous, it might still be advisable to adopt this choice when there is a small number of observations, since the node-specific variance is only weakly identifiable in this case. When the number of observations is sufficiently large one can infer node-specific error variances.

Donoho and Johnstone (1994) noted a prevalent phenomenon in many applications of inference in wavelet spaces: the "signals" – the wavelet coefficients that are large in magnitude – typically cluster in both location and scale. This phenomenon, for instance, can be clearly seen in the four test functions presented in Donoho and Johnstone (1994) (see Figure 4.1). Thus, one expects that if the coefficient $z_{j,k}$ deviates far away from zero, then the coefficients on neighboring nodes in the bifurcating location-scale tree are more likely to be deviating from zero as well. Consequently,

the appropriate amount of shrinkage for the different nodes are correlated to a varying extent depending on how close they are in the location-scale tree. To incorporate such a dependency into inference and thereby allow for effective borrowing of information in the shrinkage, Crouse et al. (1998) proposed linking the shrinkage states $S_{j,k}$ using a Markov process, resulting in a hidden Markov model evolving on the location-scale tree, called the *hidden Markov tree* (HMT).

Under the HMT, the shrinkage state of node (j,k) can depend on that of its parent:

$$\Pr(S_{j,k} = s | S_{j-1,\lfloor k/2 \rfloor} = r) = \rho_{j,k}(r,s), \tag{4.4}$$

where $\rho_{j,k}(r,s)$ is called the transition probability. Because the root node $(0,0)$ does not have a parent, the initial state of the process $S_{0,0}$ is specified by a set of initial state probabilities $\boldsymbol{\rho}_{0,0} = (\rho_{0,0}(0), \rho_{0,0}(1))$ such that:

$$\Pr(S_{0,0} = s) = \rho_{0,0}(s) \quad \text{for } s \in \mathcal{S}. \tag{4.5}$$

From now on I will refer to Eqs. (4.2), (4.3), (4.4) and (4.5) as the normal inverse-Gamma hidden Markov tree, or NIG-HMT, model.

The transition probabilities can be organized into a transition probability matrix $\boldsymbol{\rho}_{j,k}$. For simplicity I assume the following parametric structure for the transition probability matrix:

$$\boldsymbol{\rho}_{j,k} = \begin{bmatrix} 1 - \min(1, \eta 2^{-\beta j}) & \min(1, \eta 2^{-\beta j}) \\ (1-\gamma) - \min(1-\gamma, \eta 2^{-\beta j}) & \gamma + \min(1-\gamma, \eta 2^{-\beta j}) \end{bmatrix}, \tag{4.6}$$

where $0 \leq \gamma \leq 1$, $\beta \geq 0$ and $\eta > 0$. The parameter γ controls for the clustering of the wavelet signal. For large γ, if the hidden process is in the *large variance* state at a given node, then it is more likely to be in the same state at the child nodes. For $\gamma = 0$ there is no dependence between the nodes and the model proposed by Abramovich et al. (1998) is recovered. The parameter η controls the number of zero

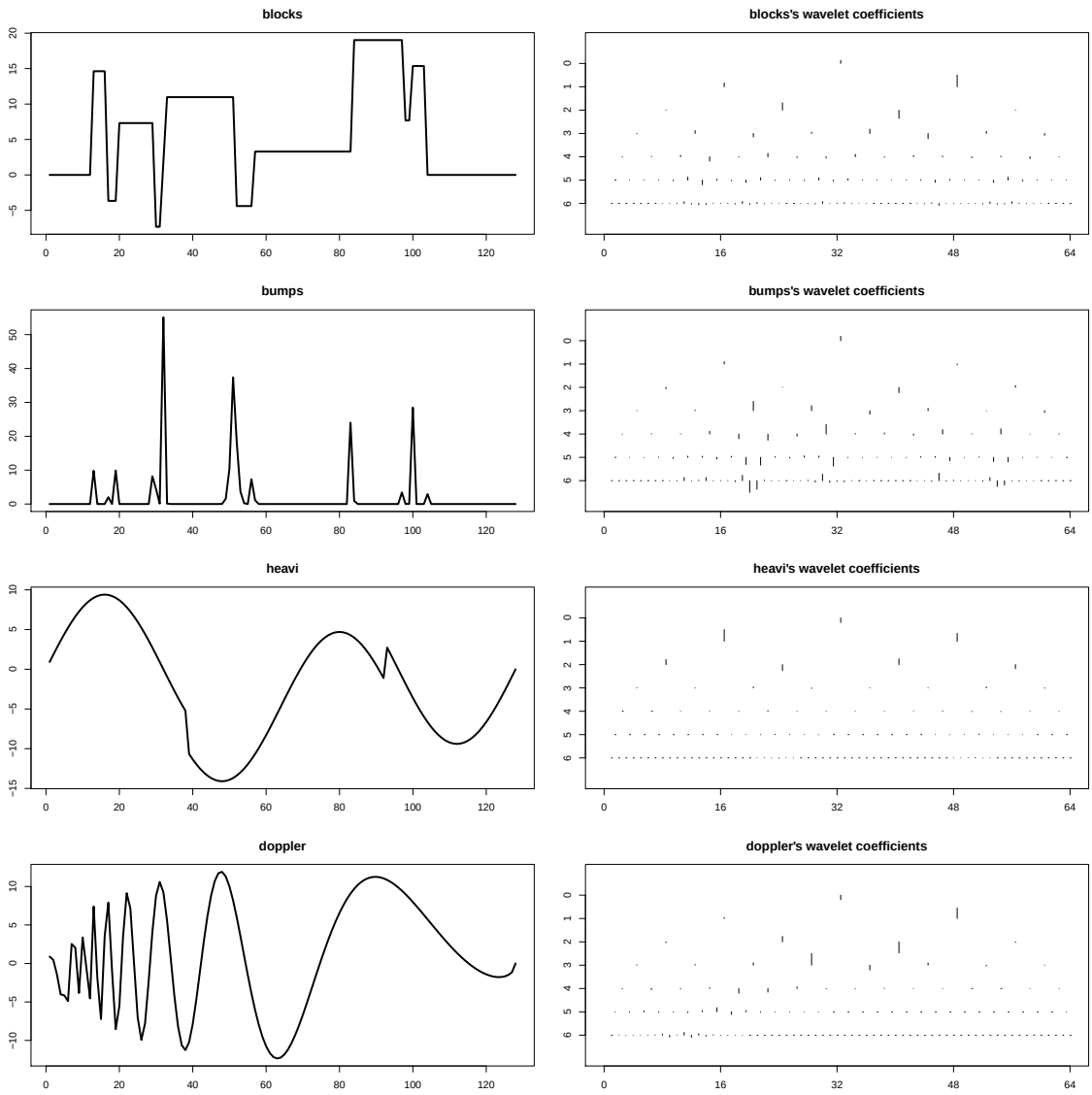

FIGURE 4.1: The four test functions from Donoho and Johnstone (1994) and the associated wavelet coefficients. The coefficients for the test functions *blocks*, *bumps* and *doppler* are concentrated in location and scale.

wavelet coefficients, while the parameter β controls the decay of the number of zero coefficients in function of the scale j. For large β the number of zero coefficients increases rapidly as j gets larger, while $\beta = 0$ corresponds to the prior belief that the probability of a coefficient being zero is not scale dependent. These hyperparameters can be fixed a priori or chosen through an empirical Bayes approach.

4.2.2 Bayesian Adaptive Shrinkage with the NIG-HMT Model

I first show how posterior inference can be carried out under the NIG-HMT model. In particular, I show that the Markov nature of the HMT combined with the Normal inverse-Gamma conjugacy, allows the joint posterior on $\{S_{j,k}, z_{j,k}, \sigma^2_{j,k} : (j,k) \in \mathcal{T}\}$ to be computed analytically through a forward-summation recursion, and can be sampled from directly through a backward-sampling recursion. Consequently, inference under this model is extremely efficient.

Now let us consider the general case with n independent functional observations $\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \ldots, \boldsymbol{y}^{(n)}$ from model (4.1). From now on, I shall use the superscript "(i)" to indicate the terms corresponding to the ith observation. The node-specific model after DWT becomes:

$$d^{(i)}_{j,k} = z_{j,k} + u^{(i)}_{j,k} \quad \text{where} \quad u^{(i)}_{j,k} \sim N(0, \sigma^2_{j,k}).$$

The interest lies in finding the posterior distribution on $\{S_{j,k}, z_{j,k}, \sigma^2_{j,k} : (j,k) \in \mathcal{T}\}$ given the observed data. To this end, let $m_{j,k}(s)$ be the marginal likelihood for the node-specific model on (j,k) given that $S_{j,k} = s$:

$$m_{j,k}(s) = \int p(\boldsymbol{d}^{(1)}, \ldots, \boldsymbol{d}^{(n)} | S_{j,k} = s, z_{j,k}, \sigma^2_{j,k}) \pi(z_{j,k}, \sigma^2_{j,k}) dz_{j,k} d\sigma^2_{j,k}.$$

Due to Normal inverse-Gamma conjugacy, the marginal likelihood is available in closed form:

$$m_{j,k}(s) = \begin{cases} C\left[\nu\sigma_0^2 + \sum_i (d^{(i)}_{j,k})^2/2\right]^{-\nu - n/2 - 1} & \text{if } s = 0 \\ C\left[\dfrac{\tau_j^{-1}}{n + \tau_j^{-1}}\right]^{1/2} \left[\nu\sigma_0^2 + \dfrac{1}{2}\left(\sum_i (d^{(i)}_{j,k})^2 - \dfrac{(n\bar{d}_{j,k})^2}{n + \tau_j^{-1}}\right)\right]^{-\nu - n/2 - 1} & \text{if } s = 1, \end{cases}$$

where $C = (2\pi)^{-n/2} (\nu\sigma_0^2)^{\nu+1} \Gamma(\nu + n/2 + 1)/\Gamma(\nu + 1)$, and $\bar{d}_{j,k} = \sum_i d^{(i)}_{j,k}/n$.

Given the marginal likelihood at each node (j,k) and for all possible states $s \in \mathcal{S}$, I can carry out a "forward-summation-backward-sampling" recursion to find the joint

posterior. The result is summarized in the next theorem. From now on, I shall use $\mathcal{D}$ to represent the totality of the observed data.

Theorem 13. *The joint posterior on* $\{S_{j,k}, z_{j,k}, \sigma^2_{j,k} : (j,k) \in \mathcal{T}\}$ *is given as follows:*

- *The posterior of the hidden states* $S_{j,k}$ *is still a HMT:*

  1. *State transition probabilities:*

  $$\Pr(S_{j,k} = s' | S_{j-1,\lfloor k/2 \rfloor} = s, \mathcal{D}) = \rho_{j,k}(s, s') \frac{\phi_{j,k}(s')}{\xi_{j,k}(s)},$$

  for $s, s' \in \mathcal{S}$ *and* $j = 1, 2, \ldots, J$*;*

  2. *Initial state probabilities:*

  $$\Pr(S_{0,0} = s | \mathcal{D}) = \rho_{0,0}(s) \frac{\phi_{0,0}(s)}{\xi_{0,0}(0)} \quad \textit{for } s \in \mathcal{S}.$$

- *The posterior of the variances* $\sigma^2_{j,k}$ *given* $S_{j,k}$ *is:*

$$[\sigma^2_{j,k} | S_{j,k} = s, \mathcal{D}] \sim \textit{Inv-Gamma}\left(\nu + 1 + \frac{n}{2}, \nu\sigma_0^2 + \frac{1}{2}\left(\sum_i (d^{(i)}_{j,k})^2 - \frac{s(n\bar{d}_{j,k})^2}{n + \tau_j^{-1}}\right)\right).$$

- *The posterior of* $z_{j,k}$ *given* $S_{j,k}$ *and* $\sigma^2_{j,k}$ *is:*

$$[z_{j,k} | \sigma^2_{j,k}, S_{j,k} = s, \mathcal{D}] \sim \begin{cases} \delta(0) & \textit{if } s = 0 \\ N\left(\dfrac{n}{n + \tau_j^{-1}} \bar{d}_{j,k}, \dfrac{1}{n + \tau_j^{-1}} \sigma^2_{j,k}\right) & \textit{if } s = 1. \end{cases}$$

The mappings $\phi_{j,k}$ *and* $\xi_{j,k} : \mathcal{S} \mapsto [0, +\infty)$ *are defined recursively in* j *as follows:*

$$\phi_{j,k}(s) = \begin{cases} m_{j,k}(s) \cdot \phi_{j+1,2k}(s) \cdot \phi_{j+1,2k+1}(s) & \textit{for } j = 0, 1, 2 \ldots, J-1 \\ m_{j,k}(s) & \textit{for } j = J, \end{cases}$$

$$\xi_{j,k}(s) = \begin{cases} \sum_{s' \in \mathcal{S}} \rho_{j,k}(s, s') \cdot \phi_{j,k}(s') & \textit{for } j = 1, 2, \ldots, J \\ \sum_{s' \in \mathcal{S}} \rho_{0,0}(s') \cdot \phi_{j,k}(s') & \textit{for } j = 0. \end{cases}$$

The recursive computation of the mappings $\phi_{j,k}$ and $\xi_{j,k}$ correspond to the "forward-summation" recursion, while the computation of the posterior parameters in terms of these mappings correspond to the "backward-sampling" recursion.

Theorem 13 provides a recipe for sampling from the exact joint posterior of $\{(S_{j,k}, z_{j,k}, \sigma^2_{j,k}\}$. An immediate corrolary is a backward recursion recipe for computing the posterior marginal probability of the states $\{S_{j,k} : (j,k) \in \mathcal{T}\}$. More specifically, the recursion proceeds as follows:

$$\Pr(S_{j,k} = s' \,|\, \mathcal{D}) = \sum_{s \in \mathcal{S}} \Pr(S_{j-1,\lfloor k/2 \rfloor} = s \,|\, \mathcal{D}) \cdot \Pr(S_{j,k} = s' \,|\, S_{j-1,\lfloor k/2 \rfloor} = s, \mathcal{D})$$

for $j = 1, 2, \ldots, J$. In functional estimation problems, the main interest often lies in the mean function $\boldsymbol{f}$, a posterior sample of which can be obtained by applying an inverse DWT to the sample of $\boldsymbol{z}$. One can also use the posterior marginal state probabilities to compute analytically the posterior mean of $\boldsymbol{z}$:

$$\tilde{z}_{j,k} := E(z_{j,k} \,|\, \mathcal{D}) = \Pr(S_{j,k} = 1 \,|\, \mathcal{D}) \cdot \frac{n}{n + \tau_j^{-1}} \bar{d}_{j,k},$$

which has an intuitive explanation in terms of shrinkage. The average of the observed wavelet coefficients $\bar{d}_{j,k}$ is shrunk toward the prior mean 0 with the amount of shrinkage being averaged over the different shrinkage states. By applying an inverse DWT to $\tilde{\boldsymbol{z}}$ I can get the posterior mean of $\boldsymbol{f}$, $E(\boldsymbol{f} \,|\, \mathcal{D}) = W^{-1}\tilde{\boldsymbol{z}}$, which is the Bayes estimator for $\boldsymbol{f}$ under L_2 loss.

4.2.3 One-Way Functional ANOVA Using NIG-HMTs

Suppose one has G groups of functional observations whose values are attained at T equidistant points $\boldsymbol{y}^{(g,i)} = (y_1^{(g,i)}, \ldots, y_T^{(g,i)})$, and

$$\begin{aligned} \boldsymbol{y}^{(g,i)} &= \boldsymbol{f}^{(g)} + \boldsymbol{\epsilon}^{(g,i)} \\ \boldsymbol{\epsilon}^{(g,i)} &\sim N(\boldsymbol{0}, \Sigma_\epsilon), \end{aligned} \tag{4.7}$$

where $i = 1, \ldots, n_g$, $g = 1, \ldots, G$, and $\Sigma_\epsilon = \text{diag}(\sigma_1^2, \ldots, \sigma_T^2)$. I wish to test the following hypothesis:

$$H_0 : \boldsymbol{f}^{(1)} = \ldots = \boldsymbol{f}^{(G)} \quad \text{vs.} \quad H_1 : \boldsymbol{f}^{(g)} \neq \boldsymbol{f}^{(g')} \quad \text{for some } g \neq g'. \tag{4.8}$$

After applying the DWT to each $\boldsymbol{y}^{(g,i)}$ I obtain:

$$\boldsymbol{d}^{(g,i)} = \boldsymbol{z}^{(g)} + \boldsymbol{u}^{(g,i)},$$

for $i = 1, \ldots, n_g$, and $g = 1, \ldots, G$. The model can be written taking the first group as reference, and modeling the differences between the other groups and the first one:

$$\boldsymbol{d}^{(g,i)} = \begin{cases} \boldsymbol{z}^{(1)} + \boldsymbol{u}^{(1,i)} & \text{if } g = 1 \\ \boldsymbol{z}^{(1)} + \boldsymbol{w}^{(g)} + \boldsymbol{u}^{(g,i)} & \text{if } g = 2, \ldots, G, \end{cases}$$

where $\boldsymbol{w}^{(g)} = \boldsymbol{z}^{(g)} - \boldsymbol{z}^{(1)}$, and $i = 1, \ldots, n_g$. The $\boldsymbol{w}^{(g)}$'s represent wavelet coefficients of the difference between the mean function $\boldsymbol{f}^{(g)}$ and the mean of the reference function $\boldsymbol{f}^{(1)}$. The hypothesis test (4.8) can be defined in the wavelet domain as follows:

$$H_0 : \boldsymbol{w}^{(g)} = \boldsymbol{0} \quad \text{for all } g > 1 \quad \text{vs.} \quad H_1 : \boldsymbol{w}^{(g)} \neq \boldsymbol{0} \quad \text{for some } g > 1. \tag{4.9}$$

I can write the model in a node-specific manner:

$$d_{j,k}^{(g,i)} = \begin{cases} z_{j,k}^{(1)} + u_{j,k}^{(1,i)} & \text{if } g = 1 \\ z_{j,k}^{(1)} + w_{j,k}^{(g)} + u_{j,k}^{(g,i)} & \text{if } g = 2, \ldots, G, \end{cases}$$

where $u_{j,k}^{(g,i)} \sim N(0, \sigma_{j,k}^2)$. I adopt spike-and-slap priors on $z_{j,k}^{(1)}$ and $w_{j,k}^{(g)}$'s:

$$[z_{j,k}^{(1)} | \sigma_{j,k}^2, S_{j,k} = s] \sim \begin{cases} \delta(0) & \text{if } s = 0 \\ N(0, \tau_j^{(1)} \sigma_{j,k}^2) & \text{if } s = 1, \end{cases}$$

and

$$[w_{j,k}^{(2)}, \ldots, w_{j,k}^{(G)} | \sigma_{j,k}^2, R_{j,k} = r] \sim \begin{cases} \delta(0) & \text{if } r = 0 \\ N(0, \tau_j^{(2)} \sigma_{j,k}^2) & \text{if } r = 1. \end{cases}$$

The hidden shrinkage variable $S_{j,k}$ controls the shrinkage state at node (j,k). The state $S_{j,k} = 0$ corresponds to total shrinkage, i.e., the wavelet coefficient $z_{j,k}^{(1)}$ is exactly zero, while the state $S_{j,k} = 1$ indicates low shrinkage which is controlled by the parameter $\tau_j^{(1)}$. The hidden rejection variable $R_{j,k}$, instead, controls the following local hypothesis:

$$H_0(j,k) : w_{j,k}^{(g)} = \mathbf{0} \quad \text{for all } g > 1 \quad \text{vs.} \quad H_1(j,k) : w_{j,k}^{(g)} \neq \mathbf{0} \quad \text{for some } g > 1.$$

If $R_{j,k} = 0$, the local null hypothesis $H_0(j,k)$ is true, while if $R_{j,k} = 1$ the local null hypothesis $H_0(j,k)$ is rejected. The global hypothesis (4.9) can be expressed in function of the hidden rejection variables $R_{j,k}$. The global null hypothesis H_0 holds true if and only if all the local hypotheses $H_0(j,k)$ are accepted:

$$H_0 : R_{j,k} = 0 \quad \text{for all } (j,k) \in \mathcal{T} \quad \text{vs.} \quad H_1 : R_{j,k} = 1 \quad \text{for some } (j,k) \in \mathcal{T}.$$

To fully specify the model I need to define the joint prior distribution for $\{S_{j,k}, R_{j,k}, \sigma_{j,k}^2 : (j,k) \in \mathcal{T}\}$. In particular, I consider a prior which is separated in three mutually independent components: $\{\sigma_{j,k}^2 : (j,k) \in \mathcal{T}\}$, $\{S_{j,k} : (j,k) \in \mathcal{T}\}$ and $\{R_{j,k} : (j,k) \in \mathcal{T}\}$.

As in the case of estimation of a single function, I consider the following prior on the variance parameters:

$$\sigma_{j,k}^2 \overset{\text{iid}}{\sim} \text{Inv-Gamma}(\nu + 1, \nu\sigma_0^2),$$

where the hyperparameters ν and σ_0^2 can be either set a priori or determined through an empirical Bayes approach. Again, the inverse-Gamma prior maintains conjugacy to the Gaussian model, and consequently the marginal likelihood can be evaluated analytically.

To incorporate dependency within the shrinkage and rejection states, I model them as a two mutually independent HMTs:

$$\Pr(S_{j,k} = s' | S_{j-1,\lfloor k/2 \rfloor} = s) = \rho_{j,k}^{(1)}(s, s')$$

$$\Pr(R_{j,k} = r' | R_{j-1,\lfloor k/2 \rfloor} = r) = \rho_{j,k}^{(2)}(r, r')$$

Because the root node $(0,0)$ does not have a parent, the initial states of the process $S_{0,0}$ and $R_{0,0}$ are specified by a set of initial state probabilities $\boldsymbol{\rho}_{0,0}^{(1)} = (\rho_{0,0}^{(1)}(0), \rho_{0,0}^{(1)}(1))$ and $\boldsymbol{\rho}_{0,0}^{(2)} = (\rho_{0,0}^{(2)}(0), \rho_{0,0}^{(2)}(1))$ such that $\Pr(S_{0,0} = s) = \rho_{0,0}^{(1)}(s)$ and $\Pr(R_{0,0} = r) = \rho_{0,0}^{(2)}(r)$. The transition probabilities can be organized into transition probability matrices $\boldsymbol{\rho}_{j,k}^{(i)}$ for $i = 1, 2$. I use the same parametric structure introduced in (4.6), where now the hyperparameters have a subscript "i" to separate those associated to the shrinkage variables ($i = 1$) from those of the rejection variables ($i = 2$).

4.2.4 Posterior Inference under the ANOVA NIG-HMT Model

Next I show how general Bayesian inference can be carried out under the ANOVA NIG-HMT model by providing a recipe for sampling draws from the corresponding posterior. The main result shows that the joint posterior $\{z_{j,k}^{(1)}, w_{j,k}^{(2)}, \ldots, w_{j,k}^{(G)}, S_{j,k}, R_{j,k}, \sigma_{j,k}^2 : (j,k) \in \mathcal{T}\}$ can be computed and sampled exactly through a forward-backward recursive algorithm.

To this end, I write the node-specific model in matrix notation:

$$\boldsymbol{d}_{j,k} = X\boldsymbol{\theta}_{j,k} + \boldsymbol{u}_{j,k},$$

where $\boldsymbol{d}_{j,k} = (d_{j,k}^{(1,1)}, \ldots, d_{j,k}^{(n_G,G)})'$ is the vector of the wavelet coefficients for all the observations at node (j,k), X is the design matrix, $\boldsymbol{\theta}_{j,k} = (z_{j,k}^{(1)}, w_{j,k}^{(2)}, \ldots, w_{j,k}^{(G)})'$ is the vector of the wavelet coefficients for the mean functions, and $\boldsymbol{u}_{j,k} = (u_{j,k}^{(1,1)}, \ldots, u_{j,k}^{(n_G,G)})'$ is the vector of the residual errors. The design matrix can be decomposed as follows:

$$X = (\mathbb{1}_n, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_G),$$

where $n = \sum_{g=1}^{G} n_g$, and $\boldsymbol{e}_g$ is a binary vector where the ith element is equal to one if the ith observation belongs to group g, and equal to zero otherwise. I also define

the following matrices:

$$X(s,r) = \begin{cases} \mathbb{1}_n & \text{if } s=1, r=0 \\ (\boldsymbol{e}_2, \ldots, \boldsymbol{e}_G) & \text{if } s=0, r=1 \\ (\mathbb{1}_n, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_G) & \text{if } s=1, r=1, \end{cases}$$

and

$$[\Lambda_{j,k}(s,r)]^{-1} = \begin{cases} \tau_j^{(1)} & \text{if } s=1, r=0 \\ \text{diag}(\tau_j^{(2)}, \ldots, \tau_j^{(2)}) & \text{if } s=0, r=1 \\ \text{diag}(\tau_j^{(1)}, \tau_j^{(2)}, \ldots, \tau_j^{(2)}) & \text{if } s=1, r=1, \end{cases}$$

Then, the marginal likelihood for the node specific model on (j,k) given that $S_{j,k} = s$ and $R_{j,k} = r$ is the following

$$m_{j,k}(s,r) = \begin{cases} C\big[\nu\sigma_0^2 + \Upsilon(s,r)\big]^{-\nu-n/2-1} & \text{if } s=0, r=0 \\ C\dfrac{|\Lambda_{j,k}(s,r)|^{1/2}}{|\Lambda^*_{j,k}(s,r)|^{1/2}}\big[\nu\sigma_0^2 + \Upsilon_{j,k}(s,r)\big]^{-\nu-n/2-1} & \text{otherwise,} \end{cases} \tag{4.10}$$

for $C = (2\pi)^{-n/2}(\nu\sigma_0^2)^{\nu+1}\Gamma(\nu+n/2+1)/\Gamma(\nu+1)$ and

$$\Upsilon_{j,k}(s,r) = \begin{cases} \boldsymbol{d}'_{j,k}\boldsymbol{d}_{j,k}/2 & \text{if } s=0, r=0 \\ \{\boldsymbol{d}'_{j,k}\boldsymbol{d}_{j,k} - [\boldsymbol{\mu}^*_{j,k}(s,r)]'\Lambda^*_{j,k}(s,r)\boldsymbol{\mu}^*_{j,k}(s,r)\}/2 & \text{otherwise,} \end{cases} \tag{4.11}$$

where $\Lambda^*_{j,k}(s,r) = X(s,r)'X(s,r) + \Lambda_{j,k}(s,r)$ and $\boldsymbol{\mu}^*_{j,k}(s,r) = [\Lambda^*_{j,k}(s,r)]^{-1}[X(s,r)'\boldsymbol{d}_{j,k}]$.

Once I have computed $m_{j,k}(s,r)$ for all $(j,k) \in \mathcal{T}$, I can find the exact joint posterior through forward-backward recursion.

Theorem 14. *The joint posterior on* $\{z^{(1)}_{j,k}, w^{(2)}_{j,k}, \ldots, w^{(G)}_{j,k}, S_{j,k}, R_{j,k}, \sigma^2_{j,k} : (j,k) \in \mathcal{T}\}$ *is given as follows.*

- *The marginal posterior of the hidden states* $\{S_{j,k}, R_{j,k} : (j,k) \in \mathcal{T}\}$ *is an HMT with*

1. *State transition probabilities:*

$$\begin{aligned}
&\rho_{j,k}\big((s,r)\to(s',r')|\mathcal{D}\big)\\
&\qquad=\Pr(S_{j,k}=s',R_{j,k}=r'|S_{j-1,\lfloor k/2\rfloor}=s,R_{j-1,\lfloor k/2\rfloor}=r,\mathcal{D})\\
&\qquad=\rho^{(1)}_{j,k}(s,s')\rho^{(2)}_{j,k}(r,r')\phi_{j,k}(r',s')/\xi_{j,k}(r,s),
\end{aligned}$$

for $j=1,\ldots,J$.

2. *Initial state probabilities:*

$$\rho_{0,0}(s,r|\mathcal{D})=\Pr(S_{0,0}=s,R_{0,0}=r|\mathcal{D})=\rho^{(1)}_{0,0}(s)\rho^{(2)}_{0,0}(r)\phi_{0,0}(r,s)/\xi_{0,0}(0,0).$$

- *The conditional posterior of* $\sigma^2_{j,k}$ *given* $S_{j,k}$ *and* $R_{j,k}$ *is:*

$$[\sigma^2_{j,k}|S_{j,k}=0,R_{j,k}=0,\mathcal{D}]\sim \textit{Inv-Gamma}\bigg(\nu+1+\frac{n}{2},\nu\sigma_0^2+\frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(d^{(g,i)}_{j,k})^2\bigg)$$

and, if $S_{j,k}\neq 0$ *and* $R_{j,k}\neq 0$,

$$[\sigma^2_{j,k}|S_{j,k}=s,R_{j,k}=r,\mathcal{D}]\sim \textit{Inv-Gamma}\bigg(\nu+1+\frac{n}{2},\nu\sigma_0^2+\Upsilon_{j,k}(s,r)\bigg).$$

- *The posterior of* $z^{(1)}_{j,k},w^{(2)}_{j,k},\ldots,w^{(G)}_{j,k}$ *given* $S_{j,k}$, $R_{j,k}$ *and* $\sigma^2_{j,k}$ *is:*

$$[z^{(1)}_{j,k},w^{(2)}_{j,k},\ldots,w^{(G)}_{j,k}|\sigma^2_{j,k},S_{j,k}=0,R_{j,k}=0,\mathcal{D}]\sim\delta(0),$$

and

$$[z^{(1)}_{j,k}|\sigma^2_{j,k},S_{j,k}=1,R_{j,k}=0,\mathcal{D}]\sim N\Big(\boldsymbol{\mu}^*_{j,k}(1,0),\sigma^2_{j,k}[\Lambda^*_{j,k}(1,0)]^{-1}\Big)$$

$$[w^{(2)}_{j,k},\ldots,w^{(G)}_{j,k}|\sigma^2_{j,k},S_{j,k}=1,R_{j,k}=0,\mathcal{D}]\sim\delta(0),$$

and

$$[z^{(1)}_{j,k}|\sigma^2_{j,k},S_{j,k}=0,R_{j,k}=1,\mathcal{D}]\sim\delta(0)$$

$$[w^{(2)}_{j,k},\ldots,w^{(G)}_{j,k}|\sigma^2_{j,k},S_{j,k}=0,R_{j,k}=1,\mathcal{D}]\sim N\Big(\boldsymbol{\mu}^*_{j,k}(0,1),\sigma^2_{j,k}[\Lambda^*_{j,k}(0,1)]^{-1}\Big),$$

and

$$[z_{j,k}^{(1)}, w_{j,k}^{(2)}, \ldots, w_{j,k}^{(G)} | \sigma_{j,k}^2, S_{j,k} = 1, R_{j,k} = 1, \mathcal{D}] \sim N\Big(\boldsymbol{\mu}_{j,k}^*(1,1), \sigma_{j,k}^2 [\Lambda_{j,k}^*(1,1)]^{-1}\Big).$$

The mappings $\phi_{j,k}(s,r)$ *and* $\xi_{j,k}(s,r) : \{0,1\} \times \{0,1\} \to [0,+\infty)$ *are defined recursively as follows:*

$$\phi_{j,k}(s) = \begin{cases} m_{j,k}(s,r) \cdot \phi_{j+1,2k}(s,r) \cdot \phi_{j+1,2k+1}(s,r) & \text{for } j = 0,1,2\ldots,J-1 \\ m_{j,k}(s,r) & \text{for } j = J, \end{cases}$$

$$\xi_{j,k}(s,r) = \begin{cases} \sum_{s',r'} \rho_{j,k}^{(1)}(s,s') \cdot \rho_{j,k}^{(2)}(r,r') \phi_{j,k}(r',s') \cdot & \text{for } j = 1,2,\ldots,J \\ \sum_{s',r'} \rho_{0,0}^{(1)}(s') \cdot \rho_{0,0}^{(2)}(r') \cdot \phi_{0,0}(s',r') & \text{for } j = 0. \end{cases}$$

4.2.5 Generalization to the Multi-Way ANOVA NIG-HMT

The one-way ANOVA NIG-HMT framework can be naturally extended to the multi-way ANOVA NIG-HMT framework. In this section I illustrate, as an example, the two-way ANOVA NIG-HMT model.

Assume one has collected functional observations $\boldsymbol{y}^{(g,h,i)} = (y_1^{(g,h,i)}, \ldots, y_T^{(g,h,i)})$ from the following model:

$$\begin{aligned} \boldsymbol{y}^{(g,h,i)} &= \boldsymbol{f}^{(A,g)} + \boldsymbol{f}^{(B,h)} + \boldsymbol{\epsilon}^{(g,h,i)} \\ \boldsymbol{\epsilon}^{(g,h,i)} &\sim N(\mathbf{0}, \Sigma_\epsilon), \end{aligned} \tag{4.12}$$

where $i = 1, \ldots, n_{g,h}$, $g = 1, \ldots, G$, $h = 1, \ldots, H$, and $\Sigma_\epsilon = \text{diag}(\sigma_1^2, \ldots, \sigma_T^2)$. I wish to test the following hypotheses:

$$H_0^{(A)} : \boldsymbol{f}^{(A,1)} = \ldots = \boldsymbol{f}^{(A,G)} \quad \text{vs.} \quad H_1^{(A)} : \boldsymbol{f}^{(A,g)} \neq \boldsymbol{f}^{(A,g')} \quad \text{for some } g \neq g',$$

and

$$H_0^{(B)} : \boldsymbol{f}^{(B,1)} = \ldots = \boldsymbol{f}^{(B,H)} \quad \text{vs.} \quad H_1^{(B)} : \boldsymbol{f}^{(B,h)} \neq \boldsymbol{f}^{(B,h')} \quad \text{for some } h \neq h'.$$

After applying the DWT to each $\boldsymbol{y}^{(g,h,i)}$ I obtain:

$$\boldsymbol{d}^{(g,h,i)} = \boldsymbol{z}^{(A,g)} + \boldsymbol{z}^{(B,g)} + \boldsymbol{u}^{(g,i)},$$

for $i = 1, \ldots, n_{g,h}$, $g = 1, \ldots, G$ and $h = 1, \ldots, H$. The model can be written taking the first level of each factor as reference, and modeling the differences between the other levels and the first one:

$$\boldsymbol{d}^{(g,h,i)} = \begin{cases} \boldsymbol{\mu} + \boldsymbol{u}^{(1,1,i)} & \text{if } g = h = 1 \\ \boldsymbol{\mu} + \boldsymbol{w}^{(A,g)} + \boldsymbol{u}^{(g,1,i)} & \text{if } g = 2, \ldots, G \text{ and } h = 1, \\ \boldsymbol{\mu} + \boldsymbol{w}^{(B,h)} + \boldsymbol{u}^{(1,h,i)} & \text{if } g = 1 \text{ and } h = 2, \ldots, H, \\ \boldsymbol{\mu} + \boldsymbol{w}^{(A,g)} + \boldsymbol{w}^{(B,h)} + \boldsymbol{u}^{(g,h,i)} & \text{if } g, h = 2, \ldots, G, \end{cases}$$

where $\boldsymbol{\mu} = \boldsymbol{z}^{(A,1)} + \boldsymbol{z}^{(B,1)}$, $\boldsymbol{w}^{(\cdot,\cdot)} = \boldsymbol{z}^{(\cdot,\cdot)} - \boldsymbol{\mu}$, and $i = 1, \ldots, n_{g,h}$. Finally, one can adopt spike-and-slap priors on $\mu_{j,k}$, $w_{j,k}^{A,g}$ and $w_{j,k}^{B,g}$ and create dependency across the associated hidden variables through HMTs.

4.3 Numerical Examples

In this section I provide three numerical examples. In the first example I compare the NIG-HMT method to Abramovich et al. (1998)'s model in single function estimation. In the second example I illustrate how inference can be carried out in a two-way functional ANOVA simulated dataset. Finally, as an illustration, I apply the proposed method to a real dataset from physiology. In all the examples I use the Daubechies least-asymmetric orthonormal compactly supported wavelet with 10 vanishing moments.

4.3.1 Single Function Estimation

In this example I generate synthetic data from the four test functions proposed by Donoho and Johnstone (1994), namely *blocks*, *bumps*, *doppler* and *heavisine*. In Figure 4.1 I plot the four functions and the associated wavelet coefficients.

For each of the four test functions I consider multiple levels of root signal to noise ratio (RSNR):

$$RSNR = \sqrt{\frac{\sum_{t=1}^{T}(f_t - \bar{f})^2/(T-1)}{\sigma^2}},$$

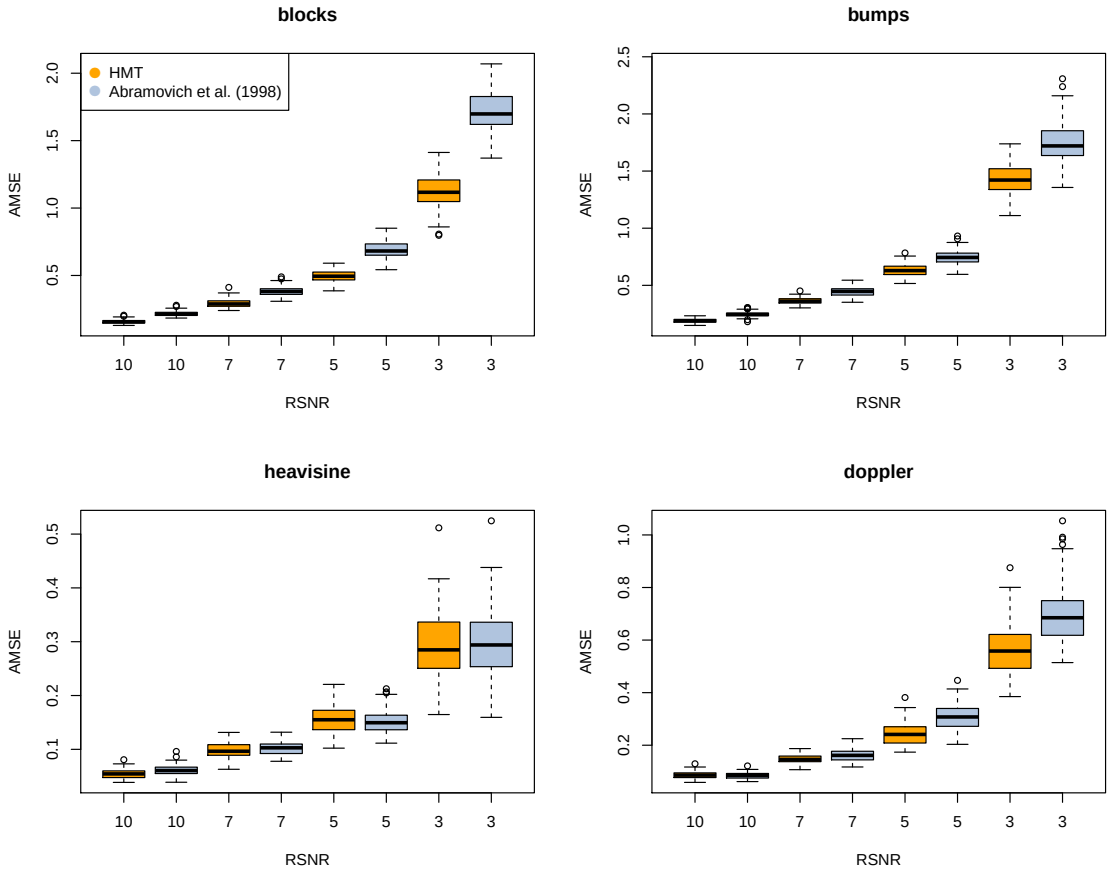

FIGURE 4.2: Boxplots of the AMSE for each of the four test functions and multiple levels of RSNR. In yellow the NIG-HMT method, and in pale blue Abramovich et al. (1998)'s model.

where $\bar{f} = \sum_t f_t/T$ and $\Sigma_\epsilon = \text{diag}(\sigma^2, \ldots, \sigma^2)$. The observations are taken at $T = 1024$ equidistant points, and for each function and each RSNR level I generate 100 datasets.

Figure 4.2 contains the average (over location) mean square error (AMSE) of the mean estimate from the NIG-HMT model and the point estimate from Abramovich et al. (1998)'s model for the various functions and noise levels. I use the R package `wavethresh` (Nason, 2012) for Abramovich et al. (1998)'s model. For both models I fix the following hyperparameters $\alpha = 0.5$ and $\beta = 1$. In the NIG-HMT model the hyperparameter η and γ are chosen through an empirical Bayes approach.

The proposed NIG-HMT method has smaller AMSE on three of the four test functions, while the two methods are comparable in the *heavisine* test function. From Figure 4.1 one can see that the wavelet coefficients are less spatially clustered in the *heavisine* function, and so the dependency introduced in the hidden states by the NIG-HMT does not improve estimation accuracy. In the other test functions, instead, the dependence structure results in more accurate estimates of the functional signals.

4.3.2 Two-Way Functional ANOVA

In this example I generate data from the following model:

$$y_t^{(g,h,i)} = f_t^{(A,g)} + f_t^{(B,h)} + \epsilon_t^{(g,h,i)},$$

where $g = 1, 2$, $h = 1, 2, 3$, $i = 1, \ldots, 4$, and $\epsilon_t^{(g,h,i)} \sim N(0, 5^2)$. The two factors are defined as follows:

$$f_t^{(A,1)} = f_t^{(A,2)} = \text{doppler}(t)$$

and

$$f_t^{(B,h)} = 10\text{bumps}(t) + 20(h-1)\big(1 + 200|t - 0.1|\big)^{-4},$$

where the functions doppler(t) and bumps(t) are defined in Donoho and Johnstone (1994). In Figure 4.3 I plot the functional observations at $T = 512$ equispaced time points grouped by the level of factor B. The observations from the second and third level of factor B have a bigger spike around $t = 0.1$.

The ANOVA NIG-HMT framework allows for posterior inference on the difference across the factor levels as well as inference on the local hypotheses. In the first plot of Figure 4.4 the pointwise 0.95 credible intervals for the difference between the two levels of factor A overlap with zero on the entire domain. In the second and third plot of Figure 4.4 I visualize the difference between the levels of factor B and the

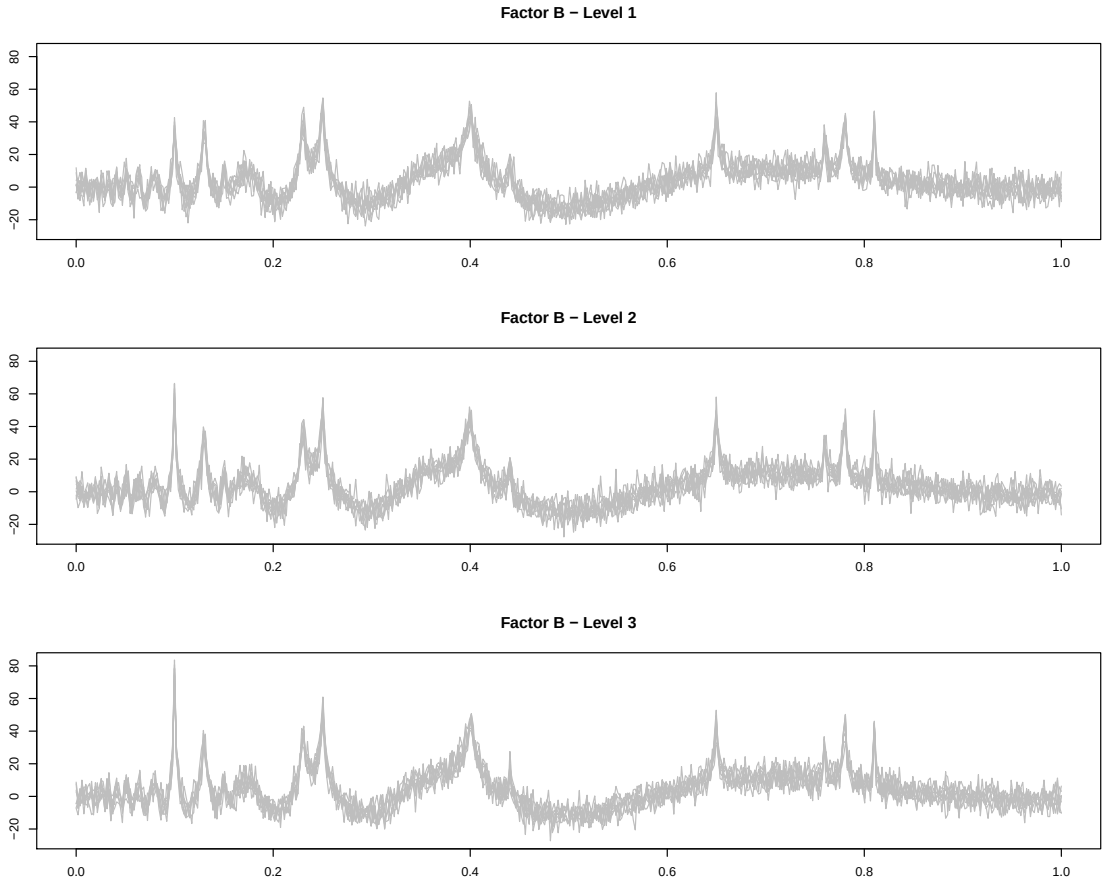

FIGURE 4.3: Functional observations from the two-way functional ANOVA example. The observations are grouped according to the level of the second factor.

reference level. The model correctly identifies the differences in the spikes around $t = 0.1$ between the levels of factor B.

In Figure 4.5 I plot the posterior probability of the hidden states for the intercept and for the two factors. In the plot associated with the intercept I can see the spatial clustering of the non-zero wavelet coefficients. In the second plot the rejection probability of the local null hypotheses is practically zero for all the local hypotheses, while in the third plot the rejection probability of the local hypotheses is practically zero in most of the domain except for the local tests at location around $t = 0.1$.

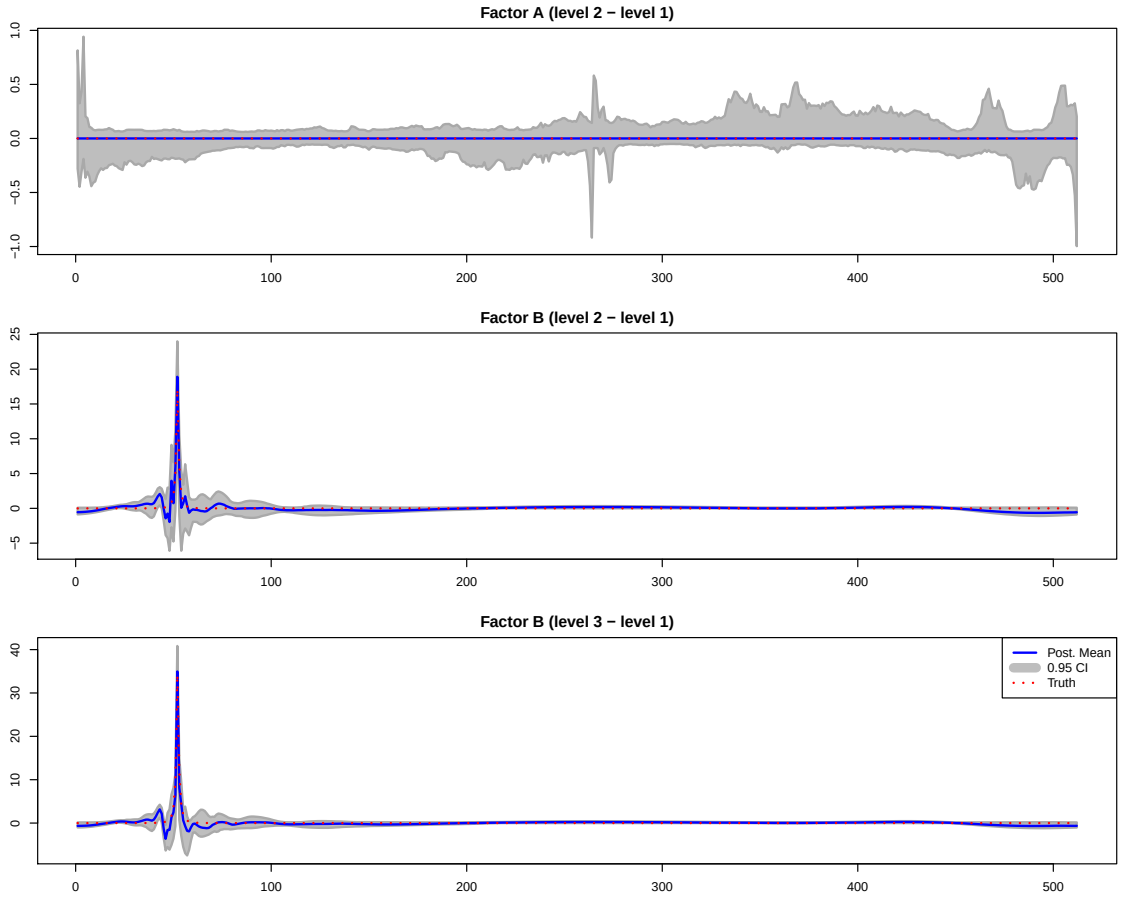

FIGURE 4.4: Posterior estimates of the difference between factors' levels from the two-way functional ANOVA example. The first plot shows the difference between the two levels of factor A. The second and third plots show the difference between level 2 and level 3 of factor B with respect to level 1 of the same factor. The solid blue line represents the posterior mean, the pointed red line is the true value, and the grey band indicates the 0.95 pointwise credible interval.

4.3.3 Orthosis Dataset

I applied the proposed two-way functional ANOVA NIG-HMT method to the *orthosis dataset*. These data were collected by Dr. David Amarantini and Dr. Luc Martin from the Laboratoire Sport et Performance Motrice, Grenoble University, France. The purpose of the study was to understand the effect of different types of constraints to the knee on movement generation. In the study 7 young men wore spring-loaded orthosis on the right knee while stepping in place. Four different ex-

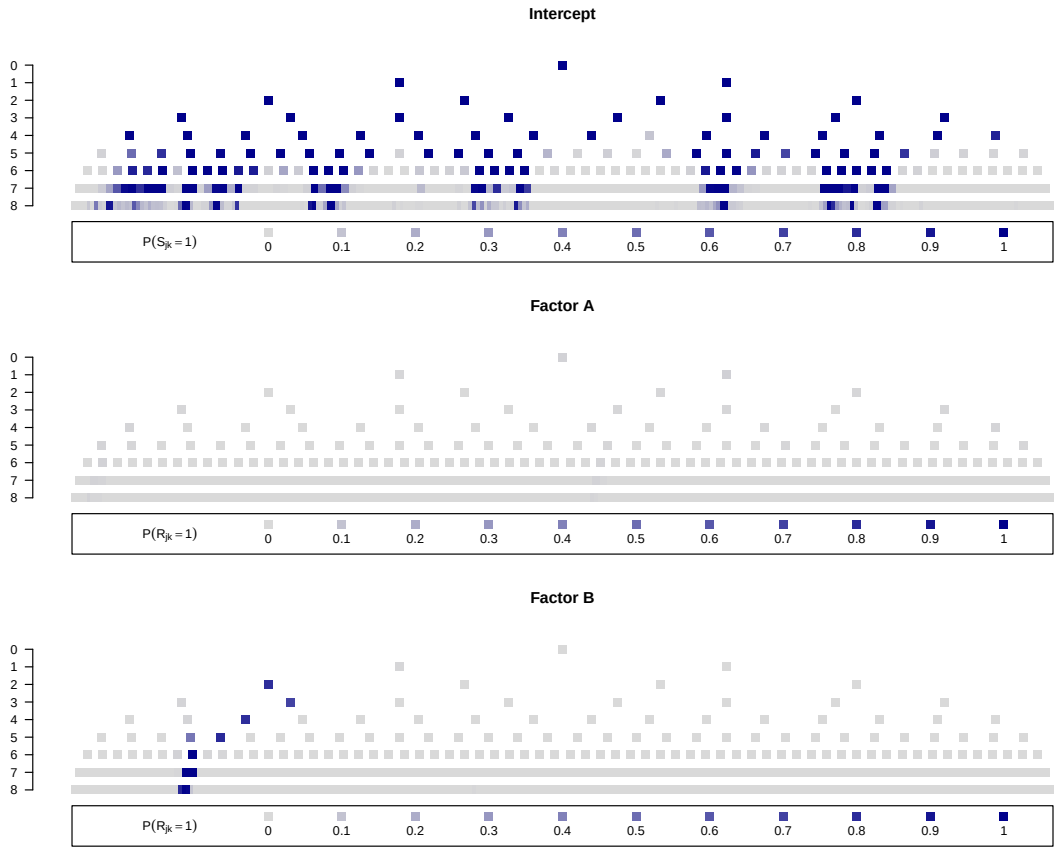

FIGURE 4.5: Marginal posterior probabilities for the hidden states $\{S_{j,k}, R^{(A)}_{j,k}, R^{(B)}_{j,k} : (j,k) \in \mathcal{T}\}$ from the two-way functional ANOVA example. Gray/blue indicates low/high probability that the associated wavelet coefficient is non-zero.

perimental conditions were considered: a control condition (without orthosis), an orthosis condition (with the orthosis only), and two spring conditions (spring 1 and spring 2). Each session lasted for 10 seconds, and the resultant moment at the knee was computed at 256 equally spaced time points. Each subject repeated the session 10 times under each of the four experimental conditions. In this analysis I focus my attention on the two types of spring. In Figure 4.6 I plot some of the functional observations. Each row corresponds to a subject, and each column corresponds to an experimental condition. I refer the reader to Cahouët et al. (2002) for further detail on the experiment as well as data collection and processing.

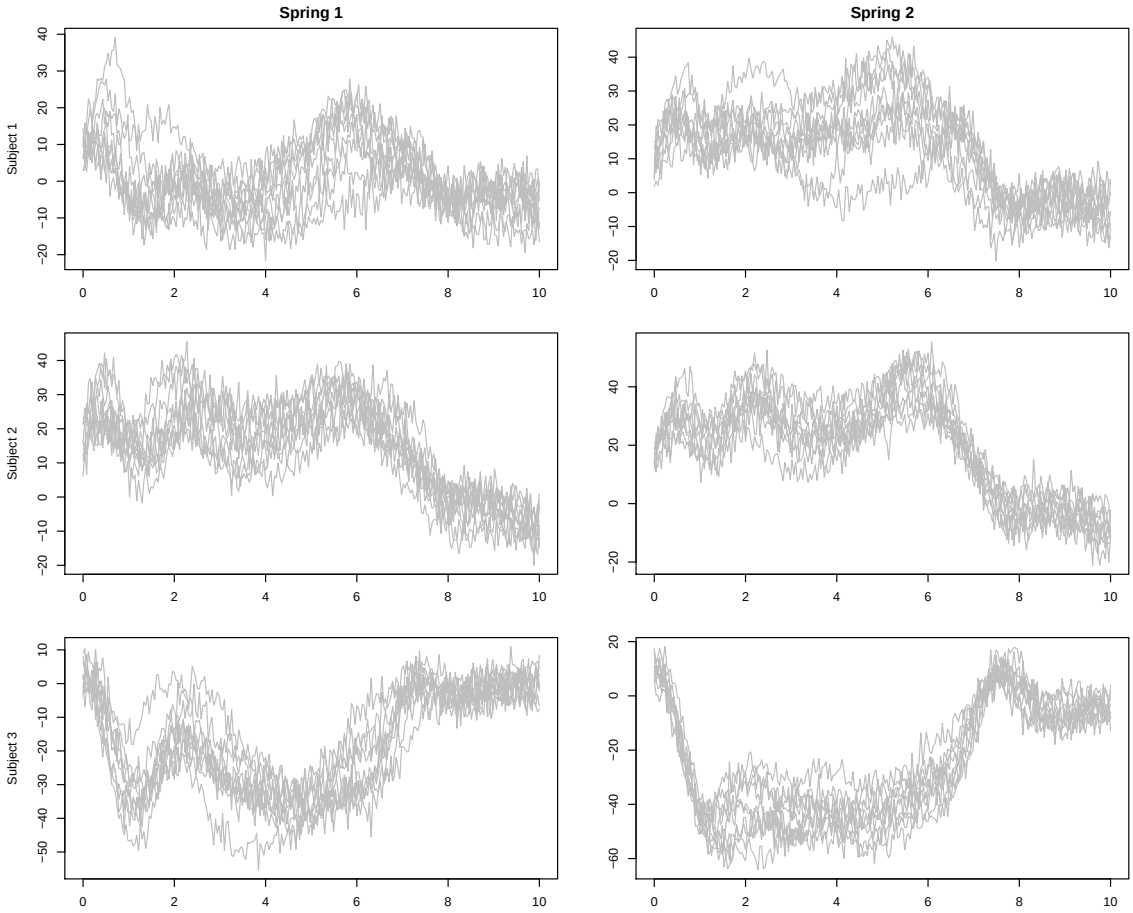

FIGURE 4.6: Functional observations from the orthosis dataset. Each row corresponds to a subject, and each column to an experimental condition.

Abramovich et al. (2004), Abramovich and Angelini (2006) and Antoniadis and Sapatinas (2007) analyzed this dataset using different functional ANOVA models based on wavelet decomposition. All these methods consider a hypothesis testing approach, but they do not provide any measure of uncertainty on the estimated effects. With the proposed method I can carry out hypothesis tests on the subjects' effect and on the conditions' effect as well as quantify the uncertainty on the different parameters of the model.

I consider a two-way functional ANOVA model, where the first factor represents the experimental condition effect and the second factor is the subject effect. In Figure

4.7 I plot the marginal posterior probability of the hidden states for the intercept and the two factors. There is no evidence of difference between the two types of spring. The marginal posterior probability for no spring effect is $\Pr(H_0^{(A)}|\mathcal{D}) = 0.98$, while the prior probability was 0.5. However, there is strong variability across subjects. The marginal posterior probability for no subject effect is practically zero, while the prior probability was 0.5. These results are consistent with previous analyses.

In Figure 4.8 the first two plots show the posterior estimates of the mean function for patient 1 under spring 1 and spring 2. The third plot shows the posterior estimate of the difference between the two experimental conditions. The 0.95 pointwise credible band overlaps with zero on the entire time domain, consistent with the results shown in Figure 4.7.

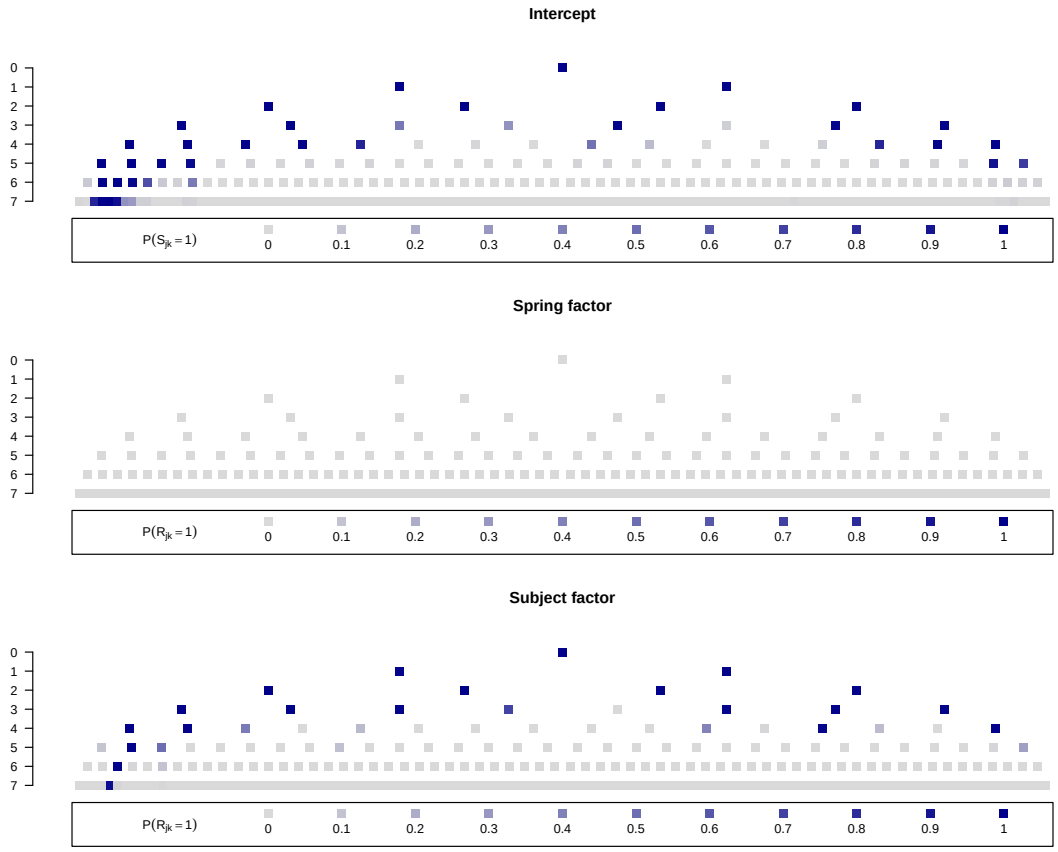

FIGURE 4.7: Marginal posterior probabilities for the hidden states $\{S_{j,k}, R_{j,k}^{(A)}, R_{j,k}^{(B)} : (j,k) \in \mathcal{T}\}$ from the orthosis dataset. Gray/blue indicates low/high probability that the associated wavelet coefficient is non-zero.

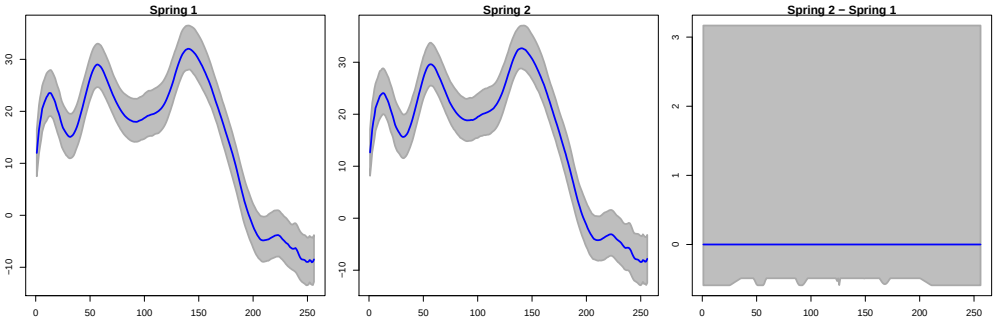

FIGURE 4.8: Posterior estimates for subject 1. The first two plots show the posterior estimates of the mean function under spring 1 and spring 2. The plot on the right shows the posterior estimate for the difference between the two experimental conditions. The blue line is the posterior mean and the gray band is the pointwise 0.95 credible interval.

5

Summary and Future Work

In this thesis I have developed novel nonparametric Bayesian methodology for two-sample comparison. Instead of directly comparing the two distributions, the two-sample comparison problem is decomposed into multiple local hypothesis tests on individual parameters of the two distributions. The evidence of difference from each local test is then combined in a probabilisitc way to obtain a global two-sample comparison test.

This general approach to two-sample comparison has several advantages. One benefit of this approach is that it leverages the flexibility of Bayesian nonparametric models and thereby adapts to the features of the underlying distributions. A second benefit is that it has high statistical power relative to existing methods. It is particularly effective when the difference across the two distributions is spatially localized. As shown via comparison with methods explicitly designed to detect global differences, the localized emphasis of the approach does not diminish the power of the approach when the difference is global. Additionally, the approach allows one to indentify the nature of the differences between the two distributions, unlike most existing methods. In real-world applications, detecting the type of difference can be

as important as identifying the existence of the difference itself.

Throughout the thesis the methodology for two-sample comparison was naturally generalized to multi-sample comparison. Furthermore, the methodology for multi-sample functional comparison in Chapter 4 was extended to multi-way functional ANOVA. Similar ideas can also be applied in the context of PTs and DP mixtures. Next I will outline some preliminary work on how nonparametric ANOVA can be carried out in the context of PTs.

Consider the following nested experimental design setting where there are K experimental conditions, J replicates for each experimental condition, and $n_{j,k}$ observations for the jth replicate under the kth condition:

$$x_{i,j,k}|Q_{j,k} \sim Q_{j,k},$$

where $i = 1, \ldots, n_{j,k}$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$. Assume the researcher is interested in understanding to which extent the variability across experiments is due to the different experimental conditions. This framework is relevant, for instance, in the context of sequencing assays such as RNA-Seq, ChIP-Seq and DNASE-Seq where the sample space is a region of the genome, and each observation represents the genome location of each individual DNA fragment. Interest lies in comparing read counts across multiple biological conditions (Anders and Huber, 2010), while correcting for across-experiment variability.

To avoid parametric assumptions on the underlying distributions $Q_{j,k}$, one can decompose the distributions $Q_{j,k}$ into collections of probability assignments $\boldsymbol{\theta}_{j,k} = \{\theta_{j,k}(B) \text{ for } B \in \mathcal{B}^{(\infty)}\}$ and define a joint prior on them. If for each region B the associated probability assignment is decomposed as follows:

$$\log(\theta_{j,k}(B)) = \beta_k(B) + \epsilon_{j,k}(B),$$

then one can consider the following local hypothesis:

$$H_0(B) : \beta_{k'}(B) = \beta_{k''}(B) \quad \text{vs} \quad H_1(B) : \beta_{k'}(B) \neq \beta_{k''}(B),$$

for $k', k'' = 1, \ldots, K$. Specifically, I assume that

$$\epsilon_{j,k}(B) \overset{\text{iid}}{\sim} N(0, \tau^2),$$

for $j = 1, \ldots, J$ and $k = 1, \ldots, K$. Under the local null:

$$\beta_0(B) = \beta_1(B) = \ldots = \beta_K(B)$$
$$\beta_0(B) \sim N(0, \sigma^2),$$

while under the local alternative:

$$\beta_k(B) \overset{\text{iid}}{\sim} N(0, \sigma^2) \quad k = 1, \ldots, K.$$

As in the case of ARM-tree, one can borrow information across the local hypotheses through a HMT. To compute the posterior probability of each local hypothesis, one must compute the marginal likelihood at each node B under both the local null and the local alternative. Unlike ARM-tree, the marginal likelihood is not available in closed form, but can be easily approximated by Laplace approximation.

In this way one can decompose the multiple sources of variability across the distributions without relying on strong parametric assumptions on the distributions. Additionally, the multi-resolution decomposition of PTs allows one to identify the location and scale of differences across the experimental conditions.

Appendix A

Appendix for Chapter 2

A.1 Proofs of Technical Results

A.1.1 Proof of Lemma 5

The likelihood restricted to B has the following recursive representation:

$$\Pr(\boldsymbol{x}(B)|\boldsymbol{\theta}_1(B),\boldsymbol{\theta}_2(B),\mathcal{T}(B))$$

$$= \begin{cases} \mu(B)^{-\sum_k n_k(B)} & \text{if } S(B)=1 \\ \prod_{k=1,2} 2^{n_k(B)} \Pr(\boldsymbol{n}_k(B)|\theta_k(B)) \prod_{i=0,1} \Pr(\boldsymbol{x}(B_i)|\boldsymbol{\theta}_1(B_i),\boldsymbol{\theta}_2(B_i),\mathcal{T}(B_i)) & \text{if } S(B)=0. \end{cases}$$

The prior $\Pi_z(\boldsymbol{\theta}_1(B),\boldsymbol{\theta}_2(B),\boldsymbol{Z}(B))$ has the following factorization:

$$\Pi_z(\boldsymbol{\theta}_1(B),\boldsymbol{\theta}_2(B),\boldsymbol{Z}(B)) = \hat{\rho}_{z,(1,0)}(B)\Pi_z(Z(B)=(1,0))$$
$$+ \sum_{z'\in\mathcal{Z}\setminus(1,0)} \hat{\rho}_{z,z'}(B)\Pi_z(\theta_1(B),\theta_2(B),Z(B)=z') \prod_{i=0,1} \Pi_{z'}(\boldsymbol{\theta}_1(B_i),\boldsymbol{\theta}_2(B_i),\boldsymbol{Z}(B_i)),$$

for $B \in \mathcal{B}^{(\infty)} \setminus \Omega$. For the root, i.e., $B=\Omega$, I have

$$\Pi(\boldsymbol{\theta}_1(B),\boldsymbol{\theta}_2(B),\boldsymbol{Z}(B)) = \hat{\pi}_{(1,0)}\Pi(Z(B)=(1,0))$$
$$+ \sum_{z'\in\mathcal{Z}\setminus(1,0)} \hat{\pi}_{z'}\Pi(\theta_1(B),\theta_2(B),Z(B)=z') \prod_{i=0,1} \Pi_{z'}(\boldsymbol{\theta}_1(B_i),\boldsymbol{\theta}_2(B_i),\boldsymbol{Z}(B_i)).$$

If I integrate the right-side of the likelihood restricted to B with respect to the prior, I obtain (2.3).

A.1.2 Proof of Lemma 6

The first claim holds because $\Phi(B, \boldsymbol{x})$ is the marginal likelihood, and therefore is equal to 1 if there are no data-points in B. The second claim follows from the ARM-tree self-similarity. For any region B that arises during the partitioning process, the ARM-tree restricted to B is a ARM-tree process with sample space $\Omega = B$, parameters restricted to B and its descendants. The condition $\alpha_k(B, 0)/\alpha_k(B, 1) = 1$ implies that $f(x) = \mu(B)^{-1}\mathbb{1}_B(x)$ is the predictive density. For a single data-point x in B, then $\Phi(B, \boldsymbol{x})$ is the predictive density at x.

A.1.3 Proof of Theorem 7

For the transition probabilities I have:

$$\begin{aligned}
\Pr(Z(B) = z'|Z(\text{parent}(B)) = z, \boldsymbol{x}) &= \Pr(Z(B) = z'|Z(\text{parent}(B)) = z, \boldsymbol{x}(B)) \\
&= \frac{\Pr(\boldsymbol{x}(B), Z(B) = z'|Z(\text{parent}(B)) = z)}{\Pr(\boldsymbol{x}(B)|Z(\text{parent}(B)) = z)} \\
&= \Pr(Z(B) = z'|Z(\text{parent}(B)) = z)\frac{\Pr(\boldsymbol{x}(B)|Z(B) = z')}{\Pr(\boldsymbol{x}(B)|Z(\text{parent}(B)) = z)} \\
&= \hat{\rho}_{z,z'}(B)\frac{\Lambda_{z'}(\boldsymbol{x}, B)}{\Phi_z(\boldsymbol{x}, B)},
\end{aligned}$$

where $B \in \mathcal{B}^{(\infty)} \setminus \Omega$ and $z, z' \in \mathcal{Z}$. Similarly, one can derive the posterior initial state probabilities.

Finally, since the Beta distribution is conjugate to the Multinomial distribution, I have:

$$\begin{aligned}
\boldsymbol{\alpha}_0(B|\boldsymbol{x}) &= \boldsymbol{\alpha}_0(B) + \boldsymbol{n}_1(B) + \boldsymbol{n}_2(B) \\
\boldsymbol{\alpha}_k(B|\boldsymbol{x}) &= \boldsymbol{\alpha}_k(B) + \boldsymbol{n}_k(B)
\end{aligned}$$

for all $B \in \mathcal{B}^{(\infty)}$ and $k = 1, 2$.

A.1.4 Notation and Lemmas for Proofs of Consistency

We observe two independent groups of i.i.d. samples $\boldsymbol{x}_1 = (x_{1,1}, \ldots, x_{1,n_1})$ and $\boldsymbol{x}_2 = (x_{2,1}, \ldots, x_{2,n_2})$ from two distributions Q_1 and Q_2 on $\Omega = [0,1)^p$, where $n = n_1 + n_2 \to \infty$, and $n_1/n \to \xi$ for some $\xi \in (0,1)$. Let (Q_1, Q_2) have a ARM-tree$(\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\lambda}, \boldsymbol{\alpha})$ distribution. The two unknown distributions are P_1 and P_2. Define $P = \xi P_1 + (1-\xi)P_2$, $p_k^j = P_k^j(B_0^j|B)$, $p^j = P(B_0^j|B)$, $q^j = Q_0(B_i^j)/Q_0(B)$, $\hat{p}_k^j = n_k(B_0^j)/n_k(B)$ and $\hat{p}^j = n(B_0^j)/n(B)$ for $k = 1, 2$ and $j = 1, \ldots, p$.

Lemma 15. *For $B \in \mathcal{B}^{(\infty)}$ and $j = 1, \ldots, p$, define*

$$\mathcal{L}_{0,1}(B, j, \boldsymbol{x}) = \frac{D(\boldsymbol{\alpha}_0(B) + \boldsymbol{n}_1^j(B) + \boldsymbol{n}_2^j(B))/D(\boldsymbol{\alpha}_0(B))}{\prod_{k=1,2} D(\boldsymbol{\alpha}_k(B) + \boldsymbol{n}_k^j(B))/D(\boldsymbol{\alpha}_k(B))},$$

where $\boldsymbol{n}_k^j = (n_k(B_0^j), n_k(B_1^j))$. Then,

1. $\log \mathcal{L}_{0,1}(B, j, \boldsymbol{x}) / \log n \xrightarrow{p} \sigma'$ *for some* $\sigma' > 0$ *if* $p_1^j = p_2^j$.

2. $\log \mathcal{L}_{0,1}(B, j, \boldsymbol{x}) / n \xrightarrow{p} \sigma''$ *for some* $\sigma'' < 0$ *if* $p_1^j \neq p_2^j$.

Proof. The case $p_1^j = p_2^j$ follows by Theorem 1 in Holmes et al. (2009), but I report the proof since it used to show the case $p_1^j \neq p_2^j$. Using Stirling's formula

$$D(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \simeq \sqrt{2\pi} \frac{x^{x-1/2} y^{y-1/2}}{(x+y)^{x+y-1/2}}, \text{ for large } x, y, \tag{A.1}$$

I can approximate the ratio

$$\begin{aligned}
\mathcal{L}_{0,1}(B, j, \boldsymbol{x}) &\simeq \frac{\prod_{k=1,2} D(\boldsymbol{\alpha}_k(B))}{D(\boldsymbol{\alpha}_0(B))} \cdot \frac{1}{\sqrt{2\pi}} \cdot \frac{(\hat{p}^j)^{\alpha_0(B,0)-1/2}(1-\hat{p}^j)^{\alpha_0(B,1)-1/2}}{\prod_k (\hat{p}_k^j)^{\alpha_k(B,0)-1/2}(1-\hat{p}_k^j)^{\alpha_k(B,1)-1/2}} \\
&\quad \cdot \sqrt{\frac{n_1(B_0^j) n_2(B_0^j)}{n(B)}} \cdot \frac{(\hat{p}^j)^{n(B_0^j)}(1-\hat{p}^j)^{n(B_1^j)}}{\prod_k (\hat{p}_k^j)^{n_k(B_0^j)}(1-\hat{p}_k^j)^{n_k(B_1^j)}} \\
&\simeq C\sqrt{n(B)} \frac{(\hat{p}^j)^{n(B_0^j)}(1-\hat{p}^j)^{n(B_1^j)}}{\prod_k (\hat{p}_k^j)^{n_k(B_0^j)}(1-\hat{p}_k^j)^{n_k(B_1^j)}},
\end{aligned}$$

for some $C > 0$. Then, $\log \mathcal{L}_{0,1}(B, j, \boldsymbol{x}) \propto \frac{1}{2} \log n(B) + \log \mathcal{L}$, where

$$\mathcal{L} = (\hat{p}^j)^{n(B_0^j)}(1-\hat{p}^j)^{n(B_1^j)} / \prod_k (\hat{p}_k^j)^{n_k(B_0^j)}(1-\hat{p}_k^j)^{n_k(B_1^j)}.$$

Then $\mathcal{L}$ is the likelihood ratio for testing composite hypotheses $H_0 : p_1^j = p_2^j = p^j$ vs $H_1 : p_1^j, p_2^j \in (0,1)$. Under the null, $-2\log \mathcal{L} \xrightarrow{d} \chi_1^2$ by Wilks (1938) and so $\log \mathcal{L}_{0,1}(B, j, \boldsymbol{x}) / \log n \xrightarrow{p} \sigma'$ for some $\sigma' > 0$.

Additionally, $\log \mathcal{L}$ can be also written as $\log \mathcal{L} = n(B) Y_{n(B)}$,

$$Y_{n(B)} \xrightarrow{p} \big(G(p^j) - \xi G(p_1^j) - (1-\xi) G(p_2^j)\big),$$

where

$$G(x) = x\log(x) + (1-x)\log(1-x) \quad \text{for } x \in (0,1).$$

If $p_1^j \neq p_2^j$, then $G(p^j) - \xi G(p_1^j) - (1-\xi) G(p_2^j) < 0$, since G is convex and $p^j = \xi p_1^j + (1-\xi) p_2^j$. Thus, $\log \mathcal{L}_{0,1}(B, j, \boldsymbol{x})/n \xrightarrow{p} \sigma''$ for some $\sigma'' < 0$. □

Lemma 16. *Consider $B \in \mathcal{B}^{(\infty)}$ such that $p_1^j = p_2^j$ for all $j = 1, \ldots, p$. If*

1. *$\eta \in (0,1)$;*
2. *$\lambda_j(B), \lambda_j(B_i^{j'}) > 0$ for $j, j' = 1, \ldots, p$ and $i = 0, 1$;*
3. *$\rho_{0,r'}(B)\rho_{0,r'}(B_i^j) > 0$ for $r' = 0, 1$, $i = 0, 1$ and $j = 1, \ldots, p$;*

then,

$$\Pr(Z(B) = (0, j', 1) | Z(\textit{parent}(B)) = (0, j, 0), \boldsymbol{x}) \xrightarrow{p} 0,$$

for any $j, j' = 1, \ldots, p$.

Proof. Define $z' = (0, j', 1)$ and $z = (0, j, 0)$, then

$$\hat{\rho}_{z,z'}(B) = \hat{\rho}_{z,z'}(B) \frac{\Lambda_{z'}(\boldsymbol{x}, B)}{\Phi_z(\boldsymbol{x}, B)} \leq \frac{\hat{\rho}_{z,z'}(B)}{\hat{\rho}_{z,z''}(B)} \frac{\Lambda_{z'}(\boldsymbol{x}, B)}{\Lambda_{z''}(\boldsymbol{x}, B)}.$$

where $z'' = (0, j', 0)$. Since $\hat{\rho}_{z,z'}(B)/\hat{\rho}_{z,z''}(B)$ is bounded for Conditions (1-3), then it is sufficient to show that

$$\frac{\Lambda_{z'}(\boldsymbol{x}, B)}{\Lambda_{z''}(\boldsymbol{x}, B)} \xrightarrow{p} 0.$$

Note that

$$\frac{\Lambda_{z'}(\boldsymbol{x}, B)}{\Lambda_{z''}(\boldsymbol{x}, B)} = \mathcal{L}_{0,1}(B, j', \boldsymbol{x})^{-1} \prod_{i=0,1} \frac{\Phi_{z'}(B_i^{j'}, \boldsymbol{x})}{\Phi_{z''}(B_i^{j'}, \boldsymbol{x})},$$

where $\mathcal{L}_{0,1}(B, j', \boldsymbol{x})^{-1} \xrightarrow{p} 0$ by Lemma 15, and for $z_i^* = \text{argmax}_{z \in \mathcal{Z}} \Lambda_z(B_i^{j'}, \boldsymbol{x})$,

$$\begin{aligned}\frac{\Phi_{z'}(B_i^{j'}, \boldsymbol{x})}{\Phi_{z''}(B_i^{j'}, \boldsymbol{x})} &= \frac{\sum_{\tilde{z}} \hat{\rho}_{z',\tilde{z}} \Lambda_{\tilde{z}}(B_i^{j'}, \boldsymbol{x})}{\sum_{\tilde{z}} \hat{\rho}_{z'',\tilde{z}} \Lambda_{\tilde{z}}(B_i^{j'}, \boldsymbol{x})} \\ &\leq \frac{\Lambda_{z_i^*}(B_i^{j'}, \boldsymbol{x})}{\sum_{\tilde{z}} \hat{\rho}_{z'',\tilde{z}} \Lambda_{\tilde{z}}(B_i^{j'}, \boldsymbol{x})} \leq \frac{1}{\hat{\rho}_{z'',z_i^*}(B_j^{j'})},\end{aligned}$$

is bounded since $\hat{\rho}_{z'',z_i^*}(B_i^{j'}) \in (0, 1)$ for Conditions (1-3). □

A.1.5 Proof of Theorem 9

First I find the tree topologies with highest marginal posterior probability under the null component of (Q_1, Q_2). To this end, assume $\boldsymbol{x}_k|\tilde{Q} \sim \tilde{Q}$ for $k = 1, 2$, and $\tilde{Q} \sim \text{OPT}(\tilde{\boldsymbol{\rho}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\alpha}}, U)$, where $\tilde{\rho}(B) = \eta$, $\tilde{\lambda}_j(B) = \lambda_j(B)$, $\tilde{\boldsymbol{\alpha}}^j(B) = \boldsymbol{\alpha}_0(B)$ for all $B \in \mathcal{B}^{(\infty)}$ and $j = 1, \ldots, p$, and U indicates the uniform distribution on Ω. Define $\mathcal{T}^{(k)}$ a tree topology where all regions are stopped at most at level k, and call $\mathbb{T}^{(k)}$ the collection of the $\mathcal{T}^{(k)}$'s with highest marginal posterior probability as $n \to \infty$. Since there exists $B \in \mathcal{B}^{(\infty)}$ where $P_1(B_i^j|B) \neq P_2(B_i^j|B)$ and $\xi P_1(B_i^j|B) + (1 - \xi) P_2(B_i^j|B) \neq 1/2$, by Theorem 4 in Wong and Ma (2010) for some large enough k, there is at least one non-stopped region C arising in $\mathcal{T}^{(k)}$ such that $P_1(C_i^j|C) \neq P_2(C_i^j|C)$, for each $\mathcal{T}^{(k)} \in \mathbb{T}^{(k)}$. Then, I show that the likelihood of rejecting the local null hypothesis on C dominates the likelihood of accepting it. In fact, define $z = (0, j, 0)$ and

$z' = (0, j, 1)$, then

$$\frac{\Lambda_z(\boldsymbol{x}, B)}{\Lambda_{z'}(\boldsymbol{x}, B)} = \mathcal{L}_{0,1}(B, j, \boldsymbol{x}) \prod_{i=0,1} \frac{\Phi_z(B_i^j, \boldsymbol{x})}{\Phi_{z'}(B_i^j, \boldsymbol{x})},$$

where $\mathcal{L}_{0,1}(B, j, \boldsymbol{x}) \xrightarrow{p} -\infty$ by Lemma 15. For $z_i^* = \text{argmax}_{z \in \mathcal{Z}} \Lambda_z(B_i^j, \boldsymbol{x})$,

$$\begin{aligned}\frac{\Phi_z(B_i^j, \boldsymbol{x})}{\Phi_{z'}(B_i^j, \boldsymbol{x})} &= \frac{\sum_{\tilde{z}} \hat{\rho}_{z,\tilde{z}} \Lambda_{\tilde{z}}(B_i^j, \boldsymbol{x})}{\sum_{\tilde{z}} \hat{\rho}_{z',\tilde{z}} \Lambda_{\tilde{z}}(B_i^j, \boldsymbol{x})} \\ &\leq \frac{\Lambda_{z_i^*}(B_i^j, \boldsymbol{x})}{\sum_{\tilde{z}} \hat{\rho}_{z',\tilde{z}} \Lambda_{\tilde{z}}(B_i^j, \boldsymbol{x})} \leq \frac{1}{\hat{\rho}_{z',z_i^*}(B_j^j)},\end{aligned}$$

is bounded since $\hat{\rho}_{z',z_i^*}(B_i^j) \in (0, 1)$. This implies that

$$\Pr(H_0|\boldsymbol{x}) \xrightarrow{p} 0 \text{ under } P_1^{(\infty)} \times P_2^{(\infty)}.$$

A.1.6 Proof of Theorem 10

For any set $B \in \mathcal{B}^l$ for $l > k$, I have that $\Psi(B, \boldsymbol{x}) = 1$, since all regions are stopped by design. For any set $B \in \mathcal{B}^{(k)}$, if $\Psi(B_i^j, \boldsymbol{x}) \xrightarrow{p} 1$ for any $j = 1, \ldots, p$, and $i \in \{0, 1\}$, then, by Slutsky's theorem,

$$\begin{aligned}\Psi(B, \boldsymbol{x}) \xrightarrow{p} &\Pr(Z(B) = (1, 0, 0)|Z(\text{parent}(B)) = (0, j', 0), \boldsymbol{x}) \\ &+ \sum_{j=1}^{p} \Pr(Z(B) = (0, j, 0)|Z(\text{parent}(B)) = (0, j', 0), \boldsymbol{x}),\end{aligned}$$

for any $j' = 1, \ldots, p$. One can write the right-hand side as follows:

$$\begin{aligned}&\Pr(Z(B) = (1, 0, 0)|Z(\text{parent}(B)) = (0, j', 0), \boldsymbol{x}) \\ &\quad + \sum_{j=1}^{p} \Pr(Z(B) = (0, j, 0)|Z(\text{parent}(B)) = (0, j', 0), \boldsymbol{x}) \\ &\quad = 1 - \sum_{j=1}^{p} \Pr(Z(B) = (0, j, 1)|Z(\text{parent}(B)) = (0, j', 0), \boldsymbol{x}),\end{aligned}$$

where $\Pr(Z(B) = (0, j, 1)|Z(\text{parent}(B)) = (0, j', 0), \boldsymbol{x}) \xrightarrow{p} 0$ for all $j, j' = 1, \ldots, p$ by Lemma 16. This implies that

$$\Psi(B, \boldsymbol{x}) \xrightarrow{p} 1 \quad \text{and} \quad \Pr(H_0|\boldsymbol{x}) \xrightarrow{p} 1.$$

A.1.7 Proof of Lemma 11

The likelihood restricted to B has the following recursive representation:

$$\Pr(\boldsymbol{x}(B)|\boldsymbol{\theta}_1(B), \boldsymbol{\theta}_2(B), \mathcal{T}(B))$$

$$= \begin{cases} \mu(B)^{-\sum_k n_k(B)} & \text{if } S(B) = 1 \\ \prod_{k=1,2} 2^{n_k(B)} \Pr(\boldsymbol{n}_k^j(B)|\theta_k^j(B)) \prod_{i=0,1} \Pr(\boldsymbol{x}(B_i^j)|\boldsymbol{\theta}_1(B_i^j), \boldsymbol{\theta}_2(B_i^j), \mathcal{T}(B_i^j)) & \text{if } S(B) = 0. \end{cases}$$

The prior $\Pi_z(\boldsymbol{\theta}_1(B), \boldsymbol{\theta}_2(B), \boldsymbol{Z}(B), \boldsymbol{J}(B))$ has the following factorization:

$$\begin{aligned} \Pi_z(\boldsymbol{\theta}_1(B), \boldsymbol{\theta}_2(B), \boldsymbol{Z}(B), \boldsymbol{J}(B)) &= \hat{\rho}_{z,(1,0)}(B)\Pi_z(Z(B) = (1, 0)) \\ &+ \sum_{z' \in \mathcal{Z} \setminus (1,0)} \sum_j \lambda_j \hat{\rho}_{z,z'}(B)\Pi_z(\theta_1(B), \theta_2(B), Z(B) = z', J(B) = j) \\ &\times \prod_{i=0,1} \Pi_{z'}(\boldsymbol{\theta}_1(B_i^j), \boldsymbol{\theta}_2(B_i^j), \boldsymbol{Z}(B_i^j)), \end{aligned}$$

for $B \in \mathcal{B}_p^{(\infty)} \setminus \Omega$. For the root, i.e., $B = \Omega$, I have

$$\begin{aligned} \Pi(\boldsymbol{\theta}_1(B), \boldsymbol{\theta}_2(B), \boldsymbol{Z}(B), \boldsymbol{J}(B)) &= \hat{\pi}_{(1,0)}\Pi(Z(B) = (1, 0)) \\ &+ \sum_{z' \in \mathcal{Z} \setminus (1,0)} \sum_j \lambda_j \hat{\pi}_{z'}\Pi(\theta_1(B), \theta_2(B), Z(B) = z', J(B) = j) \\ &\times \prod_{i=0,1} \Pi_{z'}(\boldsymbol{\theta}_1(B_i^j), \boldsymbol{\theta}_2(B_i^j), \boldsymbol{Z}(B_i^j)). \end{aligned}$$

The result follows from integration of the right-side of the likelihood restricted to B with respect to the prior.

A.1.8 Proof of Theorem 12

For the transition probabilities I have:

$$\begin{aligned}\Pr(Z(B) = z'|Z(\text{parent}(B)) = z, \boldsymbol{x}) &= \Pr(Z(B) = z'|Z(\text{parent}(B)) = z, \boldsymbol{x}(B)) \\ &= \frac{\Pr(\boldsymbol{x}(B), Z(B) = z'|Z(\text{parent}(B)) = z)}{\Pr(\boldsymbol{x}(B)|Z(\text{parent}(B)) = z)} \\ &= \Pr(Z(B) = z'|Z(\text{parent}(B)) = z) \\ &\times \frac{\sum_j \Pr(\boldsymbol{x}(B)|Z(B) = z', J(B) = j) \Pr(J(B) = j)}{\sum_j \Pr(\boldsymbol{x}(B)|Z(\text{parent}(B)) = z, J(B) = j) \Pr(J(B) = j)} \\ &= \hat{\rho}_{z,z'}(B) \frac{\Lambda_{z'}(\boldsymbol{x}, B)}{\Phi_z(\boldsymbol{x}, B)},\end{aligned}$$

where $B \in \mathcal{B}_p^{(\infty)} \setminus \Omega$ and $z, z' \in \mathcal{Z}$. Similarly, one can derive the posterior initial state probabilities.

For the direction probabilities I have:

$$\begin{aligned}\Pr(J(B) = j|Z(B) = z, \boldsymbol{x}) &= \Pr(J(B) = j|Z(B) = z, \boldsymbol{x}(B)) \\ &= \frac{\Pr(\boldsymbol{x}(B)|J(B) = j, Z(B) = z) \Pr(J(B) = j)}{\sum_j \Pr(\boldsymbol{x}(B)|J(B) = j, Z(B) = z) \Pr(J(B) = j)},\end{aligned}$$

where where $B \in \mathcal{B}_p^{(\infty)}$, $z \in \mathcal{Z}$ and $j = 1, \ldots, p$.

Finally, since the Beta distribution is conjugate to the Multinomial distribution, I have:

$$\begin{aligned}\boldsymbol{\alpha}_0(B|\boldsymbol{x}) &= \boldsymbol{\alpha}_0(B) + \boldsymbol{n}_1(B) + \boldsymbol{n}_2(B) \\ \boldsymbol{\alpha}_k(B|\boldsymbol{x}) &= \boldsymbol{\alpha}_k(B) + \boldsymbol{n}_k(B)\end{aligned}$$

for all $B \in \mathcal{B}_p^{(\infty)}$ and $k = 1, 2$.

A.2 Parameters for Numerical Example 2

1. Local shift difference:

   $(p_1, \ldots, p_5) = (0.11, 0.16, 0.25, 0.39, 0.09), \delta = (1.0, 1.0),$

$\mu_1 = (9.0, 9.9), \mu_2 = (0.0, 4.4), \mu_3 = (-2.3, -9.7),$

$\mu_4 = (3.4, 5.9), \mu_5 = (5.8, -9.5),$

$(\Sigma_1(1,1), \Sigma_1(1,2), \Sigma_1(2,2)) = (2.9, 0.5, 1.1),$

$(\Sigma_2(1,1), \Sigma_2(1,2), \Sigma_2(2,2)) = (1.2, -0.6, 2.8),$

$(\Sigma_3(1,1), \Sigma_3(1,2), \Sigma_3(2,2)) = (2.3, -1.0, 1.7),$

$(\Sigma_4(1,1), \Sigma_4(1,2), \Sigma_4(2,2)) = (1.1, -0.4, 2.9),$

$(\Sigma_5(1,1), \Sigma_5(1,2), \Sigma_5(2,2)) = (3.0, 0.2, 1.0).$

2. Local dispersion difference:

   $(p_1, \ldots, p_5) = (0.19, 0.08, 0.33, 0.27, 0.13),$

   $\mu_1 = (0.9, -7.2), \mu_2 = (-5.7, 3.3), \mu_3 = (-6.3, -2.1),$

   $\mu_4 = (7.5, -3.1), \mu_5 = (-3.1, 9.5),$

   $(\Sigma_1(1,1), \Sigma_1(1,2), \Sigma_1(2,2)) = (0.5, -0.1, 0.3),$

   $(\Sigma_2(1,1), \Sigma_2(1,2), \Sigma_2(2,2)) = (1.3, 0.7, 2.7),$

   $(\Sigma_3(1,1), \Sigma_3(1,2), \Sigma_3(2,2)) = (1.0, -0.3, 3.0),$

   $(\Sigma_4(1,1), \Sigma_4(1,2), \Sigma_4(2,2)) = (2.9, 0.5, 1.1),$

   $(\Sigma_5(1,1), \Sigma_5(1,2), \Sigma_5(2,2)) = (2.4, -0.9, 1.6).$

3. Global shift difference: $(\Sigma(1,1), \Sigma(1,2), \Sigma(2,2)) = (2.9, 0.4, 1.1).$

Appendix B

Appendix for Chapter 3

B.1 Parameters for Numerical Example 1

1. Single location shift and multiple location shifts:

   $\Sigma_1(i,i) = 1.1 \quad \text{for } i = 1,\ldots,4, \quad \Sigma_1(i,j) = 0.9 \quad \text{for } i \neq j \text{ and } i,j = 1,\ldots,4;$

   $\Sigma_2(i,i) = 2.0 \quad \text{for } i = 1,\ldots,4, \quad \Sigma_2(i,j) = 1.0 \quad \text{for } i \neq j \text{ and } i,j = 1,\ldots,4;$

   $\Sigma_3(i,i) = 0.4 \quad \text{for } i = 1,\ldots,4, \quad \Sigma_3(i,j) = -0.1 \quad \text{for } i \neq j \text{ and } i,j = 1,\ldots,4;$

   $\Sigma_4(i,i) = 0.1 \quad \text{for } i = 1,\ldots,4, \quad \Sigma_4(i,j) = 0.0 \quad \text{for } i \neq j \text{ and } i,j = 1,\ldots,4;$

   $\boldsymbol{\pi} = (0.3, 0.3, 0.2, 0.2).$

2. Local weight difference: Σ_k for $k = 1,\ldots,4$ are identical to the single location shift and multiple location shifts.

3. Global weight differences:

   $$\Sigma_1 = \text{diag}(1,1,1,1);$$

   $$\Sigma_2 = \text{diag}(2,2,2,2);$$

   $$\Sigma_3 = \text{diag}(0.2,0.2,0.2,0.2);$$

   $$\Sigma_k = \text{diag}(0.1,0.1,0.1,0.1) \quad \text{for } k = 4,\ldots,8.$$

Bibliography

Abramovich, F. and Angelini, C. (2006), "Testing in mixed-effects FANOVA models," *Journal of statistical planning and inference*, 136, 4326–4348.

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998), "Wavelet thresholding via a Bayesian approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 725–749.

Abramovich, F., Antoniadis, A., Sapatinas, T., and Vidakovic, B. (2004), "Optimal testing in a fixed-effects functional analysis of variance model," *International Journal of Wavelets, Multiresolution and Information Processing*, 2, 323–349.

Anders, S. and Huber, W. (2010), "Differential expression analysis for sequence count data," *Genome biol*, 11, R106.

Antoniadis, A. and Sapatinas, T. (2007), "Estimation and inference in functional mixed-effects models," *Computational Statistics & Data Analysis*, 51, 4793–4813.

Baringhaus, L. and Franz, C. (2004), "On a new multivariate two-sample test," *Journal of Multivariate Analysis*, 88, 190–206.

Bernardo, J. M. and Smith, A. F. (2009), *Bayesian theory*, vol. 405, John Wiley & Sons.

Boedigheimer, M. J. and Ferbas, J. (2008), "Mixture modeling approach to flow cytometry data," *Cytometry Part A*, 73, 421–429.

Cahouët, V., Luc, M., and David, A. (2002), "Static optimal estimation of joint accelerations for inverse dynamics problem solution," *Journal of Biomechanics*, 35, 1507–1513.

Chan, C., Feng, F., Ottinger, J., Foster, D., West, M., and Kepler, T. B. (2008), "Statistical mixture modeling for cell subtype identification in flow cytometry," *Cytometry Part A*, 73, 693–701.

Chen, L., Dai, P., and Dou, W. (2010), *MTSKNN: Multivariate two-sample tests based on K-nearest-neighbors*, R package version 0.0-5.

Chen, Y. and Hanson, T. E. (2014), "Bayesian nonparametric k-sample tests for censored and uncensored data," *Computational Statistics & Data Analysis*, 71, 335–346.

Chipman, H. A., Kolaczyk, E. D., and McCulloch, R. E. (1997), "Adaptive Bayesian wavelet shrinkage," *Journal of the American Statistical Association*, 92, 1413–1421.

Clyde, M., Parmigiani, G., and Vidakovic, B. (1998), "Multiple shrinkage and subset selection in wavelets," *Biometrika*, 85, 391–401.

Cron, A. and Frelinger, J. (2013), "dpmix 0.3: Python library for Bayesian inference for hierarchical and standard Dirichlet process mixtures of normals using GPU," .

Cron, A., Gouttefangeas, C., Frelinger, J., Lin, L., Singh, S. K., Britten, C. M., Welters, M. J., van der Burg, S. H., West, M., and Chan, C. (2013), "Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples," *PLoS computational biology*, 9, e1003130.

Crouse, M. S., Nowak, R. D., and Baraniuk, R. G. (1998), "Wavelet-based statistical signal processing using hidden Markov models," *Signal Processing, IEEE Transactions on*, 46, 886–902.

Daubechies, I. et al. (1992), *Ten lectures on wavelets*, vol. 61, SIAM.

Donoho, D. L. and Johnstone, I. M. (1995), "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the american statistical association*, 90, 1200–1224.

Donoho, D. L. and Johnstone, J. M. (1994), "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, 81, 425–455.

Dunson, D. B. (2010), "Nonparametric Bayes applications to biostatistics," *Bayesian nonparametrics*, 28, 223.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the american statistical association*, 90, 577–588.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230.

Franz, C. (2006), *cramer: Multivariate nonparametric Cramer-Test for the two-sample-problem*, R package version 0.8-1.

Ghosh, J. and Ramamoorthi, R. (2003), "Bayesian Nonparametrics," .

Green, P. J. and Richardson, S. (2001), "Modelling heterogeneity with and without the Dirichlet process," *Scandinavian journal of statistics*, 28, 355–375.

Griffin, J. E., Kolossiatis, M., and Steel, M. F. (2013), "Comparing distributions by using dependent normalized random-measure mixtures," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 499–529.

Henze, N. (1988), "A multivariate two-sample test based on the number of nearest neighbor type coincidences," *The Annals of Statistics*, pp. 772–783.

Holmes, C., Caron, F., Griffin, J., and Stephens, D. A. (2009), "Two-sample Bayesian nonparametric hypothesis testing," *arXiv preprint arXiv:0910.5060*.

Ishwaran, H. and James, L. F. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96.

Ishwaran, H. and Zarepour, M. (2002), "Exact and approximate sum representations for the Dirichlet process," *Canadian Journal of Statistics*, 30, 269–283.

Jara, A., Hanson, T. E., Quintana, F. A., Müller, P., and Rosner, G. L. (2011), "DPpackage: Bayesian semi-and nonparametric modeling in R," *Journal of statistical software*, 40, 1.

Jefferys, W. H. and Berger, J. O. (1992), "Ockham's razor and Bayesian analysis," *American Scientist*, pp. 64–72.

Johnstone, I. M. and Silverman, B. W. (1997), "Wavelet threshold estimators for data with correlated noise," *Journal of the royal statistical society: series B (statistical methodology)*, 59, 319–351.

Kingman, J. F. (1975), "Random discrete distributions," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–22.

Lavine, M. (1992), "Some aspects of Pólya tree distributions for statistical modelling," *The Annals of Statistics*, pp. 1222–1235.

Liu, J. S. (2008), *Monte Carlo strategies in scientific computing*, Springer.

Lo, K., Brinkman, R. R., and Gottardo, R. (2008), "Automated gating of flow cytometry data via robust model-based clustering," *Cytometry Part A*, 73, 321–332.

Lopes, H. F., Müller, P., and Rosner, G. L. (2003), "Bayesian Meta-analysis for Longitudinal Data Models Using Multivariate Mixture Priors," *Biometrics*, 59, 66–75.

Ma, L. and Wong, W. H. (2011), "Coupling optional Pólya trees and the two sample problem," *Journal of the American Statistical Association*, 106.

Morris, J. S. and Carroll, R. J. (2006), "Wavelet-based functional mixed models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 179–199.

Müller, P. and Quintana, F. A. (2004), "Nonparametric Bayesian data analysis," *Statistical science*, pp. 95–110.

Müller, P., Quintana, F., and Rosner, G. (2004), "A method for combining inference across related nonparametric Bayesian models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 735–749.

Nason, G. (2012), "wavethresh: Wavelets statistics and transforms. R package version 4.5," .

Neal, R. M. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of computational and graphical statistics*, 9, 249–265.

Papaspiliopoulos, O. and Roberts, G. O. (2008), "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models," *Biometrika*, 95, 169–186.

Pitman, J. and Yor, M. (1997), "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *The Annals of Probability*, pp. 855–900.

Porteous, I., Ihler, A. T., Smyth, P., and Welling, M. (2012), "Gibbs sampling for (coupled) infinite mixture models in the stick breaking representation," *arXiv preprint arXiv:1206.6845*.

Rosner, G. L. and Vidakovic, B. (2000), "Wavelet functional ANOVA, Bayesian false discovery rate, and longitudinal measurements of oxygen pressure in rats," .

Schilling, M. F. (1986), "Multivariate two-sample tests based on nearest neighbors," *Journal of the American Statistical Association*, 81, 799–806.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Singh, S. K., Tummers, B., Schumacher, T. N., Gomez, R., Franken, K. L., Verdegaal, E. M., Laske, K., Gouttefangeas, C., Ottensmeier, C., Welters, M. J., et al. (2013), "The development of standard samples with a defined number of antigen-specific T cells to harmonize T cell assays: a proof-of-principle study," *Cancer Immunology, Immunotherapy*, pp. 1–13.

Student (1908), "The probable error of a mean," *Biometrika*, pp. 1–25.

Teh, Y. W. and Jordan, M. I. (2010), "Hierarchical Bayesian nonparametric models with applications," *Bayesian nonparametrics*, pp. 158–207.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006), "Hierarchical dirichlet processes," *Journal of the american statistical association*, 101.

Wilks, S. S. (1938), "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, 9, 60–62.

Wong, W. H. and Ma, L. (2010), "Optional Pólya tree and Bayesian inference," *The Annals of Statistics*, 38, 1433–1459.

Biography

Jacopo Soriano was born in Rovigo, Italy on April 24, 1985. In 2004 he began his undergraduate studies at the Politecnico di Milano, Italy, where he originally pursued a degree in mechanical engineering before switching to mathematical engineering. As part of a double degree program he studied from 2006 to 2008 at the École Centrale Paris, France. He received his Bachelor of Science in 2008 from the Politecnico di Milano. In 2008 he began his graduate studies in mathematical engineering at the Politecnico di Milano, and in 2010 he received a Master of Science in mathematical engineering from the Politecnico di Milano and a Master of Science in engineering from the École Centrale Paris. In 2011 he began his graduate studies in the Department of Statistical Science at Duke University. In 2013 he earned a Master of Science in Statistical Science *en route* to the Ph.D., and in 2015 he graduated with a Doctor of Philosophy under the supervision of Prof. Li Ma. As a graduate student at Duke University he was awarded the 2014 Katherine Goodman Stern Fellowship.