University of Aix-Marseille
Department of Mathematics and Applications

Master's Thesis

Adaptive and non-adaptive test on components of mixture densities over Besov spaces

Mathieu MAISON

September 2015

Supervisor: Mr Florent AUTIN, University of Aix-Marseille, L.A.T.P.

Summary: This Master's Thesis aims to gather several recent theoretical results given in papers written by F. Autin and C. Pouet on the optimal testing procedure for the components of mixture densities in the Besov spaces. They presented solutions for the non-adaptive and the adaptive cases and when the problem is ill-posed. Their results are provided for the wavelet method for which an introduction and some practical results are given within this paper. Numerical implementations of the authors' theoretical results are given.

Table of Contents

I. Introduction 3

II. Setting of the problem 5

1. The mixture model with varying weights 5
2. Hypothesis testing and Minimax approach 6
3. Besov spaces 8

III. Wavelet framework 11

1. Overview 11
2. Wavelet system construction 11
3. Some examples of wavelets 13
4. Statistical estimation with wavelets 16

IV. Testing procedure 20

1. Test statistics 20
2. Non-adaptive case 21
   a. Fixed smallest eigenvalue of the Gram matrix 21
   b. Variable smallest eigenvalue of the Gram matrix 23
3. Adaptive case 24
   a. Fixed smallest eigenvalue of the Gram matrix 24
   b. Variable smallest eigenvalue of the Gram matrix 25

V. Numerical Experiments 27

1. Data description 27
2. Computation methodology 31
3. Results analysis 33

VI. Open question 37

1. Mixing weights behavior 37
2. Varying mixing weights results 38

VII. Conclusion 43

VIII. Appendix 44

IX. Bibliography 64

I. Introduction

The mixture model is a modeling tool used in data modeling which became very popular among several domains as economy, finance, physics, social science, marketing, etc. thanks to its ease of interpretation. For example Lodatko and Maiboroda (2007) studied the difference between atypical pneumonia patients and healthy patients.

Mixture models are based on two sets of parameters called the mixing weights and the mixing components (cf. II.1). The first papers dealt with the estimation of these parameters and especially on the weights' estimation (Hall (1981), Titterington (1983), Hall and Titterington (1984)). Later papers focused on the mixing components' estimation (Maiboroda (1996, 2000a), Pokhyl'ko (2005)). The mixing weights are divided in two groups: the fixed ones and the variable ones. We can note that Hall and Zhou (2003), proved that, in the particular setting for k-variate data, mixing weights and mixing components can both be estimated at the same time.

Recently the mixture model has been introduced in the testing problem framework. The usual problem is to know if the observations come from a trivial or a non-trivial mixture densities model. It has been done by Garel (2001, 2005) and Delmas (2003) for the fixed mixing weights and by Maidoroba (2000b) the varying ones. Their tests are based on the Kolmogorov-Smirnov test. All of these authors deal with one sample. Here we focus on a test used to determine if two samples have the same mixture densities or not. This approach has already been studied by Butucea and Tribouley (2006) in the minimax setting for the case $M = 1$. These authors also gave an automatic adaptive, with respect to the smoothness of the underlying densities, test procedure.

In this paper we summarize three articles from F. Autin and C. Pouet ([1], [2] and [3]) which deal with the construction of a test procedure designed to see if two independent samples of n independent random variables, in a mixture model with varying mixing weights, are based on the same mixture of M ($M > 1$) mixing components or not. In the first article [1] the authors focus on the non-adaptive case and give a test procedure which is asymptotically optimal in the minimax setting. In the second article [2] they analyze the adaptive case and also give an optimal test procedure. They proved that:

- the minimax rate of testing over the balls of the Besov space $B^s_{2,\infty}$ is of order $n^{-2s/(4s+1)}$
- the minimax adaptive rate of testing over such spaces get a loss of order $(\ln(\ln(n)))^{-1/2}$ which is the consequence of the adaptability of the rate. This loss is unavoidable.

While in the first two articles the smallest eigenvalue of the Gram matrix was fixed and larger than or equal to $k \in (0,1)$, the third article [3] sets the problem when the smallest eigenvalue is decreasing to zero as n increases.

In this paper, like in [1], [2] and [3], the mixing weights are assumed to be known while we deal with the mixing components.

The first part of this article describes the set-up of the problem, introducing the definition of mixture model with varying weights, the hypothesis testing and the Besov spaces.

The Besov spaces, as we will see, are very useful as they admit, contrary to Sobolev spaces, a characterization in terms of wavelet coefficients.

The second part introduces the wavelet method through the construction of the wavelet system, gives some examples of wavelet basis and we give a nonparametric statistical method which allows to estimate a function with wavelets.

The third part describes the test procedures and gives the main results regarding the minimax performance of these tests. We give results for the non-adaptive case, the adaptive case, the case where the smallest eigenvalue of the Gram matrix is fixed and where it's variable (ill-conditioned framework).

The fourth section is a numerical implementation of the theoretical results given in the preceding parts. We focus on a testing procedure which allows to determine if the mixing components of the densities of the total number of working hours in two French departments are the same or not. We analyze the departments 02 and 03 and we focus on the full time employees' working hours. For computation constraints we only choose two mixing components and we use the Haar wavelet basis framework. In this section the mixing weights are assumed to be known and thus fixed.

In the fifth part we also perform a numerical computation but we assume that the mixing weights are not exactly known and consequently are varying, which can be considered as a closer case to reality than the fixed weights case. We consider that they follow a normal distribution around their fixed values given in the preceding section. We analyze the results and compare them to the fixed mixing weights case. This part is analyzed on a purely empirical basis.

A conclusion of our results is given in the last section.

The proofs of the main results are given in the Appendix. For the section III. we refer to [5] for the demonstrations.