Statistical Methodology 10 (2012) 151–169

Contents lists available at SciVerse ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

# Testing the means of subgroups in the varying mixing weight model

F. Autin [a,*], C. Pouet [b]

[a] Université Aix-Marseille 1, Laboratoire d'Analyse, Topologie, Probabilités, 39 rue F. Joliot-Curie, 13453 Marseille cedex 13, France
[b] Ecole Centrale Marseille, Laboratoire d'Analyse, Topologie, Probabilités, 38 rue F. Joliot-Curie, 13451 Marseille cedex 20, France

ARTICLE INFO

*Article history:*
Received 8 November 2011
Received in revised form
23 July 2012
Accepted 5 August 2012

*Keywords:*
Asymptotic distribution
Hypothesis testing
Mixture models

ABSTRACT

We consider two groups divided into several subgroups and we are interested in comparing the means of two subgroups, one from each group. The samples are drawn from the two groups and the subgroup label of each observation is not defined with certainty. We show that this problem is connected to the problem of testing the expected values of mixture components with two data samples. The underlying mixture model is associated with known varying mixing weights. We provide a testing procedure which takes into account this uncertainty and performs well. Then we compare the numerical performance of this testing procedure to that of Welch's $t$-test which would have been done if true labels had been available and we assess the loss of performance of our method due to the mixing effect.



## 1. Introduction

In many cases, researchers are interested in gathering information from two populations in order to compare the parameters of subpopulations. In that setting, tests of significance are useful statistical tools for detecting a difference between parameters of two subpopulations. Related application fields are numerous. Some examples are genetics, neuronal data analysis, medicine, biology, physics, chemistry, and social sciences, among other fields.

The comparison between the means of the populations is usually carried out by using $t$-statistic and leads to the well known Student's $t$-test or Welch's $t$-test (see [15]). These $t$-tests are popular because of their ease of use and their good performances. Moreover they are robust in the sense that

* Corresponding author. Tel.: +33 413551015.
*E-mail address:* autin@cmi.univ-mrs.fr (F. Autin).

1572-3127/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2012.08.001

they still perform well when the components are not exactly Gaussian, provided that the sizes of the samples are large enough. Nevertheless, these testing methods require one to know the label of each observation, that is to say the subpopulation that it is associated with. Unfortunately, researchers sometimes do not get this information. Indeed, one can imagine some cases where the labels of data are erroneous or uncertain. To give an example of such a situation with lack of information, we consider two populations – people living in New York and in California – restricted to the people that take the bus/trolleybus or walk to go to work. Focusing on the people that take the bus/trolleybus to go to work, we are interested in checking whether the travel time of people living in New York is significantly different from that of people living in California or not. A sample of people from each state – New York or California – is available but the means of transportation (the label) associated with each data is not. This key example will be studied in detail later in this article.

We can describe two other general situations where there is also a lack of information could be very useful. The first one is the confidentiality of statistics collected by national statistics offices such as the Office for National Statistics [10], abbreviated as O.N.S., in the United Kingdom or the Institut National de la Statistique et des Etudes Economiques [4], abbreviated as I.N.S.E.E., in France. This privacy is ensured by the law and any disclosure of data is only possible if the data are made anonymous. It usually means that only aggregated data are released. For example, in France, most statistics concerning households are gathered in a statistical unit called IRIS. Each IRIS is between 1500 and 5000 households. Therefore for research or commercial purposes, one must rely on one's own customized survey in order to get the relevant variables at the microdata level, this costs time and money. The mixture model with varying mixing weights which is described in Section 2 can help to extract information from a large survey recording a few variables of interest. Another example is the formal system of time accounts for Melbourne developed by Ironmonger [5,6]. Statistical data are collected from several information systems such as the Automated Ticketing System, the Australian Bureau of Statistics or the Victorian Activity and Travel Survey. All these databases cannot be directly matched but they might share some objective information such as gender or age. These variables can help to extract information from two databases that could not be matched otherwise.

We want to address the problem of testing the means of two subpopulations when the labels of data are uncertain. More precisely, we first propose to show that this testing problem can be reformulated as a problem of testing with two samples of independent mixture variables. In our study of real data, we shall assume that the mixing weights are known. It means that the proportions of people walking or using the bus/trolleybus to work in each state (New York and California) are known, with respect to an auxiliary variable (age for instance). Then, we provide a testing procedure which takes into account this information on populations and we discuss its performance.

The testing procedure we propose is directly inspired from ideas in [1]. In this previous work, a nonparametric procedure was proposed to test whether the densities of two independent samples of independent random variables had the same mixture components or not. The value of the test statistic requires, in some sense, the mixing weight operators of the samples to be inverted (see Definition 1) as a preliminary step. This testing procedure was proved to be powerful since it is minimax over Besov spaces (more details are given in paragraph 3.1 in [1]). Here we show that a testing procedure that incorporates combinatory ideas – provided that the mixing weights are known – is more relevant than a procedure based on classification.

This paper is organized as follows. In Section 2 we present the mixture model we are interested in. We explain the connection between the testing problem for data with unknown labels and the problem of testing the expected value of the components involved in the mixture model. In Section 3, we present three testing procedures and give some theoretical results. The first procedure is called the *Oracle Procedure* and is based on Welch's $t$-test when the labels of data are exactly known. Of course this procedure is not tractable for the testing problem with missing labels but it will be used as a benchmark to assess the loss of performance of the other testing procedures. The second testing procedure we present is the *Expert Procedure*, it also uses Welch's $t$-test. Each observation is assigned the label corresponding to the largest *a priori* probability. The third and last procedure is the *Mixing Procedure* and uses combinatory properties. Section 4 deals with numerical experiments to point out the good performance of the Mixing Procedure – the one we suggest – compared to the Expert Procedure and to assess the loss of performance due to the mixing effect compared to the Oracle

Procedure. An application to real data is also presented whereas a conclusion with open problems is postponed in Section 5. Finally, the technical lemmas and the proposition used to prove our main theoretical result (see Theorem 1) together with their proofs can be found in the Appendix.

## 2. Model description and hypothesis testing problem

### 2.1. Mixture models with varying mixing weights

Let $X_1, \ldots, X_n$ be independent random variables such that, for any $1 \leq i \leq n$, the density of $X_i$ on $\mathbb{R}$, denoted by $f_{X_i}$, is a mixture density with $M$ ($M \geq 2$) components $p_1, \ldots, p_M$ and $M$ non-negative mixing weights $\omega_1(i), \ldots, \omega_M(i)$, i.e.

$$f_{X_i} = \sum_{u=1}^{M} \omega_u(i) p_u \quad \text{with} \quad \sum_{u=1}^{M} \omega_u(i) = 1.$$

We also introduce the labels attached to $X_1, \ldots, X_n$, denoted by $u_1, \ldots, u_n$. This point of view is one interpretation of mixture models among others (see Section 1.4 in [9]). The main difference lies in considering varying mixing weights in our model. This point is very important (see [1]) and first appeared in [7] to the best of our knowledge.

Similarly to the sample $X_1, \ldots, X_n$, we consider a sample of independent random variables $Y_1, \ldots, Y_n$ such that, for any $1 \leq i \leq n$, the density of $Y_i$ on $\mathbb{R}$, denoted by $f_{Y_i}$, is a mixture density with $M$ ($M \geq 2$) components $p'_1, \ldots, p'_M$ and $M$ non-negative mixing weights $\omega'_1(i), \ldots, \omega'_M(i)$, i.e.

$$f_{Y_i} = \sum_{u=1}^{M} \omega'_u(i) p'_u \quad \text{with} \quad \sum_{u=1}^{M} \omega'_u(i) = 1.$$

We also introduce the labels attached to $Y_1, \ldots, Y_n$, denoted by $v_1, \ldots, v_n$ and we assume that this second sample is independent from the first one.

If $^t$ denotes the transpose operator, the mixture model we have just introduced can be rewritten in a simpler way as follows:

$$\mathbf{f_X} = \Omega_X \mathbf{p} \quad \text{and} \quad \mathbf{f_Y} = \Omega_Y \mathbf{p}', \tag{1}$$

where

- $\mathbf{f_X} = {}^t(f_{X_1}, \ldots, f_{X_n})$, $\mathbf{f_Y} = {}^t(f_{Y_1}, \ldots, f_{Y_n})$,
- $\mathbf{p} = {}^t(p_1, \ldots, p_M)$, $\mathbf{p}' = {}^t(p'_1, \ldots, p'_M)$,
- $\Omega_X = (\omega_u(i))_{i,u}$, $\Omega_Y = \left(\omega'_u(i)\right)_{i,u}$.

**Definition 1.** The $n \times M$-matrices $\Omega_X$ and $\Omega_Y$ involved in the model (1) are called the mixing weight operators.

**Definition 2.** Any mixture model (1) such that $\Omega_X$ and $\Omega_Y$ are full rank matrices is called a mixture model with varying mixing weights.

### 2.2. Example of modeling with a mixture model for $M = 2$

Let us illustrate this theoretical set-up with the example cited in the introduction. The random variables $X_1, \ldots, X_n$ correspond to the travel time of people in the State of New York and the random variables $Y_1, \ldots, Y_n$ to the travel time of people in the State of California. The labels are the means of transportation to go to work and can be either *Bus/trolleybus* (label 1) or *Walked* (label 2). The last step to complete the description of the mixture model is to compute the mixing weights for each observation. In each state the mixing weights strongly depend on the age (over 21 or under 20 years old (abbreviated y.o.)). Table 1 illustrates this fact.

This table should be read as follows. If we consider observation $i$, the mixing weights are

- $(\omega_1(i), \omega_2(i)) = (0.5193, 0.4807)$ if the person is over 21 y.o. and lives in New York,
- $(\omega_1(i), \omega_2(i)) = (0.3465, 0.6535)$ if the person is under 20 y.o. and lives in New York,

**Table 1**
Population weights (and sizes) with respect to age.

| | Bus/trolleybus | Walked |
|---|---|---|
| New York over 21 y.o. | 51.93% (4313) | 48.07% (3993) |
| New York under 20 y.o. | 34.65% (306) | 65.35% (577) |
| California over 21 y.o. | 57.4% (4479) | 42.6% (3324) |
| California under 20 y.o. | 42.77% (497) | 57.23% (665) |

- $(\omega'_1(i), \omega'_2(i)) = (0.4260, 0.5740)$ if the person is over 21 y.o. and lives in California,
- $(\omega'_1(i), \omega'_2(i)) = (0.4277, 0.5723)$ if the person is under 20 y.o. and lives in California.

The reader can legitimately wonder why the age is assumed to be known and not the means of transportation to work. One can think of at least two good reasons. The first one is an *a priori* reason. The survey would be rather lengthy if all necessary variables were included. Therefore the survey is restricted to a small set of informative variables strongly linked with the variables of interest which are unavailable. Moreover these informative variables can be chosen to be as objective as possible and thus easily recordable. This can be called planned missing values (see [3]). The other reason is an *a posteriori* one. During the data analysis of a survey, researchers are often confronted with new hypotheses to test. In many situations, the relevant variables have not been recorded. Researchers have to plan a new survey which includes these new variables in order to check these hypotheses. This leads to a waste of time and money.

The problem of testing the means from data with undefined labels can be associated with the testing problem (2) in the mixture model (1). Indeed, it corresponds to the problem of testing the means when the labels of data are unavailable: the only information on the $X_i$'s label (resp. $Y_i$'s label) is the prior probability $\omega_l(i)$ (resp. $\omega'_l(i)$) for every $l \in \{1, \ldots, M\}$.

### 2.3. Hypothesis testing problem

We recall that two data samples $\mathbf{X} = {}^t(X_1, \ldots, X_n)$ and $\mathbf{Y} = {}^t(Y_1, \ldots, Y_n)$ are available. For a chosen label $l \in \{1, \ldots, M\}$, we are interested in testing whether the components $p_l$ and $p'_l$ have the same expected value or not. We want to address this problem in a general context. Therefore we assume that the variances $\sigma_u^2$ and $\sigma_u'^2$ of the components $p_u$ and $p'_u$ are unknown whatever $u \in \{1, \ldots, M\}$ is.

We denote by $m_l$ and $m'_l$ the expected value of the components that we are focusing on. The testing problem is divided into two hypotheses:

$$\begin{aligned} &\text{null hypothesis } \mathcal{H}_0 : m_l = m'_l, \\ &\text{alternative hypothesis } \mathcal{H}_1 : m_l \neq m'_l. \end{aligned} \tag{2}$$

We recall that to solve the testing problem (2) means to describe a decision rule (or test) $\Delta \in \{0, 1\}$ which relies on the value of a measurable function $T$ (test statistic) of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ only.

As usual, $\Delta = 1$ will mean deciding $\mathcal{H}_1$ whereas $\Delta = 0$ will mean deciding $\mathcal{H}_0$.

## 3. Description of the testing procedures

In this section we introduce three testing procedures.

### 3.1. Oracle test: $\Delta_o$

The first testing procedure we present is called the *Oracle test*. This is a two-step procedure. The first step consists in recovering the true labels of the data. The second step is the application of Welch's $t$-test to data with label $l$ in order to decide whether $m_l$ and $m'_l$ are equal or not. This test cannot be used in our context as the true labels are unknown. Nevertheless it will be used as a benchmark when we

compare the performances of the other testing procedures. It corresponds to the procedure proposed by an oracle in the statistical sense: any statistician having information on labels.

Here we describe in detail the Oracle Procedure. Let us denote

$$n_{l,o} = \sum_{i=1}^{n} \mathbf{1}\{u_i = l\} \quad \text{and} \quad n'_{l,o} = \sum_{i=1}^{n} \mathbf{1}\{v_i = l\},$$

$$\bar{X}_o^{(l)} = \frac{1}{n_{l,o}} \sum_{i=1}^{n} X_i \mathbf{1}\{u_i = l\} \quad \text{and} \quad \bar{Y}_o^{(l)} = \frac{1}{n'_{l,o}} \sum_{i=1}^{n} Y_i \mathbf{1}\{v_i = l\},$$

$$\hat{\sigma}_{l,o}^2 = \frac{1}{n_{l,o}} \sum_{i=1}^{n} (X_i - \bar{X}_o^{(l)})^2 \mathbf{1}\{u_i = l\},$$

$$\hat{\sigma}_{l,o}'^2 = \frac{1}{n'_{l,o}} \sum_{i=1}^{n} (Y_i - \bar{Y}_o^{(l)})^2 \mathbf{1}\{v_i = l\}.$$

The Oracle test $\Delta_o$ is based on the test statistic $T_o$ defined as follows

$$T_o := \frac{|\bar{X}_o^{(l)} - \bar{Y}_o^{(l)}|}{\sqrt{\frac{\hat{\sigma}_{l,o}^2}{n_{l,o}} + \frac{\hat{\sigma}_{l,o}'^2}{n'_{l,o}}}}.$$

Under the null hypothesis, the asymptotic distribution of $T_o$ is known to be the Standard Gaussian one, namely $\mathcal{N}(0, 1)$. Hence, $\Delta_o = \mathbf{1}\{T_o > q_r\}$ is a test with asymptotic type $I$ error equal to $r (0 < r < 1)$, where $q_r$ is the quantile of order $1 - \frac{r}{2}$ of the Standard Gaussian distribution.

### 3.2. Expert test: $\Delta_e$

The testing procedure we describe now relies on classification and is called the *Expert test*. It is a two-step procedure. The first step consists in assigning label $l$ to any data $X_i$ such that $\omega_\infty(i) := \max(\omega_u(i); u \in \{1, \ldots, M\}) = \omega_l(i)$ and to any data $Y_i$ such that $\omega'_\infty(i) := \max(\omega'_u(i); u \in \{1, \ldots, M\}) = \omega'_l(i)$. The second step consists in using Welch's $t$-test on the two subsamples of data that have been assigned label $l$ in order to check whether $m_l$ and $m'_l$ are different or not. Note that it means that Welch's $t$-test is applied to data having potentially wrong labels.

Put

- $n_{l,e} = \sum_{i=1}^{n} \mathbf{1}\{\omega_\infty(i) = \omega_l(i)\}$ and $n'_{l,e} = \sum_{i=1}^{n} \mathbf{1}\{\omega'_\infty(i) = \omega'_l(i)\}$,
- $\bar{X}_e^{(l)} = \frac{1}{n_{l,e}} \sum_{i=1}^{n} X_i \mathbf{1}\left\{\omega_\infty(i) = \omega_l(i)\right\}$,
- $\bar{Y}_e^{(l)} = \frac{1}{n'_{l,e}} \sum_{i=1}^{n} Y_i \mathbf{1}\left\{\omega'_\infty(i) = \omega'_l(i)\right\}$,
- $\hat{\sigma}_{l,e}^2 = \frac{1}{n_{l,e}} \sum_{i=1}^{n} (X_i - \bar{X}_e^{(l)})^2 \mathbf{1}\left\{\omega_\infty(i) = \omega_l(i)\right\}$,
- $\hat{\sigma}_{l,e}'^2 = \frac{1}{n'_{l,e}} \sum_{i=1}^{n} (Y_i - \bar{Y}_e^{(l)})^2 \mathbf{1}\left\{\omega'_\infty(i) = \omega'_l(i)\right\}$.

The Expert test $\Delta_e$ is based on the test statistic $T_e$ defined as follows

$$T_e := \frac{|\bar{X}_e^{(l)} - \bar{Y}_e^{(l)}|}{\sqrt{\frac{\hat{\sigma}_{l,e}^2}{n_{l,e}} + \frac{\hat{\sigma}_{l,e}'^2}{n'_{l,e}}}}.$$

Then, the decision rule is defined as $\Delta_e = \mathbf{1}\{T_e > q_r\}$.

### 3.3. Mixing test: $\Delta_m$

The last testing procedure we propose is based on ideas from [1]. It relies on combinatory methods and suggests to invert, in some sense, the mixing weight operators. Let us describe this new testing procedure in detail.

For any $(i, u) \in \{1, \ldots, n\} \times \{1, \ldots, M\}$, we define

$$a_u(i) = \frac{1}{\det\left(\frac{1}{n}{}^t\Omega_X\Omega_X\right)} \sum_{k=1}^{M} (-1)^{u+k} \gamma_{uk}^{(X)} \omega_k(i), \tag{3}$$

$$a'_u(i) = \frac{1}{\det\left(\frac{1}{n}{}^t\Omega_Y\Omega_Y\right)} \sum_{k=1}^{M} (-1)^{u+k} \gamma_{uk}^{(Y)} \omega'_k(i), \tag{4}$$

where $\gamma_{uk}^{(X)}$ and $\gamma_{uk}^{(Y)}$ are respectively the minor $(u, k)$ of the matrix $\frac{1}{n}{}^t\Omega_X\Omega_X$ and of the matrix $\frac{1}{n}{}^t\Omega_Y\Omega_Y$.

The quantities $a_u(i)$ and $a'_u(i)$ are called the inverse mixing weights of the mixture model (1). Following Maiboroda [8] or Pokhyl'ko [12], we have

$$\frac{1}{n}\sum_{i=1}^{n} a_u(i)\omega_u(i) = 1 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} a_u(i)\omega_v(i) = 0, \quad \forall u,\ \forall v \neq u.$$

Any couple of parameters $\left(m_u, m'_u\right)$ can be estimated by the method of moments. The involved estimators $\left(\hat{m}_u, \hat{m}'_u\right)$ are defined as follows:

$$\left(\hat{m}_u, \hat{m}'_u\right) := \left(\frac{1}{n}\sum_{i=1}^{n} a_u(i)X_i, \frac{1}{n}\sum_{i=1}^{n} a'_u(i)Y_i\right).$$

The Mixing test $\Delta_m$ is based on the test statistic $T_m$ defined as

$$T_m := \frac{|\hat{m}_l - \hat{m}'_l|}{\sqrt{\hat{\mathbb{V}}_n^{(l)}}}, \tag{5}$$

where $\hat{\mathbb{V}}_n^{(l)}$ is the estimated variance of $\hat{m}_l - \hat{m}'_l$, that is

$$\hat{\mathbb{V}}_n^{(l)} = \frac{1}{n^2}\sum_{i=1}^{n}\left[a_l^2(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)\hat{m}_u\right)^2 + a_l'^2(i)\left(Y_i - \sum_{u=1}^{M}\omega'_u(i)\hat{m}'_u\right)^2\right].$$

**Remark 1.** The random variable $\hat{m}_l$ (resp. $\hat{m}'_l$) is a good estimator of $m_l$ (resp. $m'_l$) because it is unbiased and consistent (see Lemma 4). Hence if the distance between $\hat{m}_l$ and $\hat{m}'_l$ is large, the rejection of the null hypothesis $\mathcal{H}_0$ will be more likely. This idea motivates the choice of the test statistic $T_m$ we defined above.

Under the null hypothesis $\mathcal{H}_0$, the asymptotic distribution of $T_m$ is known, according to the following theorem.

**Theorem 1.** *Assume that*

- *the components within the mixture model (1) have moments at least of order* 4,
- *the mixing weights of the mixture model (1) are such that the sequence of the smallest eigenvalue $k_n$ of the matrices $\frac{1}{n}{}^t\Omega_X\Omega_X$ and $\frac{1}{n}{}^t\Omega_Y\Omega_Y$ $(n \in \mathbb{N}^*)$ satisfies*

$$\lim_{n\longrightarrow+\infty} nk_n = +\infty, \tag{6}$$

- *the inverse mixing weights of the mixture model (1) are such that*

$$\lim_{n\to+\infty} \frac{\max(a_l^2(i);\ 1 \le i \le n)}{\sum_{i=1}^{n} a_l^2(i)} = \lim_{n\to+\infty} \frac{\max(a_l'^2(i);\ 1 \le i \le n)}{\sum_{i=1}^{n} a_l'^2(i)} = 0. \tag{7}$$

*Then, under the null hypothesis $\mathcal{H}_0$, the asymptotic distribution of $T_m$ is the Standard Gaussian one, i.e.*

$$T_m \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1). \tag{8}$$

Hence, $\Delta_m = \mathbf{1}\{T_m > q_r\}$ is a test with asymptotic type $I$ error equal to $r$ $(0 < r < 1)$.

**Remark 2.** A wide range of mixing weights of the mixture model (1) satisfy conditions (6) and (7). Examples of such mixing weights for $M = 2$ are given in (15) of Section 4.

**Proof.** Because of the independence of the two samples and the fact that under the null hypothesis $\mathcal{H}_0$, $m_l = m'_l$, it is sufficient, in order to prove Theorem 1, that

1. $\dfrac{\frac{1}{n}\sum_{i=1}^{n} a_l(i)X_i - m_l}{\sqrt{\frac{1}{n^2}\sum_{i=1}^{n} a_l^2(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)\hat{m}_u\right)^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$,
2. $\dfrac{\frac{1}{n}\sum_{i=1}^{n} a'_l(i)Y_i - m'_l}{\sqrt{\frac{1}{n^2}\sum_{i=1}^{n} a'^2_l(i)\left(Y_i - \sum_{u=1}^{M}\omega'_u(i)\hat{m}'_u\right)^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

These two results of convergence can be proved in the same way. Therefore we focus on the first one which can be rewritten as follows:

$$\frac{\sum_{i=1}^{n} a_l(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)m_u\right)}{\sqrt{\sum_{i=1}^{n} a_l^2(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)\hat{m}_u\right)^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Denote, for any $1 \le i \le n$,

$$B_n^{(l)} = \sum_{i=1}^{n} a_l^2(i)\mathbb{E}\left[\left(X_i - \sum_{u=1}^{M}\omega_u(i)m_u\right)^2\right], \tag{9}$$

$$\hat{B}_n^{(l)} = \sum_{i=1}^{n} a_l^2(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)\hat{m}_u\right)^2. \tag{10}$$

From Proposition 1, we have

$$\frac{\sum_{i=1}^{n} a_l(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)m_u\right)}{\sqrt{B_n^{(l)}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{11}$$

In the following, we prove that the same kind of result holds when the parameter $B_n^{(l)}$ is replaced by the estimator $\hat{B}_n^{(l)}$. In other words, the result we want to prove is the following:

$$\frac{\sum_{i=1}^{n} a_l(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)m_u\right)}{\sqrt{\hat{B}_n^{(l)}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{12}$$

From Slutsky's theorem, it suffices to prove that the estimator $\hat{B}_n^{(l)}$ of $B_n^{(l)}$ is consistent. We propose to divide the proof of this consistency into two steps. First we prove

$$\frac{\sum_{i=1}^{n} a_l^2(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)m_u\right)^2}{B_n^{(l)}} \xrightarrow{\text{Proba}} 1. \tag{13}$$

The second step consists in replacing $m_1, \ldots, m_M$ by consistent estimators $\hat{m}_1, \ldots, \hat{m}_M$ and in checking that the convergence in probability still holds.

From this point on, we need one more assumption, that is to say the existence of the fourth order moment for any $p_u$, $1 \le u \le M$.

We start with the first step. We apply Bienayme–Chebyshev's inequality and we have for any $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{i=1}^{n} a_l^2(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)m_u\right)^2 - B_n^{(l)}\right| > B_n^{(l)}\varepsilon\right)\\
&\quad\le (B_n^{(l)}\varepsilon)^{-2}\sum_{i=1}^{n} a_l^4(i)\mathbb{V}\mathrm{ar}\left[\left(X_i - \sum_{u=1}^{M}\omega_u(i)m_u\right)^2\right]\\
&\quad\le (B_n^{(l)}\varepsilon)^{-2}\max(a_l^2(i);\,1\le i\le n)\sum_{i=1}^{n} a_l^2(i)\mathbb{E}\left[(X_i - \mathbb{E}(X_i))^4\right]\\
&\quad\le \frac{\max(a_l^2(i);\,1\le i\le n)}{B_n^{(l)}}\left[\left(B_n^{(l)}\right)^{-1}\sum_{i=1}^{n} a_l^2(i)\right] C(p_1,\ldots,p_M)\,\varepsilon^{-2}\\
&\quad\le \left(\frac{\max(a_l^2(i);\,1\le i\le n)}{B_n^{(l)}}\right)\left(\left[\min(\sigma_u^2;\,1\le u\le M)\right]^{-1} C(p_1,\ldots,p_M)\,\varepsilon^{-2}\right)\\
&\quad\le \left(\frac{\max(a_l^2(i);\,1\le i\le n)}{\sum_{i=1}^{n} a_l^2(i)}\right)\left(\left[\min(\sigma_u^2;\,1\le u\le M)\right]^{-2} C(p_1,\ldots,p_M)\,\varepsilon^{-2}\right).
\end{aligned}$$

The last inequalities are entailed by Lemmas 3 and 2. The constant $C(p_1,\ldots,p_M)$ is given in (18). The right-hand side of the last inequality is a product of two terms. The left one tends to 0 when $n$ goes to infinity because of Assumption (7). The right one is a constant that depends only on $\varepsilon$ and the parameters of $p_u$, $1 \le u \le M$. When considering the limit to infinity with respect to $n$, we conclude that property (13) holds.

We end by proving the second step. We have

$$\begin{aligned}
&\sum_{i=1}^{n} a_l^2(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)\hat{m}_u\right)^2\\
&\quad= \sum_{i=1}^{n} a_l^2(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)m_u\right)^2 + 2\sum_{i=1}^{n} a_l^2(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)m_u\right)\\
&\qquad\times\left(\sum_{u=1}^{M}\omega_u(i)(m_u - \hat{m}_u)\right) + \sum_{i=1}^{n} a_l^2(i)\left(\sum_{u=1}^{M}\omega_u(i)(m_u - \hat{m}_u)\right)^2.
\end{aligned}$$

The first term is exactly the one appearing in the first step when divided by $B_n^{(l)}$ and also converges to 1 in probability. Now we can turn to the second term. Cauchy–Schwarz's inequality and Jensen's inequality entail that

$$\begin{aligned}
&\left|\sum_{i=1}^{n} a_l^2(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)m_u\right)\left(\sum_{u=1}^{M}\omega_u(i)(m_u - \hat{m}_u)\right)\right|\\
&\quad\le \sqrt{\sum_{i=1}^{n} a_l^2(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)m_u\right)^2}\times\sqrt{\sum_{u=1}^{M}\sum_{i=1}^{n}(m_u - \hat{m}_u)^2 a_l^2(i)\omega_u(i)}\\
&\quad\le \sqrt{\sum_{i=1}^{n} a_l^2(i)\left(X_i - \sum_{u=1}^{M}\omega_u(i)m_u\right)^2}\times\sqrt{\sum_{u=1}^{M}(m_u - \hat{m}_u)^2\sum_{i=1}^{n} a_l^2(i)}.
\end{aligned}$$

When it is divided by the square root of $B_n^{(l)}$, the first term of the right-hand side of the last inequality converges to 1 in probability due to (13). Lemma 2 entails that

$$\sum_{i=1}^{n} a_l^2(i) \leq B_n^{(l)} \left[\min(\sigma_u^2;\, 1 \leq u \leq M)\right]^{-1}.$$

Hence the second term of the right-hand side of the inequality converges to 0 in probability when divided by the square root of $B_n^{(l)}$ because of the consistency of the estimators $\hat{m}_l$ given in Lemma 4. We can proceed in the same way in order to prove that the third term converges to 0 in probability when divided by $B_n^{(l)}$.

So, we have just proved that

$$\frac{\hat{B}_n^{(l)}}{B_n^{(l)}} \overset{\text{Proba}}{\to} 1. \tag{14}$$

We conclude that the exact variance $B_n^{(l)}$ can be replaced by the consistent estimator $\hat{B}_n^{(l)}$ in order to obtain the result of convergence (11). In other words, the property (12) holds. □

## 4. Numerical experiments

In this section we perform numerical experiments and we discuss the performance of the Mixing test $\Delta_m$. For the sake of simplicity, we focus on the case $M = 2$ and we are interested in the testing problem (2) with $l = 1$. Nevertheless the same kind of experiments could be conducted with any $M > 2$ and any $l \in \{1, \ldots, M\}$.

We expect the performance of the test $\Delta_m$ to be superior to that of the test $\Delta_e$, that is to say a smaller type *II* error when the type *I* error is chosen to be $r = 0.05$. Without loss of generality, we suppose that $n$ is even.

We consider the Gaussian setting and we assume in this section that the mixing weight operators $\Omega_X$ and $\Omega_Y$ have the following form:

$$\Omega_X = \begin{pmatrix} \alpha & 1-\alpha \\ \cdots & \cdots \\ \alpha & 1-\alpha \\ 1-\beta & \beta \\ \cdots & \cdots \\ 1-\beta & \beta \end{pmatrix} \quad \text{and} \quad \Omega_Y = \begin{pmatrix} \alpha' & 1-\alpha' \\ \cdots & \cdots \\ \alpha' & 1-\alpha' \\ 1-\beta' & \beta' \\ \cdots & \cdots \\ 1-\beta' & \beta' \end{pmatrix}, \tag{15}$$

where $0.5n$ data from $\mathbf{X}$ (resp. $\mathbf{Y}$) correspond to the pair of mixing weights $(\alpha, 1-\alpha)$ (resp. $(\alpha', 1-\alpha')$) and the other $0.5n$ data from $\mathbf{X}$ (resp. $\mathbf{Y}$) correspond to the pair of mixing weights $(1-\beta, \beta)$ (resp. $(1-\beta', \beta')$). Suppose now that $\Omega_X$ and $\Omega_Y$ are full rank matrices, i.e. $\alpha + \beta \neq 1$ and $\alpha' + \beta' \neq 1$.

**Remark 3.** The reader can easily check that the properties (6) and (7) are fulfilled when the mixing weight operators $\Omega_X$ and $\Omega_Y$ satisfy (15).

### 4.1. Numerical performance of the mixing test

From Paragraphs 4.1.1 to 4.1.3 we suppose that $\alpha = \beta$ and that $\alpha' = \beta'$ for the sake of simplicity.

#### 4.1.1. Mixing test versus expert test

In this paragraph we motivate the use of the Mixing test $\Delta_m$. For any value of $(\alpha, \alpha') \in \left]\frac{1}{2}, 1\right[^2$, the performance of the Expert test can be bad even if the numbers of observations $n$ is large.

- If the two components have equal expected values, $\Delta_e$ can detect a difference between these components (wrong decision) whereas our test does not. For instance, suppose that $m_1 = m_1'$ and that $m_2'$ is far away from $(1-\alpha')^{-1}\left((\alpha-\alpha')m_1 + (1-\alpha)m_2\right)$. Since $\mathbb{E}\left(\bar{X}_e^{(1)}\right) \neq \mathbb{E}\left(\bar{Y}_e^{(1)}\right)$,

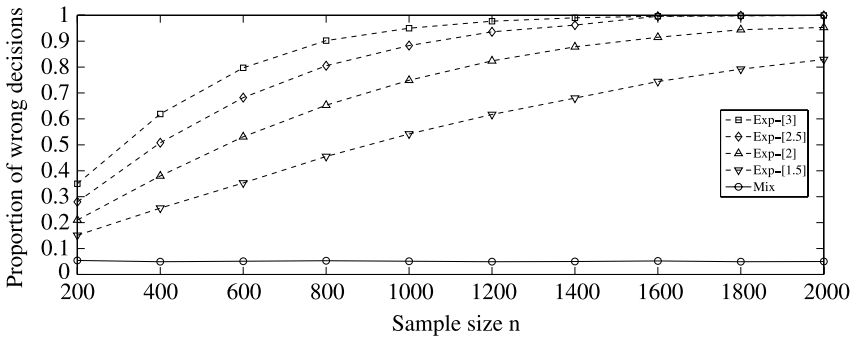


**Fig. 1.** Expert test (Exp-[$\delta$]) vs. mixing test (Mix) according to $\delta$.

**Table 2**
Proportion of correct decisions given by $\Delta_m$.

| $n$ | 500 | 1000 | 2000 | 3000 | 4000 | 5000 | 6000 |
|---|---|---|---|---|---|---|---|
| $\Delta_m$ | 0.146 | 0.242 | 0.438 | 0.595 | 0.718 | 0.810 | 0.879 |

the Expert test $\Delta_e$ is a bad choice since the event $\{T_e > 1.96\}$ also holds with high probability for $n$ large enough. Hence, the wrong decision $\mathcal{H}_1$ may often be taken.
An example of such a situation is given here in the case where $\alpha = \alpha' = 0.9$. Consider the testing problem (2) and suppose $\sigma_1 = \sigma_1' = \sigma_2 = \sigma_2' = 1$, $m_1 = m_1' = 0$, $m_2 = 1$ and $m_2' = m_2 + \delta$. For $r = 0.05$ and several values of $n$ and $\delta$, we give the proportion of wrong decisions $\mathcal{H}_1$ in Fig. 1. The results are based on 40 000 samples.
Note that the proportion of wrong decisions given by $\Delta_e$ increases as $n$ grows and can be quite large if $m_2'$ is sufficiently far away from $m_2$. Most of the time, the expert detects a difference between the components $m_1$ and $m_1'$ although there is none in that context. Comparatively, the proportion of wrong decisions given by $\Delta_m$ is around 0.05. On Fig. 1, the type $I$ error for the Mixing test is plotted only for $\delta = 3$ as they are all around the value 0.05 whatever $\delta$ is.

- The Expert test $\Delta_e$ can also fail to detect a difference between two components with different expected values whereas our test does not. For instance, suppose that $m_1 \neq m_1'$ and

$$m_2' \approx (1-\alpha')^{-1}\left(\alpha m_1 + (1-\alpha)m_2 - \alpha' m_1'\right).$$

Since $\mathbb{E}(X_i) \approx \mathbb{E}(Y_i)$, for any $1 \leq i \leq 0.5n$, using $\Delta_e$ to detect the difference between $m_1$ and $m_1'$ is a very bad choice. Indeed, according to the law of large numbers, with high probability – that increases as $n$ goes up – $\bar{X}_e^{(1)}$ and $\bar{Y}_e^{(1)}$ are very close to each other. It implies that the event $\{T_e \leq 1.96\}$ holds with high probability. The correct decision $\mathcal{H}_1$ is taken only in 5% of the cases.
An example of such a situation is given here in the case $\alpha = \alpha' = 0.9$. Consider the testing problem (2) and suppose that $\sigma_1 = \sigma_1' = \sigma_2 = \sigma_2' = 1$, $m_1 = 0$, $m_1' = 0.1$, $m_2 = 1$ and $m_2' = 2$. For $r = 0.05$ and several values of $n$ and 40 000 repetitions of $\Delta_m$, we give the proportion of correct decisions given by $\Delta_m$ in Table 2.
As expected, the proportion of correct decisions given by $\Delta_m$ increases as $n$ grows. But this is not the case for the proportion of correct decisions given by $\Delta_e$ which is always around 0.05. Most of the time, the expert is unable to detect the difference between the two components in that context.

Finally we conclude that it is much better to choose the Mixing test $\Delta_m$ for the problem we are interested in. Indeed we have seen above that the Expert test suffers from severe drawbacks in some cases that we are unable to recognize in advance.

#### *4.1.2. Mixing test versus Oracle test*

In this paragraph we compare the *empirical power* of $\Delta_m$ to that of the Oracle test $\Delta_o$. We take $r = 0.05$ and the same parameters as in the last example. We recall that the empirical power of any test $\Delta$ corresponds to the numerical evaluation of the probability to correctly decide $\mathcal{H}_1$, according to $\Delta$.

**Table 3**
Empirical powers of $\Delta_o$ and $\Delta_m$.

| Test | $n$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | 500 | 1000 | 2000 | 3000 | 4000 | 5000 | 6000 |
| $\Delta_o$ | 0.200 | 0.349 | 0.609 | 0.783 | 0.886 | 0.942 | 0.973 |
| $\Delta_m$ | 0.149 | 0.245 | 0.427 | 0.585 | 0.704 | 0.798 | 0.868 |

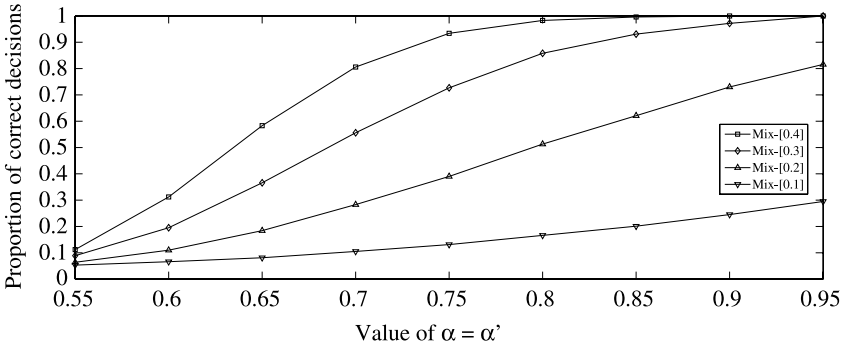


**Fig. 2.** Empirical power Mix-[$\bar{\delta}$] of the mixing test according to $\bar{\delta}$.

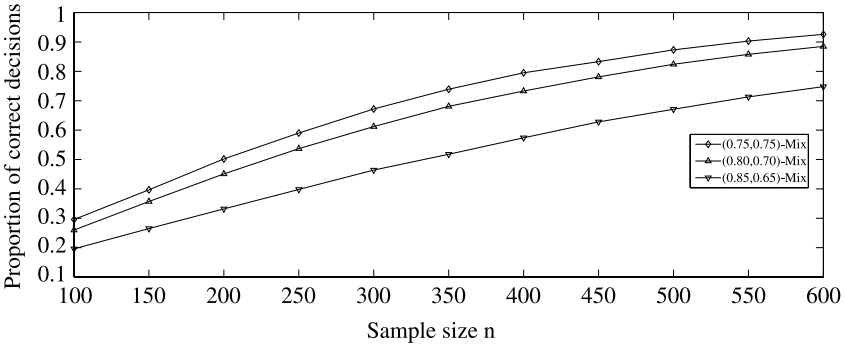


**Fig. 3.** Empirical power $(\alpha, \alpha')$-Mix of the mixing test according to $(\alpha, \alpha')$.

According to Table 3 we remark that the larger $n$ is, the better the powers of $\Delta_o$ and $\Delta_m$. Moreover we note that the empirical power of the Mixing test is smaller than that of the Oracle test but not too far.

In Fig. 2 we give the empirical power of $\Delta_m$ measured for samples of size $n = 1000$ in the case where $m_1 = 0, m_1' = \bar{\delta},\ m_2 = 1,\ m_2' = 2$ and $\sigma_1 = \sigma_1' = \sigma_2 = \sigma_2' = 1$.

Looking at Fig. 2, we can make two statements:

- the proportion of correct decisions depends on the intrinsic difficulty of the problem. Indeed the larger the absolute value of the difference $\bar{\delta} := m_1' - m_1$, the easier the problem of detection and so the more powerful the test,
- it appears than the larger the degree of certainty $\alpha$ the better the power of $\Delta_m$. This point is discussed in a more general context in the next paragraph.

*4.1.3. Comparisons on the performance of $\Delta_m$ for varying values of $(\alpha, \alpha')$*

As previously discussed, we expect that the higher the degree of certainty of the expert, the better the performance of the test $\Delta_m$. This statement is highlighted here when we consider the same parameters of components as before, $\bar{\delta} = 0.5$ and several choices of the pair $(\alpha, \alpha')$. For each choice of $(\alpha, \alpha')$, the empirical power of our test $\Delta_m$, i.e. the proportion of correct detections of a difference between $m_1$ and $m_1'$, is given in Fig. 3.

**Table 4**
Description of the population.

| | NY | CA |
|---|---|---|
| Total | 9189 | 8935 |
| Over 21 y.o. | 90.39% (8306) | 87.04% (7803) |
| Under 20 y.o. | 9.61% (883) | 12.96% (1162) |
| Walked | 49.73% (4570) | 44.5% (3989) |
| Bus/trolleybus | 50.27% (4619) | 55.5% (4979) |

The interpretation of the results presented in Fig. 3 points in the same direction as in [1]: the larger the smallest eigenvalue of both matrices $\frac{1}{n}{}^t\Omega_X\Omega_X$ and $\frac{1}{n}{}^t\Omega_Y\Omega_Y$, namely $k_n$, the better the power of the Mixing test $\Delta_m$. Note that the larger the minimum between $\alpha \in ]\frac{1}{2}, 1[$ and $\alpha' \in ]\frac{1}{2}, 1[$, the larger $k_n$. Indeed, we have

$$k_n = 2\left(\min(\alpha, \alpha') - \frac{1}{2}\right)^2.$$

*4.1.4. Brief conclusion*

Let us summarize the main facts observed in the numerical experiments. First, in some cases the Expert test can be completely wrong because of the overall design, that is to say the link between the means of the components and the mixing weights. This is a serious issue for the Expert test. The results can become even worse as the sample size increases. The test adapted to the varying mixing weights that we propose does not suffer from this drawback. The second fact is the good behavior of our test compared to the Oracle test. Although the power is smaller, it is quite satisfactory. The last important fact which has already been stressed by Autin and Pouet [1] is the effect of the mixing weights. It is known *a priori* thanks to the smallest eigenvalue of the operators $\frac{1}{n}{}^t\Omega_X\Omega_X$ and $\frac{1}{n}{}^t\Omega_Y\Omega_Y$. This point is very important as the statistician can act in order to counter this effect, e.g. he can improve the accuracy of the expert system which gives the mixing weights or increase the sample sizes.

*4.2. Application to real data*

In this section we apply our methodology to real data and we discuss the results. The real case exemplifies the whole methodology:

- identify the variable of interest and the auxiliary variable,
- model the problem and therefore compute the mixing weights according to the data,
- compute the inverse mixing weights,
- compute the test statistics and the associated $p$-value.

*4.2.1. Description of the data*

We have selected data from the US Census Bureau website, more precisely PUMS 2006 (see US Census Bureau [14]). We are interested in comparing the travel time of people living either in the State of New York (abbreviated as NY) or in the State of California (abbreviated as CA). Two means of transportation have been retained: *Bus/trolleybus* and *Walked*. We have also retained a variable related to age as it will be useful for the mixture model with varying mixing weights. This variable records the fact that a person is over 21 years old or under 20 years old.

Here are a few facts to roughly describe the PUMS sample. Table 4 gives one level information.

In Table 5 we compute the mean and the standard deviation (in parentheses) of the travel time according to the means of transportation to work.

As it can be seen in Table 5 there might be no difference between New York and California. Nevertheless if the means of transportation is unavailable, it will be perilous to decide considering the whole sample without any other information. Indeed as shown in Table 4, the difference between

**Table 5**
One-way analysis of the travel time (in minutes).

| | Walked | Bus/trolleybus | Walked and bus/trolleybus |
|---|---|---|---|
| NY | 12.25 (12.18) | 47.26 (28.79) | 29.85 (28.23) |
| CA | 11.23 (12.23) | 45.12 (28.84) | 30.04 (28.49) |

**Table 6**
Decisions for *Bus/trolleybus*.

| | Decision $n = 1000$ | $p$-value |
|---|---|---|
| Oracle test | Not rejected | 0.24 |
| Expert test | Not rejected | 0.42 |
| Mixing test | Not rejected | 0.11 |

**Table 7**
Decisions for *Walked*.

| | Decision $n = 1000$ | $p$-value |
|---|---|---|
| Oracle test | Not rejected | 0.23 |
| Expert test | Rejected | 0.04 |
| Mixing test | Not rejected | 0.48 |

New York and California is decreased because of the structure of the populations (fewer people under 20 years old in New York).

#### 4.2.2. Methodology

We assume in the following that the information about the means of transportation (the labels) is unavailable at the microdata level. The age variable is assumed to be the only auxiliary information available at the microdata level. It allows us to compute the mixing weights in the mixture model (1).

For comparison purposes we have applied the three testing procedures.

According to the notation introduced in (15) and to Table 1, we have

$$(\alpha, \beta) = (0.5193, 0.6535) \qquad (\alpha', \beta') = (0.574, 0.5723). \tag{16}$$

The samples are drawn according to the following sampling scheme: 500 persons over 21 years old and 500 persons under 20 years old are randomly sampled in each state ($n = 1000$).

We have applied the three testing procedures with the threshold value $q_r = 1.96$ ($r = 0.05$): Oracle test, Expert test and Mixing test.

First we test the equality of the expected values when the means of transportation to work is *Bus/trolleybus* (label 1) in Table 6. In this case, the other means of transportation, *Walked*, is considered as a nuisance parameter.

Next we reverse the set-up. We test the equality of the expected values when the means of transportation to work is *Walked* (label 2) in Table 7. Now the other means of transportation *Bus/trolleybus* is now a nuisance parameter.

#### 4.2.3. A tough situation

Here we are also interested in comparing the travel time of people living either in the State of Pennsylvania (abbreviated as PA) or in the State of Illinois (abbreviated as IL). Data also come from the US Census Bureau [14]. Two means of transportation to work have been retained: *Bus/trolleybus* and *Railroad*. We have also retained the gender variable as it will be useful to compute the varying mixing weights of the mixture model. As will be seen, this situation is much more involved compared to the one in the previous section.

**Table 8**
Description of the population.

| | PA | IL |
|---|---|---|
| Total | 1769 | 2899 |
| Men | 43.2% (764) | 48.2% (1398) |
| Women | 56.8% (1005) | 51.8% (1501) |
| Bus/trolleybus | 80.4% (1423) | 58.4% (1692) |
| Railroad | 19.6% (346) | 41.6% (1207) |

**Table 9**
Mixing weights.

| | Bus/trolleybus | Railroad |
|---|---|---|
| PA men | 74.6% (570) | 25.4% (194) |
| PA women | 84.9% (853) | 15.1% (152) |
| IL men | 50.8% (710) | 49.2% (688) |
| IL women | 65.4% (982) | 34.6% (519) |

**Table 10**
One-way analysis of the travel time (in minutes).

| | Bus/trolleybus | Railroad | Bus/trolleybus and railroad |
|---|---|---|---|
| Pennsylvania | 42.6 (31.05) | 59.8 (35.8) | 46 (32.7) |
| Illinois | 41.8 (26.4) | 63.1 (25.7) | 50.7 (28.2) |

**Table 11**
Decisions for *Bus/trolleybus*.

| | Decision $n = 1000$ | $p$-value |
|---|---|---|
| Oracle test | Not rejected | 0.81 |
| Expert test | Not rejected | 0.0001 |
| Mixing test | Not rejected | 0.14 |

Here are a few facts to roughly describe the PUMS sample. Table 8 gives one level information.
The mixing weights depend on the gender as illustrated in Table 9.
According to the notation introduced in (15) and Table 9, we have

$$(\alpha, \beta) = (0.746, 0.151) \qquad (\alpha', \beta') = (0.508, 0.346). \tag{17}$$

In Table 10 we compute the mean and the standard deviation (in parentheses) of the travel time according to the categorical variable means of transportation to work.

The difference between the travel times is increased if we consider the entire population. This is due to its structure. As there are more men and women who use the railroad in Illinois, the average of the travel time is increased. This is reversed in Pennsylvania.

In Table 11 we test the equality of the expected values when the means of transportation to work is *Bus/trolleybus*. We use the threshold value $q_r = 1.96$ ($r = 0.05$).

In Table 12 we reverse the set-up and we test the equality of the averages when the means of transportation to work is *Railroad*. We use the threshold value $q_r = 1.96$ ($r = 0.05$).

## 5. Conclusion and open questions

From our point of view, one of the most interesting features of the varying mixing weight model is its practical applicability. It is a versatile model which can be used in many situations with missing microdata but aggregated information. We point out that the confidentiality of statistics is such a situation. The application to real data provided above exemplifies the modeling.

**Table 12**
Decisions for *Railroad*.

| | Decision $n = 1000$ | $p$-value |
|---|---|---|
| Oracle test | Not rejected | 0.12 |
| Expert test | Unavailable | Unavailable |
| Mixing test | Not rejected | 0.08 |

The second take-away message is the excellent performance of the Mixing test we propose. It can be guessed *a priori* thanks to the smallest eigenvalue of operators involved in the mixture model. This good performance is proved both theoretically and numerically.

To conclude let us state that this work can be easily extended to mixture models in a nonparametric setting, when using the testing procedure proposed by Butucea and Tribouley [2], the Oracle test, and the one given by Autin and Pouet [1], the Mixing test.

An interesting extension which should be considered in the future is the case of mixing weights with errors. This arises when the mixing weights are computed from a model with estimated parameters or from the experts' evaluation. In this case the matrices $\Omega_X$ and $\Omega_Y$ are random. Preliminary but unpublished simulation results tend to indicate that moderate errors have a small effect. An open question is to exactly evaluate the effect of this uncertainty about the mixing weights. This problem certainly involves the study of random matrices and their properties. For example, it would be interesting to evaluate the numerical performance of the Mixing test if the mixing weights were computed with a logistic regression model.

Another interesting topic for future research is the comparison between the mixture model with varying mixing weights and imputation techniques which are popular in missing data (e.g. [13]). Indeed imputation techniques have been developed to handle missing data and this is also exactly what the mixture model with varying mixing weights does. Numerical comparisons between the Mixing test and multiple imputation are surely the next step to check the good performance of the Mixing test. It will also be very interesting to investigate the links between the mixture model and the problems encountered with missing data (Missing At Random, Missing Completely At Random, Missing Not At Random).

## Acknowledgments

The authors would like to thank the Associate Editor and the anonymous referee for their very helpful suggestions.

## Appendix

In this section we provide the technical lemmas and Proposition 1 required for the proof of Theorem 1. For the sake of simplicity, we present the lemmas with respect to $X_1, \ldots, X_n$. An analogous version of them does exist for $Y_1, \ldots, Y_n$. We recall that we assume that the inverse mixing weights of the model satisfy (7).

**Lemma 1.** *For any* $1 \leq i \leq n$,

$$\mathbb{V}\mathrm{ar}(X_i) = \sum_{u=1}^{M} \omega_u(i)\sigma_u^2 + \frac{1}{2}\sum_{u=1}^{M}\sum_{v=1}^{M} \omega_u(i)\omega_v(i)(m_u - m_v)^2.$$

**Proof.** For any $1 \leq i \leq n$,

$$\mathbb{V}\mathrm{ar}(X_i) = \sum_{u=1}^{M} \omega_u(i) \int_{\mathbb{R}} y^2 p_u(y)dy - \left(\sum_{u=1}^{M} \omega_u(i)m_u\right)^2$$

$$= \sum_{u=1}^{M} \omega_u(i)\sigma_u^2 + \sum_{u=1}^{M} \omega_u(i)m_u^2 - \left(\sum_{u=1}^{M} \omega_u(i)m_u\right)^2$$
$$= \sum_{u=1}^{M} \omega_u(i)\sigma_u^2 + \frac{1}{2}\sum_{u=1}^{M}\sum_{v=1}^{M} \omega_u(i)\omega_v(i)(m_u - m_v)^2. \quad \square$$

**Lemma 2.** *Let $B_n^{(l)}$ be defined as in (9). Then,*

$$B_n^{(l)} \geq \min\left(\sigma_u^2;\ 1 \leq u \leq M\right) \sum_{i=1}^{n} a_l^2(i).$$

**Proof.** As a direct consequence of Lemma 1, for any $1 \leq i \leq n$, one gets

$$\mathbb{V}\mathrm{ar}(X_i) \geq \sum_{u=1}^{M} \omega_u(i)\sigma_u^2.$$

Hence,

$$B_n^{(l)} = \sum_{i=1}^{n} a_l^2(i)\mathbb{V}\mathrm{ar}(X_i)$$
$$\geq \sum_{i=1}^{n} a_l^2(i) \sum_{u=1}^{M} \omega_u(i)\sigma_u^2$$
$$\geq \min\left(\sigma_u^2;\ 1 \leq u \leq M\right) \sum_{i=1}^{n} a_l^2(i). \quad \square$$

**Lemma 3.** *For any $1 \leq i \leq n$,*

$$\mathbb{E}\left[(X_i - \mathbb{E}(X_i))^4\right] \leq C(p_1, \ldots, p_M),$$

*where*

$$C(p_1, \ldots, p_M) := \max\left(\int_{\mathbb{R}} (x - m_u)^4 p_v(x)dx;\ (u, v) \in \{1, \ldots, M\}^2\right). \tag{18}$$

**Proof.** We have

$$\mathbb{E}\left[(X_i - \mathbb{E}(X_i))^4\right] = \mathbb{E}\left[\left(X_i - \sum_{u=1}^{M} \omega_u(i)m_u\right)^4\right]$$
$$= \mathbb{E}\left[\left(\sum_{u=1}^{M} \omega_u(i)(X_i - m_u)\right)^4\right]$$
$$\leq \sum_{u=1}^{M} \omega_u(i)\mathbb{E}\left((X_i - m_u)^4\right)$$
$$\leq \max\left(\int_{\mathbb{R}} (x - m_u)^4 p_v(x)dx;\ (u, v) \in \{1, \ldots, M\}^2\right). \quad \square$$

**Lemma 4.** *For any $1 \leq u \leq M$, the unbiased estimator $\hat{m}_u = \frac{1}{n}\sum_{i=1}^{n} a_u(i)X_i$ of $m_u$ is consistent, that is*

$$\hat{m}_u \overset{\text{Proba}}{\to} m_u.$$

**Proof.** Let $\varepsilon > 0$ and $u \in \{1, \ldots, M\}$. Using Bienayme–Chebyshev's inequality and Lemma 1, one gets:

$$\begin{aligned}\mathbb{P}\left(\left|\hat{m}_u - m_u\right| > \varepsilon\right) &\leq (n\varepsilon)^{-2} \sum_{i=1}^{n} a_u^2(i)\mathbb{V}\mathrm{ar}(X_i) \\ &= (n\varepsilon)^{-2} \sum_{i=1}^{n} a_u^2(i) \left( \sum_{v=1}^{M} \omega_v(i)\sigma_v^2 + \frac{1}{2} \sum_{v=1}^{M} \sum_{w=1}^{M} \omega_v(i)\omega_w(i)(m_v - m_w)^2 \right) \\ &\leq M(nk_n\varepsilon^2)^{-1} \left[\max(\sigma_v^2;\, 1 \leq v \leq M) + 2\max(m_v^2;\, 1 \leq v \leq M)\right].\end{aligned}$$

The last inequality is obtained thanks to the following bound:

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{u=1}^{M} a_u^2(i) \leq \frac{M}{k_n},$$

given in [1]. According to (6),

$$\lim_{n \longrightarrow +\infty} M(nk_n\varepsilon^2)^{-1} \left[\max(\sigma_v^2;\, 1 \leq v \leq M) + 2\max(m_v^2;\, 1 \leq v \leq M)\right] = 0$$

and we conclude that $\hat{m}_v$ is consistent.

Denote, for any $n \in \mathbb{N}^*$, any $l \in \{1, \ldots, M\}$ and any $1 \leq i \leq n$,

$$W_{ni}^{(l)} = \frac{a_l(i)}{\sqrt{B_n^{(l)}}} \left( X_i - \sum_{u=1}^{M} \omega_u(i) m_u \right). \quad \square \tag{19}$$

**Lemma 5.** *Let $B_n^{(l)}$ and $W_{ni}^{(l)}$ $(1 \leq i \leq n)$ be defined as in (9) and (19). Then, for any $\varepsilon > 0$, we have*

$$\lim_{n \longrightarrow \infty} \sum_{i=1}^{n} \mathbb{E}\left[ \left(W_{ni}^{(l)}\right)^2 \mathbf{1}\left\{ |W_{ni}^{(l)}| \geq \varepsilon \right\} \right] = 0.$$

**Proof.** Let us define

$$\kappa_n = \min\{m_u;\, 1 \leq u \leq M\} + \frac{\varepsilon\sqrt{B_n^{(l)}}}{\max(|a_l(i)|;\, 1 \leq i \leq n)},$$

$$\kappa_n' = \max\{m_u;\, 1 \leq u \leq M\} - \frac{\varepsilon\sqrt{B_n^{(l)}}}{\max(|a_l(i)|;\, 1 \leq i \leq n)}.$$

Then we have

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}\left[ \left(W_{ni}^{(l)}\right)^2 \mathbf{1}\left\{ |W_{ni}^{(l)}| \geq \varepsilon \right\} \right] &= \sum_{i=1}^{n} \int_{|y| \geq \varepsilon} y^2 \, dF_{W_{ni}^{(l)}}(y) \\ &\leq \sum_{i=1}^{n} \frac{a_l^2(i)}{B_n^{(l)}} \int_{x > \kappa_n, x < \kappa_n'} \left( x - \sum_{u=1}^{M} \omega_u(i) m_u \right)^2 dF_{X_i}(x) \\ &= \sum_{i=1}^{n} \frac{a_l^2(i)}{B_n^{(l)}} \int_{x > \kappa_n, x < \kappa_n'} \left( \sum_{u=1}^{M} \omega_u(i)(x - m_u) \right)^2 dF_{X_i}(x) \\ &\leq \sum_{i=1}^{n} \frac{a_l^2(i)}{B_n^{(l)}} \sum_{u=1}^{M} \int_{x > \kappa_n, x < \kappa_n'} (x - m_u)^2 \, dF_{X_i}(x)\end{aligned}$$

$$\leq \left(\sum_{i=1}^{n} \frac{a_l^2(i)}{B_n^{(l)}}\right) \sum_{(u,v)\in\{1,\dots,M\}^2} \int_{x>\kappa_n, x<\kappa_n'} (x-m_u)^2 p_v(x)\,dx.$$

Lemma 2 entails that

$$\sum_{i=1}^{n} \frac{a_l^2(i)}{B_n^{(l)}} \leq \left(\min(\sigma_u^2;\, 1\leq u\leq M)\right)^{-1}.$$

Since the variances under $p_u$, $1\leq u\leq M$, are finite, the integrals above tend to 0 when $n$ goes to infinity according to Lebesgue's Theorem on Dominated Convergence and Assumption (7). □

**Proposition 1.** *For any $1\leq i\leq n$, consider $W_{ni}^{(l)}$, defined as in (19). We have*

$$\sum_{i=1}^{n} W_{ni}^{(l)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

**Proof.** We apply Theorem 4.2 in [11]. It is the general setup for the Central Limit Theorem, i.e. triangular array of series $(W_{ni})_{i,n}$ of independent random variables (note that they are not identically distributed).

If the three conditions below are satisfied for any $\varepsilon>0$ and any $\tau>0$,

1. $\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}\left(|W_{ni}^{(l)}|\geq\varepsilon\right)=0$,
2. $\lim_{n\to\infty} \sum_{i=1}^{n} \int_{|y|<\tau} y dF_{W_{ni}^{(l)}}(y)=0$,
3. $\lim_{n\to\infty} \sum_{i=1}^{n} \left\{\int_{|y|<\tau} y^2 dF_{W_{ni}^{(l)}}(y) - \left(\int_{|y|<\tau} y dF_{W_{ni}^{(l)}}(y)\right)^2\right\}=1$,

then $\sum_{i=1}^{n} W_{ni}^{(l)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

Let us prove that these three conditions are satisfied. Let $\varepsilon>0$. Applying Bienayme–Chebyshev's inequality, we have

$$\sum_{i=1}^{n} \mathbb{P}\left(|W_{ni}^{(l)}|\geq\varepsilon\right) \leq \varepsilon^{-2} \sum_{i=1}^{n} \mathbb{E}\left[\left(W_{ni}^{(l)}\right)^2 \mathbf{1}\left\{|W_{ni}^{(l)}|\geq\varepsilon\right\}\right].$$

Hence, Lemma 5 clearly entails the first condition.

Let us move to the second condition. We use the same trick as above. For any $\tau>0$

$$\sum_{i=1}^{n} \int_{|y|<\tau} y dF_{W_{ni}^{(l)}}(y) = \sum_{i=1}^{n} \left(\int_{\mathbb{R}} y dF_{W_{ni}^{(l)}}(y) - \int_{|y|\geq\tau} y dF_{W_{ni}^{(l)}}(y)\right).$$

The first summand is equal to 0 as the variables $W_{ni}^{(l)}$ are centered:

$$\begin{aligned}\sum_{i=1}^{n} \left|\int_{|y|\geq\tau} y dF_{W_{ni}^{(l)}}(y)\right| &\leq \sum_{i=1}^{n} \int_{|y|\geq\tau} |y| dF_{W_{ni}^{(l)}}(y)\\ &\leq \tau^{-1} \sum_{i=1}^{n} \mathbb{E}\left[\left(W_{ni}^{(l)}\right)^2 \mathbf{1}\left\{|W_{ni}^{(l)}|\geq\tau\right\}\right].\end{aligned}$$

The second condition is also clearly entailed by Lemma 5.

We end the proof with the third condition. There are two parts (because of the two summands) in this condition. For the first part we proceed exactly as in the proof of the second condition. Indeed, we have

$$\sum_{i=1}^{n} \int_{|y|<\tau} y^2 dF_{W_{ni}^{(l)}}(y) = \sum_{i=1}^{n} \int y^2 dF_{W_{ni}^{(l)}}(y) - \sum_{i=1}^{n} \int_{|y|\geq\tau} y^2 dF_{W_{ni}^{(l)}}(y).$$

The first summand is exactly equal to 1 and the second one tends to 0 as $n$ goes to infinity, according to Lemma 5. Therefore it remains to prove that the second part tends to 0 when $n$ goes to infinity. Because the variables $W_{ni}^{(l)}$ are centered and according to Cauchy–Schwarz's inequality, we have

$$\begin{aligned}\sum_{i=1}^{n}\left(\int_{|y|<\tau} y dF_{W_{ni}^{(l)}}(y)\right)^2 &= \sum_{i=1}^{n}\left(\int_{|y|\geq\tau} y dF_{W_{ni}^{(l)}}(y)\right)^2\\ &\leq \sum_{i=1}^{n}\int_{|y|\geq\tau} y^2 dF_{W_{ni}^{(l)}}(y)\\ &= \sum_{i=1}^{n}\mathbb{E}\left[\left(W_{ni}^{(l)}\right)^2 \mathbf{1}\left\{|W_{ni}^{(l)}|\geq\tau\right\}\right].\end{aligned}$$

Still using Lemma 5, we conclude that the second part we are interested in tends to 0 when $n$ goes to infinity, as expected. □

## References


[1] F. Autin, C. Pouet, Test on components of densities mixture, Statist. & Risk Modeling 28 (2011) 389–410.
[2] C. Butucea, K. Tribouley, Nonparametric homogeneity tests, J. Statist. Plann. Inference 136 (2006) 597–639.
[3] J.W. Graham, Missing data analysis: making it work in the real world, Annu. Rev. Psychol. 60 (2009) 549–576.
[4] I.N.S.E.E. Guide du secret statistique, version du 18 October 2010. Available from: http://www.insee.fr/.
[5] D.S. Ironmonger, A system of time accounts for Melbourne, Report commissioned by the Department of Infrastructure, Victoria, 2006. Available from: http://www.transport.vic.gov.au/research/sustainability/estimating-time-use-in-melbourne.
[6] D.S. Ironmonger, A system of time accounts for Melbourne. second edition, Report Commissioned by the Department of Infrastructure, Victoria, 2008. Available from: http://www.transport.vic.gov.au/research/sustainability/estimating-time-use-in-melbourne.
[7] R.E. Maiboroda, Estimation of components distribution by mixtures with varying concentrations, Ukrainian Math. J. 48 (1996) 562–566.
[8] R.E. Maiboroda, An asymptotically effective estimate for a distribution from a sample with a varying mixture, Theory Probab. Math. Statist. 61 (2000) 121–130.
[9] G.J. McLachlan, D. Peel, Finite Mixture Models, Wiley, New York, 2000.
[10] Office for national statistics, Statistics and Registration Service Act 2007, 2007. Available from: http://www.legislation.gov.uk/.
[11] V.V. Petrov, Limit Theorems of Probability Theory, Oxford University Press, New York, 1995.
[12] D. Pokhyl'ko, Wavelet estimators of a density constructed from observations of a mixture, Theory Probab. Math. Statist. 70 (2005) 135–145.
[13] D.B. Rubin, Multiple Imputation for Nonresponse in Surveys, J. Wiley & Sons, New York, 1987.
[14] US Census Bureau, American Community Survey. Profile: California, Illinois and New York, 2006. Available from: http://www.census.gov/.
[15] B.L. Welch, The generalization of student's problem when several different population variances are involved, Biometrika 34 (1947) 28–35.