CONTAMINATION MODELS: ESTIMATION, TEST & CLUSTERING

AFRIC Conference

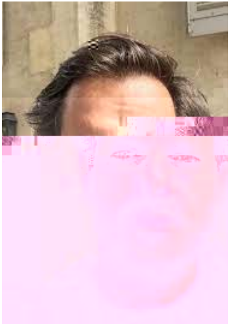
Zimbabwe, 2023/7/25

Xavier MILHAUD

Actuary and Associate Professor

Aix-Marseille University (AMU) - Department of Statistics

Joint work with



Denys Pommeret (D.P.) Yahia Salhi (Y.S.) Pierre Vandekerkhove (P.V.)

1. Motivation and framework
2. Estimation and Test (with k 2 samples)
3. Clustering (with K 2 samples)

THE CONTAMINATION MODEL FRAMEWORK

An admixture (aka contamination) model is a specific 2-component mixture model where one of the two components is known.

Consider an iid random sample $X = (X_1, \ldots, X_n)$ drawn from the admixture model with cdf L.

We have :

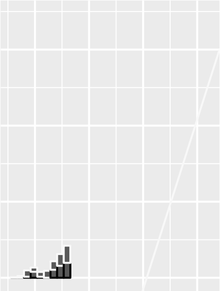
$$L(x) = p\,F(x) + (1 - p)\,G(x), \qquad x \in \mathbb{R} \tag{1}$$

with G a known cdf (gold standard), and $p \in]0, 1[$.

Goal : estimate from $(X_1, \ldots, X_n)$ the unknown component weight p and the unknown cdf F, under minimal assumptions.

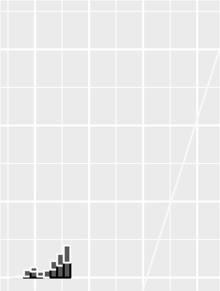
AN EXAMPLE : MORTALITY EXPERIENCE

Women (blue), men (red)



In actuarial science/finance, plenty of situations where the distribution looks like this (claim distributions, customer behaviours, ...).

PRICING WITH 'UNCAPTURED' HETEROGENEITY



Assume P1, P2 and P3 are portfolios with low exposure...and

- heterogeneous age-at-death distributions (mixture profile),
- well-known age-at-death distrib. in general pop. (gold standard).

) Could we pool them to increase exposure for pricing ?

1. Motivation and framework
2. **Estimation and Test (with k 2 samples)**
3. Clustering (with K 2 samples)

APPLICATION : SPECIFIC MORTALITY POOLING

! Back to our 3 portfolios (age-at-death densities for female here). From top left to bottom right : french national pop., P1, P2, and P3.

No obvious similar behaviour... Maybe populations 2 and 3 ?

RESULTS

	Size	Life expectancy	Weight $\hat{p}$	P1	P2	P3
P1	1 251	75.42	0.4603	—	23.28	0.717
P2	7 356	74.91	0.7003	1.4e-06	—	18.48
P3	3 456	75.56	0.6281	0.397	1.7e-05	—

! According to the test, populations 1 and 3 share a common behaviour (F_1 and F_3) characterizing their specific mortality profile...
... whereas other portfolios combinations lead to reject H_0.

) P1 and P3 could be pooled together for pricing !

Limit : pairwise comparisons instead of global test...

EXTENSION OF THE TEST TO THE k-SAMPLE CASE

Consider $k > 2$ samples, each sample $X^{(i)} = (X_1^{(i)}, \ldots, X_{n_i}^{(i)})$ follows

$$L_i(x) = p_i F_i(x) + (1 - p_i) G_i, \qquad x \in \mathbb{R}.$$

The test to perform is given by

$$H_0 : F_1 = \ldots = F_k \quad \text{against} \quad H_1 : F_i \neq F_j \text{ for some } i \neq j.$$

To do so, compare pop. i and j by defining sub-(i,j)-testing problem :

$$H_0(i,j) : \ F_i = F_j \qquad \text{against} \qquad H_1(i,j) : \ F_i \neq F_j, \tag{4}$$

Then,

! Apply IBM for each pair (i,j) & build a series of embedded r78 9.96 9.92 Tf 73.711

CLUSTER POPULATIONS INSTEAD OF INDIVIDUALS

Adapt the previous test procedure to obtain a data-driven method to cluster K unknown populations into N subgroups (characterized by a common unknown mixture component).

- N of clusters is automatically chosen by the procedure,
- Each subgroup is validated by the previous testing method.

Novelty : allows to cluster unobserved subpopulations (via unknown components).

! Not trivial because of unknown p_i's...

! Preprint : X.M., D.P., Y.S., P.V.. Contamination source based K-sample clustering , submitted, 2023. https://hal.science/hal-04129130 .

PLEASE CLUSTER THESE 5 POPULATIONS

Possible choices : [(3,4), (2,5), 1] or [(3,4), 1, 2, 5)] or [(1,2), (4,5), 3] ?

Connect to www.menti.com (code : 2732 4825)

SOLUTION

	Pop.1	Pop.2	Pop.3	Pop.4	Pop.5
Size n_i	2000	2500	2000	4500	4000
Unknown weight p_i	0.6	0.12	0.15	0.08	0.1
Known distribution G_i	E(1				

CONCLUSION

Fully implemented in R package admix!

- Fully tractable solution without shape constraints ;
- Allows for hypothesis testing and clustering;
- Clustering is made on unknown/unobserved phenomenons;
- An application to the covid-19 pandemics in our last paper (clustering countries).
- Actuarial applications whenever pooling can benefit!

Thanks for your attention

!"#$%

!&$'()'*

$!!+"

$!!,"

$!-!

./"""

!"#$%&'#()*+,-'.'/0+ .'$)*/+
1

!"#"$%$&"'()

APPENDIX 1 : 2-sample TESTING STRATEGY

- Inner model convergence regime caracterized by $Z(\;; L_1; L_2)$ and $Z({}^{c}; L$

APPENDIX 2 : k-sample test, steps of the approach

Apply the theoretical results of IBM for each pair of populations $(i; j)$, and then build a series of embedded statistics .

Then, $\forall i, j \in \{1; \ldots; k\}$,

1. Estimate b

Consider the penalization rule (mimicking Schwarz criteria) :

$$S(n) = \min \left\{ \arg \max_{1 \le r \le d(k)} \left(U_r - r \sum_{(i;j) \in S(k)} l_n(i;j) \mathbb{I}_{\{r_k(i;j) = r\}} \right) \right\}.$$

N.B. : l_n if of the form n , where should be tuned depending on our guess (H_0, H_1) to improve the test quality (further details in the paper).

⇒ Our data-driven test statistic is given by

$$\tilde{U}_n = U_{S(n)}.$$

Simulation results :

→ The test shows good empirical levels in many different situations,

→ It also has satisfactory empirical power, provided that $n_i p_i$ is high enough.