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 $k \geq 2$ samples)
3. Clustering (with $K \geq 2$ 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 $X = (X_1, ..., X_n)$ drawn from the admixture model with cdf L.

We have :

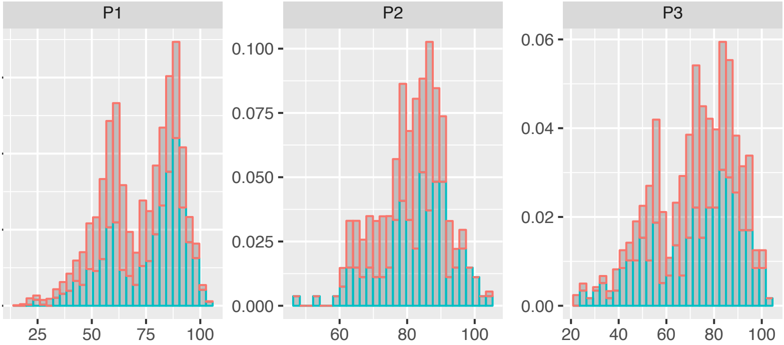
$$L(x) = p\,F(x) + (1 - p)\,G(x), \qquad x \in \mathbb{R} \tag{1}$$

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

Goal : estimate from $(X_1, ..., X_n)$ 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, ...).

HYPOTHESIS TEST ON THE UNKNOWN COMPONENT

To perform the test, we use the **decontaminated** version of the unknown component density F.

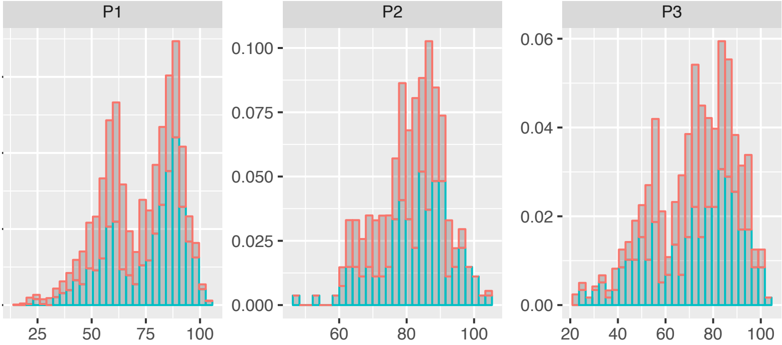
Suppose that p has been consistently estimated, then from the sample providing $\hat{L}$ the decontaminated unknown cdf follows

$$\hat{F}(x) = \frac{\hat{L}(x) - (1-\hat{p})G(x)}{\hat{p}}, \qquad x \in \mathbb{R}. \tag{2}$$

With 1 sample : test $H_0: \; F \in \mathcal{F}$ against $H_1: F \notin \mathcal{F}$.

With 2 samples, one could test $[H_0: \; F_1 = F_2$ against $H_1: F_1 \neq F_2]$

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).

$\Rightarrow$ Could we pool them to increase exposure for pricing ?

THE IBM APPROACH (2 samples)

→ Cramer-Von Mises type test : CLT for estimators of unknown p_i's and F_i's and known stochastic behavior of empirical contrast $n\hat{d}_n(\hat{\theta}_n)$.

→ Inversion / Best Matching (IBM) : with the discrepancy measure

$$d(\theta) = \int_{\mathbb{R}} (F_1(x, p_1) - F_2(x, p_2))^2 dU(x), \tag{3}$$

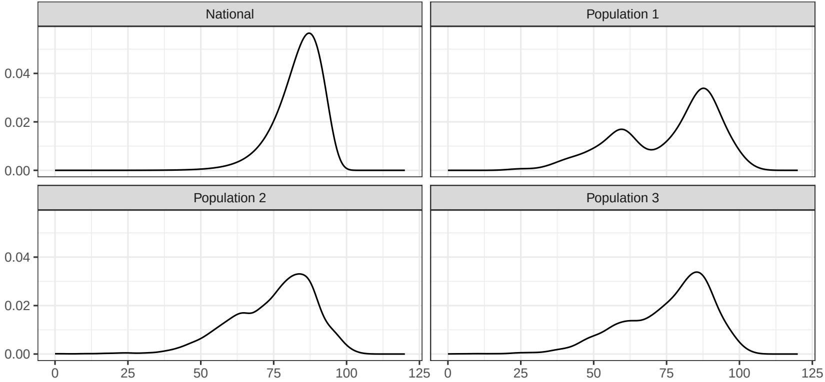
with $\theta = (p_1, p_2) \in \Theta = [\delta_1, \delta_2]^2$.

→ X.M., D.P., Y.S., P.V. **Two-sample contamination model test**. Bernoulli (2023), `https://hal.science/hal-03985733`.

→ R package admix :
`https://cran.r-project.org/web/packages/admix/index.html`

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 $X^{(i)} = (X_1^{(i)}, ..., X_{n_i}^{(i)})$ 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

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

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 statistics**.
→ Add a penalization term to select the right number of terms in the final test statistic.

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`.

ALGORITHM : A BASIC IDEA

K-sample 2-component mixture clustering (K2MC)

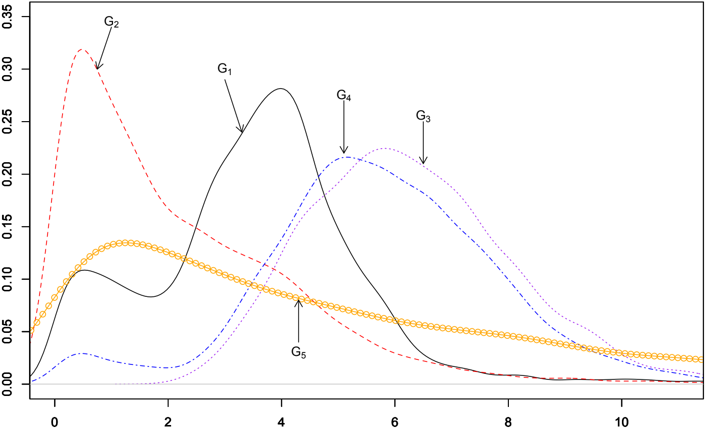
1] Initialization : create the first cluster to be filled, *i.e.* $c = 1$.
By convention, $S_0 = \emptyset$.

2] Select $\{x, y\} = \text{argmin}\{d_n(i,j); i \neq j \in S \setminus \bigcup_{k=1}^{c} S_{k-1}\}$.

3] Test H_0 between x and y.
If H_0 is not rejected then $S_1 = \{x, y\}$,
Else $S_1 = \{x\}$, $S_{c+1} = \{y\}$ and then $c = c + 1$.

4] While $S \setminus \bigcup_{k=1}^{c} S_k = \emptyset$ **do**
Select $u = \text{argmin}\{d(i,j); i \in S_c, j \in S \setminus \bigcup_{k=1}^{c} S_k\}$;
Test H_0 the simultaneous equality of all the f_j (k-sample test), $j \in S_c$:
If H_0 not rejected, then put $S_c = S_c \bigcup \{u\}$;
Else $S_{c+1} = \{u\}$ and $c = c + 1$.
End While

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	$\mathcal{E}(1/3)$	$\mathcal{E}(1/2)$	$\mathcal{G}(13,2)$	$\mathcal{G}(12,2)$	$\mathcal{E}(1/6)$
"Unknown" distribution F_i	$\mathcal{G}(16,4)$	$\mathcal{G}(16,4)$	$\mathcal{G}(15,3)$	$\mathcal{E}(1/2.5)$	$\mathcal{E}(1/2.5)$

1. Population 1 : $0.6\,\mathcal{G}(16;4) + 0.4\,\mathcal{E}(1/3)$
2. Population 2 : $0.12\,\mathcal{G}(16;4) + 0.88\,\mathcal{E}(1/2)$
3. Population 3 : $0.15\,\mathcal{G}(15;3) + 0.85\,\mathcal{G}(13;2)$
4. Population 4 : $0.08\,\mathcal{E}(1/2.5) + 0.92\,\mathcal{G}(12;2)$
5. Population 5 : $0.1\,\mathcal{E}(1/2.5) + 0.9\,\mathcal{E}(1/6)$

$\Rightarrow$ **Clusters to be found :**

<u>3 clusters</u> (pop. (1,2) ; (4,5) and 3) !

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


2024 JOINT COLLOQUIUM
ASTIN
AFIR/ERM
IAAHS
IAALS
IACA
PBSS
Subscribe to mailing list on :
Joco2024.org
RECONNECTING ACTUARIES
SEPTEMBER 22 to 26, 2024
BRUSSELS
IAA AAI
International Actuarial Association
Association Actuarielle Internationale
Institut des Actuaires en Belgique
Instituut van de Actuarissen in België
Institute of Actuaries in Belgium
iA|BE

APPENDIX 1 : 2-sample TESTING STRATEGY

- Inner model convergence regime caracterized by $Z(\theta^*, L_1, L_2)$ and $Z(\theta^c, L_1, L_2)$ under both H_0 and H_1 (closed form stochastic integrals) $\Rightarrow$ Hypothesis-free test quality !
- Possibility to sample them using $Z(\hat{\theta}_n, \hat{L}_1, \hat{L}_2)$.
- $(1-\alpha)$-quantile Monte Carlo estimation of the stochastic integral $Z(\hat{\theta}_n, \hat{L}_1, \hat{L}_2)$ denoted $\hat{q}_{1-\alpha}$.

Finally, H_0-rejection rule :

$$n\hat{d}_n(\hat{\theta}_n) \quad \geq \quad \hat{q}_{1-\alpha}. \tag{5}$$

Interpretation : *if the test statistic is too far from the inner model convergence regime we suspect that something goes wrong.*

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, $\forall i \neq j \in \{1, ..., k\}$,

1. Estimate $\widehat{\theta}_n(i,j) = \arg\min_{\theta \in \Theta_{i,j}} d_n[i,j](\theta)$,
2. Compute the statistic $T_{i,j} = n\, d_n[i,j](\widehat{\theta}_n(i,j))$.

We then obtain $d(k) = k(k-1)/2$ comparisons that we embed :

$$
\begin{aligned}
U_1 &= T_{1,2} \\
U_2 &= T_{1,2} + T_{1,3} \\
&\vdots \\
U_{d(k)} &= T_{1,2} + \cdots + T_{k-1,k},
\end{aligned}
$$

Consider the penalization rule (mimicking Schwarz criteria) :

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

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

$\Rightarrow$ Our data-driven test statistic is given by

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

Simulation results :
$\rightarrow$ The test shows good empirical levels in many different situations,
$\rightarrow$ It also has satisfactory empirical power, provided that $n_i p_i$ is high enough.