UNIVERSITÉ DE NEUCHÂTEL

Institut de statistique

Compositional analysis of a mixture distribution

with application to categorical modelling

Monique Graf[1] and Desislava Nedyalkova[2]
[1]Institut de Statistique, Université de Neuchâtel
[2]Swiss Federal Statistical Office

Séminaire de Statistiques - Institut de Statistique May, 16 2013

Outline

Introduction

Many distributions encountered in practice exhibit skewness, e.g income distributions.

Ways of addressing the issue :

- *Modify the usual normal or Student distribution as a whole.*
  Principle :
  $$f(x) = 2\phi(x)\Phi(\alpha x)$$
  O'Hagan and Leonhard (1976), Azzalini (1985), Azzalini & Genton (2008).
- *Act on the tails only.*
  Principle :
  - Use a nonparametric estimate for the bulk of data ;
  - Model the tails with e.g. the Pareto distribution :
    $$\mathbb{P}(X > x) = (x/x_{\min})^{-\alpha}, \quad x \geq x_{\min}$$
    Victoria-Feser & Ronchetti (1994), Cowell & Victoria-Feser (2006b), Van Kerm (2007).
- Choose a flexible enough parametric distribution.

Compound distributions

Many probability distributions can be represented as compound distributions, see e.g. Johnson et al. (1995) and Kleiber and Kotz (2003).

- $\boldsymbol{\theta}$: random vector of parameters with a prob. density $h(\boldsymbol{\theta})$.
- $\mathbf{y}$: vector of observations.
- $g(\mathbf{y}|\boldsymbol{\theta})$: conditional distribution.

Compound distribution = marginal distribution of $\mathbf{y}$.
We shall consider a positive range for the components of $\boldsymbol{\theta}$. Thus the marginal density of $\mathbf{y}$ is

$$f(\mathbf{y}) = \int_0^\infty g(\mathbf{y}|\boldsymbol{\theta}) h(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

The densities f, g and h may depend on some other deterministic parameters.

Mixture distribution derived from a compound distribution

Our approach can be summarized as follows :

- Compounding property : interpreted as the consequence of a mixture of populations.
- Latent parameter $\boldsymbol{\theta}$: a measure of discrepancy between populations.
- Define a partition of the domain of definition of $\boldsymbol{\theta}$ into L parts $D_1, ..., D_L$.
- Represent f as a finite mixture of densities :

$$
\begin{aligned}
f(\mathbf{y}) &= \sum_{\ell=1}^{L} \int_{D_\ell} g(\mathbf{y}|\boldsymbol{\theta}) h(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&= \sum_{\ell=1}^{L} \underbrace{\int_{D_\ell} h(\boldsymbol{\theta})\, d\boldsymbol{\theta}}_{p_{0\ell}} \underbrace{\frac{\int_{D_\ell} g(\mathbf{y}|\boldsymbol{\theta}) h(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int_{D_\ell} h(\boldsymbol{\theta})\, d\boldsymbol{\theta}}}_{f(\mathbf{y}|\boldsymbol{\theta} \in D_\ell)} \\
&= \sum^{L} p_{0\ell}\, f(\mathbf{y}|\boldsymbol{\theta} \in D_\ell) = \sum^{L} p_{0\ell}\, f_\ell(\mathbf{y}).
\end{aligned}
$$

Component densities

- The conditional densities $f_\ell(\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta} \in D_\ell)$ will be called the *component densities* in the mixture representation of $f(\mathbf{y})$.
- The weights $p_\ell = \Pr(\boldsymbol{\theta} \in D_\ell)$ of the component densities represent the probability that the underlying vector of parameters $\boldsymbol{\theta}$ belongs to domain D_ℓ, and together form a composition $\mathbf{p}$ (Aitchison, 1986).

Basic idea

Start with a parametric distribution $f(\mathbf{y})$

- possessing the compounding property,
- fitted at the overall population level.

Choose initial probabilities $p_{0\ell}, \ell = 1, ..., L$ that determine :

- the domains D_ℓ and
- the component densities $f_\ell(\mathbf{y})$.

Thus

$$f(\mathbf{y}) = \sum_{\ell=1}^{L} p_{0\ell}\, f_\ell(\mathbf{y}).$$

Keeping the components $f_\ell(\mathbf{y})$ fixed, write a model for the p_ℓ, possibly using group categories.

Generalized beta distribution of the second kind (GB2)

- The GB2 distribution (McDonald, 1984) gives an excellent fit to income distributions.
  See also McDonald and Butler (1987) ; McDonald and Xu (1995) ; McDonald and Ransom (2008) ; Jenkins (2008, 2009) ; Kleiber and Kotz (2003), Graf and Nedyalkova (2013).
- It possesses the compounding property.

For these reasons, we use it in the sequel.

Generalised Beta distribution of the second kind (GB2)

The GB2 depends on four parameters a, b, p, q.

- *Density :*

$$f(x; a, b, p, q) = \frac{a}{b\, B(p, q)} \frac{(x/b)^{ap-1}}{(1 + (x/b)^a)^{p+q}}$$

  where
  - $B(p, q)$ is the beta function,
  - $b > 0$ is a scale parameter,
  - $p > 0, q > 0$ and $a > 0$ are shape parameters.

- *Property :*
  If $X \sim GB2(a, b, p, q)$ then $1/X \sim GB2(a, 1/b, q, p)$.

Decomposition of the GB2

- The GB2 can be seen as a compound distribution with mixing parameter θ,
- where θ is the scale of a generalized gamma distribution with density :

$$g(x; a, \theta, p) = \frac{a}{\theta\,\Gamma(p)}(x/\theta)^{ap-1}\exp -(x/\theta)^a$$

- θ follows an inverse gamma distribution with density

$$h(\theta; a, b, q) = \frac{a}{b\,\Gamma(q)}(\theta/b)^{-aq-1}\exp -(\theta/b)^{-a}$$

- The GB2 density is obtained by integration over θ :

$$f(x; a, b, p, q) = \int_0^\infty h(\theta; a, b, q)\, g(x; a, \theta, p)\, d\theta$$

Use of decomposition

From this property, we define a mixture distribution by discretizing the range of the random scale and integrating out within the resulting intervals.
If f is the GB2 density and f_ℓ the conditional density in the ℓ-th interval, the probability weights $p_{0\ell}$ are determined and

$$f(x) = \sum_{\ell=1}^{L} p_{0\ell}\, f_\ell(x)$$

Keeping the densities f_ℓ fixed, we can set up a model for the probability weights p_ℓ and thus adjust the distribution to specific population groups.

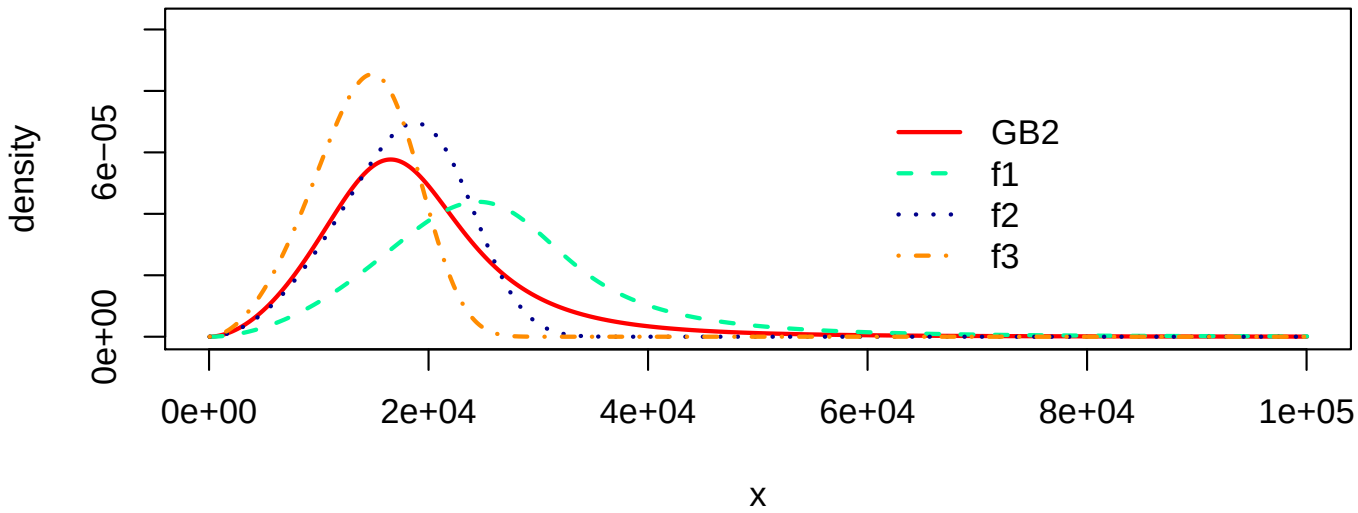
Right-, left-tail decomposition

We show that the decomposition can be made in two different ways :

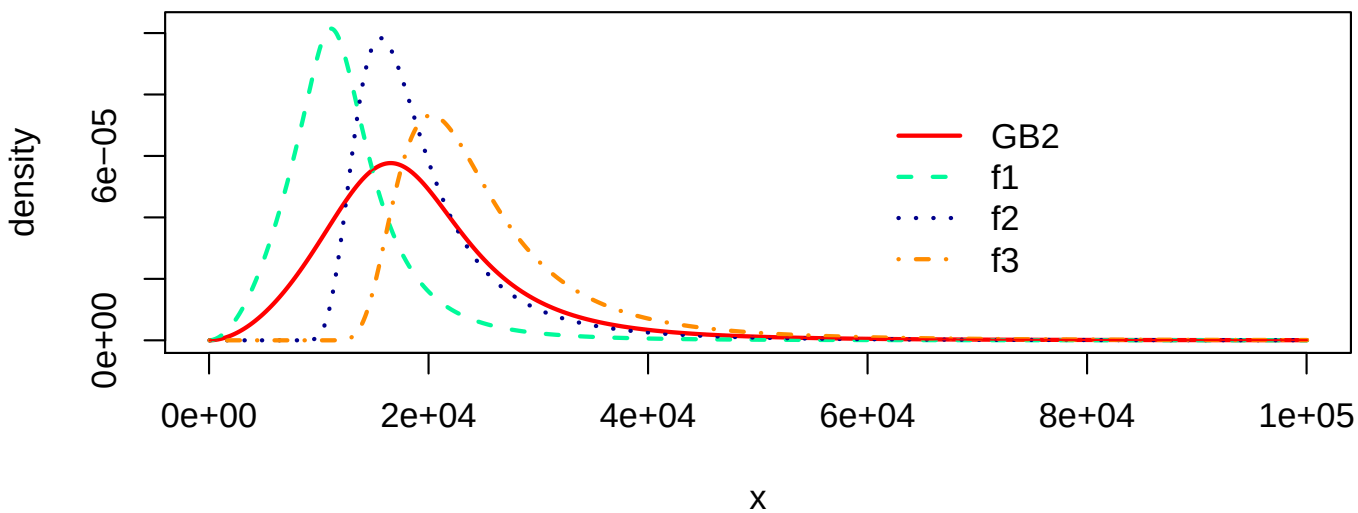
- Right tail decomposition : using the distribution of the original (income) variable
- Left tail decomposition : through the distribution of the inverse income.

The latter emphasizes the left tail and proves to give better results for poverty measures.

Right tail decomposition and GB2 density: AT 2006



Left tail decomposition and GB2 density: AT 2006



Right tail decomposition

Partition the domain of integration of the random scale parameter θ into L intervals, with limits

$$\theta_0 = 0 < \theta_1 < ... < \theta_L = \infty.$$

Set $u_\ell = (\theta_{L-\ell}/b)^{-a}, \ell = 0, ..., L, \qquad (u_{\ell-1} < u_\ell).$

$$p_{0\ell} = P(u_\ell, q) - P(u_{\ell-1}, q)$$

where $P(\cdot, q)$ is the CDF of the standard gamma distribution with shape parameter q.

Right tail Set $t = (x/b)^a + 1$.

$$f_\ell(x) = f(x)\frac{P(tu_\ell, p+q) - P(tu_{\ell-1}, p+q)}{p_{0\ell}},$$

In practice, the $p_{0\ell}$ are chosen and determine the u_ℓ.

Left tail decomposition

Left tail Set $t' = (x/b)^{-a} + 1$ and
$u'_\ell = (\theta_\ell/b)^a, \ell = 0, ..., L, \quad (u'_{\ell-1} < u'_\ell)$

$$\tilde{p}_{0\ell} = P(u'_\ell, p) - P(u'_{\ell-1}, p)$$

$$\tilde{f}_\ell(x) = f(x)\frac{P(t'u'_\ell, p+q) - P(t'u'_{\ell-1}, p+q)}{\tilde{p}_{0\ell}},$$

Model

From now on, $f_\ell(x)$ denotes a component density, whether right- or left-tail.
To each unit k in the population, we associate a density

$$\boxed{f_k(x) = \sum_{\ell=1}^{L} p_{k,\ell} f_\ell(x)}$$

Denote by

- $\mathbf{z}_k$ be a $I \times 1$ - vector of auxiliary information for unit k,
- $p_{k,\ell}$ the weight (to be estimated) of the density f_ℓ for unit k.

We pose a linear model for the log-ratios $v_{k,\ell}$:

$$\log(p_{k,\ell}/p_{k,L}) = v_{k,\ell} = \sum_{i=1}^{I} \lambda_{\ell i} z_{ki} = \mathbf{z}_k^T \boldsymbol{\lambda}_\ell$$

Log-likelihood

$$\log L(\lambda_1, ..., \lambda_{L-1}) = \sum_{k=1}^{n} w_k \log \left(\sum_{\ell=1}^{L} p_{k,\ell} f_\ell(x_k) \right)$$

$$\sim \sum_{k=1}^{n} w_k \log \left(\sum_{\ell=1}^{L} p_{k,\ell} \frac{P(t_k u_\ell, p+q) - P(t_k u_{\ell-1}, p+q)}{p_{0\ell}} \right)$$

Maximum likelihood equations

$$\sum_{k=1}^{n} w_k U_{k,\ell} \mathbf{z}_k = 0,$$

where

$$U_{k,\ell} = p_{k,\ell} \left(\frac{\left[P(t_k u_\ell, p+q) - P(t_k u_{\ell-1}, p+q) \right] / p_{0\ell}}{\sum_{j=1}^{L} p_{k,j} \left[P(t_k u_j, p+q) - P(t_k u_{j-1}, p+q) \right] / p_{0j}} - 1 \right).$$

For each $\ell = 1, ..., L-1$, the number of equations is equal to the dimension of $\mathbf{z}_k$.
The probabilities $p_{0\ell}$ are the initial (GB2) probabilities.

Initial values

Consider I categories $c_1, ..., c_I$.
Then $p_{k,\ell} = p_{i,\ell}$ if household $k \in c_i$. We have used three different starts :

1. $p_{k,\ell}^{(0)} =$
   $\sum_{k \in c_i} w_k \left(P(t_k u_\ell, p+q) - P(t_k u_{\ell-1}, p+q) \right) / \sum_{k \in c_i} w_k$,
2. $p_{k,\ell}^{(0)} = \hat{p}_\ell$, where $\hat{p}_\ell$ is the fit with $I = 1$ and $z_k = 1, \forall k$,
3. $p_{k,\ell}^{(0)} = p_{0\ell}$.

Variance estimation

Second derivatives of the pseudo-loglikelihood :

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\lambda}_\ell \partial \boldsymbol{\lambda}_i^T} = -\sum_{k=1}^{n} w_k A_{k,\ell i} \mathbf{z}_k \mathbf{z}_k^T ,$$

where

$$A_{k,\ell i} = \begin{cases} p_{k,i} U_{k,\ell} + p_{k,\ell} U_{k,i} + U_{k,\ell} U_{k,i} & \text{if } i \neq \ell, \\ -(1 - 2p_{k,\ell}) U_{k,\ell} + U_{k,\ell}^2 & \text{if } i = \ell, \end{cases}$$

Variance estimation II

Vector of all parameters of dimension $I(L-1)$:

$$\boldsymbol{\nu} = (\boldsymbol{\lambda}_1^T, ...\boldsymbol{\lambda}_{L-1}^T)^T.$$

Hessian : Matrix of second derivatives of the pseudo-loglikelihood expressed at $\hat{\boldsymbol{\nu}}$:

$$\ell''(\hat{\boldsymbol{\nu}}) = -\sum_{k=1}^{n} w_k \widehat{\mathbf{A}}_k \otimes \mathbf{z}_k \mathbf{z}_k^T,$$

where $\otimes$ is the Kronecker product and $\widehat{\mathbf{A}}_k$ is the square matrix with elements $\hat{A}_{k,\ell i}$.

Variance estimation III

Sandwich variance estimate :

- $\hat{V}(\hat{\boldsymbol{\nu}})$: design based variance of the sum of scores (=1st derivative of the log likelihood).
- Covariance matrix of parameter estimates :

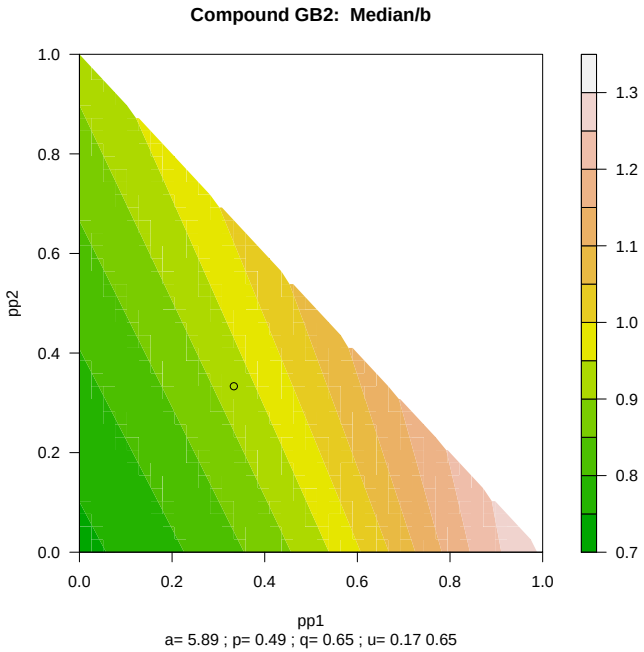
$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\nu}}) = \ell''(\hat{\boldsymbol{\nu}})^{-1}\hat{V}(\hat{\boldsymbol{\nu}})\ell''(\hat{\boldsymbol{\nu}})^{-1}.$$

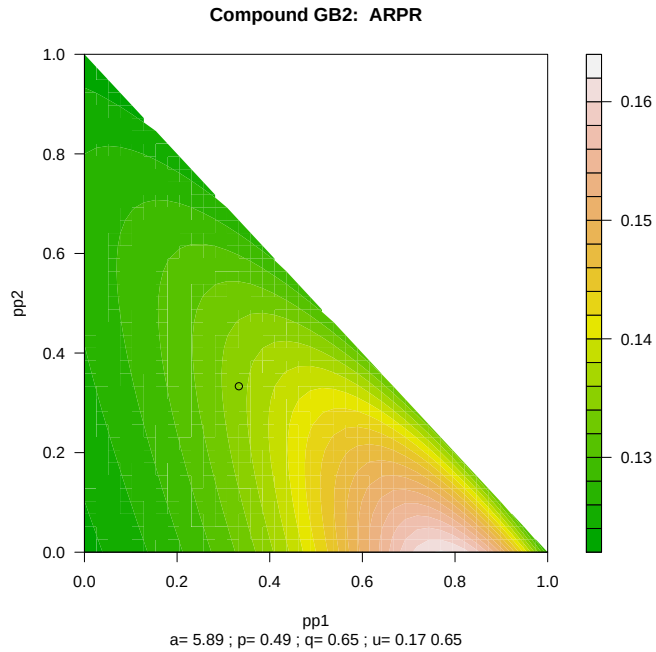
Poverty and inequality indicators under the compound GB2

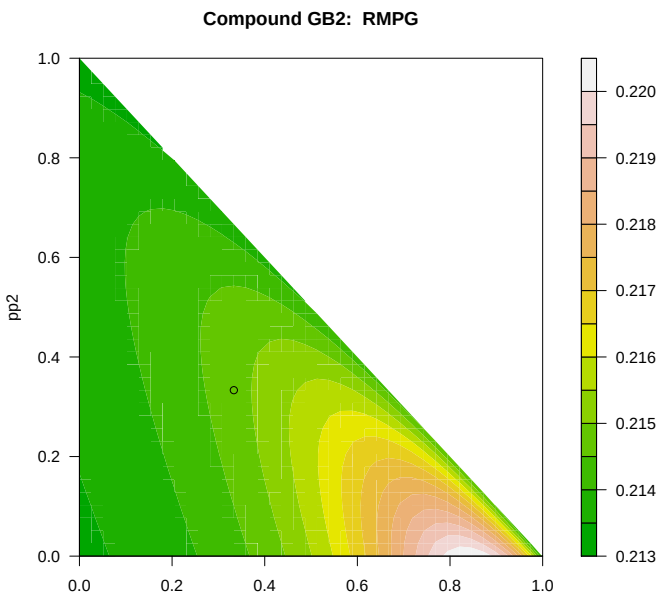
- Moments of the compound GB2 can be obtained with R built in functions.
- Quantiles and derived quantities must be found by a numerical integration.

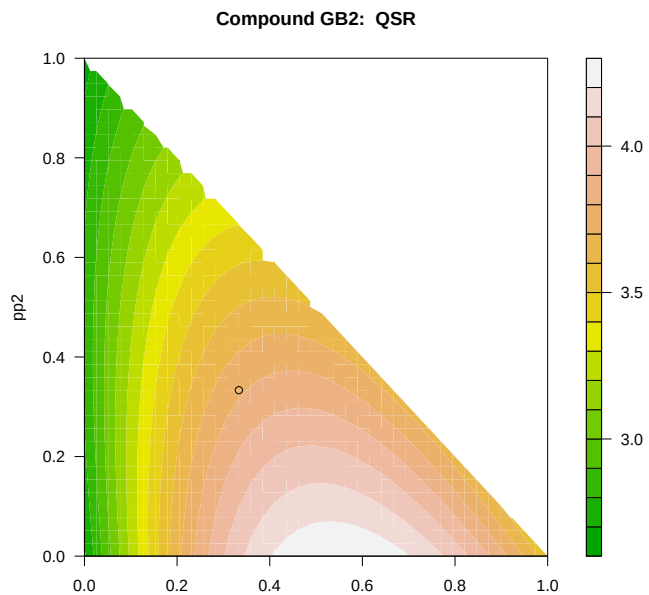
All of them can be easily computed with the R package GB2 (Graf and Nedyalkova, 2012).

Indicators under the right tail decomposition

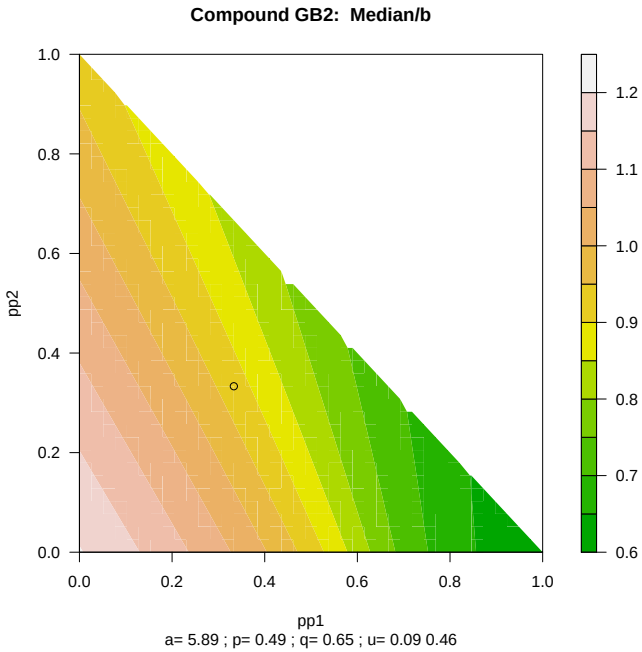


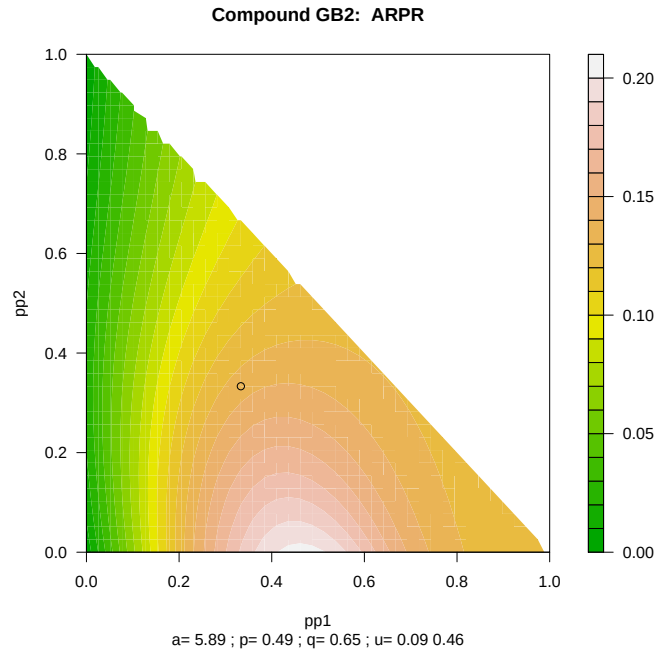


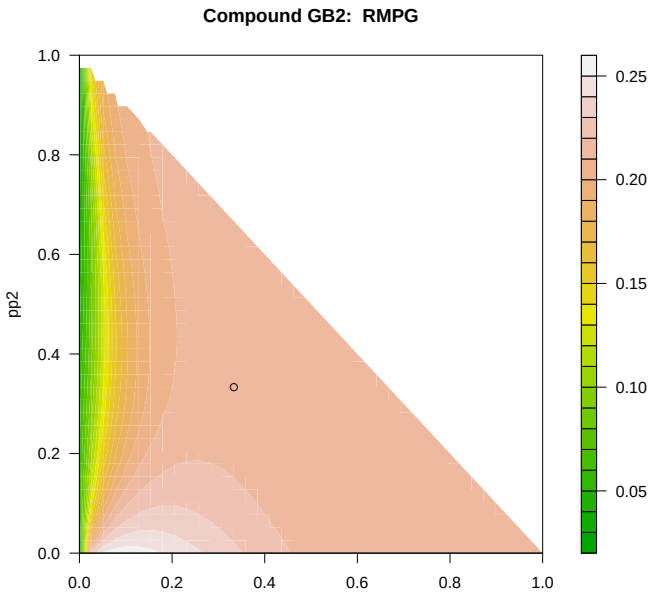


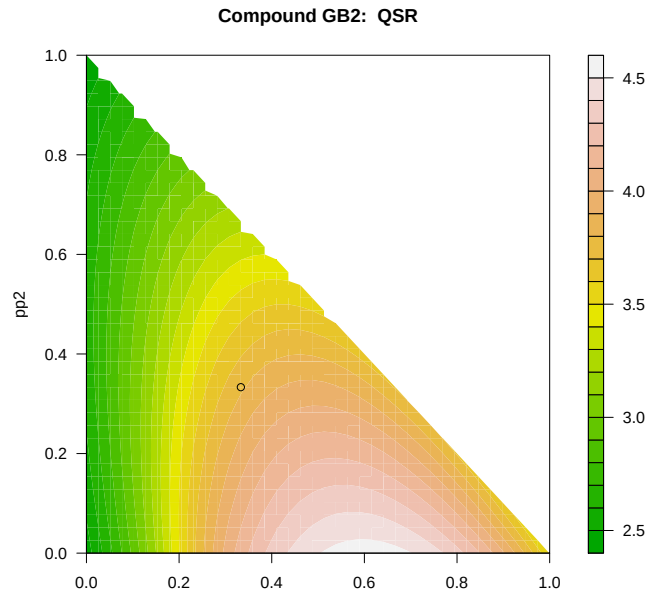


Indicators under the left tail decomposition









The EU-SILC survey

The European Union Statistics on Income and Living Conditions (EU-SILC) is described on the EUROSTAT website :

- Aims : collecting timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions. This instrument is anchored in the European Statistical System (ESS).
- The EU-SILC project launched in 2003 on the basis of a 'gentleman's agreement' in six Member States (Belgium, Denmark, Greece, Ireland, Luxembourg and Austria), as well as in Norway.
- The 2006 dataset we use incorporate 26 countries.
- The survey is also implemented in Bulgaria, Romania, Turkey and in Switzerland as from 2007. Implementation in Croatia is being discussed.

Swiss dataset CH-SILC

- In Switzerland, the SFSO is responsible of the survey.
- Meta data on the Swiss SILC survey are published by FORS (Univ. of Lausanne).
- We use the Swiss 2009 dataset.

Two applications :

1. Satisfaction scores on financial situation : how are they related to the actual financial situation ?
2. Income distribution per household composition category (Households with no child, two or more adults with children, one adult with children).

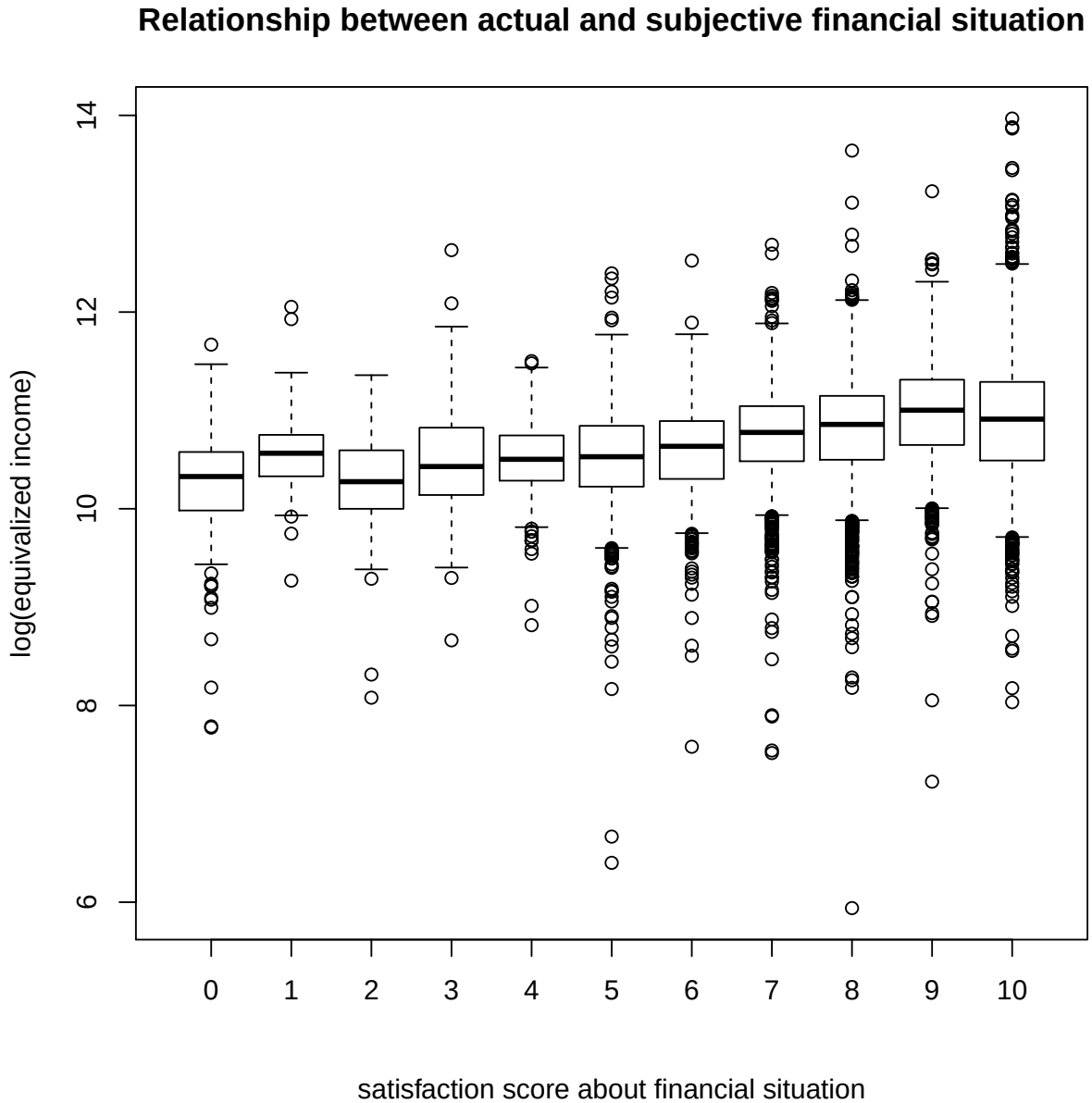
Satisfaction scores on financial situation

(in the Swiss SILC survey 2009)

Analysis made with the households data file (unit=household).
Let us first have a look at the data. Next page :

- No of HH in sample per satisfaction scores.
- Weighted box plots using the HH extrapolation weights.

sc	mis	0	1	2	3	4	5	6	7	8	9	10
no	29	102	22	50	87	176	765	612	1217	1890	737	1669



Auxiliary variables

Because of the small number of observations, we group the scores into 4 categories :

$$\boxed{0-4\,;\,5-6\,;\,7-8\,;\,9-10}$$

Different possible choices for auxiliary variables :

- Four indicators of HH categories,
- One intercept, three indicators for HH categories.

Results are similar.
The subsequent analyses are processed with the first option.

Model

A priori choices :

- Left tail decomposition.
- $L = 3$ and $p_{0\ell} = 1/3$.

Then the probability that a HH with satisfaction category i stems from component f_ℓ is :

$$\hat{p}_{i,\ell} = \begin{cases} \exp \hat{\lambda}_{i,\ell} / (1 + \sum_{j=1}^{L-1} \exp \hat{\lambda}_{i,j}) & \ell = 1, ..., L-1 \\ \\ 1/(1 + \sum_{j=1}^{L-1} \exp \hat{\lambda}_{i,j}) & \ell = L \end{cases}$$

These probabilities are the same for all HH in category i.

Parameter estimates

score i	$\hat{p}_{i,1}$	$\hat{p}_{i,2}$	$\hat{p}_{i,3}$	$b_{i,1}$	$b_{i,2}$	$s_{i,1}$	$s_{i,2}$
0-4	0.73	0.26	0.00	4.01	1.48	0.29	0.25
5-6	0.49	0.51	0.00	3.56	2.01	0.38	0.23
7-8	0.25	0.41	0.35	-0.24	0.27	0.05	0.10
9-10	0.29	0.07	0.63	-0.54	-1.42	0.04	0.35

Balances

Recall that $\lambda_{i,\ell} = \log(p_{i,\ell}/p_{i,3})$.
Denoting by $\mathbf{p}_i = (p_{i,1}, p_{i,2}, p_{i,3})^t$, we have defined

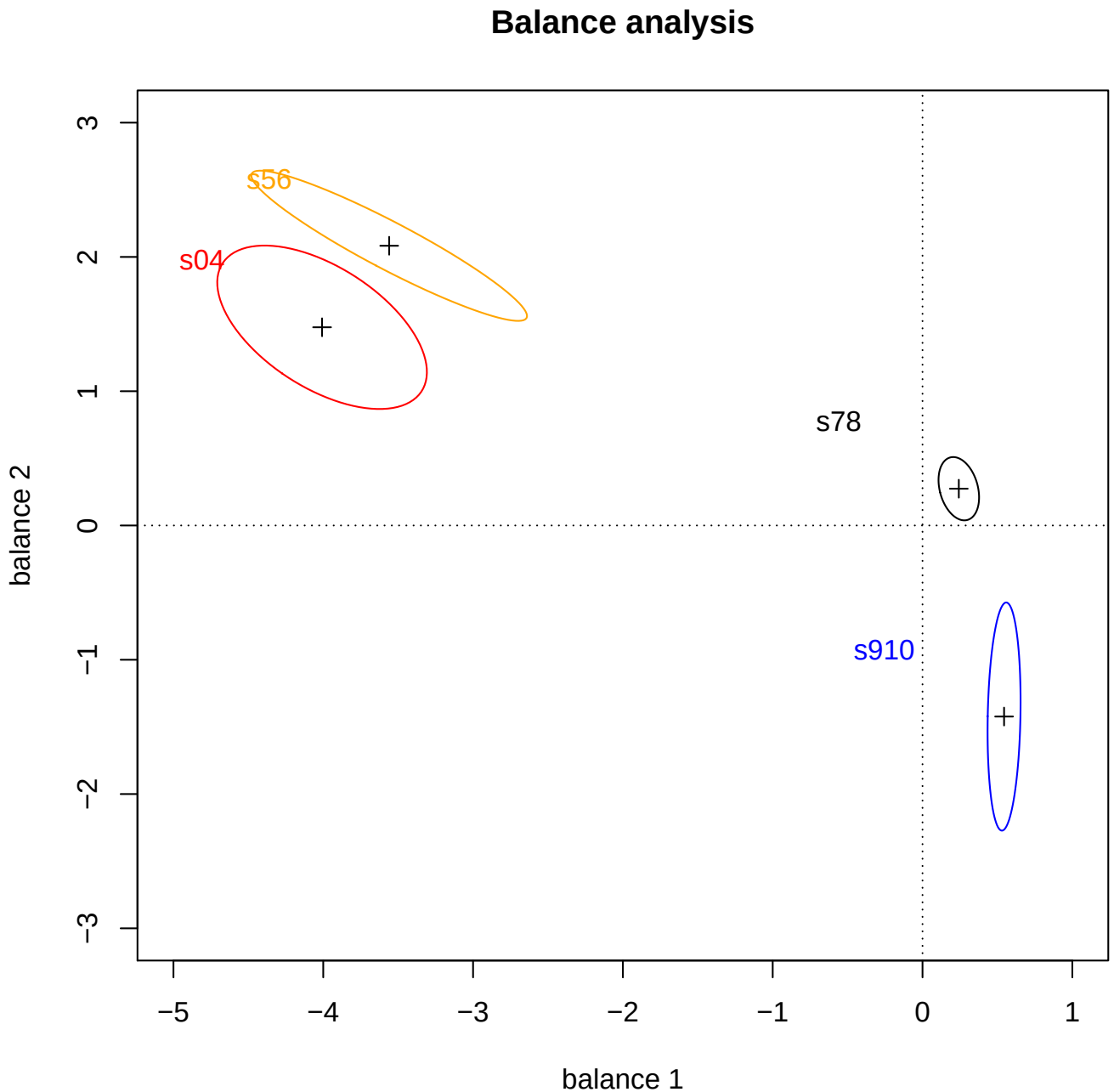
$$\boldsymbol{\lambda}_i = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \log(\mathbf{p}_i)$$

The two row vectors are not orthogonal and not unit-normed.
Egozcue and Pawlowsky-Glahn (2005, 2011) suggest to replace the parameters by orthonormal contrasts in $\log(\mathbf{p}_i)$.
We choose :

$$\mathbf{b}_i = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 0 & 1 \\ -\frac{1}{2} & 1 & -\frac{1}{2} \end{pmatrix} \log(\mathbf{p}_i)$$

The $\mathbf{b}_i$ are linearly related to the $\boldsymbol{\lambda}_i$, so their covariance matrix can be obtained from the estimated cov. matrix $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\nu}})$.

Balance analysis with 95% confidence domains



Income distribution per household composition categories

(in the Swiss SILC 2009 and the EU-SILC 2006)

Child : aged 14 years old or less.

Considering three categories of households,

- **no.child** : without children
- **sa.ch** : with a single adult with children
- **ma.ch** : with two or more adults with children

Aim : to evaluate the predictive power of the HH categories on the equivalized income distribution.
Auxiliary variables : indicator variables of the 3 household categories.
Dataset : person's level.

Model

We work with a left tail decomposition into 5 components and $p_{0\ell} = \frac{1}{5}, \ell = 1, ..., 5$.
Computations are made with the GB2 R-package (except for the variance computation).

Parameter estimates and corresponding standard errors, Swiss SILC survey 2009

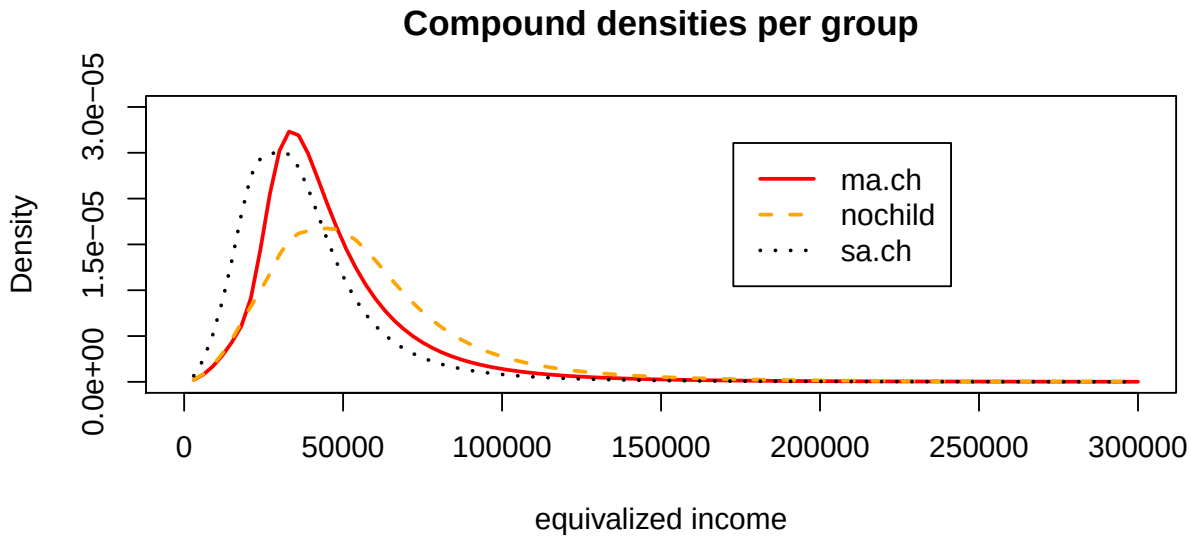
Row 'total' : compound GB2 model without explanatory variables.

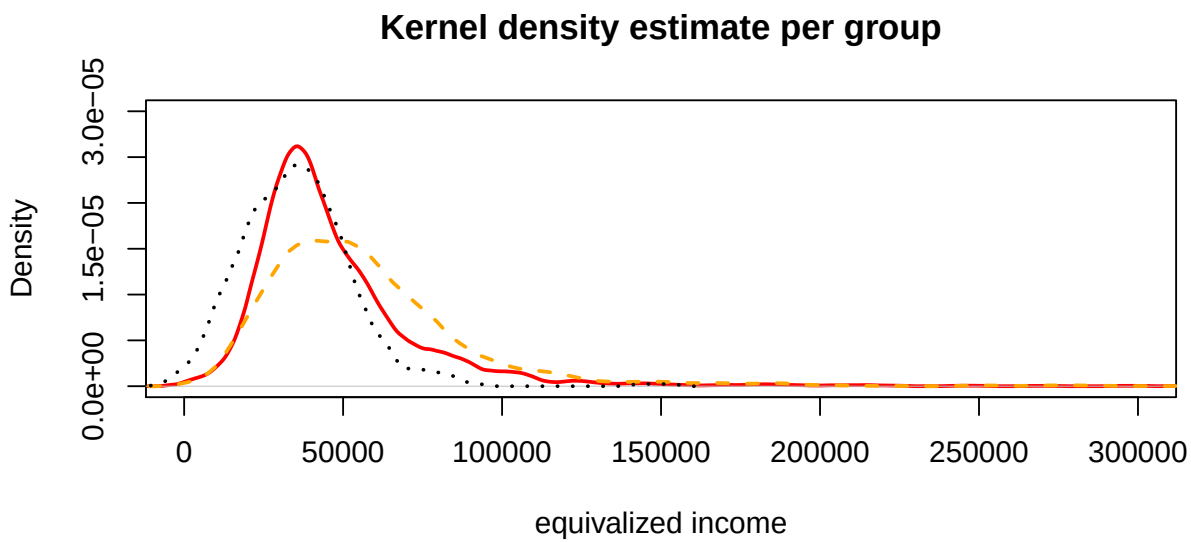
group i	$\hat{\lambda}_{i1}$	$\hat{\lambda}_{i2}$	$\hat{\lambda}_{i3}$	$\hat{\lambda}_{i4}$	s_{i1}	s_{i2}	s_{i3}	s_{i4}
total	-0.46	-0.16	-0.02	-2.62	0.10	0.18	0.22	2.96
ma.ch	0.58	1.55	1.18	-2.11	0.34	0.28	0.52	2.95
nochild	-0.85	-1.06	-0.47	-3.50	0.11	0.34	0.30	8.48
sa.ch	4.45	3.47	3.94	-14.48	0.34	1.14	0.64	0.21

Probabilities :

group i	$\hat{p}_{i,1}$	$\hat{p}_{i,2}$	$\hat{p}_{i,3}$	$\hat{p}_{i,4}$	$\hat{p}_{i,5}$
total	0.179	0.240	0.278	0.021	0.282
ma.ch	0.165	0.433	0.299	0.011	0.092
nochild	0.177	0.142	0.257	0.012	0.412
sa.ch	0.504	0.188	0.302	0.000	0.006

Swiss Silc data : Compound fit with auxiliary variables (top), Kernel estimate per household category (bottom).





EU-SILC data 2006

The same model was fitted to the 26 countries involved in the EU-SILC 2006.

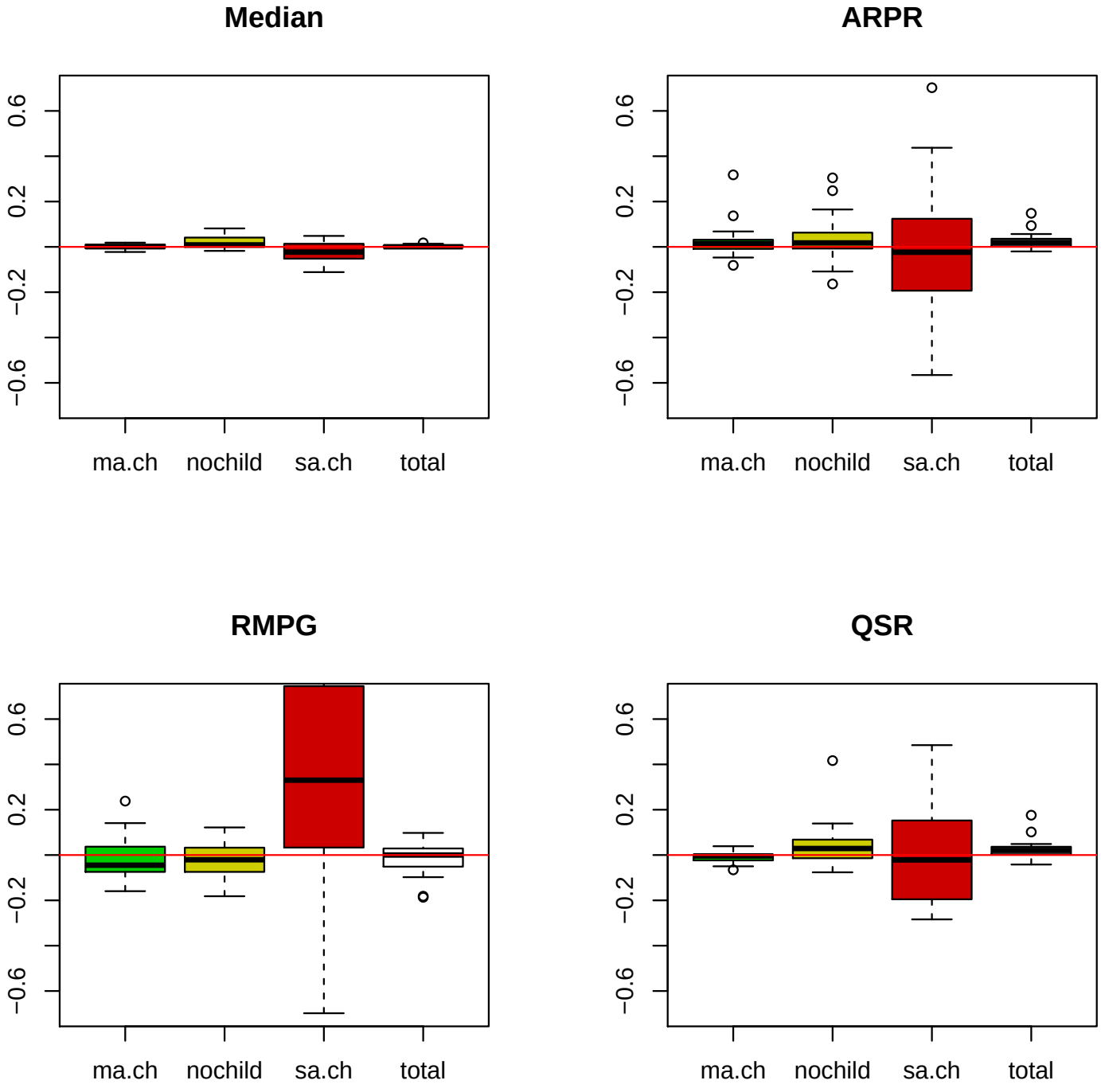
Average and minimun no of persons in sample by country

group	average	min	# RMPG outside plot	# QSR outside plot
ma.ch	8185	3937	0	0
nochild	11737	3708	0	0
sa.ch	628	165	7	1

Next slide : for median, ARPR, RMPG, QSR :

(direct estimate - CGB2)/CGB2

Relative difference of empirical to compound GB2 estimate



Perturbations

Model without auxiliary variables also fitted at the country level.
How do we have to modify the probabilities to obtain the fit per group ?
The answer is the so called perturbation (Aitchison,1986).

- $(p_1, p_2, ..., p_L)$: overall probabilities without auxiliary variables.

We can find $a_{i,\ell} > 0$ such that

$$(p_{i,1}, p_{i,2}, ..., p_{i,L}) = \frac{1}{\sum_{\ell=1}^{L} a_{i,\ell} p_\ell}(a_{i,1}p_1, a_{i,2}p_2, ..., a_{i,L}p_L)$$

Then $(a_{i,1}, , a_{i,2}, , ..., a_{i,L})$ is the perturbation to apply.
Perturbations can be scaled to 1.

Perturbations II

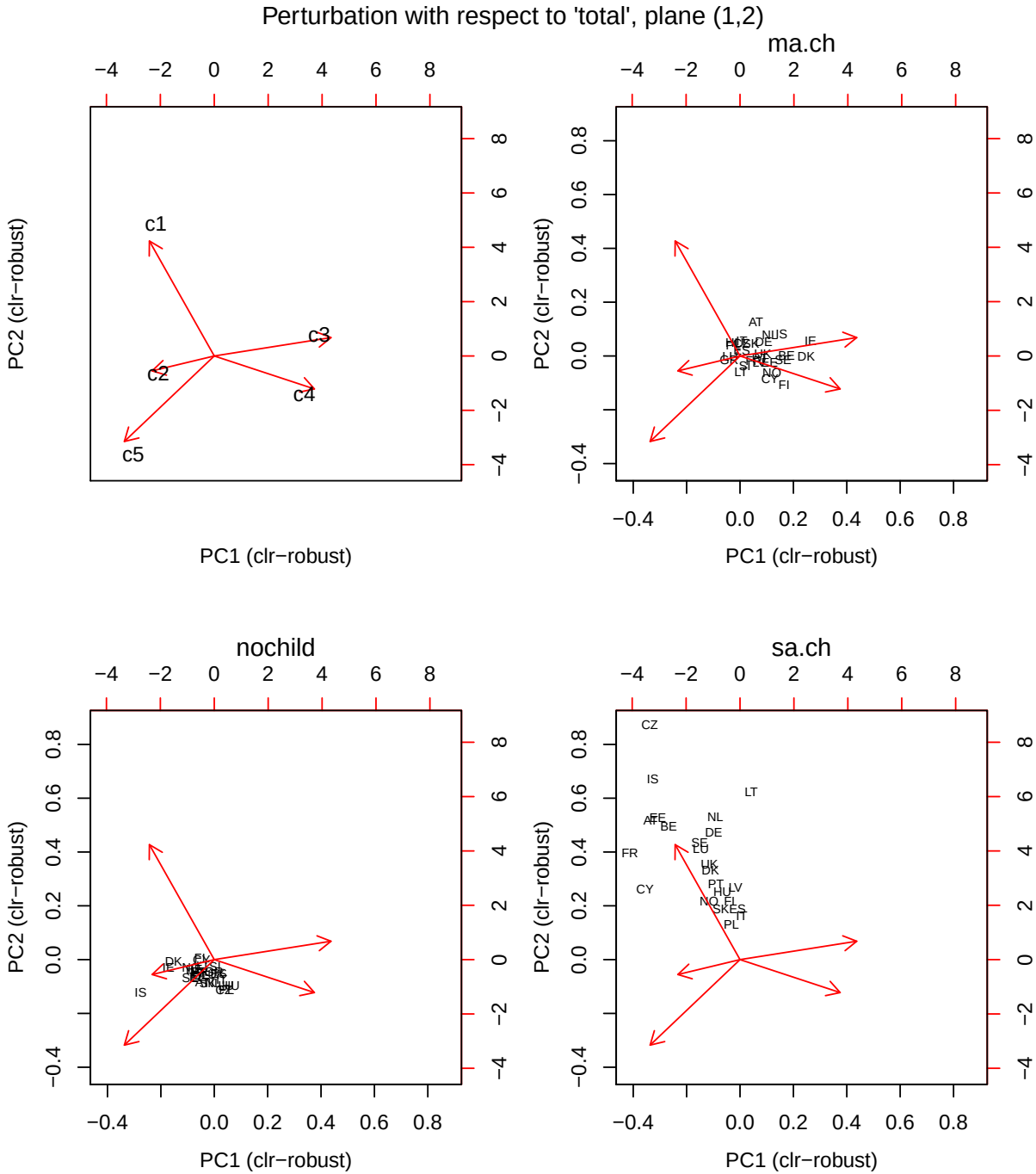
- Perturbation $(a_{i,1},, a_{i,2},, ..., a_{i,L})$: conditional probability vector of group i, given the overall mixture probabilities $(p_1, p_2, ..., p_L)$.
- If $p_\ell = p_{0\ell}$, then overall GB2 fit is perfect $\Rightarrow$ perturbation for group i = fitted probabilities to group i.
- If not, makes the fit comparable between countries.

Compositional principal components analysis

- Positivity and sum constraint on probability vectors make ordinary PCA not meaningful.
- Remedy : work with the centred log-ratio transform (Aitchison, 1986) :

$$\text{clr}(\mathbf{p}) = \left(\mathbf{I}_L - \frac{1}{L}\mathbf{1}_L\mathbf{1}_L'\right)\log(\mathbf{p})$$

- Compositional biplots have been defined by Aitchison and Greenacre (2002)
- A robust version is implemented in the R-package robCompositions (Templ et al., 2011)


Perturbation with respect to 'total', plane (1,2)
ma.ch
nochild
sa.ch
c1
c2
c3
c4
c5
PC1 (clr−robust)
PC2 (clr−robust)

Summary and discussion

- GB2 : good first approximation of the distribution of interest.
- Compound GB2 : natural way of refining the fit.
- Left tail decomposition proves better than right tail in the context of poverty measurement.
- When analysing mixture probabilities, compositional methods are a must.
- To compare different populations : a useful tool is to consider the perturbation of the mixture probabilities with auxiliary information, with respect to the probabilities computed overall, i.e. with a constant model.
- In our methodology : number of components and initial probabilities are given a priori.
  Methods to discriminate among decompositions, like Kullback-Leibler information, have not been implemented for the while.