Model-based clustering via Gaussian mixture models: a compositional sensitivity analysis

M. Comas-Cufí G. Mateu-Figueras J.A. Martín-Fernández
Universitat de Girona

July 18, 2013

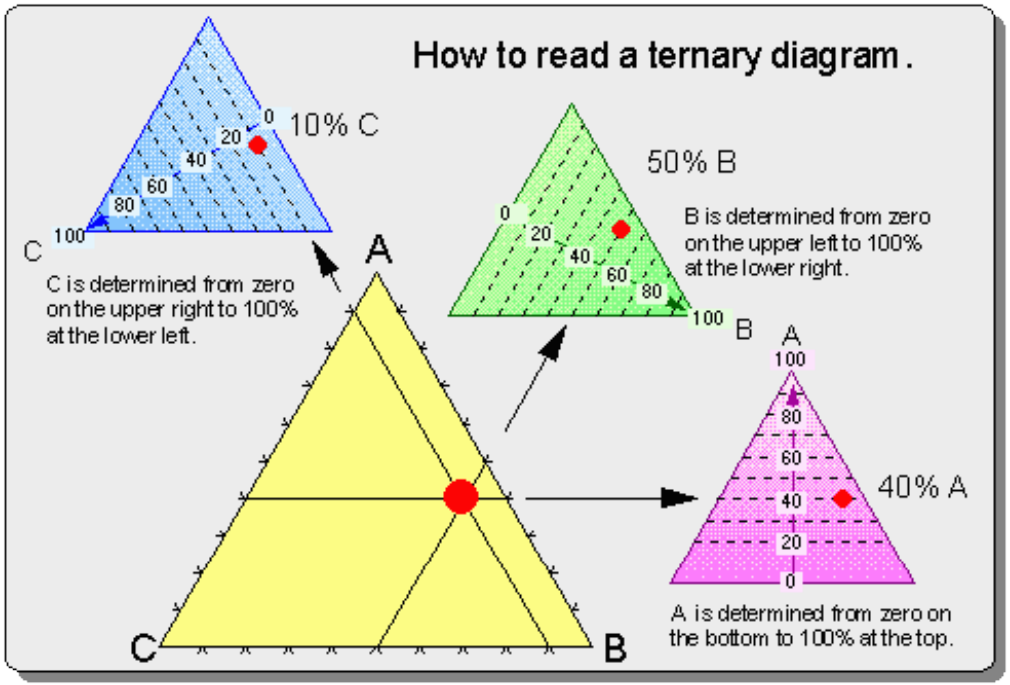
What are compositional data?

Definition

Compositional data are vectors of non-negative components showing the relative importance of a set of parts in a total.

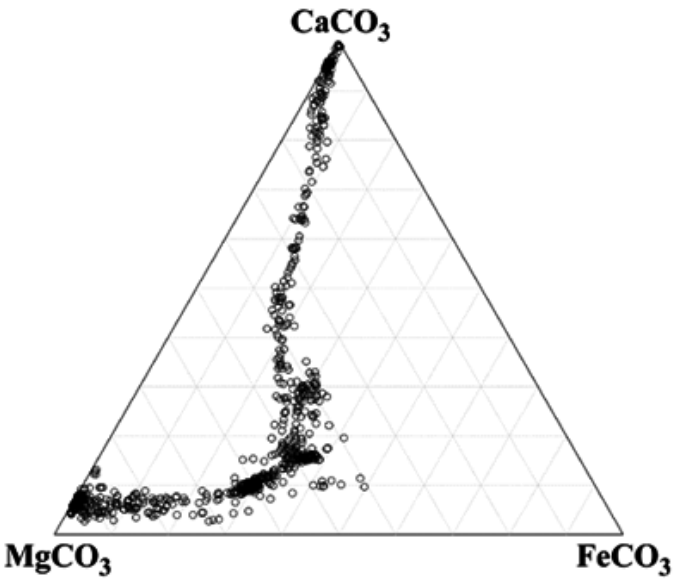
For convenience, the components of a composition are usually rescaled to sum a constant. These data are represented in a D-part simplex, *i.e.*

$$S^D = \{x \in {\mathbb{R}_+}^D; x_1 + ... + x_D = c\}.$$



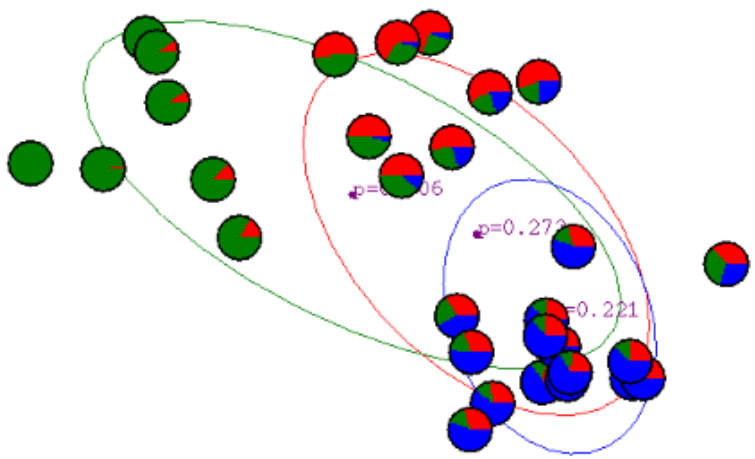
A compositional data approach to mixtures

- Compositional samples



Sample in $\mathcal{S}^D$

- Posterior probabilities τ



Posterior τ in $\mathcal{S}^D$

τ's from Moore tutorials at www.cs.cmu.edu/∼awm/tutorials.

GMM: Compositional samples

Incoherency with component elimination

Typical approach using Gaussian mixtures

	A	B	C
1	0.19	0.06	0.75
2	0.11	0.11	0.78
3	0.24	0.07	0.69
4	0.18	0.12	0.69
⋮	⋮	⋮	⋮

- Given a compositional sample

  $\rightarrow$ Colinerity between components

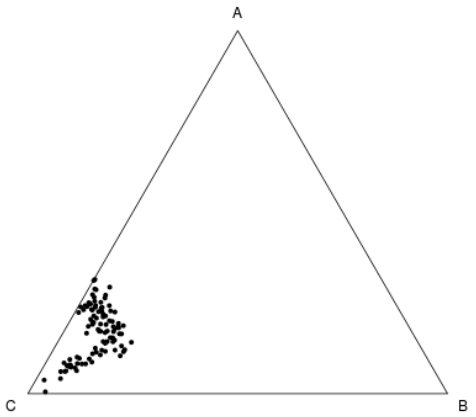
Incoherency with component elimination

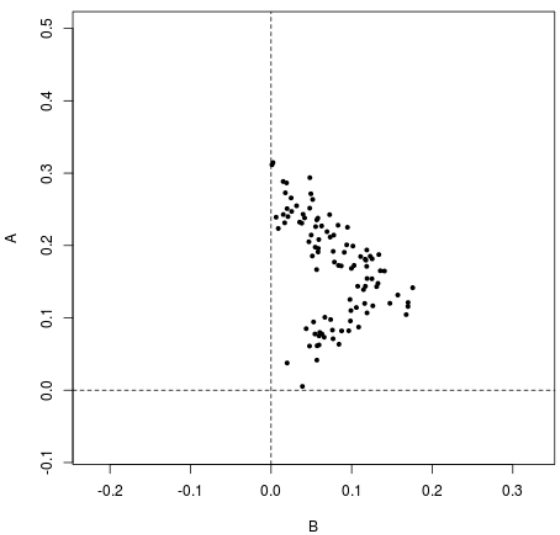
Typical approach using Gaussian mixtures

	A	B	C
1	0.19	0.06	~~0.75~~
2	0.11	0.11	~~0.78~~
3	0.24	0.07	~~0.69~~
4	0.18	0.12	~~0.69~~
⋮	⋮	⋮	⋮

- Given a compositional sample
  - → Colinerity between components
  - → Third component elimination

Incoherency with component elimination





Typical approach using Gaussian mixtures

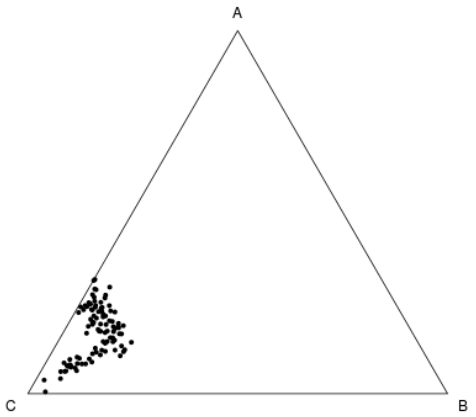
	A	B	C
1	0.19	0.06	~~0.75~~
2	0.11	0.11	~~0.78~~
3	0.24	0.07	~~0.69~~
4	0.18	0.12	~~0.69~~
⋮	⋮	⋮	⋮

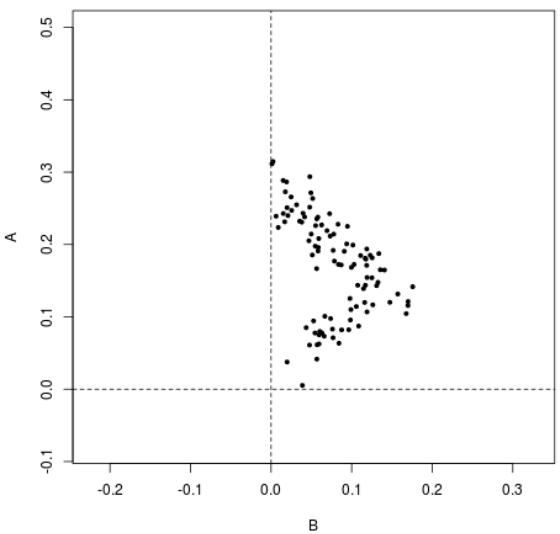
- Given a compositional sample

  → Colinerity between components

  → Third component elimination

Incoherency with component elimination



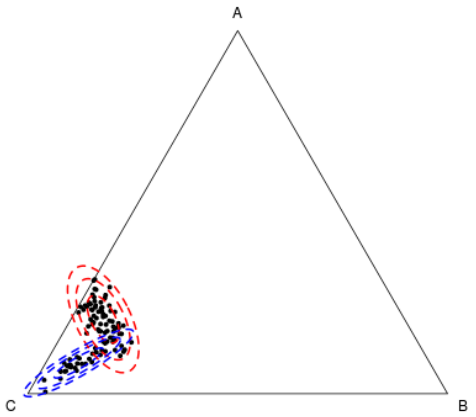


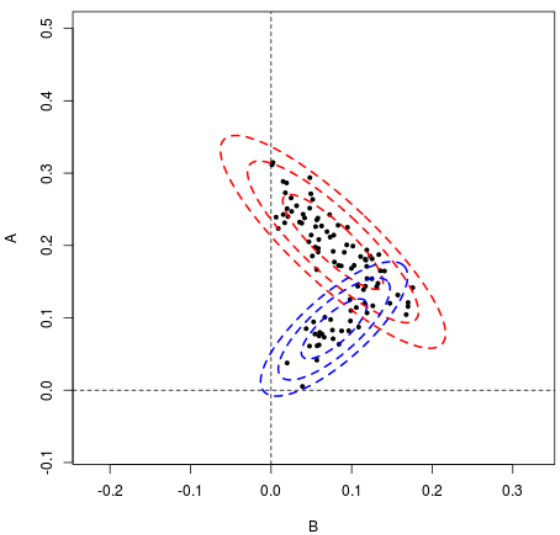
Typical approach using Gaussian mixtures

	A	B	C
1	0.19	0.06	~~0.75~~
2	0.11	0.11	~~0.78~~
3	0.24	0.07	~~0.69~~
4	0.18	0.12	~~0.69~~
⋮	⋮	⋮	⋮

- Given a compositional sample
  - → Colinerity between components
  - → Third component elimination
  - → Now we can proceed normally

Incoherency with component elimination





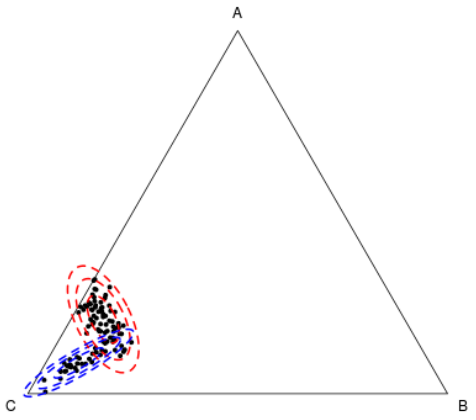
Typical approach using Gaussian mixtures

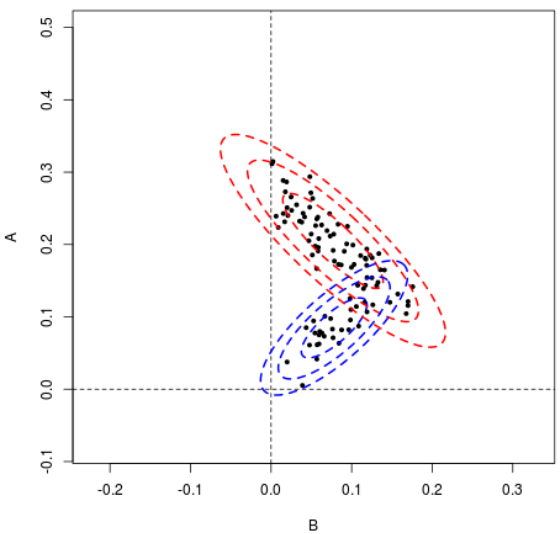
	A	B	C
1	0.19	0.06	~~0.75~~
2	0.11	0.11	~~0.78~~
3	0.24	0.07	~~0.69~~
4	0.18	0.12	~~0.69~~
⋮	⋮	⋮	⋮

- Given a compositional sample

  → Colinerity between components
  → Third component elimination
  → Now we can proceed normally

Incoherency with component elimination





Typical approach using Gaussian mixtures

	A	B	C
1	0.19	0.06	~~0.75~~
2	0.11	0.11	~~0.78~~
3	0.24	0.07	~~0.69~~
4	0.18	0.12	~~0.69~~
⋮	⋮	⋮	⋮

- Given a compositional sample

  → Colinerity between components
  → Third component elimination
  → Now we can proceed normally

Impossible events occurs in our model!!!

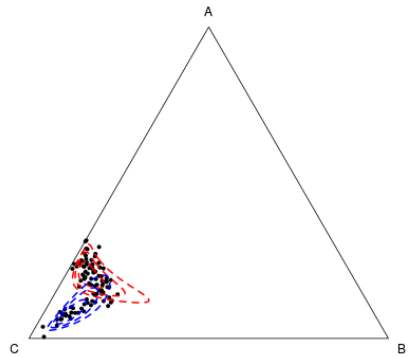
$$P(\{B < 0\}) > 0 \quad !!!!$$

Dealing with compositional data: logratio approach

Aitchison 1986, The Statistical Analysis of Compositional Data

→ Only relative information. Ratios (logratios) between components are informative:

- scale invariance and
- subcompositional coherence

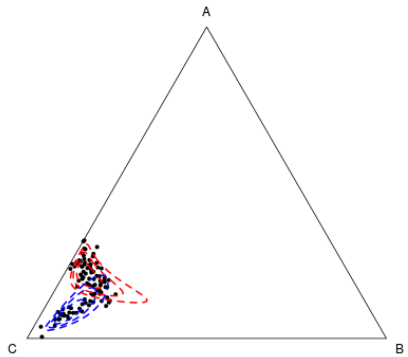


Dealing with compositional data: logratio approach

Aitchison 1986, The Statistical Analysis of Compositional Data

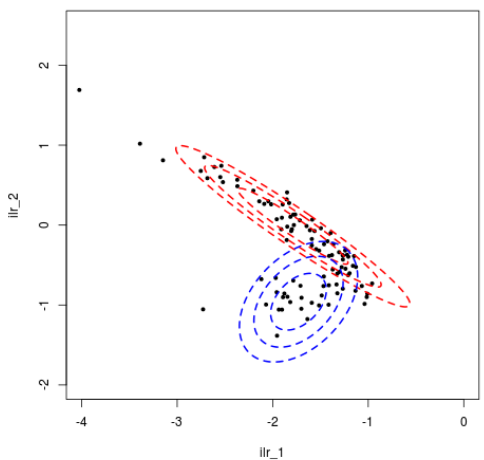
→ Only relative information. Ratios (logratios) between components are informative:

- scale invariance and
- subcompositional coherence



Egozcue et al. 2003, isometric logratio transformations (ilr transformation) for compositional data analysis

→ Isometry between D-part compositional space and $\mathbb{R}^{D-1}$



Using the isometry, the normal on the simplex is easy to define and so the gaussian mixture on the simplex.

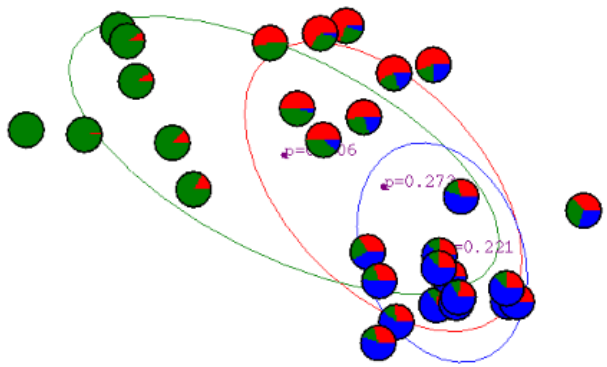
GMM: Posterior probabilities

Posteriori probabilities in a compositional framework

- Assume a GMM fixed for a dataset $X = \{x_1, ..., x_n\}$
- Let $(\tau_{i1}, \ldots, \tau_{iD})$ be the posterior probability of element x_i.

Posteriori probabilities in a compositional framework

- Assume a GMM fixed for a dataset $X = \{x_1, ..., x_n\}$
- Let $(\tau_{i1}, \ldots, \tau_{iD})$ be the posterior probability of element x_i.



Question:

What should we do if we are interested in doing any analysis with posterior probabilities?

Why we are interested in τ?

- Entropy criterias
  - Deciding the number of groups
  - Combine mixtures elements
- Describe an element according to its pertenence into a group
  - As explanatory variable.
  - As explained variable

Logratio of posterior probabilities $\log(\tau_i/\tau_j)$

A relation between τ_ℓ and $x^{(k)}$ depending on μ_ℓ, Σ_ℓ

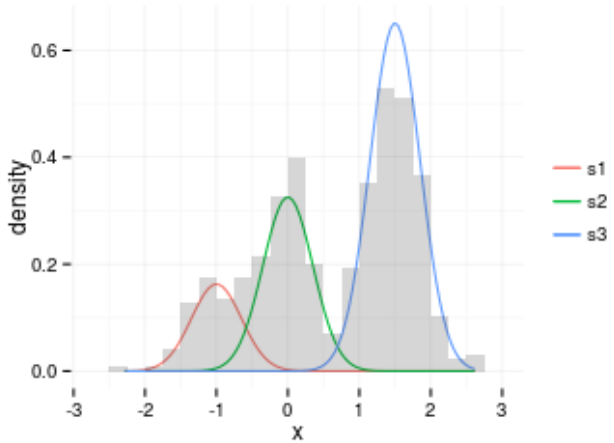
$$\begin{aligned}
\log \frac{\tau_{ki}}{\tau_{kj}} &= \log \frac{\pi_i \phi(x^{(k)};\mu_i,\Sigma_i)}{\pi_j \phi(x^{(k)};\mu_j,\Sigma_j)} = \log \frac{\phi(x^{(k)};\mu_i,\Sigma_i)}{\phi(x^{(k)};\mu_j,\Sigma_j)} + \dots \\
&= \log \frac{a_{\Sigma_i} e^{-\frac{1}{2}(x^{(k)}-\mu_i)^T \Sigma_i^{-1}(x^{(k)}-\mu_i)}}{a_{\Sigma_j} e^{-\frac{1}{2}(x^{(k)}-\mu_j)^T \Sigma_j^{-1}(x^{(k)}-\mu_j)}} + \dots \\
&= \frac{1}{2} {x^{(k)}}^T (\Sigma_j^{-1} - \Sigma_i^{-1}) x^{(k)} + C_{[\Sigma_i,\Sigma_j,\mu_i,\mu_j]} x^{(k)} + \\
&\quad C_{[\Sigma_i,\Sigma_j,\mu_i,\mu_j,\pi_i,\pi_j]}
\end{aligned}$$

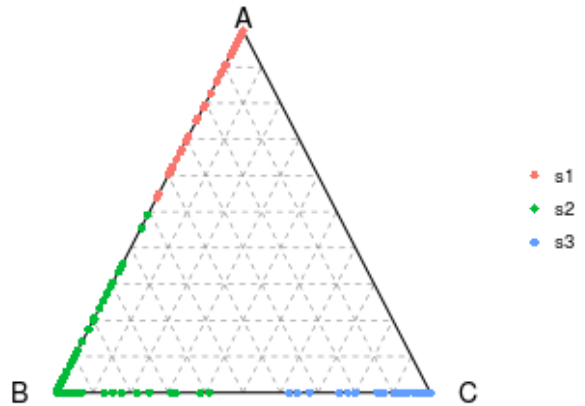
Logratio of posterior probabilities $\log(\tau_i/\tau_j)$

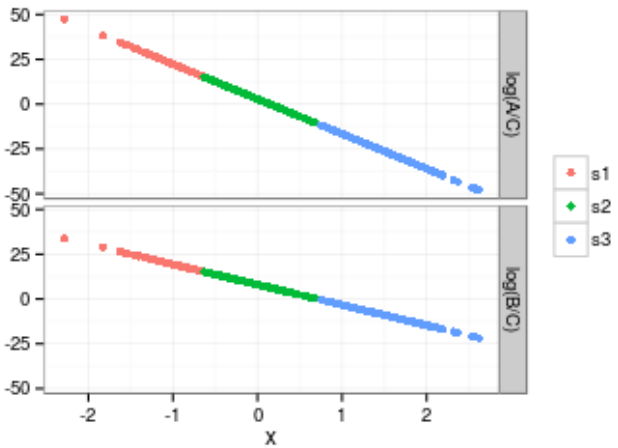
A relation between τ_ℓ and $x^{(k)}$ depending on μ_ℓ, Σ_ℓ

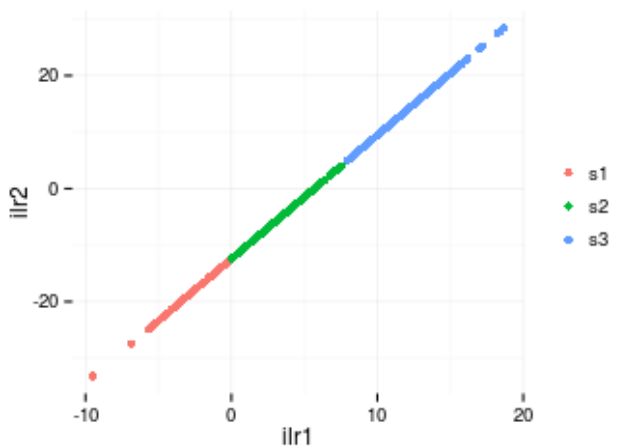
$$
\begin{aligned}
\log \frac{\tau_{ki}}{\tau_{kj}} &= \log \frac{\pi_i \phi(x^{(k)};\mu_i,\Sigma_i)}{\pi_j \phi(x^{(k)};\mu_j,\Sigma_j)} = \log \frac{\phi(x^{(k)};\mu_i,\Sigma_i)}{\phi(x^{(k)};\mu_j,\Sigma_j)} + \dots \\
&= \log \frac{a_{\Sigma_i} e^{-\frac{1}{2}(x^{(k)}-\mu_i)^T \Sigma_i^{-1}(x^{(k)}-\mu_i)}}{a_{\Sigma_j} e^{-\frac{1}{2}(x^{(k)}-\mu_j)^T \Sigma_j^{-1}(x^{(k)}-\mu_j)}} + \dots \\
&= \frac{1}{2} {x^{(k)}}^T (\Sigma_j^{-1} - \Sigma_i^{-1}) x^{(k)} + C_{[\Sigma_i,\Sigma_j,\mu_i,\mu_j]} x^{(k)} + \\
&\quad C_{[\Sigma_i,\Sigma_j,\mu_i,\mu_j,\pi_i,\pi_j]} \\
\rightarrow &\quad \text{If } \Sigma_i = \Sigma_j = \Sigma \\
&= C_{[\Sigma,\mu_i,\mu_j]} x^{(k)} + C_{[\Sigma,\mu_i,\mu_j,\pi_i,\pi_j]}
\end{aligned}
$$

Data dimensionality: equal variances

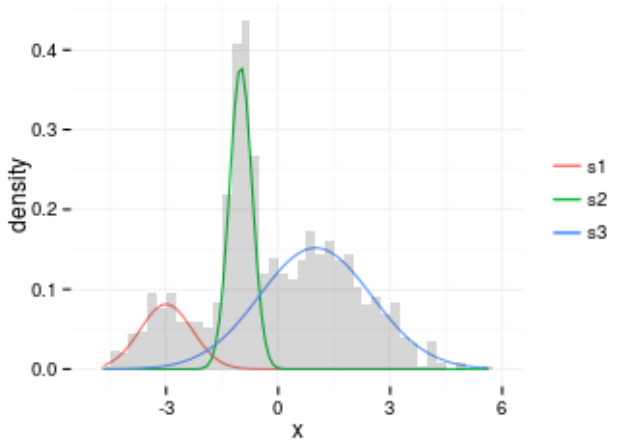


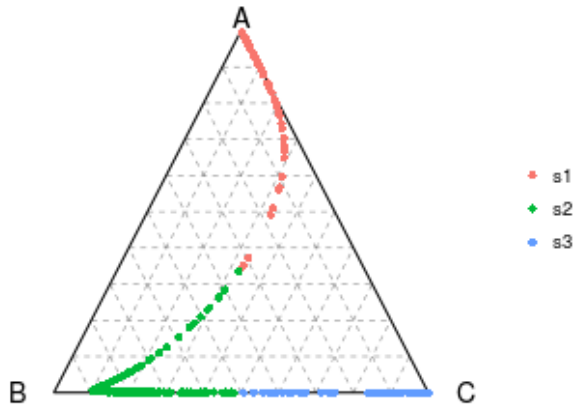


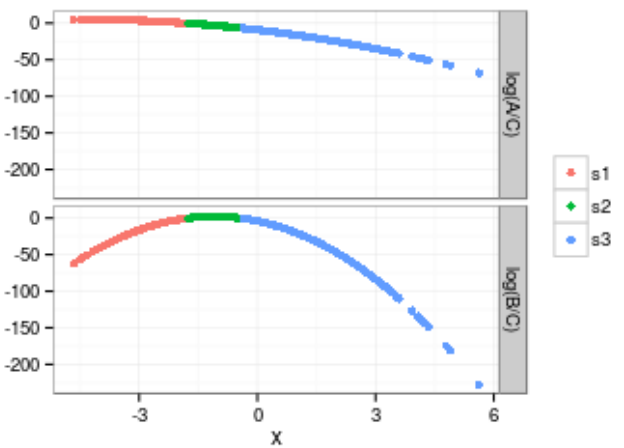


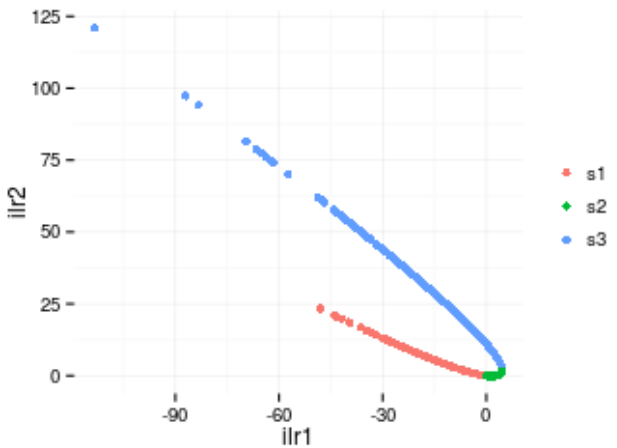


Data dimensionality: different variances

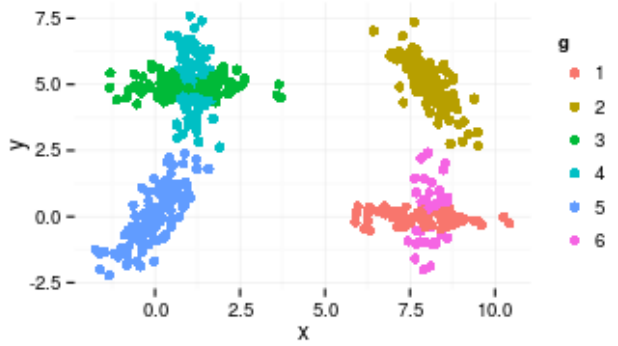


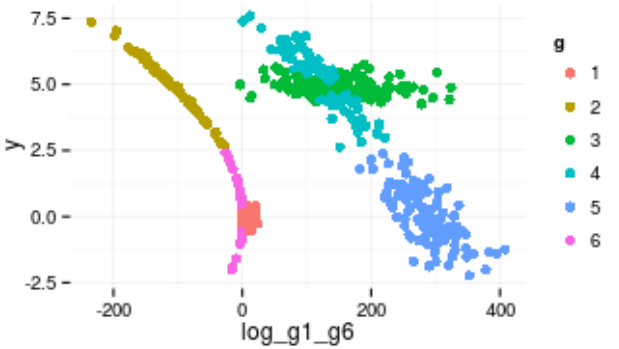


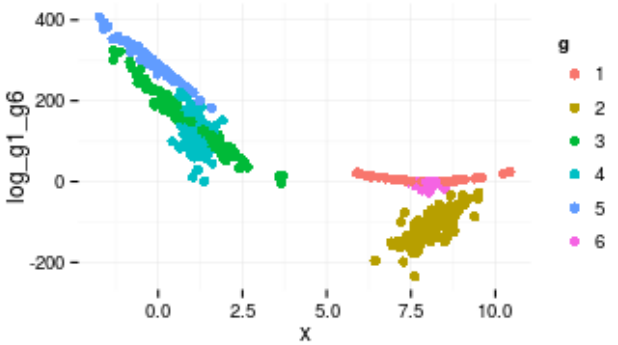




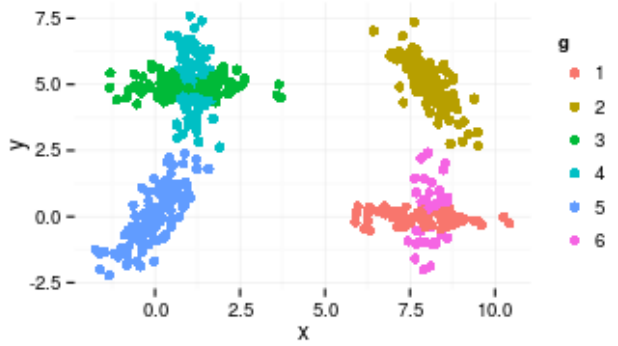
Data dimensionality: different variances

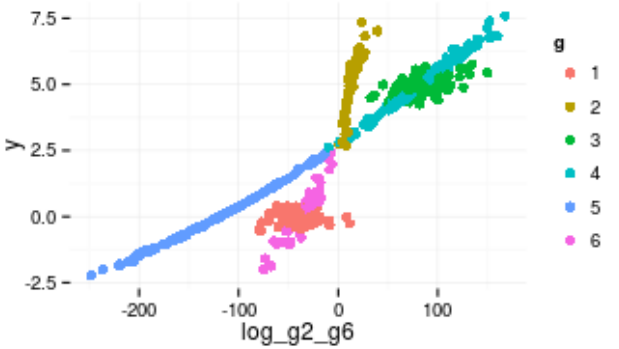


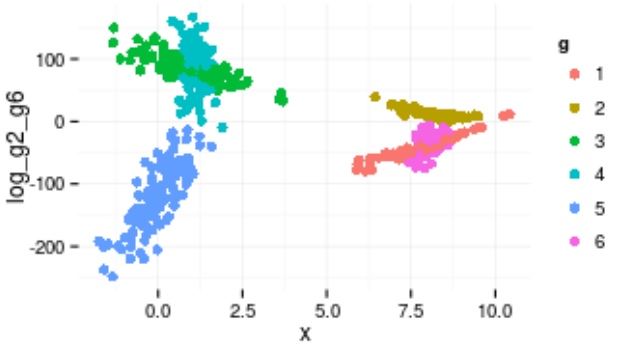




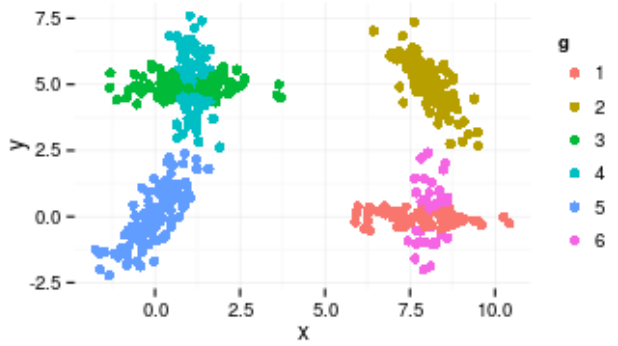
Data dimensionality: different variances

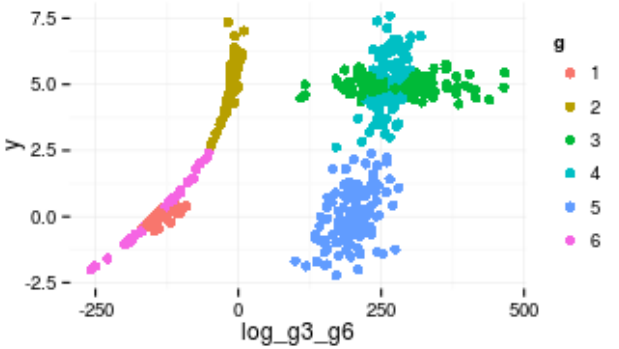


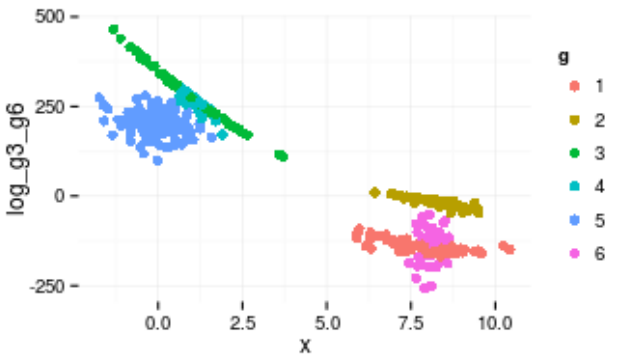




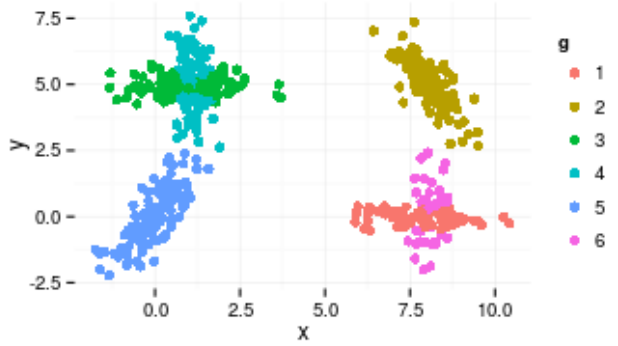
Data dimensionality: different variances

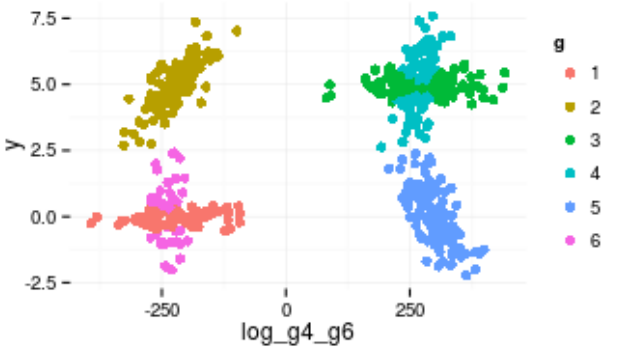


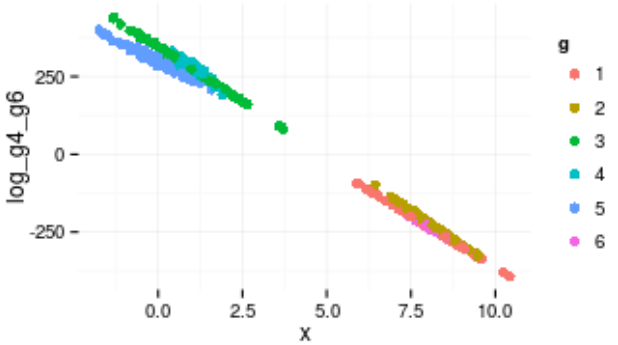




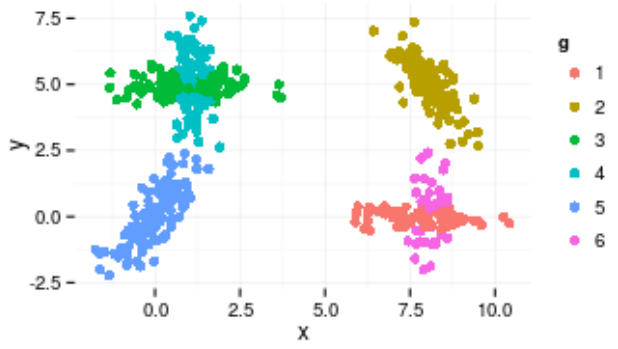
Data dimensionality: different variances

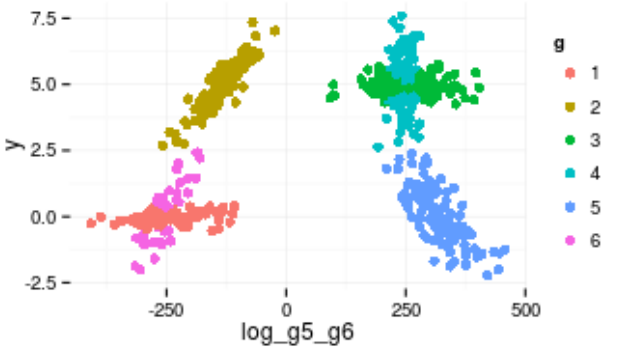


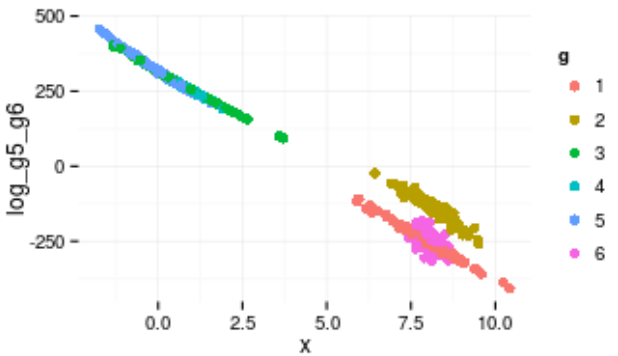




Data dimensionality: different variances

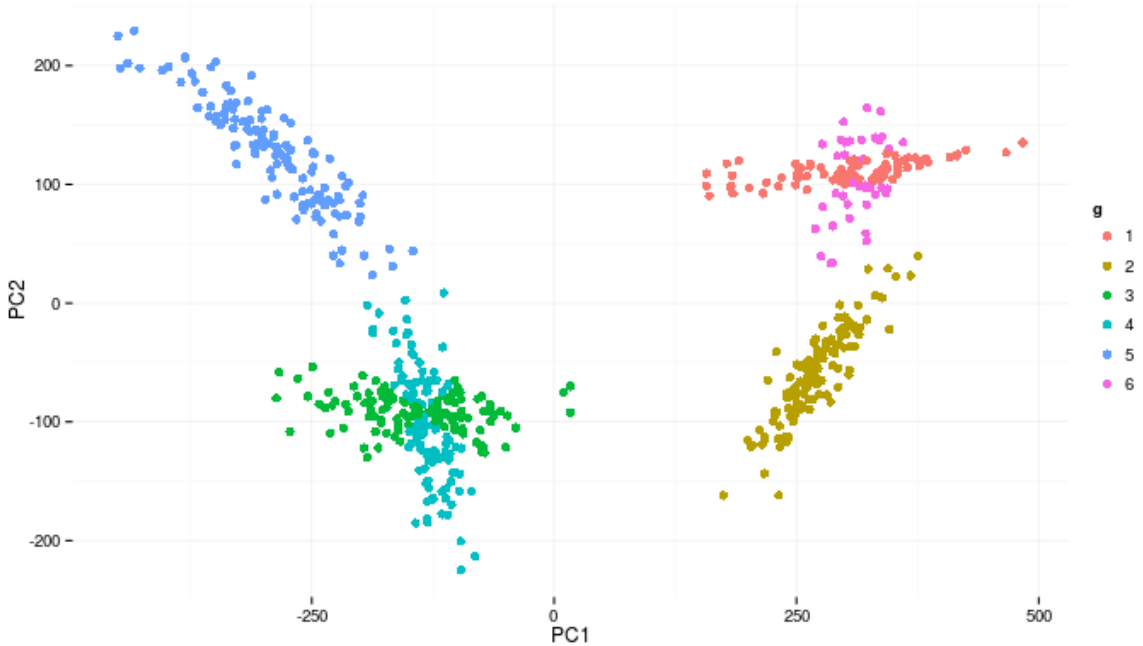






Data dimensionality: different variances

Principal components:



Conclusions

- When data are compositional:
  - Technical problems are expected and misleading conclusions may be drawn when standard data analysis techniques are used,
  - using log-ratio gaussian mixture models the sample space nature is taken into account.
- Dealing with posterioris:
  - Can be considered compositional,
  - GMM posterioris have a linear/quadratic relation with original data.