Clustering transformed compositional data using `coseq`

**Andrea Rau**, Antoine Godichon-Baggioni, and Cathy Maugis-Rabusseau

July 5, 2017 @ UseR! Bruxelles
What’s compositional data?

- $q$-tuple of nonnegative numbers that can be represented in the simplex:

$$S^q := \left\{ \mathbf{p}_i = (p_{i1}, \ldots, p_{iq}) \in \mathbb{R}^q \mid \sum_{j=1}^{q} p_{ij} = 1, p_{ij} > 0, \forall i, j \right\}$$

```r
## [1,] 0.120 0.093 0.129 0.188 0.076 0.167 0.229
## [2,] 0.208 0.178 0.081 0.034 0.152 0.260 0.088
## [3,] 0.019 0.309 0.048 0.309 0.127 0.056 0.132
```

```r
rowSums(some_data)
```

```r
## [1] 1 1 1
```

- **Our goal here**: identify clusters of observations $\mathbf{p}_i$
Example 1: RNA-seq expression profiles

- **RNA-seq** = Application of high-throughput sequencing technology to study gene expression

  ![RNA-seq Diagram](image)

- After pre-processing + alignment to reference + quantification, **count-based measures** of gene expression

- Counts affected by level of transcription + gene length + technical biases (library size, GC content)
Example 1: RNA-seq expression profiles

- Note: some genes may have condition- or developmental-stage specific expression

- We’re interested in *relative* expression throughout an experiment ⇒ use (normalized) gene expression *profiles* to identify co-expression clusters
Example 2: Velib’ bicycles

- Bicycle sharing system in Paris started in 2007
- ~ 14,500 bicycles in 1230 rental stations around the metropolitan area, subscribers check out/return from any rental station

Hourly Velib’ station occupancy data (# of available bicycles and docks) from September 1-5, 2014 ⇒ use profiles of relative occupancy per station to identify station clusters
Why transformations for compositional data?

- Unit sum constraint on $p_i$ often problematic...
- Distance between two compositional vectors typically based on transformations like the Centered Log Ratio:

$$CLR(p_i) = \left( \log \left( \frac{p_{i1}}{g(p_i)} \right), \ldots, \log \left( \frac{p_{iq}}{g(p_i)} \right) \right)$$

where $g(\cdot)$ is the geometric mean
- CLR can be sensitive to small fluctuations close to 0 $\Rightarrow$ we propose log-CLR:

$$LCLR(p_{ij}) = \begin{cases} - \left[ \log (1 - \log [p_{ij}/g(p_i)]) \right]^2 & \text{if } p_{ij}/g(p_i) \leq 1, \\ \left( \log [p_{ij}/g(p_i)] \right)^2 & \text{otherwise}. \end{cases}$$
On to the clustering model

- Cluster transformed profiles with a Gaussian mixture model:
  \[
  f(\tilde{p}_i; K, \theta_K) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \Phi(\tilde{p}_i; \mu_k, \Sigma_k),
  \]

- If variables can be assumed to be independent, a fast alternative to the GMM is $K$-means algorithm (assuming all $\pi_k = \frac{1}{K}$ and $\Sigma_k = \sigma^2 I$)

- Standard EM algorithm for parameter estimation
- Model selection using standard asymptotic criteria like BIC/ICL or non-asymptotic slope heuristics
- Assign observations to clusters using the maximum a posteriori rule
**coseq** for clustering compositional data

```r
source("http://www.bioconductor.org/biocLite.R")
biocLite("coseq")
library(coseq)

## S4 method for signature 'matrix', 'data.frame',
## and 'DESeqDataSet'
coseq(object, K, model = "kmeans",
       transformation = "logclr", ...)
```

Several options:

- Clustering model (kmeans' or Normal) and transformation (logclr, clr, ...)
- Type of normalization for RNA-seq data (TMM, DESeq, ...)
- Parallel computation using BiocParallel
- Output: S4 object of class coseqResults, with corresponding methods (plot, summary, show, clusters, likelihood, ...)
coseq for clustering RNA-seq data

data(fietz)
counts <- exprs(fietz) ## RNA-seq from mice
conds <- pData(fietz) ## 3 neocortex regions x 5 reps

library(DESeq2)
deseq <- DESeqDataSetFromMatrix(counts, conds, ~tissue)
deseq <- DESeq(deseq)

coexp <- coseq(deseq, K=2:15, alpha=0.01)

## Co-expression analysis on DESeq2 output:
## 6461 DE genes at p-adj < 0.01

## coseq analysis: kmeans approach & logclr
## K = 2 to 15

## Running K = 2 ...
Cluster visualization with `coseq`

```r
# plot(coexp)
p <- plot(coexp, conds=conds$tissue, graphs="boxplots", average_over_conds=TRUE)$boxplot
```

![Cluster visualization graph](image-url)
Cluster visualization with `coseq`

```
library(ggplot2)
p + ggtitle("RNA-seq profiles") + theme_bw()
```
Clustering diagnostics with *coseq*

```r
p2 <- plot(coexp, graphs="probapost_boxplots", order=TRUE)
p2$probapost_boxplots
```
**coseq for clustering Velib’ data**

cobike <- coseq(bike_counts, K=2:15, model="kmeans", transformation="clr")

plot(cobike, conds=hour, average_over_conds=TRUE, graphs="boxplots")
Velib’ cluster visualization with coseq

ggmap(Mymap) +
geom_point(data=data.frame(velib$position, clusters(cobike)),
aes(longitude,latitude, Cluster))
Thank you!

- MixStatSeq ANR-JCJC grant
- R Forwards
- [http://bioconductor.org/packages/coseq/](http://bioconductor.org/packages/coseq/)