Outlier Detection with Dirichlet Process Mixtures

Matthew S. Shotwell[1] Elizabeth H. Slate[2]

Vanderbilt University[1]

Medical University of South Carolina[2]

August 3, 2011

Dirichlet Process Mixture (DPM)

$$\begin{aligned} y_i|\theta_i &\sim L(\theta_i; y_i) \qquad i = 1 \,..\, n \\ \theta_i &\sim G \\ G &\sim DP(\alpha, G_0) \end{aligned}$$

- DP is a distribution over distributions
- G is discrete $\Rightarrow P(\theta_j = \theta_k) > 0$
- if $\theta_j = \theta_k$, then y_j and y_k are clustered

VANDERBILT UNIVERSITY

Product Partition Model (PPM)

$$
\begin{aligned}
y_i | z_i = k, \phi_k &\sim L(\phi_k; y_i) & i = 1 \mathrel{..} n \\
\phi_k &\sim G_0(\phi_k) & k = 1 \mathrel{..} r \\
P(\boldsymbol{z}) &\propto \prod_{k=1}^{r} \alpha \Gamma(n_k)
\end{aligned}
$$

- $\boldsymbol{z}$ is the *data partition* parameter
- estimating $\boldsymbol{z}$ 'partitions', or 'clusters' the data
- if $z_j = z_k$, then y_j and y_k are clustered
- [Hartigan, 1990]

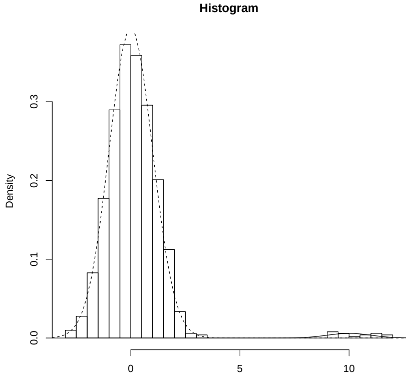
VANDERBILT UNIVERSITY

Outlier Detection Using Partitioning

Steps:

1. set a "small" cluster threshold (*e.g.* 1% of n)
2. estimate the data partition (*i.e.* cluster the data)
3. "small" clusters are considered outlying

- an *outlier partition* contains one or more small *outlier clusters*



Quantifying Evidence to Detect Outliers: Questions

If the data partition (z) is estimated, and outlier clusters are discovered, how much evidence suggests that these clusters are truely different from the others?

Can the partition estimate be restricted such that a minimum level of evidence is required to identify outlier clusters? Yes!

A Criterion for Outlier Detection: Setup

Consider an *outlier partition* $\boldsymbol{z}_o$ $(n = 10)$:

$$\begin{aligned} \boldsymbol{z}_o &= [1,1,1,1,1,1,1,1,2,3] \\ \boldsymbol{z}_{m1} &= [1,1,1,1,1,1,1,1,1,3] \\ \boldsymbol{z}_{m2} &= [1,1,1,1,1,1,1,1,2,1] \\ \boldsymbol{z}_{m3} &= [1,1,1,1,1,1,1,1,2,2] \\ \boldsymbol{z}_{m4} &= [1,1,1,1,1,1,1,1,1,1] \end{aligned}$$

- $\boldsymbol{z}_{m\cdot}$ are formed by merging the outliers in $\boldsymbol{z}_o$.
- outlier detection is a decision between $\boldsymbol{z}_o$ and $\boldsymbol{z}_{m\cdot}$.
- denote the collection $\boldsymbol{z}_{m\cdot}$ as M_o

A Criterion for Outlier Detection: The Trick

$\boldsymbol{z}_o$ is favored if, for all $\boldsymbol{z}_m \in M_o$

$$\begin{aligned} P(\boldsymbol{z}_o|\boldsymbol{y}) &> P(\boldsymbol{z}_m|\boldsymbol{y}) \\ \frac{P(\boldsymbol{y}|\boldsymbol{z}_o)}{P(\boldsymbol{y}|\boldsymbol{z}_m)} &> \frac{P(\boldsymbol{z}_m)}{P(\boldsymbol{z}_o)} \\ BF_{om} &> \frac{P(\boldsymbol{z}_m)}{P(\boldsymbol{z}_o)} \\ BF_{om} &> \frac{1}{\alpha^\nu}\beta_{om} \end{aligned}$$

- ν is the number of clusters merged to arrive at $\boldsymbol{z}_m$
- β_{om} (a ratio involving $\Gamma(\cdot)$) is always ≥ 1 for $\boldsymbol{z}_m \in M_o$
- to favor $\boldsymbol{z}_o$, BF_{om} must exceed $\frac{1}{\alpha^\nu}$
- BF_{om} must increase $1/\alpha$ fold for each outlier

VANDERBILT UNIVERSITY

A Criterion for Outlier Detection: How to Fix α

- set the criteria by fixing α
- use Jeffrey's scale of evidence for Bayes factors
- [Efron and Gous, 2001]

			Evidence for z_o
	$1/\alpha$	< 1	negative
$1 \leq$	$1/\alpha$	< 3	barely worth a mention
$3 \leq$	$1/\alpha$	< 20	positive
$20 \leq$	$1/\alpha$	< 150	strong
$150 \leq$	$1/\alpha$		very strong

VANDERBILT UNIVERSITY

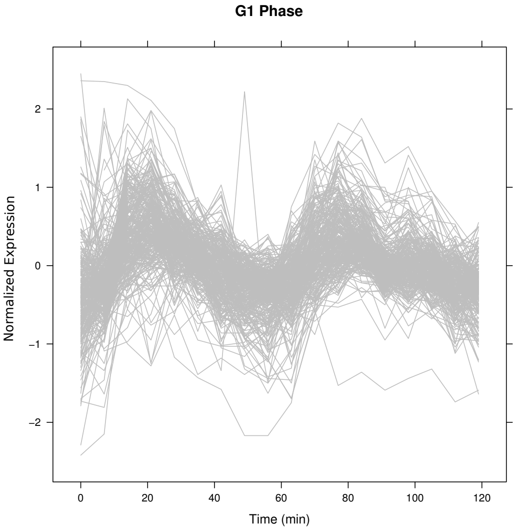
A Criterion for Outlier Detection: Nice Properties

MAP partition estimates automatically satisfy the criterion for fixed α*. Hence, no special or novel computational methods are required.*

Because the DPM accommodates any data likelihood, outlier detection with Dirichlet process mixtures is possible with any statistical model that specifies a likelihood function.

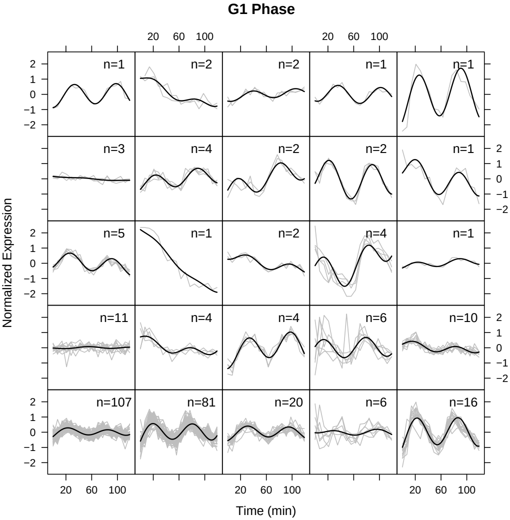
Microarray Time Series in Cell Cycle Synchronized Yeast

- [Spellman et al., 1998]
- 66 minute period, 2 cycles



Microarray Time Series in Cell Cycle Synchronized Yeast

$\alpha = \frac{1}{150}$



VANDERBILT UNIVERSITY

MAP Estimation for z

- Agglomeration [Ward, 1963]
- Polya Urn Gibbs Sampler [MacEachern, 1994]
- Split-Merge Sampler [Jain and Neal, 2004]
- SUGS [Wang and Dunson, 2010]

- sampling is overkill for MAP estimation
- we proposed a stochastic algorithm:
  - consists of 'Explode' and 'Merge' steps
  - consistent for the MAP estimate
  - avoids complexity of sampling
  - facilitates parallel search of partition space
- R package `profdpm`

Outlier Detection with Finite Mixtures

- [Fraley and Raftery, 2002]
- select $\boldsymbol{z}$ that maximizes the BIC
- requires $BF_{om} > n^{\frac{\rho}{2}\nu}$
- *i.e.* BF_{om} must increase $n^{\frac{\rho}{2}}$ fold for each outlier
- DPM outlier detection is generally more conservative.

VANDERBILT UNIVERSITY

Efron, B. and Gous, A. (2001).
Scales of evidence for model selection: Fisher versus jeffreys.
In Lahiri, P., editor, *Model Selection: Lecture Notes–Monograph Series*, volume 38, pages 208–246. Institute of Mathematical Statistics, Beachwood, OH.

Fraley, C. and Raftery, A. E. (2002).
Model-based clustering, descriminant analysis, and density estimation.
Journal of the American Statistical Association, 97:611–631.

Hartigan, J. A. (1990).
Partition models.
Communications in Statistics, Theory and Methods, 19:2745–2756.

Jain, S. and Neal, R. M. (2004).
A split-merge markov chain monte carlo procedure for the dirichlet process mixture model.
Journal of Computational and Graphical Statistics, 13(1):158–182.

MacEachern, S. N. (1994).
Estimating normal means with a conjugate style dirichlet process prior.
Communications in Statistics B, 23:727–741.

Spellman, P. T., Sherlock, G., Zhang, M. Q. Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998).
Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization.
Molecular Biology of the Cell, 9:3273–3297.

Wang, L. and Dunson, D. B. (2010).
Fast bayesian inference in dirichlet process mixture models.
Journal of Computational and Graphical Statistics, In Press.

Ward, J. H. (1963).
Hierarchical grouping to optimize an objective function.
Journal of the American Statistical Association, 58(301):236–244.