A Bayesian Test for Multimodality with Applications to DNA and Economic Data

Nalan Baştürk
joint with
Lennart Hoogerheide, Peter de Knijf, Herman K. van Dijk

Erasmus University Rotterdam, VU University of Amsterdam, Leiden University Medical Center

June 15, 2012

My research mainly focuses on mixture distributions:

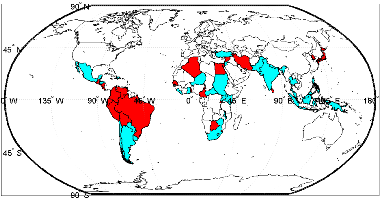
- Model-based clustering to capture heterogeneity in the data
- Mixtures as universal approximators

Focus on Bayesian methods:

- Straightforward to estimate such complex models using Bayesian techniques
- Intuitive to have distributions for parameters in this complex structures rather than assuming fixed parameters
- Effective number of observations can be quite small: refraining from asymptotic theory is important

Modeling economic growth:

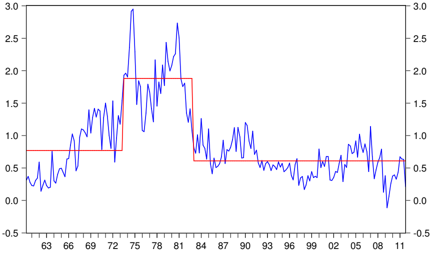
- Heterogeneity across countries, not necessarily explained by conditioning factors
- Different effects of conditioning factors (e.g. investment rate) on economic growth over time
- Changing time-series properties: composition of 'rich' and 'poor' can change over time



(joint work with Richard Paap & Dick van Dijk)

Mixture distributions for accurate inflation forecasting:

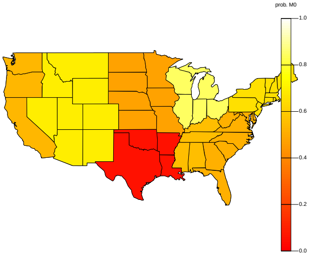
- Standard models for this do not take possible shifts over time into account
- Introducing a ‘switching’ average inflation alters the results substantially



(joint work with Cem Cakmakli, Pinar Ceyhan & Herman van Dijk)

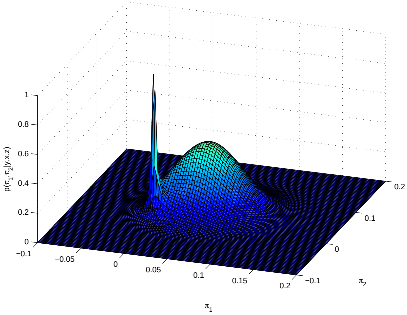
Mixtures in the ‘model space’:

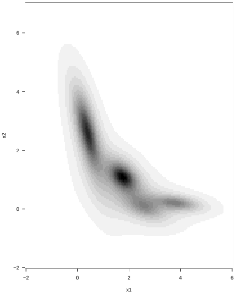
- Averaging over models when choosing one alternative is not straightforward



(joint work with Lennart Hoogerheide & Herman van Dijk)

Mixtures in the ‘parameter space’: Obtaining densities that we can ‘simulate from’





(joint work with Lennart Hoogerheide, Anne Opschoor & Herman van Dijk)

Motivation for this work

Goal:

- Assessing the number of modes in data with non-standard distribution

Details:

- Descriptive analysis (limited theory for modeling these differences)
  This 'descriptive work' on differences can later be used by specialists to find linkages between these differences and (for example) genetic diseases
- Number of 'modes' in the genetic structure is of interest
  (differences in the number of MSR sequences in DNA)
- Large dataset but quite some heterogeneity:
  Subsets of data we can claim to be 'homogenous' are small
- Count data: standard tests relying on continuous data may not be appropriate
  - We can 'treat' this data as a continuous process
  - We can develop appropriate tests for count data
- Bayesian testing method we propose is novel, to the best of our knowledge

A ‘direct’ estimate of the number of modes

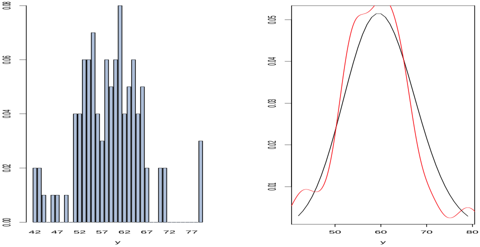
Estimating L modes $y_l \in [\min(y), \max(y)]$, $l = 1, \ldots, L$:

$$\hat{p}(y) = \frac{1}{n}\sum_{i=1}^{n} I[y_i = y] \qquad \text{(pdf estimate)}$$

$$\hat{p}(y_l) > \hat{p}(y_l - 1),\ \ \hat{p}(y_l) < \hat{p}(y_l^\star) \qquad \text{(mode definition)}$$

$$y_l^\star = \min_{y_i; y_i > y_l} \hat{p}(y_i) \neq \hat{p}(y_l)$$

Unimodal ‘true’ dist.
Multiple modes in $\hat{p}(y)$
(Izenman & Sommer, 1988;
Hall & York, 2001)



$n = 100$, $\lambda = 60$

Silverman test (Silverman, 1981)

- Applicable to continuous data
- Tests hypothesis ‘a single mode’ versus ‘at least two modes’ in the data
- Relies on Gaussian kernel estimates with window size h:

$$\hat{f}(y; h) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \phi\left(\frac{y - y_i}{h}\right)$$

  with $h \in (0, \infty)$, $\phi()$ is the std. normal density function.
- Estimated number of modes decreases with h

Bootstrap test:

1. Approximate $\hat{f}(x, h^\star)$ with minimum $h^\star$ leading to a unimodal density
2. Simulate $x^{(m)}$ from $\hat{f}(x; h^\star)$ for $m = 1, \ldots, M$ (inverse CDF technique), count number of modes $L^{(m)}$ in $x^{(m)}$ using $f(x^{(m)}; h^\star)$
3. Calculate p-value (Efron & Tibshirani, 1994)

$$\text{p-value} = \frac{1}{M} \sum_{m=1}^{M} I(L^{(m)} > 1)$$

Other related work

Testing for ‘multimodality’

- DIP test (Hartigan and Hartigan, 1985),
  ‘one mode’ versus ‘at least two modes’
  applicable to continuous data

Tests for number of mixtures in count data (mixtures of Poisson)

- Hellinger distance estimator (Karlis & Xekalaki, 1998)
- Woo & Sriram (2006), Umashanger & Sriram, 2009

Main idea of this work

- Approximating the distribution of count data using a ‘flexible’ mixture distribution
  - A finite/infinite number of mixtures to be used to approximate the distribution
  - Distributions for each mixture components should be suitable for count data, such as the Poisson distribution or negative binomial distributions can be used
- Defining the number of modes as a random variable
  - Straightforward in Bayesian context
  - From the estimated posterior distribution, we can retrieve the posterior distribution for the number of modes
- Mixture of shifted Poisson distributions
  - applicable for modeling ‘non-standard’, possibly multimodal data distribution
  - ‘shifted’ distributions overcome the ‘overdispersion/underdispersion’ problem

Finite mixture of 'shifted' Poisson distributions

y_i for $i = 1, \dots, n$ are independent realizations from a mixture of J shifted Poisson distributions:

$$y_i - \kappa_j \sim \text{Poisson}(\lambda_j) \text{ if } z_{ij} = 1 \text{ for } i = 1, \dots, n; j = 1, \dots, J,$$

where $z_{ij} = 1$ if y_i belongs to cluster j, and 0 otherwise.
Latent variable distribution:

$$\Pr[z_{ij} = 1] = \pi_j, \text{ for } i = 1, \dots, n; j = 1, \dots, J,$$

with $\pi_j > 0$ for $j = 1, \dots, J$ and $\sum_{j=1}^{J} \pi_j = 1$.
The (augmented) likelihood:

$$\ell\left(y, z|\theta\right) = \begin{cases} \prod_{i=1}^{n} \prod_{j=1}^{J} \left[\exp(-\lambda_j) \frac{\lambda_j^{y_i - \kappa_j}}{(y_i - \kappa_j)!} \right]^{z_{ij}} \pi_j^{z_{ij}}, & \text{if } y_i \geq \kappa_j \ \forall i, j \text{ with } z_{ij} = 1 \\ 0, & \text{otherwise} \end{cases}.$$

where $y = (y_1, \dots, y_n)'$, $z_i = (z_{i1}, \dots, z_{iJ})'$, $z = \{z_i, \dots, z_n\}$, $\pi = (\pi_1, \dots, \pi_J)$ and $\theta = \{\lambda, \kappa, \pi\}$.

Prior specifications

Uninformative but proper priors:

$$\begin{aligned}\lambda_j &\sim \text{unif}(\lambda_{\min}, \lambda_{\max})\\ \kappa_j &\sim \text{unif}(\kappa_{\min}, \kappa_{\max})\\ (\pi_1, \ldots, \pi_J) &\sim \text{Dirichlet}(1, \ldots, 1)\\ [\lambda_{\min}, \lambda_{\max}] &= [\kappa_{\min}, \kappa_{\max}] = [0, \max(y_i | y_i = 1, \ldots, n)]\end{aligned}$$

Possible label switching constraints:

$$\begin{aligned}\kappa_l &< \kappa_j, \text{ for } l < j\\ \kappa_l + \lambda_l &< \kappa_j + \lambda_k, \text{ for } l < j\\ \pi_l &< \pi_j, \text{ for } l < j\end{aligned}$$

(label switching is not an issue for estimating the number of modes)

Gibbs sampling scheme & the number of mixture components

For $j = 1, \dots, J$, under the condition that $y_i \geq \kappa_j \ \forall i, j$ with $z_{ij} = 1$

$$p\left(\kappa_j | y, z, \theta_{-\kappa_j}\right) \propto \frac{\lambda_j^{\sum_{i|z_{ij=1}} y_i - n_j \kappa_j}}{\prod_{i|z_{ij=1}} (y_i - \kappa_j)!}$$

$$p\left(\lambda_j | y, z, \theta_{-\lambda_j}\right) \propto \mathrm{Gamma}_{[\lambda_{\min}, \lambda_{\max}]}\left(\frac{1}{n_j}, 1 + \sum_{i|z_{ij=1}} (y_i - \kappa_j)\right)$$

$$p\left(\pi | y, z, \theta_{-\pi}\right) \propto \mathrm{Dirichlet}\left(n_1 - 1, \dots, n_J - 1\right),$$

where $n_j = \sum_{i=1}^{n} z_{ij}$ is the number of observations in component j and κ_j is an integer in $[max\{\kappa_{\text{min}}, \min_{i|z_{ij}=1}(y_i)\}, \kappa_{\max}]$.

Assessing the number of mixture components:

- AIC and BIC criteria for the number of mixtures (possible straightforward extensions)

Posterior distribution of the number of modes

Each posterior draw, $m = 1, \ldots, M$ leads to a posterior density:

$$p(\tilde{y}|\lambda^{(m)}, \kappa^{(m)}, \pi^{(m)}) = \sum_{j=1}^{J} \mathrm{pdf}_{\mathrm{Poisson}(\lambda_j^{(m)})}\left(\tilde{y} - \kappa_j^{(m)}\right).$$

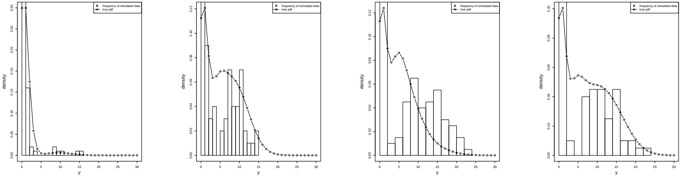
Calculation of posterior modes for integers $y = \{\tilde{y}_1, \ldots, \tilde{y}_L\}$ on the range $[\min(y), \max(y)]$.
Modes $\hat{y}_{1^{(m)}}, \ldots, \hat{y}_{\hat{J}^{(m)}}$ satisfy:

$$p(\tilde{y}_{j^{(m)}}) > p(\tilde{y}_{j^{(m)}} - 1)$$
$$p(\tilde{y}_{j^{(m)}}) < p(\tilde{y}_{t^\star})$$

where $t^\star = \min_{t;t>j^{(m)}}(p(\tilde{y}_{j^{(m)}}) \neq p(\tilde{y}_t))$, $j = 1, \ldots, \hat{J}$.

Simulated data experiments

- Simulation study follows examples in Umashanger & Sriram, 2009.
- Different number of modes and number of Poisson mixture components and Poisson parameters
- $n = 100$ observations in each sample
- Estimates of number of modes only (known number of mixtures)



# mixtures	2	3	4	4
# L (modes)	2	2	2	2
# $p(\hat{L} = L)$ (post. mean)	0.98	1.00	1.00	1.00

MSR sequences

- 270 unrelated human DNA samples from Asian, African and Caucasian origin:
  - Yoruba individuals from Ibadan, Nigeria (African),
  - Han Chinese individuals from Beijing, China (CHB), Japanese individuals from Tokyo, Japan (JPT),
  - Utah residents with ancestry from Northern and Western Europe (Caucasian)
- Effort to eliminate 'selection problems': Subjects in the sample are not from the same family

MSR	**Primer sequences**	**P. size**	**Location**	**Washing conditions**
RS447	F: ATCCAGGCAGCTCAGAGTGT			
	R: GCTCTTTCCACCAAGTGCTC	604	internal	2x 0.3xSSC, 0.1% SDS 1x 0.1xSSC, 0.1% SDS
MSR5p	F: CGATCTGCTGTCTTCATCCA			
	R: GGAAGGTGAGCTCAGGAGTG	644	distal	1x 0.3xSSC, 0.1% SDS 2x 0.1xSSC, 0.1% SDS
FLJ40296	F: TTTGGATGCTTTCCTTGACC			
	R: GCAGGCGTTTGATGTACCTT	749	internal	2x 2xSSC, 0.1% SDS 1x 1xSSC, 0.1% SDS
RNU2	F: TAAGGGCTAGGAAGGGGGTA			
	R: AATGCCAATGACAACGATGA	650	distal	3x 2xSSC, 0.1% SDS
DXZ4	F: ACTAGCCTGCCTTCCTGACA			
	R: CCAGTAGAAGTGGGCGAGAG	940	internal	1x 2xSSC, 0.1% SDS 2x 1xSSC, 0.1% SDS
CT47	F: CTGCTGCTTGATCATTTCCA			
	R: AGAGGGTAAGGAACGGGCTA	710	internal	1x 2xSSC, 0.1% SDS 2x 1xSSC, 0.1% SDS

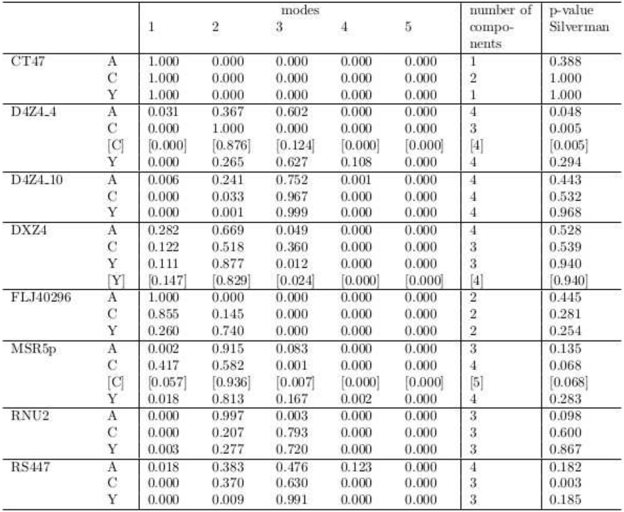
Number of mixture components for DNA data

BIC (AIC) based number of mixture components:

	Asian	Caucasian	African
CT47	1	2	1
D4Z4 4	4	3 (4)	4
D4Z4 10	4	4	4
DXZ4	4	3	3 (4)
FLJ40296	2	2	2
MSR5p	3	4 (5)	4
RNU2	3	3	3
RS447	4	3	3

- In case of different results, estimates are based on BIC
- This is still a ‘rough’ comparison, natural extensions such as a Dirichlet Process prior are to be done
- The number of mixtures is not the main purpose, we rather try to find a good approximation to the empirical distribution

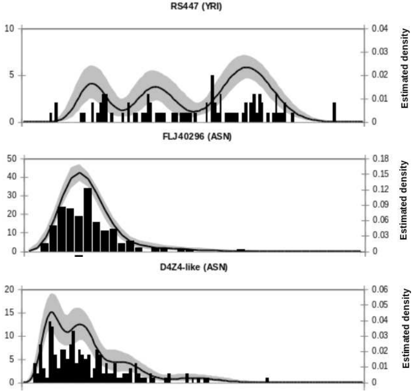
Estimated posterior probabilities of number of modes



		modes					number of components	p-value Silverman
		1	2	3	4	5		
CT47	A	1.000	0.000	0.000	0.000	0.000	1	0.388
	C	1.000	0.000	0.000	0.000	0.000	2	1.000
	Y	1.000	0.000	0.000	0.000	0.000	1	1.000
D4Z4_4	A	0.031	0.367	0.602	0.000	0.000	4	0.048
	C	0.000	1.000	0.000	0.000	0.000	3	0.005
	[C]	[0.000]	[0.876]	[0.124]	[0.000]	[0.000]	[4]	[0.005]
	Y	0.000	0.265	0.627	0.108	0.000	4	0.294
D4Z4_10	A	0.006	0.241	0.752	0.001	0.000	4	0.443
	C	0.000	0.033	0.967	0.000	0.000	4	0.532
	Y	0.000	0.001	0.999	0.000	0.000	4	0.968
DXZ4	A	0.282	0.669	0.049	0.000	0.000	4	0.528
	C	0.122	0.518	0.360	0.000	0.000	3	0.539
	Y	0.111	0.877	0.012	0.000	0.000	3	0.940
	[Y]	[0.147]	[0.829]	[0.024]	[0.000]	[0.000]	[4]	[0.940]
FLJ40296	A	1.000	0.000	0.000	0.000	0.000	2	0.445
	C	0.855	0.145	0.000	0.000	0.000	2	0.281
	Y	0.260	0.740	0.000	0.000	0.000	2	0.254
MSR5p	A	0.002	0.915	0.083	0.000	0.000	3	0.135
	C	0.417	0.582	0.001	0.000	0.000	4	0.068
	[C]	[0.057]	[0.936]	[0.007]	[0.000]	[0.000]	[5]	[0.068]
	Y	0.018	0.813	0.167	0.002	0.000	4	0.283
RNU2	A	0.000	0.997	0.003	0.000	0.000	3	0.098
	C	0.000	0.207	0.793	0.000	0.000	3	0.600
	Y	0.003	0.277	0.720	0.000	0.000	3	0.867
RS447	A	0.018	0.383	0.476	0.123	0.000	4	0.182
	C	0.000	0.370	0.630	0.000	0.000	3	0.003
	Y	0.000	0.009	0.991	0.000	0.000	3	0.185

(A: Asian, C: Caucasian, Y: African)

Estimated empirical distributions for DNA data



- Estimated density and 95% interval
- Interval estimates for posterior modes can also be extracted

Conclusion and future work

Summary:

- We propose a method to asses number of modes in count data using a flexible distribution that could a priori take several shapes
- The proposed tests is more appropriate for the analysis of count data compared to the alternative test
- The proposed method is in particular of interest for DNA analysis, explaining the differences in number of modes across gene compositions and populations

Future work:

- Simulated data experiments in order to assess the proposed test's performance
- Comparison with other tests to detect multimodality
- Applications in economic data, such as income distribution data