# Mixtures of equispaced Normal distributions and their use for testing symmetry in univariate data

Silvia Bacci*[1], Francesco Bartolucci*

* Dipartimento di Economia, Finanza e Statistica - Università di Perugia

University of Naples "Federico II", Naples, 17-19 May 2012

[1] silvia.bacci@stat.unipg.it

# Outline

# Starting point

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution $F(x)$ with density $f(x)$
- Let $\mu$ be the mean or the median of $f(.)$
- Problem of testing symmetry:

$$H_0: \quad F(\mu - x) = 1 - F(\mu + x) \quad \forall x$$

against (hypothesis of skewness)

$$H_1: \quad F(\mu - x) \neq 1 - F(\mu + x) \quad \text{for at least one } x$$

- Aim: to propose a test of symmetry based on Normal finite mixture (NM) models (Lindsay, 1996; McLachlan and Peel, 2000)

# Why testing symmetry?

- many parametric statistical methods are robust to the violation of the normality assumption of $f(x)$, being the symmetry often sufficient for their validity
- knowledge about the symmetry of $f(x)$ is relevant to choose which location parameter is more representative of the distribution, being mean, median, and mode not coincident in case of skewness
- in case-control studies the exchangeability is required for the joint distribution of observations of treated and controlled individuals: as exchangeability implies the symmetry of the distribution, knowing that a distribution is skewed allows to exclude its exchangeability
- nonparametric methods assume the symmetry of the distribution rather than its normality

# How testing symmetry?

- Traditional test based on the third sample standardised moment (Gupta, 1967)

$$b_1 = \frac{m_3}{m_2^{3/2}},$$

where $m_r = 1/n \sum_{i=1}^{n} (x_i - \bar{x})^r$, $r = 2, 3$

  - $b_1$ is commonly used to estimate the third standardised population moment

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}, \qquad \mu_r = E[(X - \mu)^r]$$

  - for samples from a symmetric distribution with finite sixth order central moment,

$$b_1 \to \mathcal{N}(0, \sigma^2), \qquad \sigma^2 = \frac{\mu_6 - 6\mu_2\mu_4 + 9\mu_2^3}{n\mu_2^3}$$

  - $\sigma^2$ is consistently estimated by substituting $\mu_j, j = 2, 4, 6$, with the appropriate sample moments
  - under $H_0$,

$$S_1 = \frac{n^{1/2} b_1}{\hat{\sigma}} \to \mathcal{N}(0, 1)$$

- Drawbacks of Gupta's test
  - $\gamma_1$ is sensitive to outliers
  - $\gamma_1$ can be undefined for heavy-tailed distributions (e.g., Chauchy)
  - $\gamma_1 = 0$ not necessarily means that $f(x)$ is symmetric
- Other tests based on alternative measures of skewness
  - Randles et al. (1980) for a triples test
  - McWilliams (1990), Modarres and Gastwirth (1996) for a runs test
  - Cabilio and Masaro (1996), Miao et al. (2006) for a test based on the Yule's skewness index
  - Mira (1999) for a test based on the Bonferroni's index
- Non-parametric tests based on the kernel estimation method
  - Fan and Gencay (1995), Ngatchou-Wandji (2006), Racine and Maasoumi (2007)
  - pros: a better goodness of fit is allowed with respect to parametric methods
  - cons: high number of unknown parameters

# Our proposal

We know that:

- NM densities (with common variance) allow to approximate arbitrarily well any continuous (symmetric or skewed) distribution
- NM densities provide a convenient semi-parametric framework in which to model unknown distributions, by keeping
  - a parsimony close to that of full parametric methods as represented by a single density
  - the flexibility of nonparametric methods as represented by the kernel method

Therefore, we propose the use of NM densities for testing symmetry about an unknown value

# The NM model

- Density of a mixture of $k$ normal components (NM$_k$)

$$f(x) = \sum_{j=1}^{k} \pi_j \phi(x; \nu_j, \sigma^2),$$

  - $\pi_j$ $(j = 1, \ldots, k)$ denotes the weight of the $j$-th component
  - $\nu_j = \alpha + \beta\delta_j$ $(j = 1, \ldots, k)$ denotes the support points of the mixture
  - $\alpha$ is the centre of symmetry
  - $\beta$ is a scale parameter
  - $\delta_1, \ldots, \delta_k$ is a grid of equispaced points between $-1$ and $1$

# Maximum likelihood estimation

- Log-likelihood of $\text{NM}_k$

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{j=1}^{k} \pi_j \phi(x_i; \nu_j, \sigma^2)$$

  - $\boldsymbol{\theta} = (\alpha, \beta, \pi_1, \ldots, \pi_k)$
  - $\ell(\boldsymbol{\theta})$ is maximised through an EM algorithm (Dempster et al., 1977)

- complete data log-likelihood

$$\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \log \phi(x_i; \nu_j, \sigma^2) + \sum_{j} z_{\cdot j} \log \pi_j$$

  - $z_{ij}$ is a dummy variable equal to $1$ if the $i$-th observation belongs to the $j$-th component and to $0$ otherwise
  - $z_{\cdot j} = \sum_i z_{ij}$

# EM algorithm

- Step E: compute the expected value of $z_{ij}$, $i = 1, \ldots, n$ and $j = 1, \ldots, k$, given the observed data $\boldsymbol{x} = (x_1, \ldots, x_n)$ and the current value of the parameters $\boldsymbol{\theta}$

$$\hat{z}_{ij} = \frac{\phi(x_i; \nu_j, \sigma^2)\pi_j}{\sum_h \phi(x_i; \nu_h, \sigma^2)\pi_h}$$

- Step M: maximise $\ell_c(\boldsymbol{\theta})$ with any $z_{ij}$ substituted by $\hat{z}_{ij}$. The solution is reached when:

$$\begin{aligned}
\beta &= \frac{\sum_i \sum_j z_{ij}(x_i - \bar{x})\delta_j}{\sum_j z_{\cdot j}(\delta_j - \bar{\delta})\delta_j}; \quad \bar{x} = \sum_i x_i/n; \quad \bar{\delta} = \sum_j z_{\cdot j}\delta_j/k \\
\alpha &= \bar{x} - \beta\bar{\delta} \\
\sigma^2 &= \sum_i \sum_j z_{ij}[x_i - (\alpha + \beta\delta_j)]^2/n \\
\hat{\pi}_j &= \frac{\hat{z}_{\cdot j}}{n} \quad j = 1, \ldots, k
\end{aligned}$$

# Selection of $k$

A crucial point with NM models concerns the choice of the number $k$ of mixture components

- coherently with the main literature we suggest to use AIC and BIC indices
  - note that AIC tends to overestimate the true number of components
- we select $k$ as an odd number
  - in this way there is one mixture component, the $[(k+1)/2]$-th, which corresponds to the centre of the distribution and its mean directly corresponds to the parameter $\alpha$

# Proposed test of symmetry

- in a symmetric density the components specular with respect to the centre of symmetry are represented in equal proportions, whereas in a skewed density they are mixed in different proportions
- therefore, if the sample observations come from a symmetric distribution, then the weights of mixture components equidistant from the centre of symmetry are equal, being different otherwise
- the hypothesis of symmetry may be formulated as

$$H_0 : \pi_j = \pi_{k-j+1}, \quad j = 1, \ldots, [k/2],$$

where $[z]$ is the largest integer less or equal than $z$ and $k$ is fixed

- the $\text{NM}_k$ model with constrained $\pi_j$ (i.e., under $H_0$) is nested in the $\text{NM}_k$ model with unconstrained $\pi_j$
- for testing symmetry we may use a likelihood ratio test, based on the deviance

$$\text{LR} = 2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_0)]$$

  - $\hat{\boldsymbol{\theta}}$ is the unconstrained maximum likelihood estimator of $\boldsymbol{\theta}$
  - $\hat{\boldsymbol{\theta}}_0$ is the maximum likelihood estimator under the constraint $H_0$
- under $H_0$, LR is asymptotically distributed as a Chi-square with a number of degrees of freedom equal to $[k/2]$ (the number of constrained weights)
- when $k = 1$ the NM degenerates to a single normal distribution and, therefore, the null hypothesis of symmetry results automatically accepted
- $k$ depends both on the number of groups characterising the population and on the level of skewness: therefore, there is not a one-to-one correspondence between the mixture components and the groups

# Monte Carlo study

- We compare
  - the NM-based test with $k$ selected through AIC
  - the NM-based test with $k$ selected through BIC
  - traditional test of Gupta (1967)
- 1000 samples with a given size $n$ and coming from a given density $f(x)$
- $n = 20, 50, 100$
- $f(x)$: $\mathcal{N}(0,1)$, $t_5$, Laplace ($Lap$), symmetric NM$_3$, $\chi^2_1$, $\chi^2_5$, $\chi^2_{10}$, standard log-normal ($logN$)
- nominal level $\alpha = 0.01, 0.05, 0.10$
- all analyses are implemented in R software

# Empirical significance levels from symmetric distributions

| | $n$ | $N(0,1)$ | $t_5$ | $Lap$ | $NM_3$ |
|---|---|---|---|---|---|
| | | $\alpha = 0.05$ | | | |
| Mixture test (AIC) | 20 | 0.059 | 0.061 | 0.069 | 0.093 |
| | 50 | 0.069 | 0.076 | 0.075 | 0.079 |
| | 100 | 0.078 | 0.083 | 0.096 | 0.060 |
| Mixture test (BIC) | 20 | 0.019 | 0.012 | 0.030 | 0.062 |
| | 50 | 0.010 | 0.014 | 0.031 | 0.058 |
| | 100 | 0.005 | 0.027 | 0.047 | 0.048 |
| Gupta's Test | 20 | 0.038 | 0.030 | 0.044 | 0.037 |
| | 50 | 0.038 | 0.029 | 0.035 | 0.045 |
| | 100 | 0.043 | 0.032 | 0.037 | 0.045 |

- the mixture-based test shows a performance very similar to that of Gupta's test when the number $k$ of components is selected by means of BIC
- when AIC is used for the model selection, an empirical level is observed constantly higher than the nominal one (the type-I error is committed too often)

# Empirical power levels from skewed distributions

| | $n$ | $\chi^2_1$ | $\chi^2_5$ | $\chi^2_{10}$ | $logN$ |
|---|---|---|---|---|---|
| | | $\alpha = 0.05$ | | | |
| Mixture test (AIC) | 20 | 0.566 | 0.229 | 0.140 | 0.421 |
| | 50 | 0.868 | 0.700 | 0.457 | 0.712 |
| | 100 | 0.984 | 0.949 | 0.787 | 0.878 |
| Mixture test (BIC) | 20 | 0.422 | 0.115 | 0.059 | 0.305 |
| | 50 | 0.825 | 0.335 | 0.147 | 0.649 |
| | 100 | 0.968 | 0.690 | 0.326 | 0.834 |
| Gupta's Test | 20 | 0.359 | 0.153 | 0.089 | 0.272 |
| | 50 | 0.496 | 0.541 | 0.373 | 0.341 |
| | 100 | 0.661 | 0.798 | 0.713 | 0.423 |

- the tendency of the AIC method to choose a relatively high number of mixture components results in an empirical power better than that obtained with the variant using the BIC method and the Gupta's test
- also the variant of mixture-based test using BIC is almost always more powerful than Gupta's test
- for all the three types of test, as the sample size increases, the empirical significance level remains constant and the empirical power increases

# Empirical example

- $n = 10$ observations about the process of tomato roots initiation
- number of mixture components selection:

| | $H_0$ false | | | | $H_0$ true | | | |
|---|---|---|---|---|---|---|---|---|
| $k$ | # par | $\hat{\ell}$ | AIC | BIC | # par | $\hat{\ell}$ | AIC | BIC |
| 1 | 2 | -47.58 | 99.17 | 102.54 | 2 | -47.58 | 99.17 | 102.54 |
| 3 | 5 | -40.55 | 91.11 | **99.55** | 4 | -43.39 | 94.79 | 101.55 |
| 5 | 7 | -37.65 | **89.29** | 101.11 | 5 | -42.56 | 95.12 | 103.56 |
| 7 | 9 | -37.85 | 93.70 | 108.90 | 6 | -42.76 | 97.51 | 107.65 |

- we perform the deviance test on the basis of models $NM_3$ and $NM_5$:

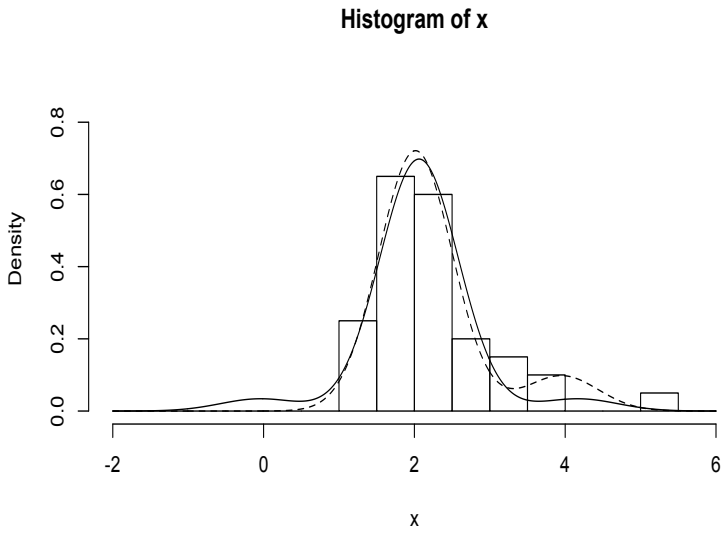
| | $k = 3$ | $k = 5$ |
|---|---|---|
| deviance | 5.68 | 9.82 |
| df | 1 | 2 |
| $p$-value | 0.0172 | 0.0074 |

In both cases the hypothesis of symmetry is rejected

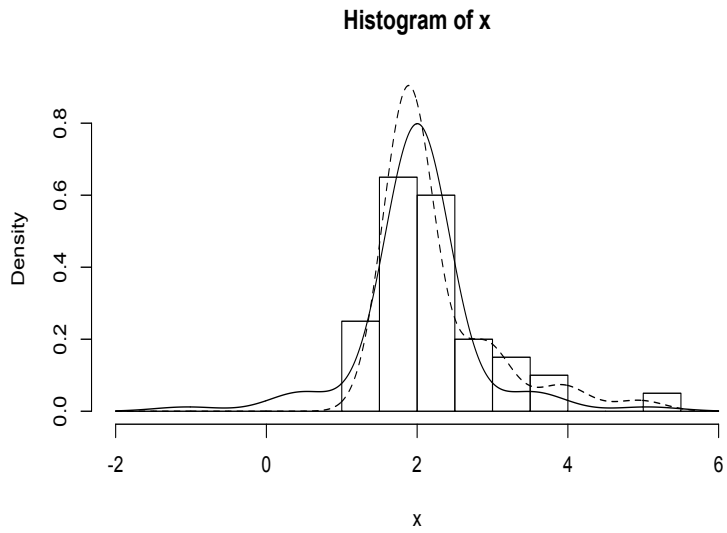
- weights estimates:

| | $k = 3$ | $k = 5$ |
|---|---|---|
| $\hat{\pi}_1$ | 0.0000 | 0.0000 |
| $\hat{\pi}_2$ | 0.8804 | 0.7569 |
| $\hat{\pi}_3$ | 0.1196 | 0.1578 |
| $\hat{\pi}_4$ | – | 0.0602 |
| $\hat{\pi}_5$ | – | 0.0251 |

- Gupta's test does not reject the hypothesis of symmetry ($S_1 = 1.782$, $p$ = 0.0748)




$NM_3$




$NM_5$

Figure: *Histogram of tomato roots data with the estimated density under the unconstrained (dashed line) and constrained (solid line) $NM_3$ and $NM_5$ models.*

# Conclusions

- In this contribution we propose the use of normal mixture (NM) models for testing symmetry about an unknown value
- The proposed likelihood ratio test is based on formulating the hypothesis of symmetry in terms of constraints on weights characterizing the NM model
- A Monte Carlo study outlined how the performance of the proposed test depends on the criterion used to select the number of mixture components: using BIC
  - a good empirical level of significance is obtained, comparable with that of the traditional Gupta's test
  - the empirical power resulted usually better than that observed with the Gupta's test

# Further developments

- comparing the performance of the mixture-based test with non-parametric symmetry tests (e.g., triples test)
- studying the dependence between the empirical levels of the test and the selected set of grid points $\delta_j$
- studying more in detail the relation between the empirical levels of the test and the selected number of mixture components $k$

# Main references

Cabilio, P. and Masaro, J. (1996). A simple test of symmetry about an unknown median. The Canadian Journal of Statistics, 24: 349 - 361.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39: 1 - 38.

Fan, Y. and Gencay, R. (1995). A consistent nonparametric test of symmetry in linear regression models. Journal of American Statistical Association, 90(430): 551 - 557.

Gupta, M. (1967). An asymptotically non parametric test of symmetry. The Annals of Mathematical Statistics, 38(3): 849 - 866.

Lindsay, B. (1996). Mixture Models: Theory, Geometry and Applications. Institute of Mathematical Statistic.

McLachlan, G. and Peel, D. (2000). Finite mixture models. Wiley.

McWilliams, T. (1990). A distribution-free test for symmetry based on a runs statistic. Journal of the American Statistical Association, 85(412): 1130 - 1133.

Miao, W., Gel, Y., and Gastwirth, J. L. (2006). Random Walk, Sequential Analysis and Related Topics - A Festschrift in Honor of Yuan-Shih Chow, chapter A new test of symmetry about an unknown median. World Scientific.

Mira, A. (1999). Distribution-free test of symmetry based on bonferroniÕs measure. Journal of Applied Statistics, 26(8): 959 - 971.

Modarres, R. and Gastwirth, J. (1996). A modified runs test for symmetry. Statistics & Probability Letters, 31: 107 - 112.

Ngatchou-Wandji, J. (2006). On testing for the nullity of some skewness coefficients. International Statistical Review, 74(1): 47 - 65.

Racine, J. and Maasoumi, E. (2007). A versatile and robust metric entropy test of time-reversibility, and other hypotheses. Journal of Econometrics, 138: 547 - 567.

Randles, R., Fligner, M., Policello II, G., and Wolfe, D. (1980). An asymptotically distribution-free test for symmetry versus asymmetry. Journal of the American Statistical Association, 75(369): 168 - 172.