Practical Testing for the Normal Mixture

JIN SEO CHO

School of Economics, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, 03722, , Korea

Abstract

The current study provides a Gaussian version used to test for the normal mixture with a single mean and two distinct variances. We derive the Gaussian versions for the model by associating its score function with the generalized Laguerre polynomial. The Gaussian version is analytical, so that it can be simulated to obtain the asymptotic critical values of the likelihood-ratio test.

Keywords: Gaussian version, Asymptotic critical value, LR test, Generalized Laguerre polynomial.
JEL: C12; C46.

1. Introduction

Mixture models are popular for empirical analysis, and testing for the mixture hypothesis is crucial for many purposes. For instance, the regime-switching model assumes an autocorrelated mixture model as a model for business cycle.

Nevertheless, testing for the mixture hypothesis is nonstandard. When testing for the mixture hypothesis naturally, a nuisance parameter is introduced that does not exist under the null of a single distribution (e.g., Davies, 1977, 1987). The null limit distribution of a standard test, such as the likelihood-ratio (LR) test, diverges from a chi-squared distribution due to the presence of the nuisance parameter (e.g., Cho and White, 2007, 2010). The null limit distribution of the test is characterized by a Gaussian process whose covariance kernel is model dependent, leading to different null limit distributions for different models.

The primary objective of this study is to provide a version of the Gaussian process that can be simulated straightforwardly. We investigate a normal mixture with a shared mean and two distinct variances.

We achieve the goal by representing the Gaussian process as a series of functions with independent Gaussian random coefficients. When a Gaussian process is represented in this format, it is easy to simulate and can be used to obtain the asymptotic critical values of LR test. For the desired representation, we demonstrate that the score function can be expressed as a sequence of orthogonal generalized Laguerre polynomials.

The existing research has provided Gaussian versions for assessing the mixture hypothesis. Cho and White (2007) have consider the mixture with distinct means and variance and provide a Gaussian version for the Gaussian process determining the null limit distribution of the LR test. Cho and White (2010) investigate the LR test for testing the exponential or Weibull mixture hypothesis and provide Gaussian versions of the Gaussian processes characterizing the null limit distributions of the LR tests. To our knowledge, no prior literature exists on a Gaussian version for testing the normal mixture with a shared mean and two distinct variances. As a related study, the EM test is proposed as an alternative test by Chen and Li (2009) to handle unbounded LR test. We impose a bounded parameter space so that the LR test is bounded under the hypothesis.

This study is structured as follows. In Section 2, we describe the mixture models and derive the version of the Gaussian process analytically. Section 3 provides simulation evidence, and we conclude in Section 4. Mathematical proofs are collected in the Appendix.

2. The Gaussian Version

We suppose that Y_t follows the normal mixture:

$$Y_t \sim \text{ IID} \begin{cases} \mathcal{N}(\mu_*, \sigma_{1*}^2), & \text{w.p. } \pi_*; \\ \mathcal{N}(\mu_*, \sigma_{2*}^2), & \text{w.p. } 1-\pi_*, \end{cases}$$

where σ_{1*}^2 and $\sigma_{2*}^2 \in [L, U]$, and we suppose L and U are sufficiently small and large, respectively so that the mixture is included in the model. This compact space condition is imposed for a bounded null limit distribution of the LR test (see Hartigan, 1985; Chen and Li, 2009). The hypothesis is that $Y_t \sim$ IID $\mathcal{N}(\mu_*, \sigma_*^2)$. That is, the null hypothesis can be constructed as follows:

$$H_0 : \pi_* = 1 \text{ and } \sigma_{1*}^2 = \sigma_*^2; \;\; \pi_* = 0 \text{ and } \sigma_{2*}^2 = \sigma_*^2; \;\text{ or }\; \sigma_{1*}^2 = \sigma_{2*}^2 = \sigma_*^2.$$

The joint hypothesis involves an identification problem. If $\pi_* = 1$, then σ_{2*}^2 is not identified. Similarly, if $\pi_* = 0$, then σ_{1*}^2 is not identified. Conversely, if $\sigma_{1*}^2 = \sigma_{2*}^2 = \sigma_*^2$, then π_* is not identified.

We can apply the LR test principle to test the hypothesis. If we let the LR test be

$$\mathcal{LR}_n := 2\left\{ \underset{\pi,\mu,\sigma_1^2,\sigma_2^2}{\arg\max}\, L_n(\pi, \mu, \mu, \sigma_1^2, \sigma_2^2) - \underset{\mu,\sigma^2}{\arg\max}\, L_n(1, \mu, \mu_2, \sigma^2, \sigma_2^2) \right\},$$

where $L_n(\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) := \sum_{t=1}^n \ell_t(\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ and

$$\ell_t(\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) := \log\left[\frac{\pi}{\sqrt{2\pi\sigma_1^2}}\exp\left[-\frac{(Y_t-\mu_1)^2}{2\sigma_1^2}\right] + \frac{1-\pi}{\sqrt{2\pi\sigma_2^2}}\exp\left[-\frac{(Y_t-\mu_2)^2}{2\sigma_2^2}\right]\right],$$

then we can obtain the following null limit distribution by applying Cho and White (2007):

$$\mathcal{LR}_n \Rightarrow \sup_{\gamma\in\Gamma} \overset{2}{\max}[0, \mathcal{U}(\gamma)],$$

where $\gamma := 1 - \sigma^2/\sigma_*^2$, $\Gamma := [1 - U/\sigma_*^2, 1 - L/\sigma_*^2]$ is the space of γ, and $\mathcal{U}(\cdot)$ is a Gaussian process with the following covariance kernel: for each γ_1 and $\gamma_2 \in \Gamma$,

$$\mathbb{E}[\mathcal{U}(\gamma_1)] = 0, \quad \mathbb{E}[\mathcal{U}(\gamma_1)\mathcal{U}(\gamma_2)] = \frac{W(\gamma_1, \gamma_2)}{\sqrt{W(\gamma_1, \gamma_1)}\sqrt{W(\gamma_2, \gamma_2)}},$$

and

$$W(\gamma_1, \gamma_2) := \frac{1}{\sqrt{1-\gamma_1\gamma_2}} - 1 - \frac{1}{2}\gamma_1\gamma_2.$$

Due to the identification problem, the null limit distribution of the LR test is now represented as a functional of the Gaussian process $\mathcal{U}(\cdot)$, and it is obtained from the null limit of the score function obtained while testing $\pi_* = 1$ and $\sigma_{1*}^2 = \sigma_*^2$; or $\pi_* = 0$ and $\sigma_{2*}^2 = \sigma_*^2$.

We next obtain a version of $\mathcal{U}(\cdot)$ analytically. If we let $D_n(\cdot)$ be the score function obtained while testing $\pi_* = 1$, the next theorem provides its null weak limit:

Theorem 1. *Given the assumptions, as a function of* γ*,*

$$D_n(\gamma) = \sum_{j=2}^{\infty}(\gamma)^j\left[-\frac{1}{\sqrt{n}}\sum_{t=1}^{n} L_j^{(-1/2)}\left(\frac{\widehat{Y}_t^2}{2}\right)\right] \Rightarrow \widetilde{\mathcal{V}}(\gamma) := \sum_{j=2}^{\infty}(\gamma)^j\sqrt{\frac{\Gamma(j+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(j+1)}}Z_j,$$

under the hypothesis, where $L_j^{(\alpha)}(\cdot)$ *is the* j*-th degree generalized Laguerre polynomial (e.g., Hochstrasser, 1964, p.775, 22.3.9), and* $Z_j \sim$ *IID* $N(0,1)$.

Remarks. (a) The score function under the null is obtained as

$$\begin{aligned} D_n(\gamma) &:= \frac{1}{\sqrt{n}}\frac{\partial}{\partial\pi}L_n(1, \widehat{\mu}_{n0}, \widehat{\mu}_{n0}, \widehat{\sigma}_{0n}^2, \widehat{\sigma}_{0n}^2) \\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left\{1 + \frac{\gamma}{2}(1 - \widehat{Y}_t^2) - \frac{1}{\sqrt{1-\gamma}}\exp\left[-\left(\frac{\gamma}{1-\gamma}\right)\frac{\widehat{Y}_t^2}{2}\right]\right\}, \end{aligned}$$

where $\widehat{Y}_t := (Y_t - \widehat{\mu}_{0n})/\sqrt{\widehat{\sigma}^2_{0n}}$.

(b) If we let $\widetilde{\mathcal{U}}(\gamma) := \widetilde{\mathcal{V}}(\gamma)/\sqrt{W(\gamma,\gamma)}$, Theorem 1 implies that $\mathcal{LR}_n \Rightarrow \sup_{\gamma\in\Gamma}\min^2[0,\widetilde{\mathcal{U}}(\gamma)]$ under the null, so that

$$\mathbb{E}[\widetilde{\mathcal{U}}(\gamma_1)\widetilde{\mathcal{U}}(\gamma_2)] = \sum_{j=2}^{\infty}(\gamma_1\gamma_2)^j\frac{\Gamma(j+1/2)}{\Gamma(\frac{1}{2})\Gamma(j+1)} = \frac{1}{\sqrt{1-\gamma_1\gamma_2}}-1-\frac{1}{2}\gamma_1\gamma_2 = W(\gamma_1,\gamma_2)$$

by noting that

$$\sum_{j=0}^{\infty}(\gamma_1\gamma_2)^j\frac{\Gamma(j+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(j+1)} = \frac{1}{\sqrt{1-\gamma_1\gamma_2}} \tag{1}$$

as shown in the Appendix. This also implies that both $\mathcal{U}(\cdot)$ and $\widetilde{\mathcal{U}}(\cdot)$ have the same covariance kernel, so that the asymptotic critical values of the LR test can be obtained by simulating $\sup_{\gamma\in\Gamma}\min^2[0,\widetilde{\mathcal{U}}(\gamma)]$.

(c) The standard normal random variables $Z_2, Z_3, \ldots$ are obtained by applying central limit theorem (CLT) to the generalized Laguerre polynomials. That is, for each j,

$$-\frac{1}{\sqrt{n}}\sum_{t=1}^{n}L_j^{(-1/2)}\left(\frac{\widehat{Y}_t^2}{2}\right) \Rightarrow \sqrt{\frac{\Gamma(j+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(j+1)}}Z_j$$

by noting that for each j,

$$\int_0^\infty \left\{L_j^{(-1/2)}(x)\right\}^2\frac{1}{\sqrt{\pi}}\exp(-x)x^{-1/2}dx = \frac{\Gamma(j+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(j+1)}.$$

Here, $\pi^{-1/2}\exp(-(\cdot))(\cdot)^{-1/2}$ denotes the asymptotic probability density function of $\widehat{Y}_t^2/2$. The independence between Z_j and $Z_{j'}$ ($j \neq j'$) follows from the orthogonality of the generalized Laguerre polynomials. That is, for $j \neq j'$,

$$\int_0^\infty L_j^{(-1/2)}(x)L_{j'}^{(-1/2)}(x)\frac{1}{\sqrt{\pi}}\exp(-x)x^{-1/2}dx = 0.$$

(d) The current approach can be used to test for a mixture of conditional normals as well. When testing for a mixture of conditional normals with two distinct scale parameters, $\mathcal{U}(\cdot)$ emerges as the resultant Gaussian process.

3. Simulation Evidence

Using the Gaussian version $\widetilde{\mathcal{U}}(\cdot)$, we conduct simulations to affirm Theorem 1.

For the simulation, we generate $Y_t \sim \text{IID}\mathcal{N}(0,1)$ and test the hypothesis that Y_t follows a normal distribution. We let σ^2_{1*}, $\sigma^2_{2*} \in [1/2, 3/2]$ but do not restrict the parameter space for μ_*. Given that $\sigma^2_* = 1$ under the null, this specification implies that $\Gamma = [-1/2, 1/2]$.

We first determine the asymptotic critical values of the LR test. We repeat 100,000 independent experiments to obtain the empirical distribution of $\sup_{\gamma\in\Gamma} \min^2[0, \widetilde{\mathcal{U}}(\gamma)]$. Here, we have approximated $\widetilde{\mathcal{U}}(\cdot)$ by

$$\widetilde{\mathcal{U}}_k(\cdot) := \frac{1}{\sqrt{W(\cdot,\cdot)}} \sum_{j=2}^{k} (\cdot)^j \sqrt{\frac{\Gamma(j+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(j+1)}} Z_j$$

with $k = 500$. We report the critical values of the LR test in Table 1.

Table 2 reports the empirical rejection rates of the LR test under the null. We let $Y_t \sim \text{IID}\mathcal{N}(0,1)$ and test the normal distribution hypothesis. The total number of iterations is 10,000. As the sample size increases, the empirical rejection rates of the LR test converge to the nominal significance levels. This implies that the Gaussian version $\widetilde{\mathcal{U}}(\cdot)$ consistently delivers the null limit distribution of the LR test.

For the power simulation, we let

$$Y_t \sim \text{ IID} \begin{cases} \mathcal{N}(0, 0.6), & \text{w.p. 1/2;} \\ \mathcal{N}(0, 1.4), & \text{w.p. 1/2,} \end{cases}$$

so that the alternative hypothesis is valid. Table 3 reports the power simulation results. The empirical rejection rates of the LR test converge to 100% as the sample size increases. This fact implies that the LR test has a consistent power.

The simulation results imply that the LR test is useful when testing for the mixture of normals driven by two different variances.

4. Conclusion

The current study analytically derives a version of the Gaussian processes associated with testing for the normal mixture with two different variances and a single mean. We obtain the Gaussian version by relating the score function to the series of the generalized Laguerre polynomials. Due to its analytical form, it can be simulated to obtain the asymptotic critical values of the LR test. The null limit distribution is model dependent. Therefore, if the normal mixture with more than two components is tested, the null limit distribution of the LR test differs from that given in the current study. We leave this as a future research topic.

5. Appendix

Before proving Theorem 1, we first prove (1). We note that

$$\sum_{j=0}^{\infty} s^j \frac{\Gamma(j+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(j+1)} = \frac{1}{\Gamma(\frac{1}{2})} \sum_{j=1}^{\infty} s^{j-1} \frac{\Gamma(j-\frac{1}{2})}{\Gamma(j)} = \frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{1}{2})} \sum_{j=1}^{\infty} s^{j-1} \frac{\Gamma(j-\frac{2}{2})}{(j-1)!(-\frac{1}{2})!}$$
$$= \sum_{j=1}^{\infty} s^{j-1} \binom{j-\frac{3}{2}}{j-1} = \sum_{j=0}^{\infty} s^j \binom{j-\frac{1}{2}}{j} = \sum_{j=0}^{\infty} (-s)^j \binom{-\frac{1}{2}}{j} = \frac{1}{\sqrt{1-s}}.$$

This establishes (1).

We now prove the main theorem.

Proof of Theorem 1: Using the formula of the generalized Laguerre polynomial generating function, we note that

$$\frac{1}{\sqrt{1-\gamma}} \exp\left[-\left(\frac{\gamma}{1-\gamma}\right)\frac{\widehat{Y}_t^2}{2}\right] = \sum_{j=0}^{\infty} \gamma^j L_j^{(-1/2)}\left(\frac{\widehat{Y}_t^2}{2}\right),$$

where $L_0^{(-1/2)}\left(\frac{\widehat{Y}_t^2}{2}\right) = 1$ and $L_1^{(-1/2)}\left(\frac{\widehat{Y}_t^2}{2}\right) = \frac{1}{2} - \frac{\widehat{Y}_t^2}{2}$ by noting that $L_0^{(-1/2)}(x) = 1$, $L_1^{(-1/2)}(x) = \frac{1}{2} - x$, and so on (e.g., Hochstrasser, 1964, p. 779, 22.5.38), implying that $\sum_{j=0}^{1} \gamma^j L_j^{(-1/2)}\left(\frac{\widehat{Y}_t^2}{2}\right) = 1 + \frac{\gamma}{2}(1 - \widehat{Y}_t^2)$, so that

$$D_n(\gamma) := \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left\{1 + \frac{\gamma}{2}(1-\widehat{Y}_t^2) - \frac{1}{\sqrt{1-\gamma}} \exp\left[-\left(\frac{\gamma}{1-\gamma}\right)\frac{\widehat{Y}_t^2}{2}\right]\right\}$$
$$= -\sum_{j=2}^{\infty} \frac{\gamma^j}{\sqrt{n}} \sum_{t=1}^{n} L_j^{(-1/2)}\left(\frac{\widehat{Y}_t^2}{2}\right).$$

We here note that if n is sufficiently large, $\widehat{Y}_t \overset{\mathrm{A}}{\sim}$ IID $\mathcal{N}(0,1)$, so that $\widehat{Y}_t^2 \overset{\mathrm{A}}{\sim}$ IID $\mathcal{X}_1^2$, and the asymptotic probability density function of $X_t := \widehat{Y}_t^2/2$ can be given as follows: $f(x) := \frac{x^{-1/2}}{\sqrt{\pi}} \exp(-x)$. We further note that for each j, $\int_0^\infty L_j^{(-1/2)}(x) f(x) = 0$ and $\int_0^\infty \{L_j^{(-1/2)}(x)\}^2 f(x) = \Gamma(j+\frac{1}{2})/\{\Gamma(\frac{1}{2})\Gamma(j+1)\}$ that is uniformly bounded by $\frac{1}{2}$ with respect to j (e.g., Hochstrasser, 1964, p.775, 22.2.12). Therefore, we can apply CLT for each j, so that for each j,

$$-\frac{1}{\sqrt{n}}\sum_{t=1}^{n} L_j^{(-1/2)}\left(\frac{\widehat{Y}_t^2}{2}\right) \Rightarrow \sqrt{\frac{\Gamma(j+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(j+1)}}Z_j.$$

Furthermore, we note that for any $j \neq j'$, $\int_0^\infty L_j^{(-1/2)}(x)L_{j'}^{(-1/2)}(x)f(x)dx = 0$ by the orthogonality of the generalized Laguerre polynomials (e.g., Hochstrasser, 1964, pp. 773-775, 22.1.1 and 22.2.12), implying that $\mathbb{E}[Z_j Z_{j'}] = 0$, meaning that Z_j and $Z_{j'}$ are independent. Therefore,

$$D_n(\cdot) = \sum_{j=2}^{\infty}(\cdot)^j\left[-\frac{1}{\sqrt{n}}\sum_{t=1}^{n} L_j^{(-1/2)}\left(\frac{\widehat{Y}_t^2}{2}\right)\right] \Rightarrow \sum_{j=2}^{\infty}(\cdot)^j\sqrt{\frac{\Gamma(j+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(j+1)}}Z_j.$$

This completes the proof. ■

6. Acknowledgments

The co-editor, Brendan Kline, and an anonymous referee provided very helpful comments for which I am most grateful. Cho acknowledges the research grant provided by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF2020S1A5A2A0104-0235).

References

Chen, J. and P. Li. Hypothesis test for Normal Mixture Models: The EM Approach, *Annals of Statistics*, 37, 2523–2542, 2009.

Cho, J. S. and H. White, Testing for Regime Switching, *Econometrica*, 75, 1671–1720, 2007.

Cho, J. S. and H. White, Testing for Unobserved Heterogeneity in Exponential and Weibull Duration Models, *Journal of Econometrics*, 157, 458–480, 2010.

Davies, R. B., Hypothesis Testing when a Nuisance Parameter is Present only under the Alternative, *Biometrika*, 64, 247–254, 1977.

Davies, R. B., Hypothesis Testing when a Nuisance Parameter is Present only under the Alternative, *Biometrika*, 74, 33–43, 1987.

Hartigan, J., Failure of Log-Likelihood Ratio Test, in L. Le Cam and R. Olshen, eds. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, 2. Berkeley: University of California Press, pp. 807 810.

Hochstrasser, U. W., Orthogonal Polynomials, in *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, ed. by M. Abramowitz and I. A. Stegun, Washington D.C.; New York: United States Department of Commerce, National Bureau of Standards; Dover Publications, vol. 55 of *Applied Mathematics Series*, chapter 22, 1964.

Test \ Level	1.00%	2.50%	5.00%	7.50%	10.00%
$\mathcal{LR}_n$	6.64	4.94	3.69	2.97	2.47

Table 1: ASYMPTOTIC CRITICAL VALUES OF THE LR TEST. Figures show the asymptotic critical values of the LR obtained by simulating $\widetilde{\mathcal{U}}_k(\cdot)$. The critical values are obtained by repeating independent experiments 100,000 times.

Test	Level \ n	1,000	5,000	10,000	20,000
$\mathcal{LR}_n$	1.00%	0.59	0.56	0.84	1.01
	2.50%	1.69	1.75	2.20	2.44
	5.00%	3.28	3.80	4.40	4.74
	7.50%	5.28	5.94	6.51	6.85
	10.0%	7.25	8.02	8.72	9.14

Table 2: EMPIRICAL REJECTION RATES OF THE LR TEST UNDER THE NULL (IN PERCENT). Figures show the empirical rejection rates of the LR test. DGP: $Y_t \sim \text{IID}\mathcal{N}(0,1)$. The parameter space of σ_*^2 is $[1/2, 3/2]$. The total number of replications is 10,000.

Test	Level \ n	200	500	800	1,000	1,500
$\mathcal{LR}_n$	1.00%	4.50	17.90	45.65	54.85	76.65
	2.50%	9.80	30.00	59.85	68.40	85.35
	5.00%	17.00	40.00	71.45	78.65	91.25
	7.50%	22.80	47.20	78.00	84.30	93.85
	10.0%	28.40	52.55	81.70	87.40	95.35

Table 3: EMPIRICAL REJECTION RATES UNDER THE ALTERNATIVE (IN PERCENT). Figures show the empirical rejection rates of the LR test. DGP: $Y_t \sim \text{IID}\mathcal{N}(0, 0.6)$ with probability 1/2; and $\mathcal{N}(0, 1.4)$ with probability 1/2. The parameter space of σ_*^2 is $[1/2, 3/2]$. The total number of replications is 2,000.