JETEM

Journal of Economic Theory and Econometrics

The Journal of the Korean Econometric Society

(계 량 경 제 학 보)

September 2009

Volume 20. Issue 3. pp.31-55

Testing for the Mixture Hypothesis of Geometric Distributions*

Jin Seo Cho† · Chirok Han‡

Abstract Use of the likelihood ratio (LR) statistic is examined to test for the mixture assumption of geometric distributions. As the asymptotic null distribution of the LR statistic is not a standard chi-square due to the fact that there are a boundary parameter problem and a nuisance parameter not identified under the null, we derive it separately and also provide a method to obtain the asymptotic critical values. Further, the finite sample properties of the LR test are evaluated by Monte Carlo simulations by examining the levels and powers of the LR test. Finally, using Kennan's (1985) strike data in labor economics, we conclude that unobserved heterogeneity is present in the data, which cannot be captured by specifying a geometric distribution.

Keywords mixture, geometric distribution, likelihood ratio statistic, unobserved heterogeneity, strike data

JEL Classification C12, C41, C80, J52

* The authors thank the co-editor, Myoung-jae Lee, and an anonymous referee. Taul Cheong provided excellent research assistance. The first author also appreciates research support from Korea University under Grant No. K0823641.

† Associate Professor, Department of Economics, Korea University, E-mail address: jinseocho@korea.ac.kr

‡ Associate Professor, Department of Economics, Korea University, E-mail address: chirokhan@korea.ac.kr

Received December 10, 2008, Revised February 8, 2009, Accepted May 1, 2009

1. Introduction

Use of mixtures of geometric distributions is popular for data analysis. For example, early literature such as Daniels (1961) advocates geometric mixtures to approximate discrete probability distributions. As another example, van den Berg and Ridder (1998) specify mixtures of exponential distributions for unemployment duration data. If observed data were grouped at integer levels, then their model would have been a mixture of geometric distributions.

This popularity is not restricted only to economics. In demographic literature, fecundability studies, risk managements, wireless communication studies and etc., whenever inter-arrival times of the events of interests are discretely grouped, mixture of geometric distributions becomes one of the most popular candidate models for data (see Heckman and Walker, 1990; Ecochard and Clayton, 2000; Crowder, 2001; and Ji, Liu, Towsley, and Kurose, 2004).

Challenges arise when specifying mixtures. Probability laws of discrete variables may be approximated too poorly if no mixture is specified, or redundant approximation can result from allowing for too many components for the mixture. Similarly to what Lancaster (1979) points out for continuous duration models, specifying a geometric distribution can result in biased parameter estimates if two or more component based mixtures are correct specifications. Conversely, if a geometric model is already a correct model, then two or more component based mixtures contain unidentified parameters. Thus, inference method should be different from the standard ones.

The main goal of the current study is therefore in selecting a proper mixing distribution for a finite geometric mixture. Specifically, we focus on testing for two component based geometric mixtures against a geometric distribution by the likelihood ratio (LR) statistic.

The LR statistic does not asymptotically follow a standard chi-square distribution under the null in the present context. As pointed out by Cho and White (2007, 2008), it has a nuisance parameter present only under the alternative (Davies, 1977; 1987) and also the boundary parameter problem. We overcome these and obtain the asymptotic null distribution as a function of Gaussian process by applying the methodology in Andrews (1999, 2001). Also, a simulation method is provided, which can deliver asymptotic critical values systematically.

Another goal of the present study is to conduct a formal testing procedure for Kennan's (1985) strike data. Kennan (1985) specifies a mixture based upon beta distributions in order to model unobserved heterogeneity after *assuming* its existence. We affirm the existence by our formal *testing* procedure so that his attempt to model unobserved heterogeneity can be justified. This application also illustrates appropriate use of LR statistic in the mixture model context.

The literature on testing the mixture hypothesis is expanding. Hartigan (1985) and Ghosh and Sen (1985) first show that the LR statistic weakly converges to a function of a Gaussian process under the null after assuming a null model without any unknown parameter. Chernoff and Lander (1995) focus on the mixture of binomials and obtain a function of Gaussian process as the weak limit of the LR statistic under the null. They also provide another Gaussian process which has the same covariance structure as the one obtained under the null, and which can be easily generated by a simulation method. Nevertheless, the Gaussian processes contained therein cannot be used for other mixtures because the LR statistic is not distribution-free. Cho and White (2007, 2008) consider the mixtures of normals, exponentials, and Weibulls and show that the asymptotic null distributions are determined by Gaussian processes with different covariance structures. They also provide simulation methods to obtain the asymptotic critical values as in Chernoff and Lander (1995). Further, Chen and Chen (2001) consider the mixtures of Poissons and normals with known variance and obtain similar conclusions, though they do not discuss any method to obtain the asymptotic critical values. Finally, Dacunha-Castelle and Gassiat (1999) examine the mixture of two regular conditional distributions and derive results similar to the mixture of marginal distributions. To our best knowledge, testing for the geometric mixture has not yet been examined in the literature.

The plan of this paper is as follows. Section 2 considers data, mixture models, and the LR statistic together with the asymptotic null distribution. We conduct Monte Carlo experiments and empirical data analyses on Kennan's (1985) data in Sections 3 and 4 respectively. Conclusion is given in Section 5, and mathematical proofs are collected in the Appendix.

2. The Mixture of Geometric Distributions

2.1. The Data Generating Process

Let $\{Y_t\}$ be a sequence of identically and independently distributed (IID) random variables. We consider an economic hypothesis associated with $\{Y_t\}$, which suggests a geometric distribution function as its marginal probability density function (PDF). That is, for $y = 1, 2, 3, \ldots$, and unknown $p_* \in (0, 1)$,

$$f(y; p_*) = p_*(1 - p_*)^{y-1} \tag{1}$$

is suggested as the marginal PDF of Y_t. Given this environment, we further suppose that the researcher wishes to test the geometric distribution hypothesis.

Daniels (1961) considers representing many discrete PDFs as a weighted mixture of geometric distributions by assuming that the marginal PDF of Y_t can be written as

$$\int f(y;p)dH(p), \tag{2}$$

where dH is a density function on $(0,1)$. In this framework, the hypothesis in (1) corresponds to the case in which dH has a unit mass at p_*. In other cases, the structure in (2) is often interpreted as the presence of unobserved heterogeneity in the literature.

Here, we follow Daniels' (1961) advocation and particularly pay attention to a finite geometric mixture. Specifically, we consider the two component mixture whose marginal PDF is given as

$$\pi_* f(y;p_{1*}) + (1-\pi_*)f(y;p_{2*}), \tag{3}$$

for $y = 1,2,3,\ldots$ and $(\pi_*, p_{1*}, p_{2*}) \in [0,1] \times (0,1) \times (0,1)$.

The model in (3) has useful aspects for applications. First, economic theory may directly indicate geometric distribution as a data generating process (DGP) for economic data. As an example, van den Berg and Ridder (1998) postulate a finite mixture of exponentials as a DGP for unemployment duration, so that a geometric mixture turns out to be a DGP corresponding to discretely (e.g., daily or weekly) grouped duration data. Second, we can even exploit the simple structure in (3) to distinguish the DGPs in (1) and (2). As Ferguson (1983) suggests, discrete mixtures can approximate arbitral distributions to the level of desired precision by adding more components under certain regularity conditions. Thus, the model in (3) may be understood as an approximation to (2). This aspect is also considered by Stinchcombe and White (1998) in terms of artificial neural networks.

2.2. Discrete Mixtures and the Log-Likelihood Ratio Statistic

We specify a model $\mathcal{M}$ as a collection of DGPs as in White (1994). Our null model is

$$\mathcal{M}_o \equiv \{f(\,\cdot\,;p) : p \in \mathbb{P}\},$$

where f is given in (2), and $\mathbb{P}$ is a convex subset of $[0,1]$ containing p_* as an interior element. The unrestricted model is specified as a collection of DGPs:

$$\mathcal{M}_u \equiv \{f_u(\,\cdot\,;\pi,p_1,p_2) : (\pi,p_1,p_2) \in [0,1]\times\mathbb{P}\times\mathbb{P}\},$$

where for $y = 1, 2, \ldots,$

$$f_u(y;\pi,p_1,p_2) \equiv \pi f(y;p_1) + (1-\pi)f(y;p_2). \tag{4}$$

Given these models, for each p, we denote the expected log-likelihoods as

$$L_o(p) \equiv E[\log f(Y_t;p)] \quad \text{and} \quad L_u(\pi,p_1,p_2) \equiv E[\log f_u(Y_t;\pi,p_1,p_2)] \tag{5}$$

respectively. Under the null, $L_o(p_*) = L_u(\pi_*,p_{1*},p_{2*})$, whereas $L_o(p_*) < L_u(\pi_*,p_{1*},p_{2*})$ under the alternative stated below. Here, p_* and (π_*,p_{1*},p_{2*}) maximize L_0 and L_u respectively. From this, we can relate the difference between $L_u(\pi_*,p_{1*},p_{2*})$ and $L_o(p_*)$ to testing the null hypothesis by Kullback-Leibler information criterion, as it ensures that $L_u(\pi_*,p_{1*},p_{2*}) = L_0(p_*)$ if and only if the DGP is indeed identical to (1). We thus consider the asymptotic null behavior of the following LR statistic:

$$LR_n \equiv 2\left\{\sum_{t=1}^{n}\log[f_u(Y_t;\hat{\pi}_n,\hat{p}_{1n},\hat{p}_{2n})] - \sum_{t=1}^{n}\log[f(Y_t;\hat{p}_{on})]\right\},$$

where n is the sample size, and $\hat{p}_{on}$ and $(\hat{\pi}_n, \hat{p}_{1n}, \hat{p}_{2n})$ are the maximum-likelihood estimators (MLEs) obtained under the null and unrestricted model assumptions respectively.

The null hypothesis of no-mixture encompasses three distinct null hypotheses:

$$H_o: \ \pi_* = 1 \text{ and } p_{1*} = p_*; \ p_{1*} = p_{2*}; \text{ or } \pi_* = 0 \text{ and } p_{2*} = p_*,$$

whereas the alternative can be rephrased as

$$H_u : \pi_* \in (0, 1) \text{ and } p_{1*} \neq p_{2*}.$$

Analyzing the LR statistic is not standard as pointed out many times in the literature. If $\pi_* = 1$ (resp. $\pi_* = 0$), then p_{2*} (resp. p_{1*}) is not identified, so that there is a nuisance parameter present only under the alternative (Davies, 1977; 1987). Further, $\pi_* = 1$ (resp. $\pi_* = 0$) is on the boundary of the parameter space, violating the interior parameter condition for standard asymptotic analysis. Alternatively, if $p_{1*} = p_{2*}$, then π_* is not identified. From these, it follows that the asymptotic null distribution of the LR statistic is not a standard chi-square.

Cho and White (2007, 2008) consider the same problem in the context of mixtures of normals, exponentials, and Weibull distributions. Applying the methodology in Andrews (1999, 2001), they show that for the same hypothesis, the LR test weakly converges to a function of a Gaussian process under the null. This is still effective even for the geometric mixtures as confirmed in Theorem 1.

Theorem 1: *If we let* $\underline{p} \equiv \inf \mathbb{P}$ *and* $\bar{p} \equiv \sup \mathbb{P}$ *be such that* $(1-\underline{p})^2(1-p_*)/(1-\bar{p})^2 < 1$ *and* $0 < \underline{p} < p_* < \bar{p} < 1$, *then*

$$LR_n \Rightarrow \mathcal{LR} \equiv \sup_{p \in \mathbb{P}} \left(\max[0, \mathcal{G}(p)]\right)^2 \tag{6}$$

under H_o, *where* $\mathcal{G}$ *is a Gaussian process such that for each* p *and* $p' \in \mathbb{P}$,

$$E[\mathcal{G}(p)\mathcal{G}(p')] = \frac{[(1-p_*)-(1-p)^2]^{1/2}[(1-p_*)-(1-p')^2]^{1/2}}{(1-p_*)-(1-p)(1-p')}. \tag{7}$$

The consequence of Theorem 1 is coherent with other results in the literature. Chernoff and Lander (1995) analyze the LR statistic for the finite mixture of binomials and obtain the results similar to (6), but their limit Gaussian process has covariance structure different from (7). Chen and Chen (2001) examine both discrete mixtures of normals with known variance and discrete mixtures of Poissons and also obtain the results similar to (6) using different Gaussian processes. Dacunha-Castelle and Gassiat (1999) further generalize these analyses by studying the mixtures of conditional distributions and show that the LR statistic weakly converges to the squared max function of a different Gaussian process.

In general, the limit Gaussian process associated with the LR statistic has a different distribution if different models and DGPs are assumed. This implies that the LR statistic is not distribution-free, thus we need to compute different asymptotic critical values for different models and DGPs.

In general, it is hard to obtain the analytical formula for the distribution of the maximum of a Gaussian process. However, the limit Gaussian process in Theorem 1 can be easily simulated by the orthonormal bases stated in Theorem 2.

Theorem 2: *Let* $\{Z_k : k = 0, 1, 2, \cdots\}$ *be an IID sequence of* $N(0,1)$ *random variables. Then* $\mathcal{G} \stackrel{d}{=} \bar{\mathcal{G}}_\infty$, *where for each* $p \in \mathbb{P} \equiv [\underline{p}, \bar{p}]$ *such that* $\underline{p} > 1 - \sqrt{1-p_*}$,

$$\bar{\mathcal{G}}_m(p) \equiv \left\{ \frac{(1-p_*)-(1-p)^2}{1-p_*} \right\}^{1/2} \sum_{k=0}^{m} \left\{ \frac{1-p}{\sqrt{1-p_*}} \right\}^k Z_k \tag{8}$$

for $m = 1, 2, \ldots$.

Note that the Gaussian process $\bar{\mathcal{G}}_\infty$ is the sum of an infinite number of functions defined on $\mathbb{P}$. Each function is multiplied by IID standard normal coefficient and converges to zero uniformly in p as k gets large. Thus, for each p, the difference between $\bar{\mathcal{G}}_m(p)$ and $\mathcal{G}(p)$ can be controlled at the level of desired precision by letting m be large enough, and we can use the simulated empirical distribution of $\bar{\mathcal{G}}_m$ as an approximation of $\bar{\mathcal{G}}$.

There are a couple of remarks relevant to Theorem 2. First, the simulation method has been used in the previous literature. See Chernoff and Lander (1995) and Cho and White (2007, 2008) among others. If no Gaussian process like $\bar{\mathcal{G}}_\infty$ is available, then conservative critical values can be applied. Davies (1977) provides a formula for this using the probability law of up-crossings. Piterbarg (1996) also suggests to obtain conservative critical values by the comparison method. These values may not be as precise as those obtained by our simulation method. Second, different parameter space $\mathbb{P}$ leads to different critical values, as pointed out by Cho and White (2007, 2008). For example, if $\mathbb{P}_2$ is a proper subset of $\mathbb{P}_1$, then the LR statistic assuming $\mathbb{P}_2$ cannot be greater than the LR statistic assuming $\mathbb{P}_1$, so that their asymptotic null distributions cannot be identical. We also note that the condition on the parameter space in Theorem 2 is implied by Theorem 1. That is,

$$\text{if } p_* < \bar{p} \text{ and } (1-\underline{p})^2(1-p_*)/(1-\bar{p})^2 < 1, \text{ then } \underline{p} > 1 - \sqrt{1-p_*}. \tag{9}$$

See the Appendix for the proof of this. Therefore, if $\mathbb{P}$ is selected to satisfy the condition in Theorem 1, then $\mathcal{G}$ can be generated without too much concern on the conditions in Theorem 2. From now, the LR test and its asymptotic limit will be denoted by $LR_n(\mathbb{P})$ and $\mathcal{LR}_m(\mathbb{P})$ respectively to accommodate the influence of parameter space and the finite number m of standard normal random variables.

In practice, as p_* is usually unknown, the asymptotic null distribution should be simulated using Theorem 2 with $\hat{p}_{on}$ instead of p_*. The proof of this is identical to Theorem 2 in Cho and White (2008). We thus omit it for the brevity.

3. Monte Carlo Experiments

In this section, we conduct Monte Carlo experiments to examine the level and power properties of the LR test statistic. For the level property, we consider the DGP: $Y_t \sim$ IID $\mathbb{G}(1/2)$, where $\mathbb{G}(p_*)$ denotes the geometric distribution with parameter p_*. We obtain the null parameter estimate by specifying two different models: $Y_t \sim$ IID $\mathbb{G}(p)$, where p is an element of $\mathbb{P}_1 \equiv [0.3965, 0.5732]$ and $\mathbb{P}_2 \equiv [0.4000, 0.5500]$ respectively. These spaces are selected to satisfy the parameter space condition in Theorem 1.

Table 1 provides critical values of the LR test. They are computed by simulating $\bar{\mathcal{G}}_m$ 10,000 times with $m = 50$. The maxima of these functions are computed by grid search with equal grid distances partitioning each of $\mathbb{P}_1$ and $\mathbb{P}_2$. In principle, the associated p_* needs to be estimated, as $\mathcal{G}_m$ depends on the value of p_*, but for simplicity the figures in Table 1 are obtained under the assumption that it is known, and this treatment makes practically no significant difference to the critical values.

Table 2 provides finite sample size properties. We consider three cases separately. First, we assume that p_* is known and apply the critical values obtained from Table 1. The results are reported on the rows indexed by $LR_n(\mathbb{P}_1)$ and $LR_n(\mathbb{P}_2)$. Second, we estimate p_* using the null model and obtain the critical values by this estimate and Theorem 2. More precisely, for each data set, we estimate $\hat{p}_{on}$ by the MLE method using the null model and obtain the empirical null distribution by generating $\widehat{\mathcal{LR}}_m \equiv \sup_{p\in\mathbb{P}} \max[0, \widehat{\mathcal{G}}_m(p)]^2$ many times, where

$$\widehat{\mathcal{G}}_m(p) \equiv \left\{ \frac{(1-\hat{p}_{on}) - (1-p)^2}{1-\hat{p}_{on}} \right\}^{1/2} \sum_{k=0}^{m} \left\{ \frac{1-p}{\sqrt{1-\hat{p}_{on}}} \right\}^k Z_k.$$

Under the null, as $\hat{p}_{on}$ is a consistent estimator for p_*, this method can provide the asymptotic null distribution with probability converging to 1 as n increases. The simulation results are found in the rows indexed by $LR_n(\hat{\mathbb{P}}_1)$ and $LR_n(\hat{\mathbb{P}}_2)$. Finally, we also apply the parametric bootstrap, which is justified by the fact that the weak limit of the LR test exists by Theorem 1. Specifically, after obtaining $\hat{p}_{on}$ from the observed data, we resample data points from $\mathbb{G}(\hat{p}_{on})$ 2,000 times to obtain the empirical distribution and compute the empirical rejection rate. The simulation results are contained in the rows indexed by $LR_n^*(\mathbb{P}_1)$ and $LR_n^*(\mathbb{P}_2)$.

Table 2 shows that the sizes approach the associated nominal levels (5% and 10%) as n increases. Moreover, for finite samples, the empirical rejection rates (i.e., the simulated type I errors) are smaller than the nominal level. Figure 1 shows the empirical and asymptotic distributions of the associated LR statistics when $\mathbb{P}_1$ is selected. As the sample size increases, the empirical distribution of the LR statistic approaches the distribution of $\mathcal{LR}_m(\mathbb{P}_1)$. We also obtained similar figures for $\mathbb{P}_2$ but omit it for brevity. Other Monte Carlo experiments not reported here suggest that the finite sample size distortion becomes large as $\underline{p}$ gets close to $1-\sqrt{1-p_*}$, whereas the size distortion does not sensitively respond to the selection of $\bar{p}$. This is mainly because the approximation of $\mathcal{G}(\cdot)$ by (8) may not be accurate when $p \simeq 1-\sqrt{1-p_*}$ due the fact that $(1-p)/\sqrt{1-p_*}$ is close to 1 though still less than 1. For such a case, a much larger m would be required to achieve the approximation at the level of desired precision.

The parametric bootstrap also provides approximately correctly sized tests, but the finite-sample type I errors are bigger than nominal levels as reported in Table 2. Thus, the parametric bootstrap and LR test may be used as complementary devices in testing the hypotheses. We leave specific examination of this as a future research topic.

For the power properties, we consider the following DGPs:

- DGP I: $Y_t \sim$ IID $\frac{1}{2}\mathbb{G}(0.3) + \frac{1}{2}\mathbb{G}(0.7)$;

- DGP II: $(Y_t|X_t = x) \sim$ IID $\mathbb{G}(x)$, $X_t \equiv \sqrt{U_t}$, and $U_t \sim$ IID $U(0,1)$.

Our unrestricted model is correctly specified for DGP I but not for DGP II. In particular, we consider DGP II to examine the response of the LR statistic to misspecified heterogeneity. They both are designed to have $E[Y_t] = 2$, so that $\hat{p}_{on}$ estimates 1/2 consistently. We present our simulation results in Table 3. For these DGPs, we conclude that first, the LR statistics are consistent; second, the power properties of the LR test statistics are similar whether or not p_* is estimated; and finally, the power of the parametric bootstrap is better than the LR test. Nevertheless, obtaining empirical null distributions by the parametric bootstrap is a time-consuming process particularly because the unrestricted model is not identified by the resampled data drawn under the null model. In such a case, iterative procedures such as the Newton-Ralphson method have to be tried with multiple initial values, and maxima obtained by these initial values need to be compared, which can consume a considerable amount of time. This becomes pertinent even to the cases with moderately large samples.

4. Empirical Application

In this section, we apply our method to the data of Kennan (1985), who aims to provide stylized facts for the duration of contract strike in U.S. manufacturing industry. As an attempt to characterize the distribution of strike duration, Kennan (1985) specifies a model for unobserved heterogeneity as a mixture of geometric distributions weighted by beta distribution and explains this heterogeneity by other explanatory variables such as powered strike ages s^m and industrial production level X_t. Specifically, he considers strike duration D_t and specifies its conditional likelihood function as

$$\log[1 - q_t(D_t; \alpha, \beta_0, \cdots, \beta_m)] + \sum_{s=0}^{D_t - 1} \log[q_t(s; \alpha, \beta_0, \cdots, \beta_m)], \tag{10}$$

where

$$q_t(s) \equiv \frac{1+\alpha s[1+\exp(\beta_0+\beta_X X_t)]}{(1+\alpha s)[1+\exp(\beta_0+\beta_1 s+\cdots+\beta_m s^m+\beta_X X_t)]}.$$

This specification is closely related to the geometric distribution. If $\alpha = \beta_1 = \beta_2 = \cdots = \beta_m = \beta_X = 0$, then (10) reduces to the log-likelihood for a simple geometric distribution. Otherwise, the polynomials in s and X_t can serve as explanatory variables for unobserved heterogeneity.

We illustrate our methodology using Kennan's (1985) data and affirm the presence of unobserved heterogeneity, which justifies his attempt to model (10). Notably this result is not obtained by Kolmogorov's goodness-of-fit (GOF) test, as detailed below.

Table 4 contains our estimation results and relevant test statistics. The associated parameter space $\mathbb{P}$ is selected in a way to satisfy the parameter constraint in Theorem 1. Specifically, after estimating the null parameter $\hat{p}_{on}$, we select $\mathbb{P}$ so that $\hat{p}_{on}$ is an interior element of $\mathbb{P}$ and generate the asymptotic null distribution using the simulation method described in Section 3. We repeat 10,000 times to simulate the asymptotic distribution by (8). The p-value associated with the LR test statistic is approximately 4%, so that the presence of unobserved heterogeneity should be admitted at the 5% significance level.

This result can be further corroborated by the size-distortion adjusted p-values, which can be obtained by applying the parametric bootstrap. For this, we first obtain fine grid points in the 95% confidence interval of p_* and resample 560 observations from $\mathbb{G}(p)$ for each grid point in the interval, where 560 is the sample size in Kennan (1985). Next, using these resampled observations, we compute the LR test and iterate this process 2,000 times for each grid point so that an empirical distribution function can be generated for each point. Finally, we compare the LR test from the original data with these empirical distributions to acquire the p-value for each grid point. The range of these p-values retrieved in this way is reported in Table 4 as [0.40%, 3.60%]. Note that the largest p-value obtained in this way is smaller than the p-value obtained by the asymptotic null distribution based upon Theorem 2, which is consistent with the illustrated under-sized empirical distributions in Figure 1. This result also affirms the presence of unobserved heterogeneity.

We also attempt to detect unobserved heterogeneity by the GOF statistic given as

$$\mathrm{GOF}_n \equiv \sup_y \sqrt{n}|\hat{F}_n(y) - \{1 - (1 - \hat{p}_{on})^y\}|,$$

where $\hat{F}_n(\cdot)$ is the empirical distribution function, and $\{1 - (1 - \hat{p}_{on})^y\}$ is the asserted null distribution function influenced by $\hat{p}_{on}$. Unfortunately, we cannot apply the critical values derived for the GOF statistic for continuous random variables, because our data are discretely distributed and the GOF statistic is not distribution-free (see Henze, 1996). Thus, we conduct our data inference by applying the parametric bootstrap again. As above, for each grid point $\tilde{p}$ in the confidence interval of p_*, we resample 560 geometric random variables from $\mathbb{G}(\tilde{p})$ and compute the GOF statistic. We iterate this computing process 10,000 times and obtain the null distribution of the GOF statistic, from which the p-value of the GOF test is computed for each grid point. These inference results are presented in Table 4. Surprisingly, the minimum p-value is far greater than 5%, and the GOF test does not identify unobserved heterogeneity, whereas the LR statistic does.

5. Conclusion

Mixture of geometric distributions is popular for empirical data analysis, as Daniels (1961) advocates approximating the distributions of discrete random variables by a mixture of geometric distributions.

We consider testing for the mixture hypothesis of geometric distributions by the LR statistic. As its asymptotic null distribution is not standard, we represent it as a function of a particular Gaussian process. Also, a representative version of this process is provided to simulate its asymptotic null distribution. Our Monte Carlo experiments affirm the theories on the LR statistic in the text. Finally, we conduct a formal testing procedure to capture unobserved heterogeneity in Kennan's (1985) strike duration data. From this, we justify his attempt to model unobserved heterogeneity, which cannot be detected by Kolmogorov's GOF test.

References

Andrews, D.W.K., 1999, Estimation When a Parameter is on a Boundary, Econometrica 67, 543–563.

Andrews, D.W.K., 2001, Testing When a Parameter is on the Boundary of the Maintained Hypothesis, Econometrica 69, 683–734.

Chen, H. and Chen, J., 2001, The Likelihood Ratio Test for Homogeneity in the Finite Mixture Models, Canadian Journal of Statistics 29, 201–216.

Chernoff, H. and Lander, E., 1995, Asymptotic Distribution of the Likelihood Ratio Test that a Mixture of Two Binomials is a Single Binomial, Journal of Statistical Planning and Inference 43, 19–40.

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

Cho, J. and White, H., 2008, Testing for Unobserved Heterogeneity in Exponential and Weibull Duration Models, UCSD Department of Economics Discussion Paper.

Crowder, M., 2001, Classical Competing Risks, CRC Press, Boca Raton.

Daniels, H., 1961, Mixtures of Geometric Distributions, Journal of the Royale Statistical Society. Series B 23, 409–413.

Dacunha-Castelle, D. and Gassiat, E., 1999, Testing the Order of a Model Using Locally Conic Parametrization: Population Mixtures and Stationary ARMA Processes, Annals of Statistics 27, 1178–1209.

Davies, R., 1977, Hypothesis Testing When a Nuisance Parameter is Present Only under the Alternative, Biometrika 64, 247–254.

Davies, R., 1987, Hypothesis Testing When a Nuisance Parameter is Present Only under the Alternative, Biometrika 74, 33–43.

Ecochard, R. and Clayton, D., 2000, Multivariate Parametric Random Regression Models for Fecundability Studies, Biometrics 56, 1023–1029.

Ferguson, T., 1983, Bayesian Density Estimation by Mixtures of Normal Distributions, Recent Advances in Statistics, edited by M. Rizvi, J. Rustagi, and D. Siegmund, Academic Press, NY, 287–302.

Ghosh, J. and Sen, P., 1985, On the Asymptotic Performance of the Log Likelihood Ratio Statistic for the Mixture Model and Related Results, Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer 2, edited by L. Le Cam and R. Olshen, University of California Press, Berkeley, CA, 789–806.

Hartigan, J., 1985, Failure of Log-Likelihood Ratio Test, Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer 2, edited by L. Le Cam and R. Olshen, University of California Press, Berkeley, CA, 807–810.

Heckman, J. and Walker, J., 1990, Estimating Fecundability from Data on Waiting Times to First Conception, Journal of the American Statistical Association 85, 283–294.

Henze, N., 1996, Empirical Distribution Function Goodness-of-Fit Tests for Discrete Models, Canadian Journal of Statistics 24, 81–93.

Ji, P., Liu, B., Towsley, D., and Kurose, J., 2004, Modeling Frame-Level Errors in GSM Wireless Channels, Performance Evaluation 55, 165–181.

Kennan, J., 1985, The Duration of Contract Strikes in U.S. Manufacturing, Journal of Econometrics 28, 5–28.

Lancaster, T., 1979, Econometric Methods for the Duration of Unemployment, Econometrica 47, 939–956.

Piterbarg, V., 1996, Asymptotic Methods in the Theory of Gaussian Processes and Fields. Translations of Mathematical Monographs, 148, American Mathematical Society, Providence.

Stinchcombe, M. and White, H., 1998, Consistent Specification Testing with Nuisance Parameters Present Only under the Alternative, Econometric Theory 14, 295–324.

van den Berg, G. and Ridder, G., 1998, An Empirical Equilibrium Search Model of the Labor Market, Econometrica 66, 1183–1221.

White, H., 1982, Maximum Likelihood Estimation of Misspecified Models, Econometrica 50, 1–25.

White, H. 1994, Estimation, Inference and Specification Analysis, Cambridge University Press, NY.

Appendix: Proofs

Proof of Theorem 1: To show the weak convergence, we verify the conditions of theorem 6(a) of Cho and White (2007). Our assumptions on IID condition and the model and the null H_o are sufficient for their assumption A1, A2(i, iii), and A3. Second, we note that for each $(\pi, p_1, p_2) \in [0,1]\times\mathbb{P}\times\mathbb{P}$ and $y = 1, 2, \cdots$, $\pi p_1(1-p_1)^{y-1}+(1-\pi)p_2(1-p_2)^{y-1} \leq (1-p_*)^{(y-1)/2}\{\pi p_1+(1-\pi)p_2\} \leq (1-p_*)^{(y-1)/2}$, so that we obtain that $E\{\log[\pi p_1(1-p_1)^{Y_t-1}+(1-\pi)p_2(1-p_2)^{Y_t-1}]\} \leq \log(1-p_*)(1-p_*)/(2p_*) < \infty$ because $p_* \in (0,1)$. Further, for each $y = 1, 2, \cdots$, and $p_1, p_2 \in \mathbb{P}$, $\pi p_1(1-p_1)^{y-1}+(1-\pi)p_2(1-p_2)^{y-1} \geq \underline{p}\bar{p}^{y-1}$, implying that $E\{\log[\pi p_1(1-p_1)^{Y_t-1}+(1-\pi)p_2(1-p_2)^{Y_t-1}]\} \geq \log(\underline{p})+(1-p_*)\log(1-\bar{p})/p_* > -\infty$. This verifies their assumption A4. Third, we verify their assumption A5(ii). For this, we can let their δ be zero as our data is identically and independently distributed. We note that some tedious but simple algebras show that all $|\partial \log f_u(Y_t;\pi,p_1,p_2)/\partial\pi|$, $|\partial \log f_u(Y_t;\pi,p_1,p_2)/\partial p_1|$, and $|\partial \log f_u(Y_t;\pi,p_1,p_2)/\partial p_2|$ are uniformly bounded by

$$\frac{Y_t+1}{\underline{p}(1-\bar{p})}\left(\frac{1-\underline{p}}{1-\bar{p}}\right)^{Y_t-1}$$

with respect to (π, p_1, p_2). Further,

$$E\left[(Y_t+1)^2\left(\frac{1-\underline{p}}{1-\bar{p}}\right)^{2(Y_t-1)}\right] = p_*\sum_{y=1}^{\infty}(y+1)^2\left\{\left(\frac{1-\underline{p}}{1-\bar{p}}\right)^2(1-p_*)\right\}^{y-1} < \infty, \quad (11)$$

because of the assumption that $(1-\underline{p})^2(1-p_*)/(1-\bar{p})^2 < 1$. Using Cauchy-Shwarz inequality, it straightforwardly follows that

$$E\left[\sup_{i,j}\left|\frac{\partial}{\partial i}\log f_u(Y_t;\pi,p_1,p_2)\frac{\partial}{\partial j}\log f_u(Y_t;\pi,p_1,p_2)\right|\right] < \infty,$$

where $i, j \in \{\pi, p_1, p_2\}$. Also, we note that for every $i, j \in \{\pi, p_1, p_2\}$, similar simple and tedious algebras show that

$$\left| \frac{\partial^2}{\partial i \partial j} \log f_u(Y_t; \pi, p_1, p_2) \right| \leq \frac{2(Y_t+1)^2}{\underline{p}^2(1-\bar{p})^4} \left(\frac{1-\underline{p}}{1-\bar{p}} \right)^{2(Y_t-1)}$$

uniformly with respect to (π, p_1, p_2). The given bound also has a finite first moment by (11). Thus,

$$E\left[\sup_{i,j} \left| \frac{\partial^2}{\partial i \partial j} \log f_u(Y_t; \pi, p_1, p_2) \right| \right] < \infty,$$

and we can let their M_t be defined as

$$M_t \equiv \frac{2(Y_t+1)^2}{\underline{p}^2(1-\bar{p})^4} \left(\frac{1-\underline{p}}{1-\bar{p}} \right)^{2(Y_t-1)}.$$

This verifies their condition A5(ii). Fourth, we verify their assumption A5(iii) and let their δ be zero using the fact that our data are identically and independently distributed as before. Note that

$$\frac{1}{f(Y_t;p)} \frac{\partial}{\partial p} f(Y_t;p) = \frac{(Y_t p - 1)}{p(1-p)}; \quad \frac{1}{f(Y_t;p)} \frac{\partial^2}{\partial p^2} f(Y_t;p) = \frac{(Y_t-1)(Y_t p - 2)}{p(1-p)^2};$$

$$\frac{1}{f(Y_t;p)} \frac{\partial^3}{\partial p^3} f(Y_t;p) = \frac{(Y_t-1)(Y_t-2)(Y_t p - 3)}{p(1-p)^3}; \text{ and}$$
$$\frac{1}{f(Y_t;p)} \frac{\partial^4}{\partial p^4} f(Y_t;p) = \frac{(Y_t-1)(Y_t-2)(Y_t-3)(Y_t p - 4)}{p(1-p)^4},$$

so that

$$\sup_{p \in \mathbb{P}} \left| \frac{1}{f(Y_t;p)} \frac{\partial}{\partial p} f(Y_t;p) \right|^4 \leq \{\alpha Y_t\}^4; \quad \sup_{p \in \mathbb{P}} \left| \frac{1}{f(Y_t;p)} \frac{\partial^2}{\partial p^2} f(Y_t;p) \right|^2 \leq \{\alpha(Y_t+1)\}^4;$$

$$\sup_{p \in \mathbb{P}} \left| \frac{1}{f(Y_t;p)} \frac{\partial^3}{\partial p^3} f(Y_t;p) \right|^2 \leq \{\alpha(Y_t+2)\}^6; \text{ and } \sup_{p \in \mathbb{P}} \left| \frac{1}{f(Y_t;p)} \frac{\partial^4}{\partial p^4} f(Y_t;p) \right| \leq \{\alpha(Y_t+3)\}^4$$

respectively, where $\alpha = 1/\underline{p}(1-\bar{p})$. Noting that

$$E[Y_t^6] = \{720 - 2520p_* + 3360p_*^2 - 2100p_*^3 + 602p_*^4 - 63p_*^5 + p_*^6\}/\{p_*^6(1-p_*)\} < \infty,$$

it is not hard to obtain M_t given in their assumption A5(iii). Finally, their assumption A6(iv) is not necessary for the weak convergence of the LR test. Instead, after letting

$$r_t(p) \equiv \frac{f(Y_t;p)}{f(Y_t;p_*)};\;\; r_t^{(1)}(p) \equiv \frac{1}{f(Y_t;p_*)}\frac{\partial}{\partial p}f(Y_t;p);\;\; \text{and}\;\; r_t^{(2)}(p) \equiv \frac{1}{f(Y_t;p_*)}\frac{\partial^2}{\partial p^2}f(Y_t;p),$$

we show that for every $\epsilon > 0$, $\lambda_{\min}\{C^{(u)}(p,p)\} > 0$ and $\lambda_{\min}\{C^{(u)}(p,p)\} < \infty$ uniformly in $p \in \{p \in \mathbb{P} : |p - p_*| \geq \epsilon\}$, where

$$C^{(u)}(p,p') \equiv \begin{bmatrix} E[r_t^{(2)}(p_*)^2] & -E[r_t^{(2)}(p_*)r_t(p)] & E[r_t^{(2)}(p_*)r_t^{(1)}(p)] \\ -E[r_t^{(2)}(p_*)r_t(p')] & E[(1-r_t(p))(1-r_t(p'))] & E[(1-r_t(p))r_t^{(1)}(p')] \\ E[r_t^{(2)}(p_*)r_t^{(1)}(p')] & E[(1-r_t(p'))r_t^{(1)}(p)] & E[r_t^{(1)}(p)r_t^{(1)}(p')] \end{bmatrix}.$$

For this, we let $C_1^{(u)}(p) \equiv E[r_t^{(2)}(p_*)^2]$,

$$C_2^{(u)}(p) \equiv \begin{bmatrix} E[r_t^{(2)}(p_*)^2] & -E[r_t^{(2)}(p_*)r_t(p)] \\ -E[r_t^{(2)}(p_*)r_t(p)] & E[(1-r_t(p))(1-r_t(p))] \end{bmatrix},$$

and note that

$$\det[C_1^{(u)}(p)] = \{(1-p_*)^2p_*^4\}^{-1}4(2-p_*);$$

$$\det[C_2^{(u)}(p)] = \frac{4\{4p^3p_*^3 - p_*^4 + 2pp_*^3(1+p_*) - p^2p_*^3(5+p_*) - p^4[2+p_*(-5+4p_*)]\}}{p^4p_*^5(p-p_*)^{-2}(1-p)^4[(-2+p)p+p_*]};$$

and $\det[C^{(u)}(p)]$ is identical to

$$\frac{4\{4p_*^4 - 4pp_*^3(3+p_*) + p^2p_*^2[6+17p_*+p_*^2] + 2p^3p_*[2-7p_*-3p_*^2] + p^4[2-7p_*+9p_*^2]\}}{p^6p_*^6(p-p_*)^{-6}(1-p_*)^4[(-2+p)p+p_*]^4}.$$

All these are strictly greater than zero and less than infinity uniformly on $\{p \in \mathbb{P} : |p - p_*| \geq \epsilon\}$. Thus, the LR statistic satisfies the conditions sufficient for the stated weak convergence.

Next, we derive the covariance structure (7). For this, we use the formula given in lemma 1(b) of Cho and White (2007). For each p and p',

$$E[r_t(p)r_t(p')] - 1 = \frac{(p - p_*)(p' - p_*)}{p_*\{(1 - p_*) - (1 - p)(1 - p')\}},$$

$$E[r_t(p)s_t] = \frac{p_* - p}{pp_*(1 - p_*)}, \quad \text{and} \quad E[s_t s_t'] = \frac{1}{p_*^2(1 - p_*)},$$

where $s_t = (Y_t p_* - 1)/[p_*(1 - p_*)]$. Also, $E[r_t(p)r_t(p')]$ cannot be computed without having $1 - p < \sqrt{1 - p_*}$, which is assumed by our condition that $\underline{p} > 1 - \sqrt{1 - p_*}$. Plugging these into $\rho(p, p') \equiv E[r_t(p)r_t(p')] - 1 - E[r_t(p)s_t]'\{E[s_t s_t']\}^{-1}E[r_t(p')s_t]$ yields

$$\rho(\alpha, \alpha') \equiv \frac{(p - p_*)^2(p' - p_*)^2}{pp'p_*(1 - p_*)\{(1 - p_*) - (1 - p)(1 - p')\}},$$

which implies the standardized covariance structure

$$\frac{\rho(p, p')}{\sqrt{\rho(p, p)}\sqrt{\rho(p', p')}} = \frac{[(1 - p_*) - (1 - p)^2]^{1/2}[(1 - p_*) - (1 - p')^2]^{1/2}}{(1 - p_*) - (1 - p)(1 - p')}.$$

Note that this is now identical to (7). This completes the proof. ■

Proof of Theorem 2: (i) We verify that for each p, $p' \in \mathbb{P}$, $E[\mathcal{G}(p)\mathcal{G}(p')] = E[\bar{\mathcal{G}}(p)\bar{\mathcal{G}}(p')]$.

Note that

$$
\begin{aligned}
E[\bar{\mathcal{G}}(p)\bar{\mathcal{G}}(p')] &= \sum_{k=0}^{\infty} \gamma_k(p)\gamma_k(p')E[Z_k^2] \\
&= \left[\frac{(1-p_*)-(1-p)^2}{1-p_*}\right]^{1/2} \left[\frac{(1-p_*)-(1-p')^2}{1-p_*}\right]^{1/2} \left[\frac{1-p_*}{(1-p_*)-(1-p)(1-p')}\right] \\
&= \frac{[(1-p_*)-(1-p)^2]^{1/2}[(1-p_*)-(1-p')^2]^{1/2}}{(1-p_*)-(1-p)(1-p')} = E[\mathcal{G}(p)\mathcal{G}(p')],
\end{aligned}
$$

where $\gamma_k(p) \equiv \left\{\frac{(1-p_*)-(1-p)^2}{1-p_*}\right\}^{1/2} \left\{\frac{1-p}{\sqrt{1-p_*}}\right\}^k$ for each k and p, so that $\bar{\mathcal{G}}(p) = \sum_{k=0}^{\infty} \gamma_k(p) Z_k$, and the second equality follows from the fact that

$$
\sum_{k=0}^{\infty} \left\{\frac{(1-p)(1-p')}{1-p_*}\right\}^k = \frac{1-p_*}{(1-p_*)-(1-p)(1-p')}
$$

using the formula for infinite sum of geometric series and $\inf \mathbb{P} > 1-\sqrt{1-p_*}$. From the Gaussian process property, their identical covariance structure shows their equivalence in distribution. ■

Proof of Equation (9): From that $(1-\underline{p})^2(1-p_*)/(1-\bar{p})^2 < 1$, we note that $1-\sqrt{1-p_*}(1-\underline{p}) > \bar{p} > p_*$, where the last inequality follows from the given assumption. Therefore, $\underline{p} > 1-\sqrt{1-p_*}$ as desired. ■

Table 1: Asymptotic Critical Values of the LR Test Statistics
Number of Repetitions: 10,000
DGP: $Y_t \sim$ IID $\mathbb{G}(1/2)$
Model: $Y_t \sim$ IID $\pi\mathbb{G}(p_1) + (1-\pi)\mathbb{G}(p_2)$

Nominal Level $\setminus \mathbb{P}$	$\mathbb{P}_1 \equiv [0.3965, 0.5732]$	$\mathbb{P}_2 \equiv [0.4000, 0.5500]$
1.00 %	6.1010	5.9416
2.50 %	4.4735	4.4145
5.00 %	3.3128	3.2366
7.50 %	2.6194	2.5720
10.0 %	2.1164	2.0974

Table 2: Levels of the LR Test
Number of Repetitions: 2,000
DGP: $Y_t \sim$ IID $\mathbb{G}(1/2)$
Model: $Y_t \sim$ IID $\pi\mathbb{G}(p_1) + (1-\pi)\mathbb{G}(p_2)$

Statistics	Levels $\setminus n$	100	500	1,000	5,000
$LR_n(\mathbb{P}_1)$	5%	0.25	2.15	2.85	4.00
	10%	1.40	5.50	6.35	8.15
$LR_n(\widehat{\mathbb{P}}_1)$	5%	0.25	2.15	2.90	3.75
	10%	1.40	5.40	6.10	7.80
$LR_n^*(\mathbb{P}_1)$	5%	6.30	5.05	4.75	4.71
	10%	12.25	9.15	9.35	9.40
$LR_n(\mathbb{P}_2)$	5%	0.20	1.20	2.60	4.00
	10%	0.50	4.75	6.80	7.75
$LR_n(\widehat{\mathbb{P}}_2)$	5%	0.20	1.30	2.55	4.00
	10%	0.50	4.55	6.50	7.40
$LR_n^*(\mathbb{P}_2)$	5%	8.10	5.60	5.70	5.55
	10%	13.75	10.85	10.10	9.95

Notes: 1. When $n =$ 50,000 and the level = 5%, the empirical levels are 4.95, 4.75, 4.70, and 4.65 for $LR_n(\mathbb{P}_1)$, $LR_n(\widehat{\mathbb{P}}_1)$, $LR_n(\mathbb{P}_2)$, and $LR_n(\widehat{\mathbb{P}}_2)$ respectively.
2. When $n =$ 50,000 and the level = 10%, the empirical levels are 9.70, 9.50, 9.90, and 9.70 for $LR_n(\mathbb{P}_1)$, $LR_n(\widehat{\mathbb{P}}_1)$, $LR_n(\mathbb{P}_2)$, and $LR_n(\widehat{\mathbb{P}}_2)$ respectively.

Table 3: Power of the LR Test (Nominal Level: 5%)
Number of Repetitions: 2,000
Model: $Y_t \sim$ IID $\pi\mathbb{G}(p_1) + (1-\pi)\mathbb{G}(p_2)$

DGP: $Y_t \sim$ IID $0.5\mathbb{G}(0.3) + 0.5\mathbb{G}(0.7)$						
Statistics \ Sample Size	100	200	300	400	500	600
$LR_n(\mathbb{P}_1)$	11.35	46.25	73.80	88.50	96.05	98.10
$LR_n(\widehat{\mathbb{P}}_1)$	12.35	47.85	74.35	89.15	96.20	98.25
$LR_n^*(\mathbb{P}_1)$	74.25	95.25	99.05	100.0	100.0	100.0
$LR_n(\mathbb{P}_2)$	4.70	33.75	61.30	77.20	88.90	94.35
$LR_n(\widehat{\mathbb{P}}_2)$	5.05	35.00	62.25	78.80	89.05	94.35
$LR_n^*(\mathbb{P}_2)$	80.00	96.60	99.85	100.0	100.0	100.0
DGP: $(Y_t \mid X_t = x) \sim$ IID $\mathbb{G}(x)$, $X_t \equiv \sqrt{U_t}$, and $U_t \sim$ IID $U(0,1)$						
Statistics \ Sample Size	25	50	100	150	200	250
$LR_n(\mathbb{P}_1)$	7.40	33.30	67.55	83.70	92.35	95.55
$LR_n(\widehat{\mathbb{P}}_1)$	7.60	32.90	66.60	82.55	92.00	95.15
$LR_n^*(\mathbb{P}_1)$	44.15	65.10	82.25	91.25	95.95	97.80
$LR_n(\mathbb{P}_2)$	4.75	25.50	60.90	74.70	84.00	90.60
$LR_n(\widehat{\mathbb{P}}_2)$	4.90	26.00	60.35	74.10	83.20	90.00
$LR_n^*(\mathbb{P}_2)$	44.55	59.15	76.55	84.20	91.30	94.60

Table 4: Strike Data Analysis Using the LR Test
Null Model: $Y_t \sim$ IID $\mathbb{G}(p)$
Unrestricted Model: $Y_t \sim$ IID $\pi\mathbb{G}(p_1) + (1-\pi)\mathbb{G}(p_2)$

Sample Size		560
Null Model		
	$\hat{p}_{on}$	0.0230
		(0.0010)
Unrestricted Model		
	$\hat{\pi}_n$	0.0200
	$\hat{p}_{1n}$	0.0320
	$\hat{p}_{2n}$	0.6426
LR Statistics		2.6564
p-values		4.05%
Size adj. p-values by bootstrapping		[0.40%, 3.65%]
Selected $\mathbb{P}$		[0.020, 0.032]
GOF Statistics		1.0869
Size adj. p-values by bootstrapping		[13.93%, 14.00%]

Note: Figures in the parentheses are standard errors, and the number of repetitions for generating the asymptotic null distribution is 10,000. The range of size distortion adjusted p-value is obtained by generating LR statistics using the grid parameters in 95% confidence interval of p_* by repeating 2,000 times. The range of p-value for the GOF statistic is also obtained in the same way. Their number of repetitions is 10,000.

Figure 1: Empirical and Asymptotic Distribution Functions of the LR Statistic