Overview

The objective of the application is to identify the parametric distribution (from a given list of standard distributions) that produces the best match for the underlying sample data. The sample data is assumed to be stationary.

Implementation

The application is designed as follows.

1. **Inputs**. The input data is a $k \times n$ matrix X that represents k estimated samples (where each sample size equals n). The analysis is performed separately for each individual sample;
2. **Calculator**. The distribution estimation is performed by implementing the steps below.
   a. For each sample from the list of k samples and each distribution from the list of standard distributions, estimate the parameters of the distribution;
   b. For each sample and each estimated parametric distribution, construct Kolmogorov-Smirnov and chi-squared statistics (and corresponding p-values), which is used to test the sample distribution. The Kolmogorov-Smirnov and chi—squared test details and related statistics estimation details are described below.
   c. Identify the distribution with the smallest chi-squared statistics and match it to the related sample.
   d. Estimate the histogram for each sample and compare it to the matched distribution.
3. **Output**. The output is represented by the following objects.
   a. List of k **histogram** objects constructed for each of k samples.
   b. Table of Kolmogorov-Smirnov and chi-squared **statistics** and **p-values** constructed for each sample and each standard distribution. The distribution with the highest p-value is matched to the related sample.
   c. The list of **distributions** matched to each related sample based on the estimated p-values.
   d. The table with the **distribution parameters** estimated for each sample and each standard distribution.

List of standard distributions

1. **Normal** distribution $N(\mu, \sigma)$ where μ is the mean and σ is the standard deviation of the distribution;
2. **Log-Normal** distribution $lnN(\mu, \sigma)$where μ is the mean and σ is the standard deviation of the natural logarithm of the distribution. The log-normal distribution has support $\{x > 0\}$ so that the logarithm of the values is distributed normally with parameters μ and σ;
3. **Beta** distribution: $p(x) = \frac{x^{a-1} \times (1-x)^{\beta-1}}{B(a,\beta)}$; where $B(a,\beta) = \frac{\Gamma(a)\Gamma(\beta)}{\Gamma(a+\beta)}$. The support of the distribution is $x \in [0,1]$.
4. **Cauchy** distribution: $p(x) = \frac{1}{\pi\gamma\left(1+\frac{(x-x_0)^2}{\gamma}\right)}$;

5. **Chi-squared** distribution: $p(x) = \frac{x^{\frac{k}{2}-1} \times e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \times \Gamma\left(\frac{k}{2}\right)}$;

6. **Exponential** distribution: $p(x) = \lambda e^{-\lambda x}$;

7. **F** (Fisher-Snedecor) distriution: $p(x) = \frac{1}{x \times B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \times \sqrt{\frac{(d_1 x)^{d_1} \times {d_2}^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}$;

8. **Gamma** distribution: $p(x) = \frac{1}{\Gamma(a)} \times \beta^a \times x^{a-1} e^{-\beta x}$;

9. **Geometric** distribution: $p(k) = p \times (1-p)^k$;

10. **Laplace** distribution: $p(x) = \frac{1}{2b} \times e^{-\frac{|x-\mu|}{b}}$;

11. **Logistic** distribution: $p(x) = \frac{q(x)}{s \times (1+q(x))^2}$, where $q(x) = e^{-\frac{(x-\mu)}{s}}$;

12. **Poisson** distribution: $p(k) = \frac{1}{k!} \times \lambda^k \times e^{-\lambda}$;

13. **T** (Student) distribution: $p(x) = \frac{\Gamma\left(\frac{1+\nu}{2}\right)}{\sqrt{\pi\nu} \times \Gamma\left(\frac{\nu}{2}\right)} \times \left(1 + \frac{x^2}{\nu}\right)^{\frac{1+\nu}{2}}$;

14. **Triangular** distribution: $p(x) = \frac{2 \times (x-a)}{D}$ for $a \leq x < c$ and $p(x) = \frac{2 \times (b-x)}{D}$ for $c \leq x \leq b$, where $D = (b-a) \times (b-c)$;

15. **Uniform** distribution: $p(x) = \frac{1}{b-a}$ for $a \leq x \leq b$;

16. **Weibull** distribution: $p(x) = \frac{k}{\lambda} \times \left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}$, $x \geq 0$.

Distribution parameters

1. **Normal**: $\mu = E[x]$, $\sigma = stdev[x]$;
2. **Log-Normal**: $\mu = E[\ln x]$, $\sigma = stdev[\ln x]$;
3. **Beta**: $\frac{\alpha}{\alpha+\beta} = E[x]$; $\frac{\alpha-1}{\alpha+\beta-2} = Mode$; $\Rightarrow \alpha = \max\left[0, \frac{1-2 \times Mode}{1-\frac{Mode}{\mu}}\right]$; $\beta = \max\left[0, \alpha \times \left(\frac{1}{\mu} - 1\right)\right]$
4. **Cauchy**: $x_0 = Median\ (= Mode)$; $x_0 + \gamma \times \tan\left[\pi \times \left(F - \frac{1}{2}\right)\right] = quantile$;
5. **Chi-square**: $k = \max[0, E[x]]$;
6. **Exponential**: $\lambda = \max\left[0, \frac{1}{E[x]}\right]$;
7. **F**: $\frac{df_1}{df_1 - 2} = E[x]$; $\frac{df_2 - 2}{df_2} \times \frac{df_1}{df_1 + 2} = Mode[x]$; $\Rightarrow$

   $df_1 = \max\left[1, \frac{2 \times E[x]}{E[x]-1}\right]$; $df_2 = \max\left[1, \frac{2}{1 - Mode[x] \times \frac{df_1 + 2}{df_1}}\right]$;

8. **Gamma**: $\frac{\alpha}{\beta} = E[x]$; $\frac{\alpha}{\beta^2} = var[x]$; $\Rightarrow \beta = \frac{\max[0,E[x]]}{var[x]}$; $\alpha = \frac{E^2[x]}{var[x]}$;
9. **Geometric**: $p = \min\left[1, \max\left[0, \frac{1}{E[x]}\right]\right]$;
10. **Laplace**: $\mu = E[x]$; $2b^2 = var[x]$; $\Rightarrow\ b = \frac{stdev[x]}{\sqrt{2}}$;
11. **Logistic**: $\mu = E[x]$; $\frac{s^2\pi^2}{3} = var[x]$; $\Rightarrow s = \sqrt{3} \times \frac{stdev[x]}{\pi}$;
12. **Poisson**: $\lambda = \max[0, E[x]]$;
13. **T** (Student): $\frac{\nu}{\nu-2} = var[x]$; $\Rightarrow \nu = \max\left[1{,}2 \times \frac{var[x]}{var[x]-1}\right]$;
14. **Triangular**: $a = \min x$; $b = \max x$; $\frac{a+b+c}{3} = E[x]$; $\Rightarrow c = 3 \times E[x] - a - b$;
15. **Uniform**: $a = \min x$; $b = \max x$;
16. **Weibull**: $\lambda \times \Gamma\left(1 + \frac{1}{k}\right) = E[x]$; $\lambda \times (\ln 2)^{\frac{1}{k}} = Mode[x]$, where λ is scale parameter and k is shape parameter.

Application of different distributions

Testing for the distribution

Histogram estimation