Usage of parameter estimation tool for offset mixture of Gaussian distributions and Laplace distributions

Masato Fujita *

October 11, 2013

1 Introduction

The enclosed software implements the algorithm discussed in Reference [1]. This paper summarizes the usage of this software.

We first define our terminology. We will consider the sum of a discrete random variable with a random variable following the finite mixture of an arbitrary number of 1-dimensional Gaussian distributions and Laplace distributions with zero mean. We call it an *offset mixture distribution* (of Gaussian distributions and Laplace distributions) in this paper. The discrete random variable corresponds to the offset of the aircraft. On the other hand, the finite mixture distribution represents the lateral navigation performance of the aircraft.

The probability density function of an offset mixture distribution is given by the following equation.

$$p(x|\boldsymbol{\omega},\boldsymbol{\pi},\boldsymbol{\sigma},\boldsymbol{\lambda}) = \sum_{l=1}^{L} \omega_l \left(\sum_{k=1}^{m} \pi_k \mathcal{N}(x-o_l,\sigma_k) + \sum_{k=1}^{n} \pi_{m+k} \mathcal{D}(x-o_l,\lambda_k) \right) \quad (1)$$

Here, $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$ are finite non-negative numbers satisfying $\sum_{k=1}^{K} \pi_k = 1$, where $K = m+n$. $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_m)$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ denote positive numbers. $\boldsymbol{o} = (o_1, o_2, \ldots, o_L)$ are real numbers. $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_L)$ are non-negative numbers satisfying $\sum_{l=1}^{L} \omega_l = 1$. We call ω_l an *offset mixing coefficient* in this paper. The notation $\mathcal{N}(x,\sigma)$ denotes the probability density function of a 1-dimensional Gaussian distribution with zero mean, and $\mathcal{D}(x,\lambda)$ denotes the probability density function of a Laplace distribution with zero mean.

The functions of the software are as follows:

*Electronic Navigation Research Institute

1. Generate pseudo-random numbers following the given offset mixture distribution model.

2. Estimate the parameter values of distribution model from the given data set. The EM algorithm and the variatioal Bayesain methods are both implemented.

3. Estimate the applied offset when a offset mixture distribution model is given.

We discussed about offset mixture distributions, but this software can handle non-offset mixture distributions, namely normal mixture distributions as well. Section 2 treats the case where offset is not allowed. Section 3 describes the case where offset is allowed.

The source codes enclosed eith the software also contain some useful classes. The program can be used as a library as well. See the eclosed Javadoc for the structure of this codes.

2 Offset is not allowed.

We discuss about the case where offset is not allowed. The probability distribution model considered in this section is given by the following equation. It is a spacial case given in the equation (2), namely the case where $L = 1$ and $\boldsymbol{\omega} = \{0\}$.

$$p(x|\boldsymbol{\pi}, \boldsymbol{\sigma}, \boldsymbol{\lambda}) = \sum_{k=1}^{m} \pi_k \mathcal{N}(x, \sigma_k) + \sum_{k=1}^{n} \pi_{m+k} \mathcal{D}(x, \lambda_k) \tag{2}$$

2.1 Generating psudo-random numbers.

The software can generate psudo-random numbers. The usage of this fuction is as folllows:

Usage

java -jar OffLatDistEst.jar u nde rg <Path to distribution definition file> <Number of generated random numbers> <Path to output file>

The $1,000$ samples following the distrbution defined in the file 'tmp/NDE.properties' are generated in the folllowing example. The outputs are recorded in the file 'tmp/NDEdata.csv.'

Example

java -jar OffLatDistEst.jar u nde rg tmp/NDE.properties 1000 tmp/NDEdata.csv

The format employed in the distribution definition file is the 'properties' format intoroduced in Reference [2]. Each line in a .properties file normally

stores a single property. Several formats are possible for each line, including key=value, key = value, key:value, and key value. The description $m = 1$ means the value of the variable 'm' is equal to one.

The variable m denotes the number of Gaussian components of mixture distribution. The variable n denotes the number of Laplace components defined in the mixture distribution. The variable 'pi_Ni' is the i-th mixing coefficient of the Gaussian component. Here, i denotes the number. The variable 'pi_DEi' is the i-th mixing coefficient of the Laplace component. The variables 'sigma$i$' and 'lambda$i$' are the standard deviation of the i-th Gaussian component and the scale parameter of the i-th Laplace components, respectively. The following is an example of distribution definition. The following example means that $m = 1$, $n = 1$, $\pi_1 = 0.95$, $\pi_2 = 0.05$, $\sigma_1 = 1$ and $\lambda_1 = 10$ in the equation (2).

Example

```
m=1
n=1
pi_N1=0.95
sigma1=1
pi_DE1=0.05
lambda1=10
```

The output of the program in the above example situation is given in the following format. Each line stores a single sample. The first number in the line is the generated random number and the second one shows that the number is generated from which component. The value 0 means the first component of the mixture. The value 1 means the second component. The first component corresponds to the first Gaussian component and the second componet corresponds to the first Laplace component.

Example

```
-0.6748209837410951,0.0
-0.4277818997347839,0.0
-28.151872997991628,1.0
0.8039747832470926,0.0
-0.2489646290545688,0.0
```

2.2 Estimating the model parameters.

The following command is used when the model parameters are estimated. from the data files

Usage

java -jar OffLatDistEst.jar u nde em/vb <Path to initial condition setting file> <Path to data file> <Path to output file>

The description 'em/vb' in the above usage means 'em' or 'vb.' The EM al-

gorithm is used when 'em' is chosen and the variational Bayesian method is emplyed when 'vb' is chosen.

The initial condition is saved in 'tmp/NDEEM.properties,' and the data file is 'tmp/NDEdata.csv' in the following example. The EM algorithm is employed in this example.

Example

```
java -jar OffLatDistEst.jar u nde em tmp/NDEEM.properties
tmp/NDEdata.csv tmp/NDEEM_out.properties
```

The data file is assumed to follow the csv format. The first number in each line is stored as a sample datum, and the remainings are ignored. Both input and output files employ the properties format. The variables which should be defined depends on the employed estimation algorithm. We first consider the EM algorithm case, then treat the variational Bayesian methodcase afterward.

2.2.1 Initial Condition/Output file format of EM algorithm

In the EM algorithm, the initial values of model parameters in the equation (2) should be given. The format of the initial condition setting file is the same as the format of the distribution definition file defined in Section 2.1. The format of the output file is also almost same as the distribution definition file. A new variable 'logLiklihood' is defined in the output file other than the variables defined in the distribution definition file. It is the log likelihood value.

2.2.2 Initial Condition/Output file format of variatioal Bayesian method

When the variational Bayesian methodlogy is employed, the prior distribution of the model parameters should be defined.

We define the sequence of real numbers $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_K)$ by the following equation. Set $K = m + n$.

$$\eta_k = \begin{cases} \dfrac{1}{\sigma_k^2} & (1 \leq k \leq m) \\ \dfrac{1}{\lambda_{k-m}} & (m+1 \leq k \leq K) \end{cases} \tag{3}$$

This paper introduces the variational Bayesian methodology for finding an approximation of the posterior distribution of $\boldsymbol{\pi}$ and $\boldsymbol{\eta}$. The prior distribution of $\boldsymbol{\pi}$ is a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha_0})$, where $\boldsymbol{\alpha_0} = (\alpha_{0,1}, \alpha_{0,2}, \ldots, \alpha_{0,K})$, and the prior distribution of η_k is a Gamma distribution $\mathrm{Gam}(\eta_k|a_{0,k}, b_{0,k})$. The posterior distributions of $\boldsymbol{\pi}$ and η_k are again Dirichlet and Gamma distributions.

The initial condition file defines the parameters of prior distributions. The variables m and n are the numbers of Gaussian and Laplace components, respectively. The 'alpha$i$' is the parameter of Dirichlet distribution. The 'ai' and 'bi' are the parameters of i-th Gamma distribution. The parameters are defined

as $m = 1$, $n = 1$, $\alpha_{0,1} = 0.95$, $\alpha_{0,2} = 0.05$, $a_{0,1} = 1$, $b_{0,1} = 1$, $a_{0,2} = 10$ and $b_{0,2} = 1$ in the following example.

Example

```
m=1
n=1
alpha1=0.95
a1=1
b1=1
alpha2=0.05
a2=10
b2=1
```

The format of the output file is similiar to that of the initial condition file. The output file gives the parameters of the posterior distributions. It also contains the model parameters of the mixture distribution defined in the equation (2). These are estimated by means of the MAP estimation. The format for defining the model parameters of the mixture distribution is the same as the format of the distribution definition file defined in Section 2.1. A new variable 'lowerbound' is the value defined in Reference [1].

3 Offset is allowed.

We consideer the case where offset is allowed. The distribution model considered in this section is given in the equation (1).

3.1 Generating psudo-random numbers.

The software can generate psudo-random numbers. The usage of this fuction is as folllows:

Usage

java -jar OffLatDistEst.jar u **onde** rg <Path to distribution definition file> <Number of generated random numbers> <Path to output file>

The second argument was 'nde' in the case where offset is not allowed. However, the magic word 'nde' is replaced by 'onde' in this usage.

The $1,000$ samples following the distrbution defined in the file 'tmp/ONDE.properties' are generated in the folllowing example. The outputs are recorded in the file 'tmp/ONDEdata.csv.'

Example

java -jar OffLatDistEst.jar u onde rg tmp/ONDE.properties 1000 tmp/ONDEdata.csv

The format of the input file is almost the same as the case where offset is not allowed. The new variables 'L', 'offset$i$' and 'omega$i$' are introduced in this format. The variable 'L' corresponds to the number of possible offset which is denoted by L in the equation (1). The variables 'offset$i$' and 'omega$i$' are o_i and ω_i in the equation (1). The following example means that $m = 1$, $n = 1$, $\pi_1 = 0.95$, $\pi_2 = 0.05$, $\sigma_1 = 1$, $\lambda_1 = 10$ $L = 3$, $o_1 = 0$, $o_2 = 100$, $o_3 = 200$, $\omega_1 = 0.6$, $\omega_2 = 0.3$ and $\omega_3 = 0.1$ in the equation (1).

Example

```
m=1
n=1
pi_N1=0.95
sigma1=1
pi_DE1=0.05
lambda1=10
L=3
offset1=0
offset2=100
offset3=200
omega1=0.6
omega2=0.3
omega3=0.1
```

The output of the program in the above example situation is given in the following format. Each line stores a single sample just like the non-offset case. The first two numbers in the line are the same as the non-offset case. The last number is the applied offset.

Example

```
98.7234011320356,0.0,100.0
99.6129842355946,0.0,100.0
99.75406291903305,0.0,100.0
101.20332639212567,0.0,100.0
-1.8654843885172798,0.0,0.0
0.32696081002411587,0.0,0.0
```

3.2 Estimating the model parameters.

The following command is used when the model parameters are estimated. from the data files

Usage

java -jar OffLatDistEst.jar u **onde** em/vb <Path to initial condition setting file> <Path to data file> <Path to output file>

The description 'em/vb' in the above usage means 'em' or 'vb.' The EM algorithm is used when 'em' is chosen and the variational Bayesian method is emplyed when 'vb' is chosen. Just like the case of random sample generation, the second argument becomes 'onde' and the remaining is the same as the non-offset case.

The initial condition is saved in 'tmp/ONDEEM.properties,' and the data file is 'tmp/ONDEdata.csv' in the following example. The EM algorithm is employed in this example.

Example

java -jar OffLatDistEst.jar u onde em tmp/ONDEEM.properties tmp/ONDEdata.csv tmp/ONDEEM_out.properties

3.2.1 Initial Condition/Output file format of EM algorithm

The format of the initial condition setting file is the same as the format of the distribution definition file defined in Section 3.1. The format of the output file is also almost same as the distribution definition file with the variable 'logLiklihood.'

3.2.2 Initial Condition/Output file format of variatioal Bayesian method

References

[1] Fujita, M. (2013). Estimation of Navigation Performance and Offset by the EM Algorithm and the Variational Bayesian Methods, *Advances and applications in statistics*, 35(1), 1–27.

[2] `http://docs.oracle.com/cd/E23095_01/Platform.93/ATGProgGuide/html/s0204propertiesfileformat01.html`