MIXTURE DECOMPOSITION OF CENSORED PARETO-II DENSITY WITH NORMAL ONE BY EM-ALGORITHM

V.S. Stepanov
CEMI, www.cemi.rssi.ru
Moscow, Russia
e-mail: stepanov@cemi.rssi.ru

Abstract

We tackle a problem of decomposition of mixture of the censored Pareto-II density with Gaussian one having its mean value nearby the Pareto density mode. It was used the EM-algorithm that has been applied to Monte-Carlo sample by the mixture (1). The model can be useful to detect a Pareto-type signal under Gaussian noise in bioinformatics, economy.

1 Gene Expression and Its Measurement

According to the traditional concept accepted in molecular genetics [1, 2], the DNA molecule is a matrix for RNA synthesis, and RNA - a matrix for synthesis of proteins in a cell. In turn, proteins is a stuff of which organisms are constructed and which provides their operation. The information on protein as for a way of its synthesis is concluded in a piece of DNA (i.e. in a site of the molecule stranded like a spiral staircase). Rungs of the last serve pairs of nitrogen bases (*A, T, G* and *C*). The piece of DNA coding certain protein is a gene.

The gene finds out itself (express) during synthesis of protein. The gene expression generates a primary series of amino acids of the corresponding protein. The gene expression is a two-phase process which intermediate product is messenger RNA (mRNA). On the 1st stage of "transcription" there is a synthesis of a mRNA molecule and on the 2nd stage of "translation" the amino acids chain of the protein is synthesized on the information stored into the mRNA. The chain further strands into 3-D space and a functional protein will result.

The major property of a DNA molecule is the principle of complementary nature for the nitrogen bases. So, the basis A by one circuit of DNA will couple only with T (complement to A) in other circuit, and G will couple only with C. The RNA molecule possesses with similar property also.

A level of gene expression is measured by number of the mRNA molecules, created by a cell; usually, the more mRNA replicas the cell makes, the more protein replicas creates. Therefore quantities of various mRNA in the specimen can indirectly specify types and quantities of proteins presented at a cell. A molecule mRNA often is named as a *transcript*.

2 DNA-chips for Gene Expression Measurement

In practice, an expression level is measured by various methods, such as SAGE, GeneChip, etc. Since the 1990-th, the biological microchips with probes for the DNA analysis, or DNA-chips, are applying to gene researches at various institutes, such as EIMB (*www.eimb.ru*), NIH (*www.nih.gov*) , etc. There are some versions of the chips - all of them consist of one string of a DNA molecule or its fragments (probes) which contact with a substrate having sizes as a thumbprint.

DNA-chips work on a property of the complementary nature of DNA and/or RNA. They can simultaneously trace tens thousand reactions of the base-pairing on a chip. It gains because each kind of a probe lays into a chip cell. The last cells form the structure similar to the hugest chessboard. DNA or RNA molecules lied into cells contain fluorescent labels. DNA (or RNA) is distinguished by a scanner with confocal microscope after their linkage with the probes in separate cells. After the data processing on computer, one can see result on the monitor where each cell on the chip is painted by individual color.

Usually one takes a pool of cells for particular organism, which influenced to some external factors. Then the expression level is measured for particular gene as a number transcripts, falling on a cell. At use of *hybridization* Ц when two various one-chained segments of a molecule of DNA with partially complement series of the bases couple, Ц intensity of a luminescence of this or that micro-point (cell of the DNA-chip) in a logarithmic scale is connected with number of mRNA transcripts, falling on a cell, under the linear law.

In the described way it is possible to receive rather easily and quickly empirical distribution of expression levels for all genes, probes for which have been placed on the chip earlier. As a result, a gene expression profile, informing about a place and time of activation of genes, turns out. After the profile interpretation, one can judge on some gene functions in a cell. For example, it is possible to try to find out, which genes have joined or switched off in a cell at influence on it by poison, for example, nicotine.

For the DNA-chips produced by GeneChip method (*www.affymetrix.com*) where the hybridization has been used, an information noise has always arised into the expression data. Its filtration is very important for revealing the genes that have been seldom activated in a cell, and therefore the useful signal is weak in a corresponding cell of the chip, In other words, the signal in a cell of the DNA-chip is commensurable on its brightness with noise level. Such genes are representing the greatest scientific interest because their functions in a cell are poorly studied in state-of-the-art of molecular biology.

The next model of a mixture of a signal with gene expression data into a cell of DNA-chip and Gaussian noise could be useful for the noise filtration. Besides, the model can be applied in economy to distinguish two groups of people on their income per capita in month. The groups are the Russian oligarchs having from $2000 up to $200,000 and the "middle class", described with the help of normal density.

3 Mixture Model for Gene Expression Data

Mixture of the normal density $f(x; m, \sigma^2)$ and the Pareto-2 density is :

$$F(x) = \alpha\ f(x; \mu, \sigma^2)\ +\ (1-\alpha)\ \Psi(x; b, k), \tag{1}$$

where $\alpha \in (0,1)$ - share of normal component, x - a value of an expression level from DNA-chip; $\mu, \sigma^2,\ b,\ k$ - the unknown parameters estimated on a sample, as well as α.

Under $x \in (c, R),\ k > 0,\ b > -c,\ c > 0$, the censored Pareto-II density is

$$\Psi(x) = \Psi(x; b, k) = a(b,k) \cdot [(c+b)/(x+b)]^{(k+1)} \tag{2}$$

and $\Psi(x) = 0$ when $x \le c$ or $x \ge R$. The real number $a(b,k)$ into (2) is $a(b,k) = k\ (c+b)^{-1}[1-((c+b)/(R+b))^k]^{-1}$. The term and formula (2) turn out after the next density normalization: $const \cdot \int_c^R (x+b)^{-(k+1)} \cdot dx = 1$.

The random value with (2) has got the probability P to hit into $(z_1, z_2) \subseteq [c, R]$:

$$P = [(c+b)^{-k} - (R+b)^{-k}]^{-1}\ [(z_1+b)^{-k} - (z_2+b)^{-k}]^{-1}. \tag{3}$$

4 Data Modeling by the Monte-Carlo

The library for cells-yeast with twenty thousand mRNA transcripts was analyzed at [3] using SAGE-method. A model of discrete Pareto-type distribution was offered for the expression level over there. The author found out that the library has contained N = 5324 expressed genes. Then the probability P_j was estimated by him that the given gene has been presented in library exactly by j transcripts; $j = 1, 2, \ldots$. Further he adjusted the continuous Pareto-2 distribution (2) to the data with the next numerical values $c = 0.85$, $R = 1000$, $b = -0.66$, $k = 0.715$.

The values were used by us under Monte-Carlo simulation of $N_1 = 5000$ observations with density (2). For this purpose the density (2) was approximated by a step function. For performance of such approximation the probability (3) was used as well as uniformed pseudo-random values within interval (z1, z2). The second sample was generated by normal density with some given values of μ, σ^2; the sample size was $N_2 = 600$ or $N_2 = 900$. Thus the mixture of these two samples was used as an initial data for operation of the EM-algorithm, described below.

5 Mixture Decomposition Using the EM-algorithm

In model (1) all parameters were estimated by maximization of logarithm of likelihood ratio $L(\Theta)$ using the EM algorithm [4]. In our task $k < 3$ and $-c < b < 1.5$, therefore the condition of limitation of $L(\Theta)$ is carried out. As is known, after introduction the posteriori probabilities $g_{i,1},\ g_{i,2}$ of class 1 or 2 under given x_i , the solved problem

$$L(\Theta) \to max; \qquad \Theta \equiv (\alpha, \mu, \sigma^2, b, k) \tag{4}$$

will be reduced to the following maximization problems: 1. On unknown α ; 2. On noise parameters $\theta_1 \equiv (\mu, \sigma^2)$; 3. On parameters $\theta_2 \equiv (b, k)$ of Pareto density.

These problems are solved in a two-stage iterative process [4], estimating all over again posteriori probabilities $g_{i,1},\ g_{i,2}$ for a point x_i , and then separately minimizing $Ln(\theta_1)$ and $Ln(\theta_2)$. On a stage of maximization, the values $g_{i,1},\ g_{i,2}$ are estimated again. The point θ_1^* in which maximum of $Ln(\theta_1)$ is reached, is known [4]. We used the robust estimators of μ and σ^2 over here: the exponentially weighted λ-estimate [4,5] or "radical" one [5].

For the decision of a problem (4), some numerical Newton's algorithm has been constructed. On its step t , the gradient vector $\nabla(\theta)$ and Hess matrix $H(\theta)$ with the 2^{nd} partial derivatives for function $Ln(\theta)$ were used (we denote $\theta \equiv \theta_2$ below). So the estimator of $\theta_{t+1} = (b, k)$ on step $t+1$ was calculated according to the formula

$$\theta_{t+1} = \theta_t - H^{-1}(\theta_t) \cdot \nabla(\theta_t). \tag{5}$$

In our lecture, the decomposition of mixture (1) using the EM-algorithm will be resulted for the data generated by Monte Carlo with the Pareto density (2) with its parameters that are representative for cells-yeast as well as with various values of parameters of noise μ, σ^2: $\mu \cong c$, $\mu = c$ and σ^2.

References

[1] Gibbs W.W.(2004) Shadow Part of Genome. *Scientific American*,**2**, pp. 20-27 .

[2] Friend S.H., Stoughton R.B. (2002) Magic of Microchips. *Scientific American*,**1** , pp. 4-10 (*www.sciam.ru* ; ISSN 0208-0621) .

[3] Kuznetsov V.A. (2001) Analysis of Stochastic Process of Gene Expression in a Single Cell . Proc. 2001 IEEE-EURASIP Workshop on Nonlinear Signals and Image Processing, University of Delaware, USA.

[4] Aivazian S.A. et al. (1989) Applied Statistics. Classification and Reduction of Dimensionality. Finansy i Statistika, Moscow (ISBN 5-279-00054-X).

[5] Shurygin A.M. (2000) Applied Stochastics. Finansy i Statistika, Moscow (ISBN 5-279-02201-2).