UDC 519.2

Statistical Analysis of Data Generated by a Mixture of Two Parametric Distributions

Yu. K. Belyaev*, D. Källberg*, P. Rydén*

* *Department of Mathematics and Mathematical Statistics, Umeå University, SE 901 87 Umeå, Sweden*

Abstract. We introduce a novel approach to estimate the parameters of a two-component mixture distribution. The method combines a grid-based approach with the method of moments and reparametrization. The grid approach enables the use of parallel computing and the method can easily be combined with resampling techniques. We derive a reparametrization for the mixture of two Weibull distributions, and apply the method on gene expression data from one gene and 409 $ER+$ cancer patients.

Keywords: mixture of parametric distributions, point process, consistency, accuracy of estimators, resampling, Palm intensities, responsibilities.

1. Introduction

Novel technologies in medicine and manufacturing industry are generating high-dimensional and complex data which have the potential to provide vital information and knowledge, but statistical analyses remain a bottle neck. In cancer research the expression of thousands of genes are measured, with the objective to search for novel disease subtypes by applying cluster analysis [1, 2]. This requires that the dimension of the problem is reduced through variable selection [1, 2]. We address variable selection in a parametric framework for the case when there are two types of sub-diseases. For each variable we assume that the observations are generated by a two-component mixture distribution, with a set of unknown parameters. This is a well-studied problem that has been addressed more than 100 years, see e.g. [3–5]. Karl Pearson carried out the first attempt to estimate the normal mixture distribution using the method of moments, here we modify this approach and introduce a grid-based method that can be applied for a wide range of distribution families. We introduce a measure that is a function of the responsibilities, which can be easily estimated and used to select informative genes for the cluster analysis.

2. Main section

We consider n_d observations $\mathbf{x}_1^{n_d} = \{x_1, \ldots, x_{n_d}\}$ from a population with two groups of individuals. Let $\mathbf{t}_1^{n_d} = \{t_1, \ldots, t_{n_d}\}$ denote the individual's unobservable group labels (1 or 2) and regard $\{x_j, t_j\}_{j=1,2,\ldots}$ as

observations of the independent and identically distributed (i.i.d.) random variables $\{X_j, T_j\}$, where X_j has the probability density function (p.d.f.)

$$p[x, \boldsymbol{\theta}] = \omega_1 p_1[x, \boldsymbol{\theta}_1] + (1 - \omega_1) p_2[x, \boldsymbol{\theta}_2].$$

Here $p_i[x, \boldsymbol{\theta}_i]$ denotes the p.d.f. of X_j given that $T_j = i$ and ω_1 denotes the proportion of individuals belonging to group 1. Let $\boldsymbol{\theta} = \{\omega_1, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ denote the k unknown parameters of the p.d.f. and suppose that the overall objective is to estimate a parameter that can be expressed as a function of the model parameters, i.e. $\varphi = g(\boldsymbol{\theta})$.

We propose a general solution to the above problem that can easily be combined with a resampling procedure in order to derive a *confidence interval* (CI) for the parameter of interest. In the first stage the k parameters are divided in two groups: l grid parameters $\boldsymbol{\theta}_g$ and $k - l$ free parameters $\boldsymbol{\theta}_f$. Through reparametrization the $(k - l)$ first moments can be expressed as algebraic functions of the parameters, i.e.

$$\mu_1 = f_1(\boldsymbol{\theta}_g, \boldsymbol{\theta}_f), \ldots, \mu_{k-l} = f_{k-l}(\boldsymbol{\theta}_g, \boldsymbol{\theta}_f), \tag{1}$$

where $\mu_m = \mathsf{E}[X_j^m], m = 1, ..., (k - l)$.

If the grid parameters are considered as known and the moments are empirically estimated then the equation system defined in (1) can be solved. We propose a grid-based approach where we for each grid-point estimate the $(k - l)$ free parameters via the reparametrization approach. Finally $\boldsymbol{\theta}$ is estimated with the grid-point estimate $\hat{\boldsymbol{\theta}}$ that maximizes the log-likelihood function (or some other fitness criterion). Below we show how reparametrization can be used for the case when the data is described by a mixture of two Weibull distributions.

Using the theory of point processes [6] we can apply the *Palm intensities* to define the *responsibilities*

$$q_i[x_j, \boldsymbol{\theta}] = \mathsf{P}[T_j = i \mid X_j = x_j] = \frac{\omega_i p_i[x_j, \boldsymbol{\theta}_i]}{\omega_1 p_1[x_j, \boldsymbol{\theta}_1] + (1 - \omega_1) p_2[x_j, \boldsymbol{\theta}_2]},$$

for $i = 1, 2$, and $j = 1, ..., n_d$, and where $\omega_2 = 1 - \omega_1$. Suppose that the responsibilities are used to predict which cancer type the patients have, then the expected number of *correctly classified* (cc) individuals can be expressed as

$$\varphi_{cc} = \sum_{j=1}^{n_d} \Big(q_1[x_j, \boldsymbol{\theta}]^2 + q_2[x_j, \boldsymbol{\theta}]^2 \Big),$$

where $\boldsymbol{\theta}$ are the true parameters. The measure φ_{cc} can be estimated by replacing $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\theta}}$. The estimated measure $\hat{\varphi}_{cc}(\mathbf{x}_1^{n_d})$ can be used for variable selection in a high-dimensional cluster analysis problem.

We now consider the special case when data is described by a mixture of two 2-parameter Weibull distributions ($Mix2W$) where X_j has the probability density function (p.d.f.)

$$p_{2W}[x, \boldsymbol{\theta}_5] = \omega_1 p_W[x, \alpha_1, \beta_1] + (1 - \omega_1) p_W[x, \alpha_2, \beta_2],$$

with

$$p_w[x, \alpha, \beta] = \frac{\beta}{\alpha} \left(\frac{x}{\alpha}\right)^{\beta-1} e^{-(x/\alpha)^\beta}, \quad x \geq 0,$$

where $\alpha > 0$ and $\beta > 0$ denote the scale and shape parameters, respectively.

The first three moments of X_j are

$$\mu_1 = \omega_1 \alpha_1 g_{11} + (1 - \omega_1) \alpha_2 g_{21}, \tag{2}$$

$$\mu_2 = \omega_1 \alpha_1^2 g_{12} + (1 - \omega_1) \alpha_2^2 g_{22}, \tag{3}$$

$$\mu_3 = \omega_1 \alpha_1^3 g_{13} + (1 - \omega_1) \alpha_2^3 g_{23}, \tag{4}$$

where $g_{ik} = \Gamma[1 + k/\beta_i]$ and $\Gamma[\cdot]$ is the gamma function, $i = 1, 2, k = 1, 2, 3$. For known values of $\boldsymbol{\psi}_5 = \{\mu_1, \mu_2, \mu_3, \beta_1, \beta_2\}$ this is a determined equation system with unknown parameters $\{\omega_1, \alpha_1, \alpha_2\}$. From (2) it follows that α_2 can be expressed as a function of α_1 and together with (3) we get that α_1 is given by the solutions of a quadratic equation with coefficients depending on ω_1. Inserting these solutions in (4) yields cubic equations that can be solved w.r.t. ω_1 if the moments μ_1, μ_2, μ_3 are replaced with the empirical moments $\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3$, where $\hat{\mu}_k = \frac{1}{n_d} \sum_{j=1}^{n_d} x_j^k$. The local estimator $\tilde{\boldsymbol{\theta}}_5$ of $\boldsymbol{\theta}_5$ given $\{\beta_1, \beta_2\}$ is defined as the relevant solution (i.e. $\alpha_1, \alpha_2 > 0$, and $0 < \omega_1 < 1$) that maximizes the log-likelihood function. Note that for some values $\psi = \{\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \beta_1, \beta_2\}$ there may not exist any relevant solutions of the equation system. Let G denote the regular grid with grid-points $\{\beta_1, \beta_2\}$. The global estimator $\hat{\boldsymbol{\theta}}_5 = \{\hat{\omega}_1, \hat{\alpha}_1, \hat{\beta}_1, \hat{\alpha}_2, \hat{\beta}_2\}$ of $\boldsymbol{\theta}_5$ is given by the local estimator $\tilde{\boldsymbol{\theta}}_5$ that maximizes the log-likelihood function, i.e.

$$\hat{\boldsymbol{\theta}}_5 = \arg\max_{\tilde{\boldsymbol{\theta}}_5} \sum_{j=1}^{n_d} \log\left(\tilde{\omega}_1 p_1[x_j, \tilde{\alpha}_1, \tilde{\beta}_1] + (1 - \tilde{\omega}_1) p_2[x_j, \tilde{\alpha}_2, \tilde{\beta}_2]\right).$$

We call this algorithm the Hybrid, Reparametrization, and Discretization (HRD) algorithm. The HRD-estimator $\hat{\boldsymbol{\theta}}_5$ can be used to estimate the responsibilities and the measure φ_{cc}. The approach can easily be combined with resampling in order to obtain the accuracy of the estimators (see [7]).

Next we consider an example where we model microarray breast cancer data obtained from the Cancer Genome Atlas (TCGA). Expression

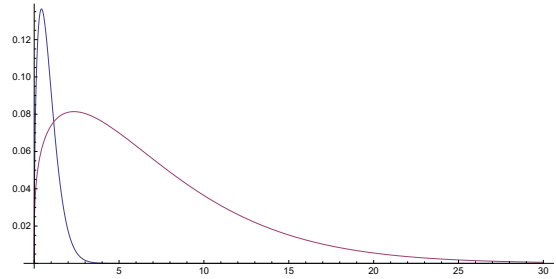

Figure 1. Estimated components of the $Mix2W$ model.

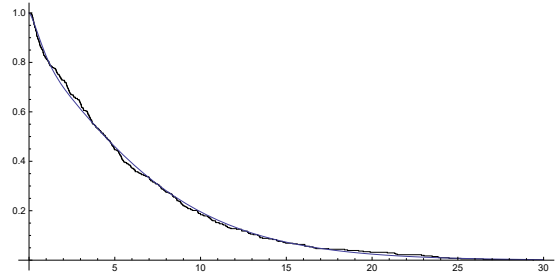

Figure 2. Empirical survival distribution function and the estimated mixture of two Weibull survival distribution functions.

levels for the gene GLNT3 were taken from 409 patients with positive estrogen receptor (ER+) status, a well known subgroup of the disease. The untransformed data were modeled with a mixture of two Weibull distributions, with the weight $\hat{\omega}_1 = 0.17$, the scale and shape parameters were $\hat{\alpha}_1 = 0.926, \hat{\beta}_1 = 1.75, \hat{\alpha}_2 = 7.514, \hat{\beta}_2 = 1.287$, see Fig. 1. The survival distribution function of the fitted model was close to the empirical survival distribution function, see Fig 2.

3. Conclusions

The problem on how to estimate the parameters in a two-component mixture distribution is an old problem that has attracted a lot of attention. The idea to lower the complexity of the problem by momentarily consider some parameters as fixed over a grid is attractive since: modern computers are powerful, the approach is well-suited for parallelization and the approach can easily be combined with resampling. Moreover, the values of

the log-likelihood function can easily be visualized for 1- or 2-dimensional grids, which can be very informative. For the Weibull example we used a 2-dimensional grid and fixed the shape parameters, but there are several alternative reparametrizations that could be considered. Generally, there is a tradeoff between the complexity of the grid and the complexity of the equation system, the higher dimension of the grid the simpler equation system. One advantage by considering a high-dimensional grid is that there will be no need to estimate higher order moments which can be difficult, in particular if the sample size is relatively small and if the data are "contaminated" with outliers. There are several open questions regarding the construction of the grid. For example: How should the boundaries be selected? How dense does the grid need to be? Can an iterative grid-based approach, allowing for uneven grid-densities, be used?

Acknowledgments

This work was supported by grants from the Swedish Research Council, Dnr 340-2013-5185 (P. R.), the Kempe Foundations, Dnr JCK-1315 (D. K., P. R.), and the Faculty of Science and Technology, Umeå University (Yu. B., D. K., P. R.).

References

1. *Freyhult E., Landfors M., Önskog J., Hvidsten T. R., Rydén P.* Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering // BMC Bioinformatics. — 2010. — Vol. 11. — Article 503.
2. *Bolón-Canedo V., Sánchez-Marono N., Alonso-Betanzos A., Benítez J. M., Herrera F.* A review of microarray datasets and applied feature selection methods-Information Sciences. — 2014. — Vol. 282, no. 1. — P. 111–135.
3. *Bordes L., Mottelet S., Vandekerkhove P.* Semiparametric estimation of a two-component mixture model // Annals of Statistics. — 2006. — Vol. 34, no. 3. — P. 1204—1232.
4. *Carta J. A., Ramirez P.* Analysis of two-component mixture Weibull statistics for estimation of wind speed distributions // Renewable Energy. — 2007. — Vol. 32, no. 3. — P. 518—531.
5. *Celeux G., Chauveau D., Diebolt J.* Stochastic versions of the EM algorithm: an experimental study in the mixture case // Journal of Statistical Computation and Simulation. — 1996. — Vol. 55, no. 4. — P. 287–314.
6. *Kallenberg O.* Foundations of modern probability. — Springer, 2006.
7. *Belyaev Yu. K., Nilsson L.* Parametric maximum likelihood estimators and resampling. Statistical Research Report, Department of Mathematical Statistics, Umeå University, 1997-15. — ISSN 1401-730X.