Liet. matem. rink., T. 42, spec. nr., 537–540

Projection-based estimation of multivariate distribution density

Mindaugas KAVALIAUSKAS, Rimantas RUDZKIS (MII)
e-mail: snaiperiui@takas.lt, rudzkis@ktl.mii.lt

In spite of the existance of a great number of distribution analysis methods, the estimation of multivariate distribution density remains a serious problem in practice so far if dimension is large enough. Even if the parametric expression of density is known, it is difficult to calculate a good estimate, using the classical methods. We will illustrate the problem by the Gaussian mixture model, very popular in the classification, which we have analysed by simulation.

Let $X \in \mathbf{R}^d$ be an observed random vector with an unknown distribution density $f(x)$ and $X(1), \ldots, X(n)$ be a sample of independent copies of X. In the Gaussian mixture model

$$f(x) = \sum_{j=1}^{q} p_i \varphi_i(x) \stackrel{def}{=} f(x, \theta), \tag{1}$$

where φ_i is the multivariate normal distribution density and θ is the vector of all model parameters.

The maximum likelihood approach could be used to estimate $f(x)$. However it is obvious that to calculate the estimate

$$\theta_{MLE} = \arg\max_{\theta} \prod_{t=1}^{n} f(X(t), \theta)$$

constructively is very difficult if the dimension d is large, since in the general case

$$\dim \theta = q\frac{d^2 + d}{2} + qd + q - 1.$$

For example, if $d = 6$, $q = 3$ then $\dim \theta = 83$ and if $d = 10$, $q = 3$ then $\dim \theta = 197$. The most popular way to calculate the approximation of θ_{MLE} is the EM algorithm. This recurrent procedure converges to θ_{MLE} only if the initial estimate $\widehat{\theta}^{(0)}$ is close enough to θ_{MLE}.

Obviously, it is much easier to estimate the distribution density f_τ of univariate projections $X_\tau = \tau' X$. Since there exists one-to-one correspondence

$$f \leftrightarrow \{f_\tau,\ \tau \in \mathbf{R}^d\}$$

it is natural to discuss the methods of estimation of f by using statistical estimates of f_τ.

So, we suggest calculating a lot of estimates $\widehat{f}_\tau$ as well as $\widehat{f}(x)$ thereby instead of estimating $f(x)$ by the complicated classical method. This alternative approach uses simpler procedures, however it requires a great amount of calculations. The simulation results we present here were obtained by calculating $\widehat{f}_\tau$, $\tau \in T$, where $cardT = 20000$. It takes about 6 hours for doing that by a personal computer. Naturally, constantly improving computers stimulates the popularity of procedures demanding voluminous calculations, e.g., methods of bootstrap, jack-knife, data mining.

The usage of the projection approach is not a new idea in the statistical analysis of multivariate distributions. For instance Friedman (1987) analyzed estimates of the following shape

$$\widehat{f}(x) = \varphi(y^{(S)}) \prod_{k=1}^{S} \frac{\widehat{f}_{\tau_k}^{(k-1)}(\tau_k' y^{(k-1)})}{\varphi(\tau_k' y^{(k)})},$$

where $\widehat{f}_{\tau_k}^{(k-1)}(\cdot)$ is the density estimate of $\tau_k' Y^{(k-1)}$. Here $Y^{(k)}$ is obtained from X after k step of structure removal (the sample projections into direction τ_k is replaced by corresponding quantiles of Gaussian distribution, $k = 1, 2, ...$). $y^{(k)}$ is calculated from x using the same structure removal transformations.

This is the method of non-parametric estimation. If we use it in case (1), then, after the first iteration already, there is no possibility to use the parametric structure.

Therefore, there is no wonder that the analysis by simulation showed that density (1) obtained by the Friedman method yields much greater errors than the estimation methods using the EM algorithm.

The projection-based method suggested, allow us to use parametric estimates for calculating projection densities.

Let us employ the inversion formula

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbf{R}^d} \mathrm{e}^{-it'x} \psi(t)\, \mathrm{d}t, \quad \psi(t) = \mathbf{E}\, \mathrm{e}^{it'x}.$$

By denoting $u = \|t\|$, $\tau = t/\|t\|$ and replacing variables we obtain

$$f(x) = \frac{1}{(2\pi)^d} \int_{\tau:\ \|\tau\|=1} \mathrm{d}s \int_0^\infty \mathrm{e}^{-iu\tau' x} \psi(u\tau) u^{d-1}\, \mathrm{d}t. \tag{2}$$

The first integral is surface integral over unit sphere.

The distribution density f_τ of the projection $X\tau$ is a mixture of univariate Gaussian densities $f_\tau(y) = f(y, \theta_\tau)$, $y \in \mathbf{R}^1$. In case (1) $\dim \theta_\tau = 3q - 1$.

Denote $\psi_\tau(u) = \mathbf{E}\, \mathrm{e}^{iuX_\tau} = \psi(u, \theta_\tau)$. Then,

$$\psi(u)\tau = \psi_\tau(u) \text{ and } \widehat{\psi}(u)\tau = \psi_{\hat{\tau}}(u). \tag{3}$$

Having selected the set T of projection directions and estimates $\{\widehat{\theta}_\tau,\ \tau \in T\}$, basing on (2) and (3), we obtain the estimate

$$\widehat{f}(x) = \frac{c}{card T} \sum_{\tau \in T} \int_0^\infty \mathrm{e}^{-iu\tau' x} \widehat{\psi}_\tau(u) u^{d-1} \mathrm{e}^{-hu}\, \mathrm{d}u, \tag{4}$$

where $c = c(d)$. The multiplier e^{-hu} is used for additional smoothing of estimates. Here h is a small quantity, which is selected so that $\widehat{f}(x)$ be non-negative or the domain area of negative values be small.

Thus, an estimation of multivariate density f is reduced to an estimation of univariate distribution densities f_τ and corresponding characteristic functions. Note that we obtain the estimate $\widehat{f}$ but we do not obtain $\widehat{\theta}$ which is necessary for the sample data classification. To obtain $\widehat{\theta}$ we can apply the same technique for estimating each component φ_j of the Gaussian mixture

$$\varphi_j \leftrightarrow \{\varphi_{\tau,j},\ \tau \in \mathbf{R}^d\}.$$

The density $\varphi_{\tau,j}$ of projection to the direction τ, corresponding to the component φ_j, is the univariate Gaussian density with mean $m_\tau(j)$ and variance $\sigma_\tau^2(j)$. The estimates of these quantities are components of the vector $\widehat{\theta}_\tau$. Consequently, $\widehat{\varphi}_{\tau,j}$ is obtained from formula (4) replacing $\psi(u, \widehat{\theta}_\tau)$ by $\exp\{iu\widehat{m}_\tau(j) - u^2\widehat{\sigma}_\tau^2(j)/2\}$.

At the end we present some simulation results. Unfortunately these results are not numerous and we cannot draw strict conclusions, though the preliminary results are rather optimistic. The Table 1 shows the typical errors of these three algorithms:

- the pseudo-estimate $\widehat{f}$ calculated using the multivariate EM algorithm with the theoretical value of the parameter θ (i.e., $\theta^{(0)} = \theta$);
- the pseudo-estimate $\widehat{f}$ found using (1), where $\forall\tau\ \widehat{\theta}_\tau$ is calculated using EM algorithm with $\theta^{(0)} = \theta$;
- the estimate $\widehat{f}$ calculated using (1), where $\forall\tau\ \widehat{\theta}_\tau$ is calculated using the software developed at the Institute of Mathematics and Informatics in Vilnius. The initial value $\widehat{\theta}_\tau^{(0)}$ of EM algorithm is chosen in an automated way. The methods is described in paper by R. Rudzkis and M. Radavicius (1995).

Table 1

The errors of $\hat{f}$. The parameters of the mixtures: a) $d = 5, q = 5, n = 500$, b) $d = 5, q = 6, n = 500$

Error	The multivariate EM pseudo-estimate	The projection-based pseudo-estimate	The projection-based estimate
$\frac{1}{n}\sum_{i=1}^{n} \lvert\widehat{f}(x_i) - f(x)\rvert$	a) 0.4876 b) 0.4390	a) 0.7375 b) 0.5426	a) 0.8089 b) 0.7314
$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(x_i) - f(x))^2}$	a) 1.0291 b) 1.1537	a) 1.0318 b) 0.9412	a) 1.2619 b) 1.1214

The first two estimates are pseudo-estimates because in practice the value θ is unknown and thus we cannot select initial value in such way. These pseudo-estimates allow us to compare the accuracy of the MLE with that of the projection-based method (1), where the projection parameters are estimated using the maximum likelihood method.

The Monte-Carlo simulation study showed that the projection based distribution density estimation errors is a little bit large than error of the EM pseudo-estimate. This is the optimistic result because the EM pseudo-estimate is an approximation of MLE.

The proposed projection-based density estimation algorithm still have some unresolved problems. One of them is selection of smoothing bandwidth h, see (4). h can be selected so that $\widehat{f}$ be non-negative. If we take h too small, the estimate gives huge errors. This result requires more detail study.

References

[1] J.H. Friedman, Exploratory projection pursuit, *Journal of the American Statistical Association*, **82**(397), 249–266 (1987).

[2] R. Rudzkis, M. Radavicius, Statistical estimation of a mixture of Gaussian distributions, *Acta Applicandae Mathematicae*, **38**, 37–54 (1995).

Daugiamačio pasiskirstymo tankio vertinimas taikant projektavimą

M. Kavaliauskas, R. Rudzkis

Straipsnyje nagrinėjamas daugiamačio pasiskirstymo tankio vertinimo būdas naudojantis vienamačių duomenų projekcijų tankio įverčius. Naudojamas pasiskirstykio parametrizavimas. Autoriai skiria daug dėmesio tankio įvertinimui daugiamačio Gauso skirstinio atveju. Aptariami preliminarūs kompiuterinio modeliavimo rezultatai.