ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 8, 2004

The projection-based multivariate density estimation

MINDAUGAS KAVALIAUSKAS, RIMANTAS RUDZKIS AND TOMAS RUZGAS

ABSTRACT. The paper discusses methods of estimation of the multivariate density function by using statistical estimates of the densities of the univariate sample projections. The Gaussian mixture model is analyzed. Density parameterization is applied, parameters calculation methods based on the parameters of projections are proposed. A simulation study is carried out and the simulation results are discussed.

1. Introduction

Both parametric and non-parametric statistical estimation of a density function becomes more and more problematic with the growth of dimensionality. Authors of this article will analyze the Gaussian mixture model, one of the most popular data models in practice. Usage of projected data to reduce the dimensionality is not a new idea. It was used by Friedman et al. (1984), for example. Usually the projection-pursuit approach has been followed. Authors of this paper use another projection-based approach to the multivariate density estimation.

Let $X \in \mathbb{R}^d$ be an observed random vector with unknown density function $f(x)$ and $X(1), ..., X(n)$ form an independent sample of X. Let us denote the density of univariate projection $X_\tau = \tau' X$ by $f_\tau(x)$. The one-to-one correspondence exists

$$f \leftrightarrow \{f_\tau, \tau \in \mathbb{R}^d\} \tag{1}$$

and it is natural to discuss the methods of estimating f using statistical estimates of f_τ. Does this approach to the estimation of a multivariate density have any sense in practice? Yes, it has.

Received September 19, 2003.
2000 *Mathematics Subject Classification.* 62G07, 62H12, 62F10.
Key words and phrases. Multivariate distribution density, density estimation, projection pursuit, Gaussian distribution density.

Let us consider a mixture of the multivariate Gaussian distributions as an example. In this case the density function is

$$f(x) = \sum_{i=1}^{q} p_i \varphi(x, M_i, R_i) \stackrel{def}{=} f(x, \theta), \tag{2}$$

where q is the number of clusters, φ is the multivariate normal density and θ is the vector of all model parameters. Thus, we must estimate parameter θ to obtain the parametric estimate of the density f. The *maximum likelihood estimate* (MLE) is asymptotically optimal in the Gaussian mixture model and it is natural to use it for the parameter estimation. Even if the dimension d and the number of clusters q are small, the dimension of parameter θ is large (for example, if $d = 10$ and $q = 5$ then $dim\,\theta = 329$) and the calculation of MLE is difficult. The recurrent *expectation maximization* (EM) algorithm could be used to calculate approximation of MLE but it converges to θ_{MLE} only if initial estimate is close enough to θ_{MLE}. So, calculation of parameter θ and corresponding multivariate density f is problematic, if dimension is large.

The calculation of density and the corresponding parameter in one-dimensional case is much easier. Therefore we suggest the projection-based approach. We are going to make many univariate projections of our sample and calculate estimates $\widehat{f}_\tau$ as well as $\widehat{f}$, instead of estimating $\widehat{f}$ by some complicated classical method.

2. Employing the inversion formula

Let us employ the inversion formula

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-it'x} \psi(t)\, dt, \;\; \psi(t) = \mathbf{E} e^{it'X}. \tag{3}$$

By denoting $u = \|t\|$, $\tau = t/\|t\|$ we obtain after a change of variables

$$f(x) = \frac{1}{(2\pi)^d} \int_{\tau:\, \|\tau\|=1} ds \int_0^\infty e^{-iu\tau' x} \psi(u\tau) u^{d-1}\, dt. \tag{4}$$

The first integral is a surface integral over the unit sphere.

Let us denote $\psi_\tau(u) = \mathbf{E} e^{iuX_\tau}$. Then,

$$\psi(u\tau) = \psi_\tau(u) \;\text{ and }\; \widehat{\psi}(u\tau) = \widehat{\psi}_\tau(u). \tag{5}$$

Having selected the set T of projection directions and using (4) and (5), we obtain the estimate

$$\widehat{f}(x) = \frac{c(d)}{card T} \sum_{\tau \in T} \int_0^\infty e^{-iu\tau' x} \widehat{\psi}_\tau(u) u^{d-1} e^{-hu}\, du. \tag{6}$$

One can see the additional multiplier e^{-hu} under the integral. This multiplier performs the additional smoothing of the estimate $\widehat{f}$ with the Gaussian kernel. The simulation study showed that this multiplier is necessary and it decreases significantly error of the estimate. Here h is a small number, which is selected so that $\widehat{f}(x)$ is non-negative.

The formula (6) could be used with different estimates of the univariate characteristic functions $\widehat{\psi}_\tau(u)$. We analyzed the case of the Gaussian mixture model. In this case, the projected data also satisfy the Gaussian mixture model

$$X_\tau \sim f_\tau(x,\theta) = \sum_{i=1}^{q} p_{i,\tau}\varphi(x, m_{i,\tau}, \sigma_{i,\tau}). \tag{7}$$

We used the parametric estimates of corresponding univariate characteristic functions, which depend on estimates of the parameters of projected data

$$\widehat{\psi}_\tau(u) = \sum_{i=1}^{q} \widehat{p}_i \exp(iu\widehat{m}_{i,\tau} - \widehat{\sigma}^2_{i,\tau}u^2). \tag{8}$$

The number of parameters in each direction is much smaller than in the multivariate case and estimation of it is not so problematic. For example, if $q = 5$, then $dim\,\theta_\tau = 14$. The same time we had $dim\,\theta = 329$ in the multivariate case. In the case of the Gaussian mixture model, the integral in (6) can be constructively calculated.

3. The least squares approach for the projected parameters

We also propose another projection-based method for estimating the multivariate density function. This method is based on the least squares approach. Its properties are different from those based on the inversion formula. The inversion formula method can be used for both, parametric and non-parametric estimates of univariate characteristic functions, but it always gives a non-parametric multivariate density (even if Gaussian mixture model is satisfied). This makes it impossible to use such an estimator for data classification. The result of the method based on the least squares approach gives a parametric expression of the density estimate. This is an advantage of the method.

So, let X, and thereby X_τ, satisfy the Gaussian mixture model. Having a set of projection directions $T = \{\tau\}$ and the estimates of the parameters of the projected density $\widehat{p}_i$, $\widehat{m}_{i,\tau}$ and $\widehat{\sigma}^2_{i,\tau}$, we can calculate the parameter of the multivariate density for each cluster, $i = 1, ..., q$, using the least squares method. Because of $p_{i,\tau} = p_i$, $m_{i,\tau} = \tau' m_i$ and $\sigma^2_{i,\tau} = \tau' R_i \tau$, we can define the estimates from the requirements:

$$\widehat{p}_i : \sum_{\tau} (\widehat{p}_i - \widehat{p}_{i,\tau})^2 \longrightarrow \min, \tag{9}$$

$$\widehat{m}_i : \sum_{\tau} \left(\tau' \widehat{m}_i - \widehat{m}_{i,\tau}\right)^2 \longrightarrow \min, \tag{10}$$

$$\widehat{R}_i : \sum_{\tau} \left(\tau' \widehat{R}_i \tau - \widehat{\sigma}^2_{i,\tau}\right)^2 \longrightarrow \min. \tag{11}$$

The calculation of the estimates $\widehat{p}_i$, $\widehat{m}_i$ and $\widehat{R}_i$ can be done using common technique, however the definition of $\widehat{R}_i$ this way does not ensure that it will be positive definite. We suggest the following to ensure $\widehat{R}_i$ to be positive definite. Let us define

$$\widehat{R}_i \stackrel{def}{=} \widehat{A}_i \widehat{A}_i^T, \tag{12}$$

where $\widehat{A}_i$ is a lower triangular matrix. The matrix $\widehat{A}_i$ can be calculated by using the recurrent procedure

$$\widehat{A}_i^{(k+1)} : \sum_{\tau} \left(\tau' \widehat{A}_i^{(k)} \widehat{A}_i'^{(k+1)} \tau - \widehat{\sigma}^2_{i,\tau}\right)^2 \longrightarrow \min. \tag{13}$$

The initial value $\widehat{A}_i^{(0)}$ can be calculated from the condition

$$\widehat{A}_i^{(0)} : \sum_{\tau} \left(\tau' \widehat{A}_i^{(0)} \widehat{A}_i'^{(0)} \tau - \widehat{\sigma}^2_{i,\tau}\right)^2 \longrightarrow \min, \tag{14}$$

under assumption that $\widehat{A}_i^{(0)}$ is diagonal. Diagonality ensures the equation to be linear and $\widehat{A}_i^{(0)}$ can be calculated using common technique.

This projection-based multivariate density estimation method gives a parametric estimate of the density, but it still has a few drawbacks. One of them is that in each direction τ the number of estimated clusters must be equal to q. In real situation, in some projections the projected clusters overlap and clusterization procedure for univariate data gives smaller number of clusters. Then the proposed least squares algorithm should not be used. The other drawback is that we still relay on some clusterization procedure which estimates parameters of the Gaussian mixture. The next section describes the method without these drawbacks.

4. The least squares approach for the projected densities

Let us minimize the sum of squared distance between densities:

$$\widehat{\theta} : \sum_{\tau} \left\| f_\tau(\cdot, \theta) - \widehat{f}_\tau(\cdot) \right\|_2^2 \longrightarrow \min, \tag{15}$$

where $f_\tau(\cdot, \theta)$ is defined by (7) and $\widehat{f}_\tau(\cdot)$ is some density estimate of projected data. This requirement can be rewritten as

$$\widehat{\theta} : \sum_{\tau} \int_{\mathbb{R}} \left(f_\tau^2(x, \theta) - 2 f_\tau(x, \theta) \widehat{f}_\tau(x) \right) dx \longrightarrow \min, \tag{16}$$

where the last component $\widehat{f}_\tau^2(x)$ is omitted because it does not depend on the variable θ. We can use the parametric estimate of $\widehat{f}_\tau(\cdot)$ or we can replace the integral of the second component by sum over projected data. In this case we obtain from (16)

$$\widehat{\theta}: \sum_\tau \int_{\mathbb{R}} f_\tau^2(x,\theta)\,dx - \frac{2}{n}\sum_{j=1}^{n} f_\tau(\tau' X_j,\theta) \longrightarrow \min. \tag{17}$$

Let us divide vector of all unknown model parameters θ into three parts

$$\theta = \begin{pmatrix} \mathbf{p} \\ \mathbf{u} \\ \mathbf{v} \end{pmatrix}, \tag{18}$$

where $\mathbf{p}$ is vector of probabilities, $\mathbf{u}$ is vector of all mean parameters and $\mathbf{v}$ is vector of all covariance parameters of clusters. Because integral of the product of the Gaussian densities remains the Gaussian density

$$\int_{\mathbb{R}} \varphi(x,\, m_1, \sigma_1^2)\,\varphi(x,\, m_2, \sigma_2^2)\,dx = \varphi(m_1 - m_2,\, 0,\, \sigma_1^2 + \sigma_2^2), \tag{19}$$

equation (17) has the following form

$$\widehat{\theta}: \; Q(\theta) = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p})\varphi_{\mathbf{i}}(\mathbf{u},\mathbf{v}) \longrightarrow \min. \tag{20}$$

Here φ is the Gaussian density function and $\mathbf{i}$ is the index-vector of sum over all projection directions τ, clusters $i,j = 1,...,q$ and sample points $l = 1,...,n$. Notice, that

$$\varphi_{\mathbf{i}} = \frac{1}{\sqrt{\beta_{\mathbf{i}}}} \exp\left(-\frac{\alpha_{\mathbf{i}}}{2\beta_{\mathbf{i}}}\right), \text{ where } \alpha_{\mathbf{i}} = \left((\mathbf{u} - a_{\mathbf{i}})^T c_{\mathbf{i}}\right)^2, \; \beta_{\mathbf{i}} = b_{\mathbf{i}}^T \mathbf{v}. \tag{21}$$

Here $a_{\mathbf{i}}$, $b_{\mathbf{i}}$ and $c_{\mathbf{i}}$ are vectors depending on the sample and the set of projection directions. If we have fixed the sample and directions of projections, these vectors are constants.

Let us calculate derivatives of the function $Q(\theta)$:

$$\frac{d\varphi_{\mathbf{i}}}{d\mathbf{u}} = -\frac{\varphi_{\mathbf{i}}}{\beta_{\mathbf{i}}}\alpha_{\mathbf{i}}' = A_{\mathbf{i}}(\mathbf{u} - a_{\mathbf{i}}), \text{ where } A_{\mathbf{i}} = -\frac{\varphi_{\mathbf{i}}}{\beta_{\mathbf{i}}}\left[c_{\mathbf{i}} c_{\mathbf{i}}^T\right], \tag{22}$$

$$\frac{d\varphi_{\mathbf{i}}}{d\mathbf{v}} = \varphi_{\mathbf{i}}\left(\frac{\alpha_{\mathbf{i}}}{2\beta_{\mathbf{i}}^2} - \frac{1}{2\beta_{\mathbf{i}}}\right)\beta_{\mathbf{i}}' = d_{\mathbf{i}} - B_{\mathbf{i}}\mathbf{v}, \tag{23}$$

$$\text{where } \; d_{\mathbf{i}} = \frac{\varphi_{\mathbf{i}}\alpha_{\mathbf{i}}}{2\beta_{\mathbf{i}}^2} b_{\mathbf{i}}, \; B_{\mathbf{i}} = \frac{\varphi_{\mathbf{i}}}{2\beta_{\mathbf{i}}^2}\left[b_{\mathbf{i}} b_{\mathbf{i}}^T\right].$$

Thus,

$$\frac{dQ}{d\mathbf{u}} = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) A_{\mathbf{i}}(\mathbf{u} - a_{\mathbf{i}}) = C\mathbf{u} - e, \tag{24}$$

$$\text{where } C = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) A_{\mathbf{i}},\ e = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) A_{\mathbf{i}} a_{\mathbf{i}},$$

$$\frac{dQ}{d\mathbf{v}} = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p})(d_{\mathbf{i}} - B_{\mathbf{i}}\mathbf{v}) = g - D\mathbf{v}, \tag{25}$$

$$\text{where } D = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) B_{\mathbf{i}},\ g = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) d_{\mathbf{i}}.$$

Here $a_{\mathbf{i}}$, $b_{\mathbf{i}}$, $c_{\mathbf{i}}$, $d_{\mathbf{i}}$, e and g are vectors and $A_{\mathbf{i}}$, $B_{\mathbf{i}}$, C and D are matrices.

Since

$$\widehat{\theta}:\ Q(\theta) \longrightarrow \min \ \Rightarrow\ \frac{dQ}{d\mathbf{p}}(\widehat{\theta}) = 0,\ \frac{dQ}{d\mathbf{u}}(\widehat{\theta}) = 0,\ \frac{dQ}{d\mathbf{v}}(\widehat{\theta}) = 0, \tag{26}$$

we can define a recurrent algorithm for calculation of $\widehat{\theta}$:

$$\mathbf{u}^{(k+1)} = C^{-1}(\theta^{(k)})\, e(\theta^{(k)}),\ \mathbf{v}^{(k+1)} = D^{-1}(\theta^{(k)})\, g(\theta^{(k)}). \tag{27}$$

Value of $\mathbf{p}$ at each step of the algorithm is calculated using common technique because $\frac{dQ}{d\mathbf{p}}$ is a linear function of parameter $\mathbf{p}$.

5. Results

The suggested methods were studied by using Monte-Carlo method. We will present some simulation results here. All simulations are done assuming the Gaussian mixture model. We used 5-dimensional data sets in our study, i.e. $d = 5$. The sample size was $n = 500$. The number of projection directions *card* T was up to 20000. We used various numbers of clusters to vary the number of parameters of the model. The methods were compared using error

$$\varepsilon = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{f}(x) - f(x) \right)^2. \tag{28}$$

The following estimates and pseudo-estimates were used for comparison:

- $\widetilde{f}_{MLE}$ – the pseudo-estimate calculated using the multivariate EM algorithm with initial theoretical value of the parameter θ (i.e., $\theta^{(0)} = \theta$). In this case EM converges to MLE.
- $\widehat{f}_{EM}$ – the multivariate parametric estimate calculated using the multivariate EM algorithm. Ideas about selection of initial parameter value and other details can be found in Rudzkis and Radavičius (1995).
- $\widetilde{f}_{INV}$ – the projection-based pseudo-estimate calculated by (6) from MLE pseudo-estimates $\theta_{\tau,MLE}$ in each direction.

- $\widehat{f}_{INV}$ – the estimate calculated by formula (6) from real estimates of θ_τ in each direction.
- $\widetilde{f}_{LSP}$ – the projection-based pseudo-estimate calculated using the parametric least squares approach (see, (9)-(14)) from MLE pseudo-estimates $\theta_{\tau,MLE}$ in each direction.

Simulation study showed that $\widetilde{f}_{MLE}$ and $\widetilde{f}_{INV}$ give similar errors, but if the number of clusters is 5 or more (number of parameters is large enough in this case) the projection-based $\widetilde{f}_{INV}$ pseudo-estimate is slightly better.

The estimate $\widehat{f}_{INV}$ is also better that $\widehat{f}_{EM}$ by up to 10%, if the number of parameters is large.

The projection-based estimate $\widetilde{f}_{LSP}$ gives very good results. It gives up to 20% smaller error than $\widetilde{f}_{MLE}$, if the number of parameters is large. But this parametric least squares estimation method has one drawback. The number of estimated clusters must be the same in all directions and also clusters must be numbered in the same order in all directions. This drawback is very difficult to overcome in a real situation because of overlapping of clusters. Therefore we were not able to get good estimates based on parametric least squares approach, but only using pseudo-estimate which uses theoretical value of projected density parameters as initial point for calculating MLE.

The method based on the least squares approach for projected densities is currently under study.

Conclusion. The projection-based estimates give similar error as classical methods if the number of parameters is small. If the number of parameters is large, projection-based estimates are better because projection to an one-dimensional space reduces the number of parameters very much. In the Gaussian mixture model the projection-based method can give smaller error than MLE. This is because MLE is only asymptotically effective estimate.

References

Friedman, J. H., Stuetzle, W. and Schroeder, A. (1984). Projection pursuit density estimation. *J. Amer. Statist. Assoc.* **79**, 599–608.

Friedman, J. H. (1987). Exploratory projection pursuit. *J. Amer. Statist. Assoc.* **82**, 249–266.

Huber, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13**, 435–475.

Rudzkis, R. and Radavičius, M. (1995). Statistical estimation of a mixture of Gaussian distributions. *Acta Appl. Math.* **38**, 37–54.

Institute of Mathematics and Informatics, Akademijos 4, LT-2600 Vilnius, Lithuania

E-mail address: snaiperiui@takas.lt

E-mail address: rudzkis@ktl.mii.lt

E-mail address: tomas.ruzgas@paspara.com