Method for Reproducing the Estimation of Heterogeneous Mixtures of Distributions based on Non-Parametric Spline for Anomaly Detection in Digital Images*

Anhelina Zhultynska[1,2,†] and Pylyp Prystavka[1,*,†]

[1] *National Aviation University, 1 Liubomyra Huzara ave., 03058 Kyiv, Ukraine*
[2] *Interregional Academy of Personnel Management, 2 Frometivska str., 03039 Kyiv, Ukraine*

Abstract

The parametric methods have limited ability to take into account the complexity and heterogeneity of data distributions, and it is the problem of this category of methods. This paper presents a nonparametric method using local polynomial splines to estimate heterogeneous mixtures of distributions. By leveraging B-splines, the method adapts to various data distributions without strict assumptions about their shapes. This approach improves parameter estimation accuracy in complex data structures where traditional parametric methods are inadequate.

Keywords

heterogeneous mixtures, polynomial splines, nonparametric methods

1. Introduction

The problem with parametric methods is their limited ability to take into account the complexity and heterogeneity of data distributions. This is especially true when the data does not follow a normal distribution, which is one of the basic assumptions of many parametric models. Real data often exhibit diverse and skewed distributions, which makes it difficult to use classical approaches. Therefore, it makes sense to consider methods that do not require a clear model and strict adherence to this model and that are not so dependent on the skewness or kurtosis of individual components of the distributions.

Non-parametric methods, in particular the method of reproducing the estimation of heterogeneous mixtures of distributions based on local polynomial splines, are the answer to these limitations. They allow us to build models that adapt to the diversity of data distributions and do not require strict assumptions about their shape. Such methods provide more flexibility, especially in conditions where parametric methods are not effective enough.

2. Literature review

The Support Vector Machine (SVM) is a supervised learning method used for classification and regression. The basic idea is to find the optimal hyperplane that best separates objects of different classes on the feature plane. One of the advantages of SVM is its effectiveness in dealing with high-dimensional data and the ability to work with heterogeneous mixtures of distributions, but it can be sensitive to large amounts of data and requires proper selection of hyperparameters, such as the regularization parameter and kernel selection [1, 2].

The k-means method is one of the most common cluster analysis methods in machine learning. It is used to divide a data set into k clusters based on the similarity of their features [3]. The basic idea is to select several cluster centers and assign each dataset to the cluster center closest to it. The

* *CPITS 2025: Workshop on Cybersecurity Providing in Information and Telecommunication Systems, February 28, 2025, Kyiv, Ukraine*

* Corresponding author.

† These authors contributed equally.

angelinaremark1@gmail.com (A. Zhultynska); chindakor37@gmail.com (P. Prystavka)
0000-0001-9178-897X (A. Zhultynska); 0000-0002-0360-2459 (P. Prystavka)

main advantages of the k-means method are its simplicity and speed, which allows you to perform clustering quickly even on large amounts of data. In addition, this method works well when clusters are spherical.

However, the k-means method has its drawbacks. It is sensitive to the choice of initial cluster centers, which can lead to different results under different initial conditions. Paper [4] shows how this drawback can be improved. In addition, the k-means method assumes that each cluster has the same variance, which may not be sufficient for some types of data where clusters may have different shapes or sizes [5–7]. It is also important to keep in mind that k-means is not able to account for heterogeneity in data distributions, which can be a problem in the context of mixture reproduction.

The Expectation-Maximization (EM) algorithm is an iterative method used to estimate the parameters of statistical models with hidden variables, such as a mixture of distributions. This method combines an Expectation step and a Maximization step to iteratively update and find the most likely model parameters. One of the advantages of the EM algorithm is the ability to work with complex data distributions and find their parameters, but it can get stuck in local maxima and require multiple runs from different initial conditions. Publications [8–10] present methods to improve the algorithm's performance.

3. Problem statement

The use of polynomial splines in the context of recreating mixtures of distributions can be an effective approach that allows you to adapt to the diversity of the data and approximate the distribution density of each component of the mixture. The basic idea is to approximate complex functions using piecewise polynomial functions that maintain smoothness and continuity over the entire data interval. They can approximate even non-homogeneous distributions well, which can be useful in cases where other methods may not give satisfactory results. In addition, polynomial splines do not require any explicit assumptions about the shape of the data distribution, making them a versatile tool for analyzing a variety of data sets [1].

4. The study materials and methods

Let an array $\Omega_{2,N}=\{X_l; l=\overline{1,N}\}$, $X_l \in R^2$ consisting of N observations be given. Suppose that this data set consists of a mixture of K distributions. Let $\eta_1(X), \eta_2(X), \ldots, \eta_k(X)$ be the distributional densities of each mixed component that make up the mixture. The goal is to reproduce the estimated distributional density of the mixture $f(X)$ without reducing the commonality $f(X) \in C^{r,\ldots,r}$, which is a linear combination of the densities of each component:

$$f(X)=\sum_{k=1}^{K} \rho_k \eta_k(X),$$

where ρ_k is the weighting factor of the k^{th} component of the mixture, $\sum_{k=1}^{K} \rho_k=1$.

The task is to find the estimates of the component densities $\eta_1(X), \eta_2(X), \ldots, \eta_k(X)$ and the corresponding weighting factors ρ_k.

One effective approach to solving this problem is to use local polynomial splines based on B-splines close to the interpolation mean, a mathematical tool that allows you to approximate functions using smooth curves, adapting to the diversity of the data. The main idea is to build a flexible model that takes into account the heterogeneity of distributions and provides accurate parameter estimates. This approach allows us to take into account the complexity of the data structure.

Consider the algorithm.

Step 1. We perform a histogram evaluation of the variation series, as a result, we get $\{(x_{1,i}, x_{2,j}), n_{i,j}, p_{i,j}; i, j \in Z\}$ based on a uniform partition of the set Δ_{h_1,h_2}, where $(x_{1,i}, x_{2,j})$ is a variant that determines the center point of the $(i, j)^{\text{th}}$ element of the partition Δ_{h_1,h_2}; h_1, h_2 is the partition step; $n_{i,j}$ is the frequency (the number of elements that fall within the boundaries of the $(i, j)^{\text{th}}$ element of the partition Δ_{h_1,h_2}); $p_{i,j}$ is the relative frequency of the variant.

Step 2. The resulting histogram is approximated by a two-dimensional local polynomial spline $S_{2,0}$ based on B-splines that are close to interpolation on average [11] (Fig. 1b).

$$S_{2,0}(p, x_1, x_2) = \sum_{i \in Z} \sum_{j \in Z} B_{2,h_1}(x_1 - i h_1) B_{2,h_2}(x_2 - j h_\sigma) p_{i,j},$$

where (with the notation accuracy up to the split step) [8]

$$B_{2,h}(x) = \begin{cases} 0, & x \notin [-3h/2; 3h/2], \\ (3+2x/h)^2/8, & x \in [-3h/2; -h/2], \\ 3/4 - (2x/h)^2/4, & x \in [-h/2; h/2], \\ (3-2x/h)^2/8, & x \in [h/2; 3h/2]. \end{cases}$$

Step 3. We search for local maxima of the spline (Fig. 1c) $M_q = \{(m i_q, m j_q); q = \overline{1,g}\}$, where g is the number of local maxima. We consider in pairs $\{M_v, M_s\}$, $v, s = \overline{1,g}, v \neq s$. For each pair of local maxima, we implement the following algorithm:

1. Draw a line $f_v(x) = k_\alpha x + b$ passing through the points $M_v(x_{1,v}, x_{2,v})$ and $M_s(x_{1,s}, x_{2,s})$, where $k_\alpha = \dfrac{x_{2,s} - x_{2,v}}{x_{1,s} - x_{1,v}} = tg\alpha$, $b = \dfrac{x_{1,s} x_{2,v} - x_{1,v} x_{2,s}}{x_{1,s} - x_{1,v}}$.
2. Draw lines $d_{v,\xi}(x) \| f_v(x)$, $d_{v,\xi}(x) = k_\alpha x + b + c_\xi$, where $c_\xi - const$, ξ is the number of parallel lines drawn.
3. Draw lines $a_{v,\psi}(x) = k_\gamma x + b$ at an angle φ to $f_v(x)$, where $\gamma = \alpha \pm \varphi$, $k_\gamma = \dfrac{tg\alpha \pm tg\varphi}{1 \mp tg\alpha \cdot tg\varphi} = \dfrac{k_\alpha \pm tg\varphi}{1 \mp k_\alpha \cdot tg\varphi}$, ψ is the number of lines drawn at an angle.
4. Form an array of lines $U = \{u_\tau, \tau = \overline{1,e}\}$, where $u_\tau = \{f_v(x), d_{v,\xi}(x), a_{v,\psi}(x)\}$, e is the number of lines drawn, $e = v \cdot (1 + \xi + \psi)$. Find the point on each line $(x_{1,\tau}, x_{2,\tau})$ for which $S_{2,0}(p_\tau, x_{1,\tau}, x_{2,\tau}) = min$. Form an array of the found minimum values (Fig. 1d):

$$U_{min} = \underset{S_{2,0}(p, x_1, x_2)}{argmin}\, U_q = \underset{S_{2,0}(p, x_1, x_2)}{argmin}\, u_{q,\tau}(x_1, x_2),$$

where $q = \overline{1,g}$, $\tau = \overline{1,e}$.

Step 4. Based on the data obtained, we find a regression. In general, any regression can be created. To describe the algorithm, let's choose the simplest example and reproduce the linear regression (Fig. 1e) $\overline{x_{2,q}} = \beta_{q,0} + \beta_{q,1} x_1$, where $x_{2,q}$ is the dependent variable, x_1 is the independent variable and $\vec{B} = \{\beta_{q,0}; \beta_{q,1}\}$ is the vector of parameters that is unknown. We find the estimates of the parameters $\hat{\beta}$ using the least squares method

$$\hat{B} = \underset{\beta}{argmin} \sum_{\tau=1}^{e} (x_{2,q,\tau} - \hat{\beta}_{q,0} - \hat{\beta}_{q,1} x_{1,\tau})^2.$$

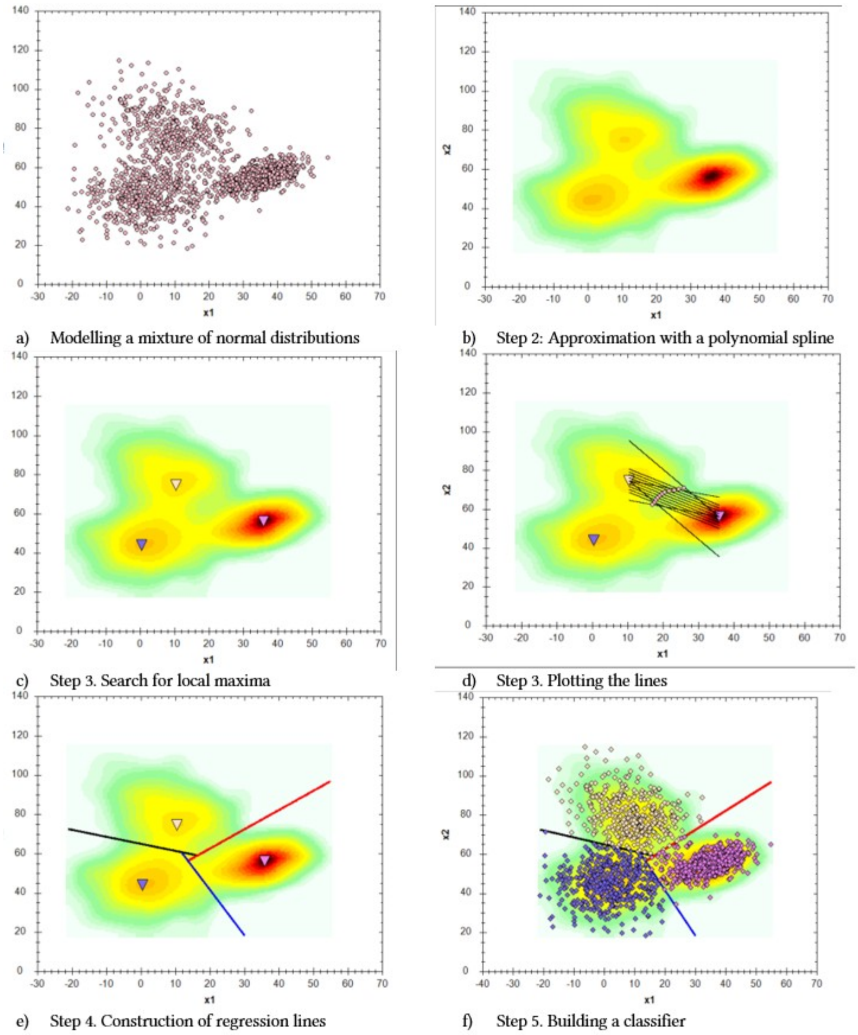
Similarly, we find the linear regression $\overline{x_{1,q}} = \lambda_{q,0} + \lambda_{q,1} x_2$, where $x_{1,q}$ is the dependent variable, x_2 is the independent variable and $(\lambda_{q,0}, \lambda_{q,1})$ is the vector of parameters that is unknown [12–14].

Next, we form the average vector of parameters $(\omega_{q,0},\omega_{q,1})$, where $\omega_{q,0}=\frac{\beta_{q,0}+\lambda_{q,0}}{2}$, $\omega_{q,1}=\frac{\beta_{q,1}+\lambda_{q,1}}{2}$. We build a linear regression $\overline{x_{2,q}}=\omega_{q,0}+\omega_{q,1}x_1$.

The resulting line is reduced to the form $z_q=\omega_{q,0}+\omega_{q,1}x_1-x_2$, where q is the number of local maxima.

Step 5. We build a classifier based on the constructed lines (Fig. 1f). Let $\widetilde{Y}=\{\widetilde{y}_k,k=\overline{1,K}\}$ be the set of classes, $Z=\{z_k(X),k=\overline{1,K}\}$ be the set of discriminant functions that separate classes $\widetilde{y}_k$. Then we define $\forall\, l:\widetilde{y}_l=I(X),\, l=\overline{1,N}$, where

$$I(X)=\begin{cases}1, & z_1(X)<0\wedge z_v(X)>0,\\ 2, & z_1(X)>0\wedge z_2(X)>0,\\ \vdots & \vdots\\ i, & z_i(X)<0\wedge z_{i+1}(X)>0,\\ \vdots & \vdots\\ K, & z_{K-1}(X)<0\wedge z_K(X)<0.\end{cases}$$



a) Modelling a mixture of normal distributions

b) Step 2: Approximation with a polynomial spline

c) Step 3. Search for local maxima

d) Step 3. Plotting the lines

e) Step 4. Construction of regression lines

f) Step 5. Building a classifier

Figure 1: Illustrated description of the algorithm

Remarks. If the lines do not intersect at one point but form a triangle, the classification of the points in the middle of the triangle can be done as follows:

- This area is not identified.
- Identification is carried out with priority, the points are classified into a mixture that has a higher empirical probability.
- If the empirical probability of several mixtures coincides, the identification can be performed in the order of bypassing the constructed boundaries [15].

For the experiment, we model K two-dimensional distributions (Fig. 1a). To further compare with existing methods, we will model a normal distribution with the parameters $\vec{\Theta}_k = \{\mu_{1,k}, \mu_{2,k}, \sigma_{1,k}, \sigma_{2,k}, r_k, \rho_k, N_k\}$, $k=\overline{1,K}$, where ρ_k is the weighting factor of the k-th component of the mixture, $\sum_{k=1}^{K} \rho_k = 1$, N_k is the number of elements of the k-th component of the mixture [16–18].

We obtain $\Omega_{2,N} = \{X_l, y_l; l=\overline{1,N}\}$, where $X=\{x_1; x_2\}, X \in R^2$, $y=\{1,\ldots,K\}$ is the class index.

Then the density of the mixture of K components is defined as:

$$f(X,\vec{\Theta}) = \sum_{k=1}^{K} \rho_k \eta_k(X,\vec{\Theta}_k),$$

$$\eta_k(X,\vec{\Theta}_k) = \frac{\exp\left(\frac{-1}{2(1-r_k{}^2)}\left[\frac{(x_1-\mu_{1,k})^2}{\sigma_{1,k}^2} + \frac{(x_2-\mu_{2,k})^2}{\sigma_{2,k}^2} - \frac{2r_k(x_1-\mu_{1,k})(x_2-\mu_{2,k})}{\sigma_{1,k}\sigma_{2,k}}\right]\right)}{2\pi\sigma_{1,k}\sigma_{2,k}\sqrt{1-r_k{}^2}}.$$

It is necessary to find $\Omega_{2,N} = \{X_l, \tilde{y}_l; l=\overline{1,N}\}$, so that $y_l = \tilde{y}_l$.

Let's illustrate each of the stages of the algorithm. Then, after applying the clustering method, comparing the obtained arrangement of elements in clusters with the initial arrangement of elements (reference), it is possible to estimate the clustering error $\varepsilon = \frac{N_\varepsilon}{N}$, where N is the total number of elements, $N_\varepsilon = \sum_{l=1}^{N} Q(X_l)$ is the number of elements that fell into a class different from the one they were in according to the modeling results, where

$$Q(X_l) = \begin{cases} 1, & y_l \neq \tilde{y}_l \\ 0, & y_l = \tilde{y}_l \end{cases}.$$

Let's compare the above algorithm with existing methods. The results are shown in Table 1.

Table 1
Comparative Analysis of Methods

Comparative factors	Presented method	Support vector method (SVM)	K-means method	EM-algorithm
Processing large amounts of data	+	−	+	+
Resilience to the choice of initial parameters	+	−	+	+
Robustness to emissions	+	−	−	−
Independence from data normalization	+	−	−	+

Independence from the choice of the number of clusters	+	+	–	–

Conclusions and future studies

According to the results of the study, one of the main advantages of the proposed method is the ability to process an arbitrary amount of data due to the preliminary histogram approximation. In contrast to the support vector method, this method demonstrates robustness to the choice of initial parameters. Compared to the k-means method and the EM algorithm, the proposed method is robust to outliers, as it can identify anomalies as unlikely events and disregard them. In addition, the proposed method does not depend on a predefined number of clusters, as it is determined based on the number of local maxima of the density function. In addition, unlike the support vector method and the k-means method, this method does not require data normalization, since the polynomial spline approximation does not require strict assumptions about the shape of the data distribution [19].

Given that the model of digital images is not well formalized, further research is planned to apply the proposed method to segment digital images. This will avoid data normalization and dependence on assumptions about their distribution. The next stage of research is to test the effectiveness of the method in the tasks of segmentation and anomaly detection in digital images.

Declaration on Generative AI

While preparing this work, the authors used the AI programs Grammarly Pro to correct text grammar and Strike Plagiarism to search for possible plagiarism. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] P. Prystavka, Polynomial splines in data processing, Dnipropetrovsk, 2004.

[2] Y. Qin, D. Li, A. Zhang, A new SVM multiclass incremental learning algorithm, Math. Problems Eng. 2015 (2015) 1–5. doi:10.1155/2015/745815

[3] J. S. Al-Azzeh, et al., Analysis of self-similar traffic models in computer networks, Int. Rev. Modelling Simulations 10(5) (2017) 328–336. doi:10.15866/iremos.v10i5.12009

[4] P. Jat, K. Jain, A revised and efficient k-means clustering algorithm, Int. J. Comput. Sci. Eng. 6(12) (2018) 118–124. doi:10.26438/ijcse/v6i12.118124

[5] O. Popov, et al., Physical features of pollutants spread in the air during the emergency at NPPs, Nuclear and Radiation Safety 4(84) (2019) 88–98. doi:10.32918/NRS.2019.4(84).11

[6] T. Tran, et al., Enabling multicast and broadcast in the 5G core for converged fixed and mobile networks, IEEE Transactions on Broadcasting 66(2) (2020) 428–439. doi:10.1109/TBC.2020.2991548

[7] S. Kotenko, et al., The mathematical modeling stages of combining the carriage of goods for indefinite, fuzzy and stochastic parameters, Int. J. Integr. Eng. 12(7) (2020) 173–180. doi:10.30880/ijie.2020.12.07.019

[8] Y. Li, Y. Chen, Research on initialization on EM algorithm based on gaussian mixture model, J. Appl. Math. Phys. 6(1) (2018) 11–17. doi:10.4236/jamp.2018.61002

[9] O. Popov, et al., The use of specialized software for liquid radioactive material spills simulation to teach students and postgraduate students, CTE Workshop Proceedings 9 (2022) 306–322. doi:10.55056/cte.122

[10] J. Tobin, C. P. Ho, M. Zhang, Reinforced EM algorithm for clustering with Gaussian mixture models, in: 2023 SIAM International Conference on Data Mining (SDM), 2023, 118–126. doi:10.1137/1.9781611977653.ch14

[11] A. Ligun, A. Shumeiko, Asymptotic methods of curve reconstruction, NASU Institute of Mathematics, 1997.

[12] I. Ostroumov, et al., Modelling and simulation of DME navigation global service volume, Adv. Space Res. 68(8) (2021) 3495–3507.

[13] V. Tkachuk, et al., Using mobile ICT for online learning during COVID-19 lockdown, Information and Communication Technologies in Education, Research, and Industrial Applications. ICTERI 2020. Communications in Computer and Information Science, vol. 1308 (2021) 46–67. doi:10.1007/978-3-030-77592-6_3

[14] S. Gnatiuk, et al., Sparse generative embeddings of handwritten digits, in: International Conference on Advanced Computer Information Technologies, ACIT, 2023, 604–607.

[15] I. Ostroumov, et al., A probability estimation of aircraft departures and arrivals delays, in: Computational Science and Its Applications, ICCSA 2021, Lecture Notes in Computer Science, vol. 12950, 2021, 363–377.

[16] Z. Hu, et al., Advanced method for compressing digital images as a part of video stream to pre-processing of UAV data before encryption, in: Advances in Computer Science for Engineering and Education VI. ICCSEEA 2023. Lecture Notes on Data Engineering and Communications Technologies, vol. 181, 2023, 371–381. doi:10.1007/978-3-031-36118-0_33

[17] I. Ostroumov, K. Marais, N. Kuzmenko, Aircraft positioning using multiple distance measurements and spline prediction, Aviation, 26(1) (2022) 1–10.

[18] N. S. Kuzmenko, I. V. Ostroumov, K. Marais, An accuracy and availability estimation of aircraft positioning by navigational aids, in: 2018 IEEE 5th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC), 2018, 36–40. doi:10.1109/MSNMC.2018.8576276

[19] S. Gnatyuk, P. Prystavka, S. Dolgikh, Fairness audit and compositional analysis in trusted AI program, Advances in Computer Science for Engineering and Education VI, ICCSEEA 2023, Lecture Notes on Data Engineering and Communications Technologies, vol. 181, 2023, 690–699. doi:10.1007/978-3-031-36118-0_62