Prediction of Multimodal Poisson Variable using Discretization of Gaussian Data

Evženie Uglickich[1] [a], Ivan Nagy[1,2] [b] and Matej Petrouš[2] [c]

[1] *Department of Signal Processing, The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod vodárenskou věží 4, 18208 Prague, Czech Republic*

[2] *Faculty of Transportation Sciences, Czech Technical University, Na Florenci 25, 11000 Prague, Czech Republic*

Keywords: Poisson Distribution Prediction, Discrete Data, Discretization, Mixture based Clustering, Bayesian Recursive Mixture Estimation.

Abstract: The paper deals with predicting a discrete target variable described by the Poisson distribution based on the discretized Gaussian explanatory data under condition of the multimodality of a system observed. The discretization is performed using the recursive mixture-based clustering algorithms under Bayesian methodology. The proposed approach allows to estimate the Gaussian and Poisson models existing for each discretization interval of explanatory data and use them for the prediction. The main contributions of the approach include: (i) modeling the Poisson variable based on the cluster analysis of explanatory continuous data, (ii) the discretization approach based on recursive mixture estimation theory, (iii) the online prediction of the Poisson variable based on available Gaussian data discretized in real time. Results of illustrative experiments and comparison with the Poisson regression is demonstrated.

1 INTRODUCTION

This paper deals with predicting a discrete variable described by the Poisson distribution. This task is highly desired in various application fields, which deal with modeling a number of random independent events observed with a constant intensity per time unit, for example, social sciences, engineering, medicine and many others (Guenni, 2011). Examples of specific applications of the Poisson models include, e.g., the description of the number of bankruptcies (Jaggia and Kelly, 2018), customer arrivals (Donnelly, 2019; Anderson et al., 2017), network failures (Levine et al., 2011), aircraft shutdowns, patients with specific diseases, file server virus attacks (Doane and Seward, 2010), boarding passengers (Petrouš et al., 2019), etc.

In this paper, the model of the Poisson target variable conditioned by continuous explanatory data is considered. In this area, traditionally, the use of the Poisson regression models (Heeringa et al., 2010; Falissard, 2012; Armstrong et al., 2014; Agresti, 2018) as well as their zero-inflated versions (Long and Freese, 2014; Diallo et al., 2018) can be met. In some sources, the application of linear regression techniques to Poisson-distributed count data due to the high number of their possible realizations is also mentioned, see, for instance, (Agresti, 2012).

As regards the description of multimodal Poisson-distributed data, the publications dealing with mixtures of Poisson distributions (Congdon, 2005), mixtures of Poisson regressions (Lim et al., 2014; Počuča et al., 2020) as well as Poisson-gamma models (Agresti, 2012) can be found in this area. The Gaussian-Poisson mixture models capturing the relationship between the Poisson-distributed and Gaussian variables are described in the papers of (Perrakis et al., 2015; Yu et al., 2016; Zha et al., 2016; Silva et al., 2019). The parameter estimation of the mentioned mixture models is solved primarily using the iterative expectation-maximization (EM) algorithm, see, e.g., (Gupta and Chen, 2011).

The studies of (Li et al., 2010; Bejleri and Nandram, 2018; Petrouš and Uglickich, 2020) consider the Poisson prediction problem close to that discussed in this paper. In the presented paper, the prediction approach is based on the description of the relationship between the target Poisson distributed variable measured for a limited period of time and continu-

[a] https://orcid.org/0000-0003-1764-5924
[b] https://orcid.org/0000-0002-7847-1932
[c] https://orcid.org/0000-0002-8585-8721

Uglickich, E., Nagy, I. and Petrouš, M.
Prediction of Multimodal Poisson Variable using Discretization of Gaussian Data.
DOI: 10.5220/0010575006000608
In *Proceedings of the 18th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2021)*, pages 600-608
ISBN: 978-989-758-522-7

ous explanatory multidimensional Gaussian variable observed permanently. Their joint model is estimated and used for the construction of the Poisson predictive model. The presented solution is based on the three key points: (i) the discretization of the Gaussian explanatory data, (ii) construction of local models of the Poisson target variables on the discretization intervals on explanatory data (i.e., their clusters), and (iii) prediction of the target variable with the help of actual discretization. The discretization of the continuous explanatory measurements is proposed with the help of the recursive mixture-based clustering (Kárný et al., 2006; Nagy and Suzdaleva, 2017) under Bayesian methodology. The similar issue was discussed, e.g., in the papers of (Gupta et al., 2010; Kianmehr et al., 2010; Dash et al., 2011; Sriwanna et al., 2019). The aim of the discretization is a search for clusters in the explanatory data space for the further construction of the Poisson local models on them. The real-time discretization is used for finding the actual learnt models to be used for the prediction.

The layout of the paper is organized as follows: Section 2 represents the preliminary part. It introduces necessary denotations and reminds the basic facts about the maximum likelihood parameter estimation of the Poisson distribution and Bayesian recursive estimation of the Gaussian probability density function. Section 3 is the main emphasis of the paper. Section 3.1 formulates the prediction problem in general. Sections 3.2 presents the discretization and prediction approach for the case of a scalar Gaussian variable, while Section 3.3 generalizes it for multidimensional variables. Section 3.4 summarizes the main steps of the solution in the form of the algorithm. Results of illustrative experiments can be found in Section 4. Section 5 provides conclusions and future plans.

2 PRELIMINARIES

The algorithms presented in this paper are based on the parameter estimation of the Poisson and Gaussian distributions. To specify the used denotations, the estimation approaches are briefly recalled below.

A single Poisson distribution describing the scalar discrete variable y has the form of the probability function (denoted by the pdf along with the probability density function)

$$f\left(y=y_t|\lambda\right)=e^{-\lambda}\frac{\lambda^y}{y!} \qquad (1)$$

with the parameter λ and realizations $y_t \in \{0,1,\dots,N_y\}$ at time $t=1,\dots,T$. The maximum likelihood estimate of the parameter λ is known to be the average of the measured realizations, see, e.g., (Sinharay, 2010)

$$\hat{\lambda}=\frac{1}{T}\sum_{t=1}^{T}y_t. \qquad (2)$$

A single Gaussian pdf describing the scalar continuous variable x has the form

$$f(x|\theta,r)=(2\pi r)^{-\frac{1}{2}}\exp\{-\frac{1}{2r}(x-\theta)^2\} \qquad (3)$$

with the expectation θ, variance r and realizations $x_t \in \mathbb{R}$ at time instants t. In this paper, the variance r is assumed to be known. The unknown expectation θ is estimated recursively according to the Bayesian methodology (Peterka, 1981), where the posterior pdf of θ is evolved in time as follows:

$$f(\theta|x(t))\propto f(x_t|\theta)f(\theta|x(t-1)), \qquad (4)$$

which uses the denotation of the form $x(t)=\{x_0,x_1,\dots,x_t\}$ with the involved prior knowledge x_0. $f(\theta|x(t-1))$ denotes the conjugate prior Gaussian pdf. The recursion starts with the expertly chosen prior pdf, enabling the update of the Gaussian pdf statistics (Peterka, 1981) for the case of the known variance in the following way:

$$V_t = V_{t-1}+x_t, \qquad (5)$$
$$\kappa_t = \kappa_{t-1}+1, \qquad (6)$$

where the initial statistics V_0 and κ_0 can be set with the help of prior or expert knowledge. The point estimate of the expectation θ giving the average of the measured realizations x_t is re-computed at each time instant t

$$\hat{\theta}_t=\frac{V_t}{\kappa_t}. \qquad (7)$$

3 POISSON PREDICTION BASED ON DISCRETIZED GAUSSIAN DATA

3.1 Problem Formulation

Let us observe a system, which changes its behavior in different working modes. The set of observations on the multimodal system includes realizations both of the variables y_t and x_t up to the time $t=T$, and only x_t for $t>T$. The task is to describe the relationship between the Poisson target variable y_t and Gaussian explanatory variable x_t and predict realizations y_t for the time $t>T$ recursively in real time based on the permanently measured data x_t only.

An example of such system can be a bus or tram station, where the number of boarding passengers can be described by the Poisson distribution. It naturally impacts the passenger demand on the stations, which is an importation issue in the transportation data analysis. However, collecting the data sets of the number of boarding passengers is an expensive task under the conditions of the missing infrastructure. Hence, the solution is seen in constructing and estimating the model linking the number of passengers and variables observed around the individual stations under assumption of their normality. The developed model can be then used for predicting the number of passengers.

3.2 Scalar Gaussian Data

First, for the more transparent presentation, the scalar case of the Gaussian explanatory variable x_t will be considered. Here, the denotations for random variables and their realizations will be identical for the simplicity.

The relationship between the Poisson variable y_t and Gaussian variable x_t is generally assumed in the form of the joint pdf

$$f(y_t, x_t|\lambda, \theta) = f(y_t|x_t, \lambda) f(x_t|\theta), \qquad (8)$$

which is decomposed according the chain rule (Peterka, 1981) and assuming the mutual independence of y_t and θ as well as x_t and λ. The marginal pdf $f(x_t|\theta)$ in (8) is the Gaussian model (3) of the explanatory data x_t, which can be estimated recursively in real time. The main problem appeared here is the pdf $f(y_t|x_t, \lambda)$ conditioned by the continuous data x_t, which needs a solution of the task close to the classification of the data x_t among the values of y_t. From this point of view, the relationship between y_t and x_t can be described by the Poisson regression (Heeringa et al., 2010), multinomial logit regression (Tang et al., 2012; Agresti, 2012) or negative binomial regression models (Agresti, 2018). However, this would require analysis of the entire data set, which is not suitable for the recursive real time performance of the prediction algorithm to be developed.

The idea is to express the relationship of y_t and x_t through the discretization of the continuous data, i.e., discretize the explanatory variable x_t so that the Poisson model (1) of the variable y_t exists for each discretization interval of x_t. This will allow to replace the discussed pdf $f(y_t|x_t, \lambda)$ in (8) by the Poisson pdf (1) in the form

$$f(y_t|\lambda_{\tilde{x}_t}), \qquad (9)$$

where $\tilde{x}_t$ is the new discretized random variable such that

$$\tilde{x}_t \in \{1, 2, \ldots, N_{\tilde{x}}\} \qquad (10)$$

and its values label the discretization intervals of the explanatory data x_t. The pdf (9) conditioned by the parameter $\lambda_{\tilde{x}_t}$ exists for each value of $\tilde{x}_t$, i.e., for each discretization interval.

This means that the unknown variables are the parameters θ and λ of the involved pdfs along with the values of the discretized variable $\tilde{x}_t$ at each time instant, which would indicate the current discretization interval where the data item x_t belongs to. These variables have to be estimated in order to use the obtained learnt model describing the relationship of y_t and x_t for the prediction of y_t.

In this paper, the task specified above is proposed to be divided in three parts: (i) the discretization of the explanatory data x_t, which focuses on the estimation of the Gaussian pdf $f(x_t|\theta)$ and resulting in the estimates of θ and $\tilde{x}_t$, (ii) the estimation of the local Poisson models $f(y_t|\lambda_{\tilde{x}_t})$ on the obtained discretization intervals of the explanatory data giving the estimates of λ and (iii) the prediction of the variable y_t based on the actually measured and discretized explanatory data x_t. These parts of the approach are presented below.

3.2.1 Explanatory Data Discretization

This part of the approach deals with the explanatory data x_t available up to the time $t = T$ only. Here, the Gaussian data discretization using the mixture-based clustering (Nagy and Suzdaleva, 2017) inspired by (Kárný et al., 1998; Kárný et al., 2006) will be used. It is explicitly suitable for the mentioned task, as it (i) runs recursively online based on permanently measured data and (ii) allows to set the number of clusters expressing intervals for the discretization of x_t beforehand.

The scheme of the recursive discretization leading to the estimation of the required variables θ and $\tilde{x}_t$ at each time instant includes the following steps:

The joint pdf construction The Bayes rule, see e.g., (Gelman et al., 2013), is applied to the joint pdf of the unknown variables θ and $\tilde{x}_t$ according to (Kárný et al., 1998; Kárný et al., 2006) in the following way:

$$f(\tilde{x}_t, \theta|x(t)) \propto f(x_t, \tilde{x}_t, \theta|x(t-1))$$
$$= f(x_t|\theta, \tilde{x}_t) f(\theta|x(t-1)) f(\tilde{x}_t|x(t-1)), \qquad (11)$$

where the pdf $f(x_t|\tilde{x}_t, \theta)$ is supposed to have a form of $f(x_t|\theta_{\tilde{x}_t})$ conditioned by $\theta_{\tilde{x}_t}$ existing for each value of $\tilde{x}_t$, i.e., $\theta = \{\theta_{\tilde{x}_t}\}_{\tilde{x}_t=1}^{N_{\tilde{x}}}$, the pdf $f(\theta|x(t-1))$ is the prior Gaussian pdf and $f(\tilde{x}_t|x(t-1))$ is a prior vector uniform distribution.

The discretized variable posterior distribution The posterior distribution of $\tilde{x}_t$ based on the current data is derived by marginalizing (11) over the parameters θ, i.e.,

$$f(\tilde{x}_t|x(t)) = \int_{\theta^*} f(x_t|\theta,\tilde{x}_t) f(\theta|x(t-1)) \times f(\tilde{x}_t|x(t-1))\, d\theta, \quad (12)$$

where θ^* denotes the entire definition space. The posterior pdf of $\tilde{x}_t$ is just a vector distribution of the dimension $N_{\tilde{x}}$, where each of its entries provides the probability of the membership of the current data item x_t to each of the $N_{\tilde{x}}$ discretization intervals at time t. These probabilities are called the proximities of the data value x_t to the models $f(x_t|\theta_{\tilde{x}_t})$, see (Nagy et al., 2016; Nagy and Suzdaleva, 2017; Jozová et al., 2021). The point estimate of the variable $\tilde{x}_t$ is a trivial argument of the maxima of the discussed distribution (12), i.e.,

$$\tilde{x}_t = \arg\max_i f(\tilde{x}_t|x(t)), \;\; i \in \{1,2,\ldots,N_{\tilde{x}}\}. \quad (13)$$

To compute the proximities to be used in (12), the realization of the explanatory variable x_t at time t is substituted along with the last available point estimate of the expectation (7) into each Gaussian pdfs (3) for all values of $\tilde{x}_t$ under assumption of the known variance and then normalized (Nagy and Suzdaleva, 2017).

The statistics update Similarly to the recursive mixture estimation (Kárný et al., 1998; Kárný et al., 2006), the normalized proximities are used for the update statistics (5)–(6)

$$V_{i;t} = V_{i;t-1} + m_{i;t} x_t, \quad (14)$$

$$\kappa_{i;t} = \kappa_{i;t-1} + m_{i;t}, \quad (15)$$

where $m_{i;t}$ denotes the i-th normalized proximity from $f(\tilde{x}_t|x(t))$ for $\tilde{x}_t = i$. The updated statistics are used to re-compute the point estimates (7) of the parameters θ for each $\tilde{x}_t$. The recursive computations are repeated until the time $t = T$, while the observations x_t are available.

The results of this part of the approach are the values of $\tilde{x}_t$ denoting the discretization intervals of continuous data at each time instant along with the estimated models of x_t.

3.2.2 Poisson Local Model Estimation

The second part of the solution is aimed at the construction of the Poisson models (9) for each discretization interval locally. Here, it should be reminded that the observations of the multimodal system contain the data sets of y_t and x_t at each time instant up to the time $t = T$. Having the pre-set number of the discretization intervals $N_{\tilde{x}}$ and point estimates of $\tilde{x}_t$ at time t, the parameters $\lambda_{\tilde{x}_t}$ of the Poisson pdfs (9) are estimated according to (2) such that to obtain the average of only those realizations y_t that were measured simultaneously with the x_t discretized to the interval labeled by $\tilde{x}_t$.

The result of this part of the solution is the estimated Poisson models (9) for each discretization interval of the Gaussian explanatory data.

3.2.3 Poisson Prediction

For the time $t > T$ the realizations of y_t are no longer available and should be predicted. For this aim, the learnt models $f(x_t|\theta_{\tilde{x}_t})$ and $f(y_t|\lambda_{\tilde{x}_t})$ are used for each value of $\tilde{x}_t$. The advantage of the approach is the possibility to determine the value of $\tilde{x}_t$ in real time. This is done according to (13) using the actually measured continuous data x_t and computing their proximities to the discretization intervals. Finally, the point prediction of the Poisson target variable is given by

$$\hat{y}_t = \arg\max_j f(y_t|\lambda_{\tilde{x}_t}), \;\; j \in \{0,1,\ldots,N_y\} \quad (16)$$

for the current value of $\tilde{x}_t$ denoted the discretization interval, where the actually measured data item x_t belongs.

Learning the models $f(x_t|\theta_{\tilde{x}_t})$ can be used in this part of the approach as well using the relations (14)-(15) and (7).

3.3 Multidimensional Gaussian Data

This section focuses on a multidimensional case of the Gaussian explanatory variable $x_t = [x_{1;t}, \;\; x_{2;t}, \;\; \ldots \;\; x_{n;t}]$ in the joint pdf (8), which is much more desired from a practical point of view. Here, the individual variables of the vector x_t should be discretized. The common discretization for all of them will lead to the loss of information in case each of them requires its own discretization intervals. This means that they should be treated separately, each with its own individual variable $\tilde{x}_{l;t} \in \{1,\ldots,N_{\tilde{x}_l}\}$, $l = \{1,\ldots,n\}$.

The individual discretization suggests that the approach based on the mixture-based clustering (Nagy and Suzdaleva, 2017) described in Section 3.2.1 should be applied to each Gaussian variable $x_{l;t}$ separately under assumption of mutual independence of the observations in their discretization intervals. The local Poisson models according to Section 3.2.2 are estimated individually for each variable $x_{l;t}$ as well.

During the Poisson prediction part of the solution according to Section 3.2.3, the normalized proximities to the discretization intervals are computed individually using the current data of each Gaussian variable $x_{l;t}$. Further, for all of these variables, the weighted average of the pdfs from all their discretization intervals is calculated

$$f(y_t|\lambda_{\tilde{x}_{l;t}}) = \sum_{i=1}^{N_{\tilde{x}_l}} m_i f(y_t|\lambda_i),\ \forall l = \{1,\ldots,n\}, \quad (17)$$

where $i \in \{1,\ldots,N_{\tilde{x}_l}\}$ is equal to the value of the individual discretized variable $\tilde{x}_{l;t}$, which can be different for each Gaussian variable $x_{l;t}$. The result of this step is n pdfs $f(y_t|\lambda_{\tilde{x}_{l;t}})$, which express the relationship between y_t and each $x_{l;t}$.

Now, using the naïve Bayes principle (Forsyth, 2019) and the Bayes rule, it can be shown that under condition of the assumed independence of individual explanatory variables $x_{l;t}$, it holds (see derivations in Appendix)

$$f(y_t|\lambda_{\tilde{x}_t}) \propto \frac{\prod_{l=1}^{n} f(y_t|\lambda_{\tilde{x}_{l;t}})}{(f(y_t))^{n-1}}, \quad (18)$$

i.e., the product of n obtained pdfs divided by the value of the marginal distribution of y_t raised to the power of $n-1$ gives the resulting predictive model taking into account all the entries of the vector x_t. The denotation $\tilde{x}_t$ as the subscript on the left side of the relation (18) means a set of all $\tilde{x}_{l;t}$.

Finally, the point prediction of the Poisson variable y_t is obtained again via (16).

The presented solution of the multidimensional case is summarized as an algorithm below.

3.4 Algorithm

{Algorithm initialization for $t = 1$}
for all $l \in \{1,2,\ldots,n\}$ **do**
  1. Set the numbers of discretization intervals $N_{\tilde{x}_l}$ for each Gaussian variable using prior or expert knowledge.
  for all $i \in \{1,2,\ldots,N_{\tilde{x}_l}\}$ **do**
    1. Set the initial statistics $V_{i;t-1}$, $\kappa_{i;t-1}$ for each discretization interval of each Gaussian variable using prior or expert knowledge.
    2. Compute the point estimates of the expectations with the help of (7) and initial statistics.
  end for
end for
{Gaussian data discretization)}
for $t = 2,3,\ldots,T$ **do**
  for all $l \in \{1,2,\ldots,n\}$ **do**
    1. Measure the value of $x_{l;t}$.
    for all $i \in \{1,2,\ldots,N_{\tilde{x}_l}\}$ **do**
      1. Substitute the previous point estimate of the expectation $\hat{\theta}_{i;t-1}$ and the actual value of $x_{l;t}$ into the scalar Gaussian pdf (3) of the corresponding explanatory variable, compute the proximity $m_{i;t}$ of this data value to the i-th discretization interval and normalize it.
      2. Update the statistics $V_{i;t}$, $\kappa_{i;t}$ according to (14) and (15).
      3. Re-compute the point estimates of the expectation $\hat{\theta}_{i;t}$ via (7).
      4. Obtain the point estimate of the discretized variable $\tilde{x}_l$ according to (13), which labels the current discretization interval of each Gaussian variable.
    end for
  end for
end for
{Poisson local model estimation}
for all $l \in \{1,2,\ldots,n\}$ **do**
  for all $i \in \{1,2,\ldots,N_{\tilde{x}_l}\}$ **do**
    1. Compute the point estimates of the Poisson pdfs applying (2) to the measurements y_t corresponding to each discretization interval of each explanatory variable $x_{l;t}$.
  end for
end for
{Poisson prediction}
for $t = T+1, T+2,\ldots$ **do**
  for all $l \in \{1,2,\ldots,n\}$ **do**
    1. Measure the value of $x_{l;t}$.
    for all $i \in \{1,2,\ldots,N_{\tilde{x}_l}\}$ **do**
      1. Compute the proximities $m_{i;t}$ using the final point estimates of the expectations and normalize them.
    end for
    2. Compute the weighted average of the Poisson pdfs from all the discretization intervals of each Gaussian variable according to (17).
    3. Obtain the predictive Poisson pdf via (18).
    4. Compute the point prediction of y_t according to (16).
  end for
end for

The algorithm was tested in a free and open source programming environment Scilab (www.scilab.org). The illustrative experiments are presented below.

4 EXPERIMENTS

The aim of the experiments was to verify the proposed approach and demonstrate the prediction of the Poisson variable using the learnt models and available Gaussian data only.

To test the presented algorithm, the simulated data sets containing 3000 values of the Gaussian vector $x_t = [x_{1;t},\ x_{2;t},\ x_{3;t},\ x_{4;t}]$ and the Poisson scalar variable y_t were used. The simulations were prepared so that to have the discretization intervals close to each other for some of the explanatory variables and far from each other for others.

For the experiments, 2800 data items from the randomized data sets were utilized during the discretization part according to Section 3.2.1 as well as the Poisson local model estimation from Section 3.2.2. The rest of 200 simulations were used for the prediction part, see Section 3.2.3.

One of the significant benefits of the proposed approach is a possibility to use the individual prior knowledge of each explanatory variable for the initialization of the mixture-based clustering used for the discretization part of the solution. This prior knowledge is obtained from histograms of the corresponding variables and substituted into the initial statistics $V_{i;t-1}$ with $t = 1$, which were then recursively updated according to Section 3.2.1. All of the four Gaussian explanatory variables had three initialized discretization intervals. For the illustration, the histogram of data of one of them used up to the time $t = 2800$ is presented in Figure 1.

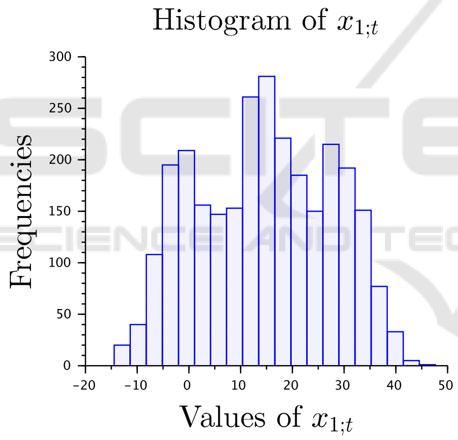

Figure 1: Histograms of one of the explanatory variable.

Three hills with the centers around -2, 19 and 32 respectively can be guessed in the figure. These values are then substituted into the initial statistics $V_{i;t-1}$ and indicate the centers of the three clusters for the discretization part. For the initialization of the counter statistics $\kappa_{i;t-1}$, the initial number of data, i.e., the value of 1, is used for all of the intervals of the variables.

The expectations of the Gaussian models are estimated using the known fixed variance, which has been set equal to 5 for all of them. This choice of the variance value allows to have the clusters of simulated data partially overlapping, which makes them closer to reality. The estimation provides twelve discretization intervals in the form of clusters located around their initially guessed and gradually updated centers. This means that twelve Poisson pdfs are estimated according to Section 3.2.2 on the obtained intervals using the data y_t measured at the same time instants as the Gaussian data belonging to the discretized intervals. The illustrative example depicting the XY graph of the data of the variable $x_{1;t}$ and y_t is demonstrated in Figure 2. In this figure, the point estimates of the Poisson pdfs obtained locally on each of the discretization intervals of $x_{1;t}$ are denoted by '•'.

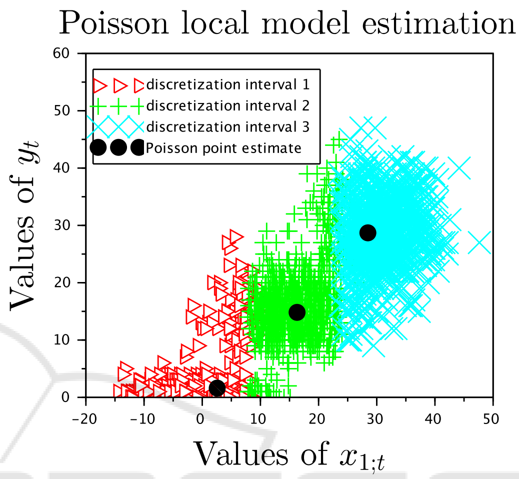

Figure 2: The Poisson local model estimation on the discretization intervals of the Gaussian variable $x_{1;t}$.

In the prediction part, the discretization intervals are determined using the real-time Gaussian data. Using their proximities, the local Poisson pdfs are united into the final predictive pdf according to Section 3.2.3. An example of the obtained prediction results is given in Figure 3.

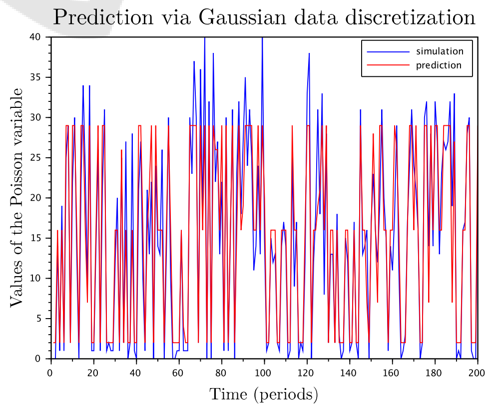

Figure 3: The Poisson variable prediction using the discretization of Gaussian explanatory data.

For a comparison, the prediction based on the Poisson regression described in (Petrouš et al., 2019) was chosen. The mentioned method includes two parts: (i) the Poisson mixture model recursive estimation and (i) the least square Poisson regression estimation, which was applied for the prediction of the Poisson variable. For this algorithm, the histogram-based initialization was set for the Poisson components. For a better visibility, a fragment of the algorithms comparison is presented in Figure 4. It can be seen that the compared results are very close visually.

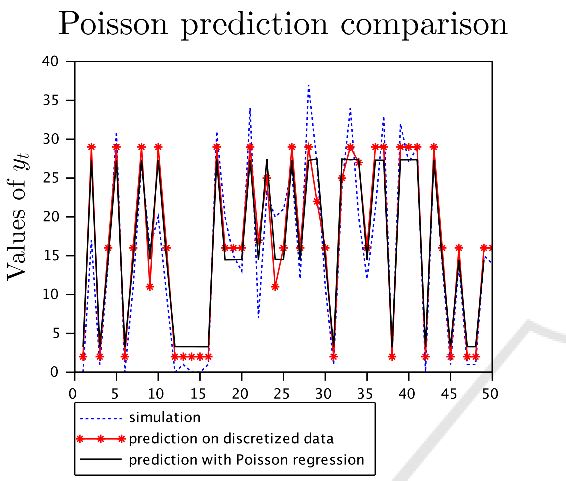

Figure 4: A fragment of the Poisson variable prediction based on the discretization of Gaussian explanatory data compared with the Poisson regression.

To evaluate the prediction accuracy for 200 tested data, the root-mean-square error was computed

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{200}(y_t - \hat{y}_t)^2}{200}}, \tag{19}$$

where $\hat{y}_t$ denotes the prediction at time t. The values of the RMSE averaged over 100 random simulated datasets can be seen in Table 1.

Table 1: Average RMSE.

	Average RMSE
The proposed approach	0.2838133
The Poisson regression	0.2974521

4.1 Discussion

The main aim of the presented study was to verify the algorithm of the prediction of the Poisson variable using real-time continuous data for the estimated models. The aim was successfully achieved. The prediction results look promising and show slight improvements in the comparison with the Poisson regression as one of the theoretical counterparts.

To highlight advantages brought by the proposed approach, it is worth noticing the modeling of the explanatory variables and estimation of the Poisson model conditioned by the results of this modeling in the form of values of the discretized variable. This allows to use available explanatory data for the Poisson prediction in real time recursively, unlike the Poisson regression estimating the entire explanatory data set offline. The use of the individual prior knowledge for the initialization of the algorithm is another significant benefit.

The potential application of the proposed prediction approach can be expected in the area of transportation passenger demand modeling.

The limitations of the approach are concerned with the assumption of the data multimodality necessary for the discretization with the help of the mixture based clustering as well as using the reproducible statistics of the involved pdfs.

5 CONCLUSIONS

The presented paper focused on the task of predicting a discrete target variable described by the Poisson distribution based on the discretized Gaussian explanatory multimodal data. For the discretization, the recursive mixture-based clustering algorithms under Bayesian methodology was used. The Poisson and Gaussian models were estimated on each of the discretization intervals using available data in order to construct the predictive Poisson model, which is used online for the prediction based on the real-time Gaussian data. The prediction results compared with the Poisson regression demonstrated minor improving in the prediction accuracy.

The future work regarding the testing of the algorithm will include (i) experiments with real data, (ii) setting the higher numbers of the discretization intervals, which would help not to loss the information during the discretization, as well as (iii) setting the different numbers of the intervals corresponding to different explanatory variables. The case studies with other continuous distributions will be also explored.

ACKNOWLEDGEMENTS

This work was supported by the project Arrowhead Tools, the project number ECSEL 826452 and MSMT 8A19009.

REFERENCES

Agresti, A., (2012). Categorical Data Analysis. 3rd Ed. John Wiley & Sons.

Agresti, A., (2018). An Introduction to Categorical Data Analysis. 3rd Ed. Wiley, 2018.

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., Cochran, J. J., (2017). Essentials of Modern Business Statistics with Microsoft Office Excel (Book Only) 7th Edition. Cengage Learning.

Armstrong, B.G., Gasparrini, A., Tobias, A., (2014). Conditional Poisson models: a flexible alternative to conditional logistic case cross-over analysis. BMC Medical Research Methodology, 14:122.

Bejleri, V., Nandram, B., (2018). Bayesian and frequentist prediction limits for the Poisson distribution. Communications in Statistics - Theory and Methods, 47:17, 4254-4271.

Congdon, P., (2005). Bayesian Models for Categorical Data. John Wiley & Sons.

Dash, R., Paramguru, R., Dash, R., (2011). Comparative analysis of supervised and unsupervised discretization techniques. International Journal of Advances in Science and Technology, 2(3): 29:37.

Diallo, A. O., Diop, A., Dupuy, J.-F., (2018). Analysis of multinomial counts with joint zero-inflation, with an application to health economics, Journal of Statistical Planning and Inference, vol. 194, p. 85-105.

Doane, D., Seward, L. (2010). Applied Statistics in Business and Economics, 3rd Edition, Mcgraw-Hill.

Donnelly Jr., R., (2019). Business Statistics 3rd Edition, Pearson.

Falissard, B., (2012). Analysis of Questionnaire Data with R. Chapman & Hall/CRC, Boca Raton.

Forsyth, D., (2019). Applied Machine Learning. Springer.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B., (2013). Bayesian Data Analysis (Chapman & Hall/CRC Texts in Statistical Science), 3rd ed., Chapman and Hall/CRC.

Guenni L.B., (2011). Poisson Distribution and Its Application in Statistics. In: Lovric M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_448

Gupta, A., Mehrotra, K., Mohan, C. K., (2010). A clustering-based discretization for supervised learning. Statistics & Probability Letters, 80(9-10): 816-824.

Gupta, M. R. and Chen, Y., (2011). Theory and Use of the EM Method. (Foundations and Trends(r) in Signal Processing). Now Publishers Inc.

Heeringa, S.G., West, B.T., Berglung, P.A., (2010). Applied Survey Data Analysis. Chapman & Hall/CRC.

Jaggia, S., Kelly, A., (2018). Business Statistics: Communicating with Numbers. 3rd Edition. McGraw-Hill Education.

Jozová, Š., Uglickich, E., Nagy, I., (2021). Bayesian mixture estimation without tears. In: Proceedings of the 18th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2021), accepted.

Kárný, M., Böhm, J., Guy, T.V., Jirsa, L., Nagy, I., Nedoma, P., Tesař, L., (2006). Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer, London.

Kárný, M., Kadlec, J., Sutanto, E.L., (1998). Quasi-Bayes estimation applied to normal mixture. Preprints of the 3rd European IEEE Workshop on Computer-Intensive Methods in Control and Data Processing, Eds: J. Rojíček, M. Valečková, M. Kárný, K. Warwick, p. 77–82, CMP '98 /3./, Prague, CZ, 07.09.1998–09.09.

Kianmehr, K., Alshalalfa, M., Alhajj, R., (2010). Fuzzy clustering-based discretization for gene expression classification. Knowledge and Information Systems, vol. 24, 441-465.

Levine, D. M., Stephan, D.F., Krehbiel, T.C., Berenson, M.L., (2011). Statistics for Managers Using Microsoft ® Excel, Sixth Edition. Boston, MA: Prentice Hall.

Li, Y., Sha, Y., Zhao, R., (2010). Poisson prediction of the loss of teachers in high schools. In: Proceedings of 2010 International Conference on Multimedia Technology, Ningbo, China, pp. 1-3.

Lim, H. K., Li, W. K., Yu, P. L.H., (2014). Zero-inflated Poisson regression mixture model, Computational Statistics & Data Analysis, vol. 71, p. 151-158.

Long, J. S., Freese, J., (2014). Regression Models for Categorical Dependent Variables Using Stata. 3rd Ed. Stata Press.

Nagy, I., Suzdaleva, E., 2017. Algorithms and Programs of Dynamic Mixture Estimation. Unified Approach to Different Types of Components, SpringerBriefs in Statistics. Springer International Publishing, 2017.

Nagy, I., Suzdaleva, E., Kárný, M., Mlynářová, T., (2011). Bayesian estimation of dynamic finite mixtures. International Journal of Adaptive Control and Signal Processing. 25(9): 765-787.

Nagy, I., Suzdaleva, E. and Pecherková, P., (2016). Comparison of various definitions of proximity in mixture estimation. In: Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2016), pp. 527-534.

Perrakis, K., Karlis, D., Cools, M., Janssens, D., (2015). Bayesian inference for transportation origin-destination matrices: the Poisson-inverse Gaussian and other Poisson mixtures. Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 178, p. 271-296.

Peterka, V., (1981). Bayesian system identification, in Eykhoff, P. (Ed.), Trends and Progress in System Identification. Oxford, Pergamon Press, pp. 239-304.

Petrouš, M., Suzdaleva, E., Nagy, I., (2019). Modeling of passenger demand using mixture of Poisson components. In: Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2019), p. 617-624.

Petrouš, M. and Uglickich, E., (2020). Modeling of mixed data for Poisson prediction. In: Proceedings of IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI 2020), p. 77-82.

Počuča, N., Jevtić, P., McNicholas, P. D., Miljkovic, T., (2020). Modeling frequency and severity of claims with the zero-inflated generalized cluster-weighted models. Insurance: Mathematics and Economics, vol. 94, p. 79-93.

Silva, A., Rothstein, S.J., McNicholas, P.D., Subedi, S., (2019). A multivariate Poisson-log normal mixture model for clustering transcriptome sequencing data. BMC Bioinformatics 20, 394.

Sinharay, S., (2010). Discrete Probability Distributions. In: International Encyclopedia of Education. 3rd Edition), Editor(s): P. Peterson, E. Baker, B. McGaw, Elsevier, pp.132-134.

Sriwanna, K., Boongoen, T., Iam-On, N., (2019). Graph clustering-based discretization approach to microarray data. Knowledge and Information Systems, vol. 60, 879-906.

Tang, W., He, H., Tu, X. M., (2012). Applied Categorical and Count Data Analysis. Chapman and Hall/CRC, 2012.

Yu, J., Gwak, J., Jeon, M., (2016). Gaussian-Poisson mixture model for anomaly detection of crowd behaviour. In: Proceedings of 2016 International Conference on Control, Automation and Information Sciences (ICCAIS), pp. 106-111.

Zha, L., Lord, D., Zou, Y., (2016). The Poisson inverse Gaussian (PIG) generalized linear regression model for analyzing motor vehicle crash data. Journal of Transportation Safety & Security, vol. 8, p. 18-35.

APPENDIX

Using the naïve Bayes principle (Forsyth, 2019) and the assumption of the independence of the measured random variables x_1 and x_t, it holds

$$\begin{aligned} f(y|x_1,x_2) &\propto f(x_1,x_2|y)f(y) \\ &= f(x_1|y)f(x_2|y)f(y). \end{aligned} \tag{20}$$

According to the Bayes rule, it can be written

$$f(x_1|y) = \frac{f(y|x_1)f(x_1)}{f(y)}, \tag{21}$$

$$f(x_2|y) = \frac{f(y|x_2)f(x_2)}{f(y)}. \tag{22}$$

Substituting (21) and (22) into (20), it is obtained

$$\begin{aligned} f(y|x_1,x_2) &\propto \frac{f(y|x_1)f(x_1)}{f(y)}\frac{f(y|x_2)f(x_2)}{f(y)}f(y) \\ &= \frac{f(y|x_1)f(x_1)f(y|x_2)f(x_2)}{f(y)} \\ &= \frac{f(y|x_1)f(y|x_2)}{f(y)}f(x_1)f(x_2), \end{aligned} \tag{23}$$

where $f(x_1)f(x_2)$ is a constant value for the measured data items x_1 and x_2.