Statistical Audit via Gaussian Mixture Models in Business Intelligence Systems

Bruno H. A. Pilon
Dep. of Electrical Engineering
University of Brasilia (UnB)
Brasilia - DF, Brazil
bhernandes@gmail.com

João Paulo C. L. da Costa
Dep. of Electrical Engineering
University of Brasilia (UnB)
Brasilia - DF, Brazil
jpdacosta@unb.br

Juan J. Murillo-Fuentes
Dep. of Signal Theory
University of Sevilla
Sevilla, Spain
murillo@us.es

Rafael T. de Sousa Júnior
Dep. of Electrical Engineering
University of Brasilia (UnB)
Brasilia - DF, Brazil
desousa@unb.br

ABSTRACT

A Business Intelligence (BI) System employs tools from several areas of knowledge for the collection, integration and analysis of data to improve business decision making. The Brazilian Ministry of Planning, Budget and Management (MP) uses a BI System designed with the University of Brasília to ascertain irregularities on the payroll of the Brazilian federal government, performing audit trails on selected items and fields of the payroll database. This current auditing approach is entirely deterministic, since the audit trails look for previously known signatures of irregularities which are composed by means of an ontological method used to represent auditors concept maps. In this work, we propose to incorporate a statistical filter in this existing BI system in order to increase its performance in terms of processing speed and overall system responsiveness. The proposed statistical filter is based on a generative Gaussian Mixture Model (GMM) whose goal is to provide a complete stochastic model of the process, specially the latent probability density function of the generative mixture, and use that model to filter the most probable payrolls. Inserting this statistical filter as a pre-processing stage preceding the deterministic auditing showed to be effective in reducing the amount of data to be analyzed by the audit trails, despite the penalty fee intrinsically associated with stochastic models due to the false negative outcomes that are not further processed. In our approach, gains obtained with the proposed pre-processing stage overcome impacts from false negative outcomes.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Probabilistic algorithms, statistical software, time series analysis.

Keywords

Business intelligence, statistical audit, Gaussian mixtures.

1. INTRODUCTION

Mixture models constitute a versatile probabilistic tool for representing the presence of subpopulations within a set of observations. They thus facilitate a much more detailed description of complex systems, describing different features of the data by inferring all the parameters of each component of the mixture and by explaining how the set of sources interact together to form a mixture model.

Evidences of the versatility of mixture models is their application in such diverse areas such as astronomy [1], ecology [2] and engineering [3]. In the context of Business Intelligence (BI) systems, mixture models can be used to represent arbitrarily complex probability density functions [4]. This characteristic makes them a reliable choice for representing complex likelihood functions in supervised learning scenarios [5], or priors for Bayesian parameter estimation [6].

In this work, mixture models techniques are applied in a real world scenario. The Human Resources Auditing Department (CGAUD)[1], subordinated to the Brazilian Ministry of Planning, Budget and Management (MP)[2], uses a specialized BI System to ascertain irregularities on the payrolls of the Brazilian federal government. The initial BI solution was presented in [7] and several improvements were proposed in [8], [9] and [10].

Methods for irregularities detection are usually classified in two categories [11]. One is knowledge-based detection, where fraudulent occurrences are previously defined and categorized. Thus, in this kind of detection, the irregularity must be known and described *a priori* and the system is usually unable to deal with new or unknown irregularities. Knowledge-based intrusion detection schemes in network and computer systems are shown in [12].

Alternatively, a behavior-based fraud detection scheme assumes that an irregularity can be detected by observing occurrences that are most dissimilar from the norm [11]. A

[1]In portuguese, *Coordenadoria Geral de Auditoria.*
[2]In portuguese, *Ministério do Planejamento, Orçamento e Gestão.*

valid and standardized behavior can be extracted from previous reference information, and this model can be compared to a fraudulent candidate in order to check for the degree of divergence between them. In [13], the authors present a credit card fraud detection method using neural networks trained with previous related data. In [14], a method based on a Gaussian mixture model of network traffic is described for intrusion detection in mobile networks.

The current CGAUD BI system is entirely based on a knowledge-based approach for irregularity detection. The existing system uses audit trails built with ontological indexation via concept maps in order to detect inconsistencies [7, 8, 10]. The audit trails comprise a set of heuristics based on complex Brazilian federal legislation, which dictate the income of each public employee according to his career and his position in the public administration organization.

In addition to a complex regulatory basis, the amount of data periodically generated regarding the payroll of federal employees is massive: Around 14GB of raw data per month, and more than 200 million rows in the financial data table each year [9]. Thus, the processing cost of auditing this amount of data is very high, since each audit trail has to go through all database performing relational statements on the search for irregularities.

In fact, whereas the monthly payroll of the Brazilian federal government is around 12.5 billion *reais*, the current BI system is capable of auditing approximately 5 billion *reais* each month [10].

This paper proposes a complementary statistical approach, with a generative Gaussian Mixture Model (GMM) filter in a pre-processing stage with the objective of compute payrolls with low probability of being irregular and exclude them of the following audit trails. By learning a mixture model that represents the most probable behavior of the payrolls of the Brazilian federal government, we are able to perform a selection on all payrolls and delivers to the audit trails only payrolls that diverges the most from the norm. This new approach significantly increases the efficiency of the BI system and its processing capacity, with a penalty of losing a few false negatives at this stage.

The remainder of this paper is organized as follows. Section 2 defines the finite mixture models framework, including the specific case of GMM. In Section 3, the current BI system that is operated by CGAUD is described. Section 4 includes the new proposal of the statistical audit module via GMM for the deterministic BI system. In Section 5, experimental results and discussions are presented and in Section 6 conclusions are drawn and future developments are suggested.

2. FINITE MIXTURE MODELS

A convex combination of two or more probability density functions is a *mixture*. The approximation of any arbitrary distribution can be achieved by the combination of the properties of a set of individual probability density functions [15], making mixture models a powerful tool for modeling complex data. While within a parametric family, mixture models offer malleable approximations in non-parametric settings and, although based on standard distributions, mixture models pose highly complex computational challenges [16].

To accompany our model, let $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L)$ be an unlabeled random sample obtained in an independent and identically distributed (iid) manner. The pdf of a mixture model is defined as

$$p(\mathbf{x}|\Theta) = \sum_{k=1}^{K} \alpha_k p_k(\mathbf{x}|\theta_k) \quad k = 1, \ldots, K, \tag{1}$$

where $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$ is a set of d observed random samples, $K \in \mathbb{Z}^+$ is the number of components (sources) in the mixture, $p_k(x|\theta_k)$ is the pdf of the k^{th} component and $\Theta = (\alpha_1, \ldots, \alpha_K, \theta_1, \ldots, \theta_K) \in \Omega$ is the set of parameters of the mixture, with Ω being the parameter space of all possible combinations of values for all the different parameters of the mixture. The collection α_k is the mixing proportion (or weighting factor) of the k^{th} component, representing the probability that a randomly selected $\mathbf{x}_i \in \boldsymbol{\mathcal{X}}$ was generated by the k^{th} component.

In the particular case of a Gaussian mixture model, (1) can be written as

$$p(\mathbf{x}|\Theta) = \sum_{k=1}^{K} \alpha_k p_k(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^d$ is the mean vector and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$ is the covariance matrix, both of them originated by the k^{th} Gaussian component. Each of those component density is a Gaussian function of the form

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\}. \tag{3}$$

Note that the Gaussian mixture model is completely parametrized by its mean vectors, covariance matrices and mixture weights from all component densities [17].

Given that (1) and (2) represent a convex combination of K distributions [15], it can be stated that

$$\alpha_k \geq 0, \text{ for } k \in \{1, \ldots, K\}, \text{ and } \sum_{k=1}^{K} \alpha_k = 1. \tag{4}$$

In addition, since each $p_k(\mathbf{x}|\theta_k)$ defines a pdf, $p(\mathbf{x}|\Theta)$ will also be a pdf [15].

One straightforward interpretation of mixture models is that (1) describes a complete stochastic model [18], thus giving us a recipe to generate new data points. Another point of view, in the mixture model context, is that any observed data sample is generated from a combination of K distinct random processes, each one modeled by the density $p_k(x|\theta_k)$, with α_k defining the proportion of a particular random process in the overall observations.

2.1 Estimation of Parametric Mixture Models

Once defined the mixture model and the particular case of a GMM, next a numerical approach that will allow the estimate of the parameters set Θ is presented.

Let $\mathbf{X} \in \mathbb{R}^{N \times L}$ be a set of N unlabeled observations, where $\mathbf{x}_{ik}$ is the value of the i^{th} observation for the k^{th} component. Since the observed set $\mathbf{X}$ is independently and identically distributed (iid), the joint pdf for $\mathbf{X}$ can be written as

$$p(\mathbf{X}|\Theta) = \prod_{i=1}^{N} p(\mathbf{x}_i|\theta_1, \ldots, \theta_k). \quad (5)$$

The likelihood function of the data, also assuming that $\mathbf{x}_i$ are independently distributed, is defined as

$$p(\mathbf{X}|\Theta) = \mathcal{L}(\Theta|\mathbf{X}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \alpha_k p_k(\mathbf{x}_i|\theta_k). \quad (6)$$

The likelihood can be thought of as a function of the parameters Θ where the observed data $\mathbf{X}$ is fixed. In the maximum likelihood problem, our goal is to find the Θ that maximizes $\mathcal{L}(\Theta|\mathbf{X})$, thus determining which parameters values are more likely for the observed values [19]:

$$\Theta^* = \arg\max_{\Theta \in \Omega} \mathcal{L}(\Theta|\mathbf{X}). \quad (7)$$

In general cases, it is often preferable to maximize $\log(\mathcal{L}(\Theta|\mathbf{X})$ instead, since it is analytically easier [19]. However, in many cases an analytical solution is not possible to develop. One alternative is to maximize the likelihood in an expectation-maximization approach.

2.2 Expectation Maximization Algorithm

The expectation maximization (EM) algorithm is an iterative method for estimating the maximum likelihood of a stochastic model where exists a dependency upon latent, or unobserved, data [20].

Throughout the remainder of this subsection, the EM algorithm is used to obtain an accurate approximation of the maximum likelihood of a mixture model which has *incomplete* data associated with it. This consideration is taken into account when optimizing the likelihood function is analytically intractable, but the likelihood function can be simplified by assuming the existence of additional but missing values [19].

Therefore, let $\boldsymbol{\mathcal{X}}$ be a random incomplete observed data set, $\boldsymbol{\mathcal{Y}}$ be a random unobserved data set and $\boldsymbol{\mathcal{Z}} = (\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$ be a *complete* data set.

To establish the notation, let $p(\mathbf{z}|\Theta) = p(\mathbf{x}, \mathbf{y}|\Theta)$ be the joint pdf of the random variables $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$, $g(\mathbf{x}|\Theta)$ be the marginal pdf of $\boldsymbol{\mathcal{X}}$ and $k(\mathbf{y}|\mathbf{x}, \Theta)$ be the conditional probability of $\boldsymbol{\mathcal{Y}}$ given $\boldsymbol{\mathcal{X}} = \mathbf{x}$.

The EM algorithm aims to maximize the incomplete data log-likelihood [20],

$$\log[\mathcal{L}(\Theta|\boldsymbol{\mathcal{X}})] = \log[g(\mathbf{x}|\Theta)] \quad \text{for } \Theta \in \Omega,$$

by using $p(\mathbf{x}, \mathbf{y}|\Theta)$ and $g(\mathbf{x}|\Theta)$. From Bayes' rule, $p(\mathbf{z}|\Theta)$ can be represented as

$$p(\mathbf{z}|\Theta) = p(\mathbf{x}, \mathbf{y}|\Theta) = k(\mathbf{y}|\mathbf{x}, \Theta) \cdot g(\mathbf{x}|\Theta), \quad (8)$$

for $\mathbf{x} \in \boldsymbol{\mathcal{X}}$ and $\mathbf{y} \in \boldsymbol{\mathcal{Y}}$.

The E-step of the EM algorithm seeks to find the expected value of the complete data log-likelihood, defined as

$$\log[\mathcal{L}(\Theta|\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})] = \log[p(\mathbf{x}, \mathbf{y}|\Theta)]. \quad (9)$$

In (9), the observed samples $\boldsymbol{\mathcal{X}}$ and some *a priori* parameter estimate $\Theta^p \in \Omega$ are given as inputs. In addition, an auxiliary function $\mathcal{Q}$ is defined such as

$$\mathcal{Q}(\Theta|\Theta^p) = \mathbb{E}[\log[p(\mathbf{x}, \mathbf{y}|\Theta)|\mathbf{x}, \Theta^p]], \quad (10)$$

where $\mathbf{x} \in \boldsymbol{\mathcal{X}}$, $\mathbf{y} \in \boldsymbol{\mathcal{Y}}$, $\Theta^p \in \Omega$ and $\mathbb{E}[\cdot]$ denotes the expectation operator. The key thing in (10) is that $\boldsymbol{\mathcal{X}}$ and Θ^p are constants, Θ is a regular variable which we want to optimize and $\boldsymbol{\mathcal{Y}}$ is a random variable governed by the distribution $k(\mathbf{y}|\mathbf{x}, \Theta)$.

The M-step of the EM algorithm intents to maximize (10) by selecting a new set of parameters $\Theta^* \in \Omega$ such that

$$\Theta^* \in \arg\max_{\Theta \in \Omega} \mathcal{Q}(\Theta|\Theta^p). \quad (11)$$

The EM algorithm presented in [20] can abstractly be summarized as follows:

1. E-Step: Calculate $\mathcal{Q}(\Theta|\Theta^p)$.
2. M-Step: Pick $\Theta^* \in \arg\max_{\Theta \in \Omega} \mathcal{Q}(\Theta|\Theta^p)$.
3. $\Theta^p \leftarrow \Theta^*$.
4. Iterate (1)-(3) until some convergence criterion is met.

At each iteration, the EM algorithm increases the log-likelihood converging to a local maximum [21]. More properties on convergence of the EM algorithm can be found at [20] and [22].

3. THE CURRENT BI SYSTEM MODEL

The Integrated System for the Administration of Human Resources (SIAPE)[3], is a national system that manages the monthly payroll of Brazilian federal employees [23].

SIAPE includes information of approximately two and half million workers among active, retired and pensioners; 14GB of raw data are generated each month, with 212 fields of personal, functional and financial data [9]. In the 2012 fiscal year, the size of the SIAPE's database file ended up with more than 27 million rows for the public workers table and about 200 million rows in tha financial data table [9].

Furthermore, there is a massive legal basis from which the payroll of the Brazilian federal staff are generated. The Federal Constitution of Brazil, laws, decrees and executive orders created more than 2,200 different rubrics [10], the basic element of the payroll, consisting of a positive or negative value according to the characteristics of the position of the employee in the public administration organization.

According to the Brazilian legislation, CGAUD is the responsible department for auditing the rubrics of every payroll aimed at fraud detection such as incompatibility of benefits, inconsistencies and irregularities. Before the initial BI solution proposed in [7], CGAUD performed the audit process in a manual fashion.

After the implementation of the BI System proposed in [7], with several improvements proposed in [8], [9] and [10], the current state of the art BI system for auditing SIAPE is based on an ontology indexation process via concept maps in order to detect irregularities on the payrolls, big data technologies such as Hadoop and Hbase for increasing the performance of the processing stage and a reimbursement

[3]In portuguese, *Sistema Integrado de Administração de Recursos Humanos*.

tracking system for monitoring the payroll of federal employees who have to reimburse the Brazilian Treasury. Fig 1 shows the architecture of the current BI system. Please refer to [10] for details on the existing BI architecture.

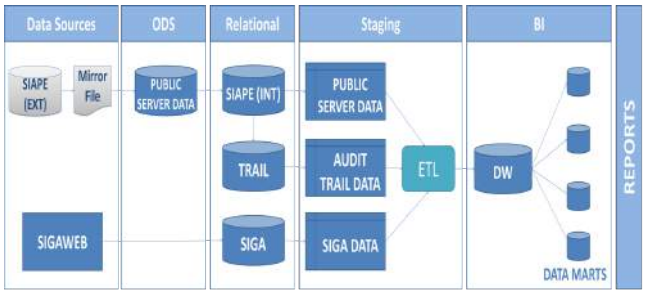


Figure 1: Architecture of the current state-of-the-art BI system [10]

Despite the fact that the audit process is made before the payroll is actually paid to the employee, the existing audit process is fully based on audit trails, *i.e.* a deterministic analysis of the complete data structure where the information is presumably encoded in the *hypothesis*. Ontological audit trails mapping summarizes a set of hypothetic rules based on Brazilian legislation, such as incompatibility of rubrics, and the real world data validates or refutes those hypothesis. Fig. 2 shows an example of audit trail concept map. Please refer to [7] for more details on the construction of audit trails.

Hence, the current audit process has no predictive component and no pre-processing of the huge amount of monthly incoming data, having to check every row of a specific rubric in order to detect any irregularity.

4. STATISTICAL ANALYSIS ON A DETERMINISTIC BI ENVIRONMENT

Considering the amount of data to be analyzed by the current BI system, and the rising trend of the number of audit trails which also causes the processing requirements of the system to rise proportionally, this work proposes a complementary statistical approach to the system based on GMM described in Section 2. Focused on the *data*, the system aims to model a pdf for each category of the Brazilian federal staff with common payroll characteristics (professors, police officers, judges, etc.), hence defining a regular behavior for the payrolls of each category.

After the definition of a standard payroll behavior for a given category of employees, the next goal is to classify each individual payroll as regular or possibly inconsistent. In other words, the principle of the proposed system can be stated as the higher the probability of a random payroll, the less likely that payroll is to be inconsistent. One way to validate this thesis is to use the current BI system as a qualitative measure, where the removal of the most probable payrolls from the audit trails should not drastically impact the result of the original audit trails.

The data used in this work consists of 101,400 payroll entries of the federal professors category, since this is one of the categories with more employees of the Brazilian federal staff [24]. The chosen month of application of the proposed technique was June/03, since federal employees in Brazil are

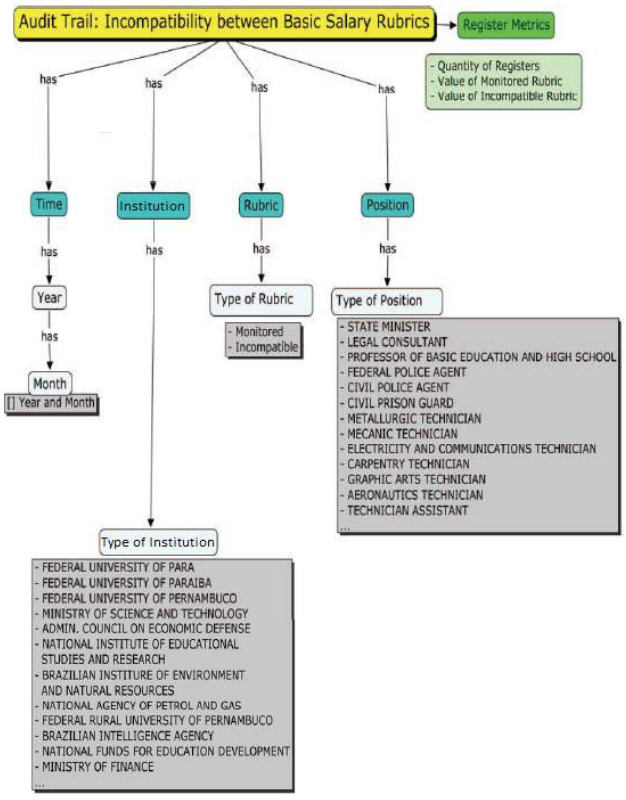


Figure 2: Example of a concept map for an audit trail [7]

monthly payed and June is one month of the year with a high variance among all the other months given that the first part of the 13rd salary is payed on that month for a substantial part of the governmental staff [25].

Each payroll is arranged in a two dimensional structure, where instead of dealing with more than 2,200 different rubrics, the whole information of the rubrics is condensed into gross income in one dimension and total deductions in the other dimension. Generalizing positive rubrics (incomes) in one dimension and negative rubrics (deductions) in another dimension resulted in the scatter diagram of the data sets shown in Fig. 3.

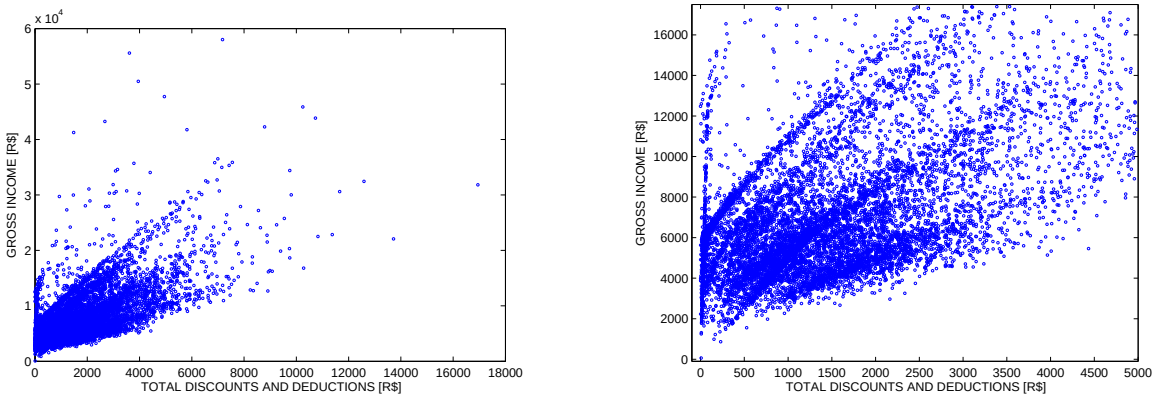
Fig. 3 shows a 10,000 points sample of the data. As expected, it can be noted in the cropped scatter plot a positive correlation between the amount of gross income and the total deductions and discounts from the payroll.

4.1 Statistical Audit Module

The proposed statistical audit module, based on a generative GMM pdf, was incorporated into the original BI system described in Section 3 between SIAPE and TRAIL databases, inside the relational stage (see Fig. 1). A block diagram in Fig. 4 shows the new audit module in the BI architecture.

The statistical audit module acts as a filter for the audit trail database, removing high probability payrolls contained in the SIAPE database and feeding the TRAIL database only with payrolls that are most dissimilar from the normal payroll behavior modeled by a GMM pdf.

Given that the TRAIL database is generated through



(a) (b)

Figure 3: Scatter plot (a) of 10,000 samples of payroll data, with gross income in one dimension (ordinate) and total discounts and deductions in the other dimension (abscissa). Both dimensions are plotted in *Reais*, the Brazilian currency. (b) shows (a) zoomed around the origin, for a better visualization of the correlation profile.

Figure 4: Block architecture of the proposed statistical audit module solution in the original BI architecture shown in Fig. 1.

computationally cost relational statements between each audit trail and the whole SIAPE database, and the GMM pdf inside the statistical audit module is generated by a one time optimization process via EM algorithm, a positive trade off between cost and effort could be established. Although any probabilistic method has a certain degree of uncertainty associated with it, which in this specific case translates into losing some false negatives (*i.e.* high probability payrolls that have some kind of irregularity) in the statistical audit module, the gain obtained in terms of computational efficiency and velocity of execution enables the whole BI system to audit a more significant portion of the overall payroll of Brazilian public employees.

4.2 GMM for Statistical Auditing

The approach chosen to model the data is a finite Gaussian mixture model (GMM), which gives a complete statistical description of the latent underlying system that generated the data. GMM is a parametric model, completely defined by its mixing weights, mean vectors and covariance matrices. Therefore, in the context of this work, GMM can be seen as a generative model capable of defining the probability of a random payroll to occur.

In order to obtain the pdf of our GMM, *i.e.* learn the parameters $\Theta = \{\alpha_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$, we apply the EM algorithm for a GMM. The derivation of closed form solutions for the equations presented in Subsection 2.2 for a mixture of multivariate Gaussians require techniques which are beyond the scope of this project. For a detailed derivation of those, please refer to [19].

Using the framework previously defined in subsection 2.2, a closed form solution for the auxiliary function $\mathcal{Q}$, the conditional expectation of the complete data, can be written as:

$$\mathcal{Q}(\Theta|\Theta^p) = \sum_{i=1}^{N}\sum_{k=1}^{K} \frac{\alpha_k^p p_k(\mathbf{x}_i|\theta_k)}{p(\mathbf{x}_i|\Theta)} \log(\alpha_k) + \sum_{i=1}^{N}\sum_{k=1}^{K} \frac{\alpha_k^p p_k(\mathbf{x}_i|\theta_k)}{p(\mathbf{x}_i|\Theta)} \log(p_k(\mathbf{x}_i|\theta_k) \quad (12)$$

The expression for $\mathcal{Q}$ derived in (12) appears in the E-step of the EM algorithm and may be maximized for a particular pdf p_k.

Assuming p_k being a multivariate Gaussian distribution in the form of (3), and recalling that

$$\Theta^p = (\alpha_1^p, \ldots, \alpha_k^p, \theta_1^p, \ldots, \theta_k^p) \in \Omega$$

is our prior set of parameters, the goal is to implement the M-step of the EM algorithm to obtain updated maximizers denoted by $\Theta^* = (\alpha_1^*, \ldots, \alpha_k^*, \theta_1^*, \ldots, \theta_k^*) \in \Omega$.

This can be achieved by maximizing $\mathcal{Q}$ with respect to α_k and $\theta_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, leading to the *updated parameter equations* of the M-step of the EM algorithm for α_k^*, $\boldsymbol{\mu}_k^*$ and $\boldsymbol{\Sigma}_k^*$ as being, respectively:

$$\alpha_k^* = \frac{1}{K}\sum_{i=1}^{N} \frac{\alpha_k^p p_k(\mathbf{x}_i|\theta_k^p)}{p(\mathbf{x}_i|\Theta^p)}; \quad (13)$$

$$\boldsymbol{\mu}_k^* = \frac{\sum_{i=1}^{N} \mathbf{x}_i \frac{\alpha_k^p p_k(\mathbf{x}_i|\theta_k^p)}{p(\mathbf{x}_i|\Theta^p)}}{\sum_{i=1}^{N} \frac{\alpha_k^p p_k(\mathbf{x}_i|\theta_k^p)}{p(\mathbf{x}_i|\Theta^p)}}; \quad (14)$$

$$\boldsymbol{\Sigma}_k^* = \frac{\sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu}_k^*)(\mathbf{x}_i - \boldsymbol{\mu}_k^*)^T \frac{\alpha_k^p p_k(\mathbf{x}_i|\theta_k^p)}{p(\mathbf{x}_i|\Theta^p)}}{\sum_{i=1}^{N} \frac{\alpha_k^p p_k(\mathbf{x}_i|\theta_k^p)}{p(\mathbf{x}_i|\Theta^p)}}. \quad (15)$$

Again, please refer to [19] for a detailed derivation of (13), (14) and (15).

4.2.1 Initialization and Convergence Issues for EM

A crucial point of the EM algorithm is the initial set of parameters Θ^p of the model. A standard way to obtain Θ^p is to choose random α_k values uniformly from $[0, 1]$ and estimate the individual source parameters with a M-step [26].

In order to deal with the effects of random initialization and a possible convergence to a local maximum, all estimations can be repeated a number of times and the solution with the highest likelihood is selected [26].

5. RESULTS AND DISCUSSION

One key aspect of modeling the payroll data set as a bidimensional GMM is the number of source components K in (1). Whereas the number of sources can be linked directly to the number of clusters of a classification algorithm, in many cases extending the finite mixture model such as $K \to \infty$ produces densities whose generalization is highly competitive with other commonly used methods [27].

Recalling that our classification proposal is not based on the number of classes, or source components, but instead is based exclusively on the pdf generated by the mixture model, where payrolls that have a probability level above a certain threshold are classified as less likely to be irregular. In this particular case, not limiting the number of classes *a priori* removes an extra parameter of the stochastic model to be estimated.

Hence, whereas the number of clusters increases with the number of sources in a classical mixture model, the underlying pdf of the mixture tends to stabilize as shown in Fig. 5. Due to computational constraints, it is not possible to extend the number of classes to infinity, but in our particular case the pdf showed to be stable with a number of sources $K \geq 30$. It is important to state that, in Fig. 5, our main interest reside at the areas inside the red and brown contours. These are the areas with highest probability values and thus these are the areas we look for stability.

Another interesting feature noticed in Fig. 5 is the decay rate of the log-likelihood function. As the number of sources increases, the resulting likelihood function tends to increase as well [26]. Taking that assumption to the limit, when the number of sources is equal to the number of observed data points, the likelihood of each point being generated by its own data source is equal to 1. Nevertheless, it can be observed in our system that, as the number of source components increase, the rate of decay of the log-likelihood function decreases, leading to the conclusion that adding more source components to the mixture does not add much more significant information about the system.

With the number of sources $K = 30$ defined in (1), the EM algorithm was applied to the set of data of SIAPE regarding the federal professors staff, originally a data set with 101,400 payroll entries.

In order to avoid a possible convergence to a local maximum, all estimations are made 15 times [26] with different random set of initial parameters $\Theta^p = (\alpha_k^p, \boldsymbol{\mu}_k^p, \boldsymbol{\Sigma}_k^p) \in \Omega$ in (13), (14) and (15). The initial set of parameters are obtained by randomly choosing K observations from $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_L)$ in (1) as initial component means. The mixing weights are uniform. The covariance matrices for all components are diagonal, where the element j on the diagonal is the variance of $\boldsymbol{\mathcal{X}}(:, j)$.

The convergence criteria adopted for the EM algorithm is the termination tolerance on the log-likelihood function in (8), where the algorithm stops when the new guesses of parameters Θ^* produce only minimal increments of the log-likelihood function given in (8), *e.g.* increments smaller than 10^{-5}. Thus, the convergence criteria is met when only negligible improvements of the solution can be achieved by performing new iterations.

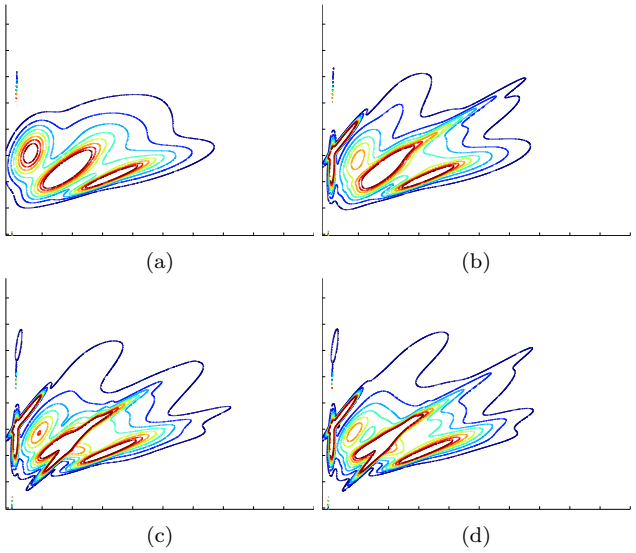


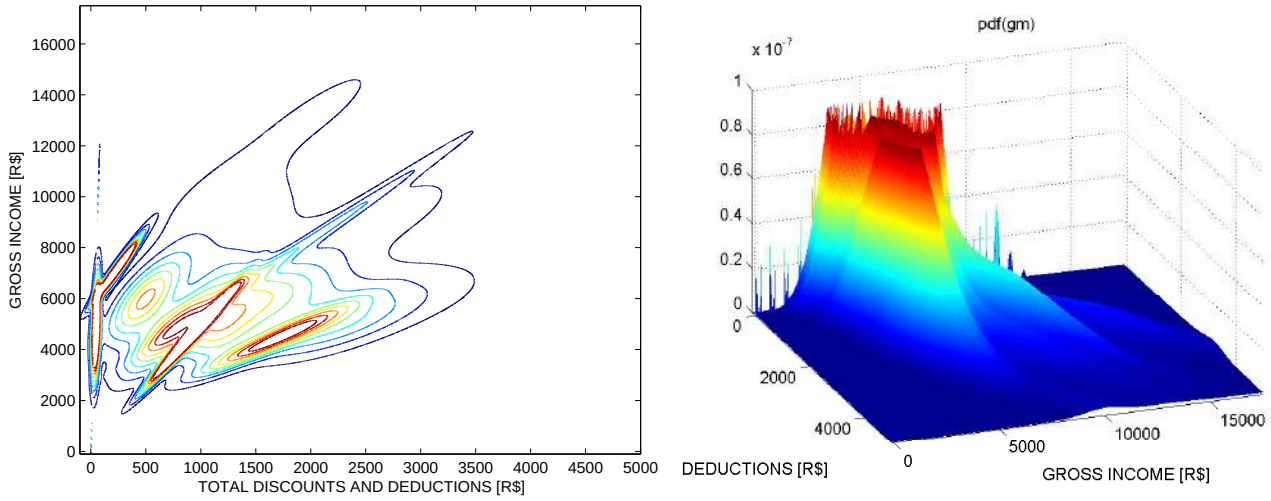
Figure 5: Contour plot of the estimated pdf of the dataset presented in Fig. 3 with (a) 8 sources (log-likelihood: -174317); (b) 16 sources (log-likelihood: -173282); (c) 24 sources (log-likelihood: -173019) and (d) 32 sources (log-likelihood: -172937). The axis in all subfigures are the same as in Fig.3b.

The resulting pdf of the GMM, optimized with the EM algorithm according to the settings previously described, is shown in Fig. 6.

It can be seen in Fig. 6, through the equi-probability contour lines, that the learned GMM pdf possess high values of probability where the set of observed data samples are more dense. Since the observed data set regards the proportion of gross income and total discounts and deductions from payrolls, the hypothesis we are seeking to test is that employees which have a proportion of gross income *versus* discounts that are far away from the normal behavior of the GMM pdf are more likely to have irregularities in their payroll.

To confirm that hypothesis, a pre-processing stage was created on the current BI system model described in Section 3. This stage consists of a statistical filter, where high probability payrolls according to the GMM pdf are presumed regular and are not processed by the deterministic BI system based on audit trails.

Table 1 shows the statistical filter results by comparison with unfiltered data. The information of the table is organized as follows: In the first line, it can be seen that if we filter, *i.e* remove, the 5% most probable payrolls off our observed data set, this implicates in an average loss of 0.92% of the total occurrences of the current audit process (false negatives). Analogously, filtering 20% of the most probable payrolls off our observed data set causes an average loss in the current audit trails of 8.16%. The audit trails chosen to populate the table are the ones that contain the most



(a) (b)

Figure 6: (a) Contour plot and (b) surface plot of the resulting pdf of the proposed GMM.

significant number of occurrences.

Given that the GMM filter proposed in this work is unique for all the audit trails, so the output of the GMM filter feeds all the audit trails in the current BI system, the gain in terms of efficiency is considerable since, with the use of the filter, it is possible to reduce the processing requirements of the system by 20% with an average audit loss of about 8.16%. In other words, it can be stated that if we submit 80% of the less probable payrolls to the audit trails, we will be able to detect 91.84% of the irregularities.

In addition, the underlying rules that dictates the behavior of payrolls are highly related to federal Brazilian legislation, thus the proposed GMM of a certain month of the year should not present severe changes if the legislation regarding public workers remain unchanged.

6. CONCLUSIONS

This paper proposed a statistical filter based on GMM applied in a BI system environment. At the current stage, the BI system uses a deterministic approach based on audit trails via concept maps to detect irregularities and inconsistencies in the payroll of the Brazilian public employees. With the insertion of the proposed statistical filter as a pre-processing stage of the BI system, it was possible to obtain a significant gain of efficiency in the overall system.

The statistical filter developed in this work models a generative underlying pdf that governs the observed data set as a mixture of Gaussians. When applied to a real world data, the filter successfully reduced in 20% the amount of data to be analyzed by the audit trails, with a penalty of losing 8.16% in false negatives. In addition, the statistical filter is unique for all the audit trails, which extends its efficiency gain since each audit trail can use the one time filtered data as input. Finally, considering that Brazilian legislation set the rules for the payroll of federal public staff, the generative underlying pdf should not change expressively if the legislation remains unchanged, which enables the GMM pdf of a certain month to be used in the subsequent years used without having to be recomputed.

Considering that, nowadays, the BI system of CGAUD is capable of auditing approximately 5 billion *reais* each month, where the total payroll of the Brazilian public employees is around 12.5 billion *reais* [10], an increase in the processing capacity of the BI system through a statistical pre-processing filter yields a more comprehensive auditing process regarding the whole payroll, even considering the penalty related to false negative outcomes intrinsic to prob-

Table 1: Fraud occurrences detected by audit trails divided according to their probability of occurrence.

Probability of Occurrence	% Reference	Audit Trail #13		Audit Trail #18		Audit Trail #31		% Average	% Cumulative
		Abs	%	Abs	%	Abs	%		
0 ∼ 5 %	5%	329	1.26%	226	0.77 %	28	0.72 %	0.92 %	0.92 %
5 ∼ 10 %	5%	847	3.25%	410	1.40 %	65	1.67 %	2.11 %	3.02 %
10 ∼ 20 %	10%	1806	6.93%	1152	3.93 %	177	4.54 %	5.13 %	8.16 %
20 ∼ 40 %	20%	4704	18.05%	5308	18.09 %	857	21.99 %	19.38 %	27.53 %
40 ∼ 60 %	20%	4074	15.63%	7147	24.36 %	1145	29.38 %	23.12 %	50.66 %
60 ∼ 80 %	20%	5785	22.20%	7245	24.69 %	983	25.22 %	24.04 %	74.70 %
80 ∼ 100 %	20%	8512	32.67%	7853	26.76 %	642	16.47 %	25.30 %	100.00 %
Total	100%	26057	100.00%	29341	100.00 %	3897	100.00 %	####	####

abilistic models.

Future developments in the area include predictive serial analytics, moving from a spacial analysis repeated every month to a predictive time series analysis, enabling the system to feedback itself and learn to track irregularities over time.

Acknowledgments

The authors wish to thank the Brazilian research, development and innovation Agencies CAPES (FORTE Project, Forensic Sciences Notice 25/2014), CNPq (Grant 303905/2014-0) and FINEP (Grant RENASIC/PROTO 01.12.0555.00), as well as the Brazilian Ministry of Planning, Budget and Management (Cooperation Agreement 26/2012), for their support to this work.

References

[1] Lee, K., Guillemot, L., Yue, Y., Kramer, M., Champion, D.: Application of the Gaussian mixture model in pulsar astronomy-pulsar classification and candidates ranking for the fermi 2fgl catalogue. Monthly Notices of the Royal Astronomical Society **424**(4), 2832–2840 (2012)

[2] Lu, D., Moran, E., Batistella, M.: Linear mixture model applied to Amazonian vegetation classification. Remote sensing of environment **87**(4), 456–469 (2003)

[3] Sönmez, M.K., Heck, L., Weintraub, M., Shriberg, E., Kemal, M., Larry, S., Mitchel, H., Shriberg, W.E.: A lognormal tied mixture model of pitch for prosody-based speaker recognition. SRI International (1997)

[4] Figueiredo, M.A., Jain, A.K.: Unsupervised learning of finite mixture models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **24**(3), 381–396 (2002)

[5] Hastie, T., Tibshirani, R.: Discriminant analysis by Gaussian mixtures. Journal of the Royal Statistical Society. Series B (Methodological) pp. 155–176 (1996)

[6] Dalal, S., Hall, W.: Approximating priors by mixtures of natural conjugate priors. Journal of the Royal Statistical Society. Series B (Methodological) pp. 278–286 (1983)

[7] Campos, S.R., Fernandes, A.A., de Sousa Jr, R.T., De Freitas, E.P., da Costa, J.P.C.L., Serrano, A.M.R., Rodrigues, D.D.C., Rodrigues, C.T.: Ontologic audit trails mapping for detection of irregularities in payrolls. In: International Conference on Next Generation Web Services Practices (NWeSP), pp. 339–344 (2012)

[8] Fernandes, A.A., Amaro, L.C., Da Costa, J.P.C.L., Serrano, A.M.R., Martins, V.A., de Sousa, R.T.: Construction of ontologies by using concept maps: A study case of business intelligence for the federal property department. In: Business Intelligence and Financial Engineering (BIFE), 2012 Fifth International Conference on, pp. 84–88. IEEE (2012)

[9] Huacarpuma, R.C., Rodrigues, D.d.C., Serrano, A.M.R., da Costa, J.P.C.L., de Sousa Jr, R.T., Holanda, M., Araujo, A.P.F.: Big data: A case study on data from the Brazilian ministry of planning, budgeting and management. IADIS Applied Computing 2013 (AC) Conference (2013)

[10] Serrano, A.M.R., Rodrigues, P.H., Huacarpuma, R.C., da Costa, J.P.C.L., de Freitas, E.P., de Assis, V.L., Fernandes, A.A., de Sousa Jr, R.T., Marinho, M.A.M., Pilon, B.H.A.: Improved business intelligence solution with reimbursement tracking system for the Brazilian ministry of planning, budget and management. 6th International Conference on Knowledge Management and Information Sharing (KMIS) (2014)

[11] Bolton, R.J., Hand, D.J.: Statistical fraud detection: A review. Statistical Science pp. 235–249 (2002)

[12] Anderson, D., Frivold, T., Valdes, A.: Next-generation intrusion detection expert system (NIDES): A summary. SRI International, Computer Science Laboratory (1995)

[13] Ghosh, S., Reilly, D.L.: Credit card fraud detection with a neural-network. In: System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on, vol. 3, pp. 621–630. IEEE (1994)

[14] Miziara, F., Puttini, R.S., Sousa Jr, R.T.: Détection d'intrusion en réseaux mobiles ad hoc utilisant un modèle de mélange gaussien pour le comportement du trafic. 2nd Joint Conference on Security in Network Architectures and Information Systems pp. 89–100 (2007)

[15] McLachlan, G., Peel, D.: Finite mixture models. John Wiley & Sons (2004)

[16] Marin, J.M., Mengersen, K., Robert, C.P.: Bayesian modelling and inference on mixtures of distributions. Handbook of statistics **25**, 459–507 (2005)

[17] Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digital signal processing **10**(1), 19–41 (2000)

[18] McLachlan, G.J., Basford, K.E.: Mixture models. inference and applications to clustering. Statistics: Textbooks and Monographs, New York: Dekker, 1988 **1** (1988)

[19] Bilmes, J.A., et al.: A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. International Computer Science Institute **4**(510), 126 (1998)

[20] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) pp. 1–38 (1977)

[21] Redner, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the EM algorithm. Society of Industrial and Applied Mathematics Review **26**(2), 195–239 (1984)

[22] Wu, C.J.: On the convergence properties of the EM algorithm. The Annals of statistics pp. 95–103 (1983)

[23] SIAPE: Sistema Integrado de Administração de Recusros Humanos (2015). URL `https://www.serpro.gov.br/conteudo-solucoes/produtos/administracao-federal/`

[24] MPOG: Boletim Estatístico de Pessoal (2013). URL `http://www.planejamento.gov.br/secretarias/upload/Arquivos/servidor/publicacoes/boletim_estatistico_pessoal/2013/Bol207_Jul2013_2.pdf`

[25] Brasil, Legal Regime of the Federal Public Employee. Law number 8112 of December 11, 1990

[26] McLachlan, G.J., Bean, R., Peel, D.: A mixture model-based approach to the clustering of microarray expression data. Bioinformatics **18**(3), 413–422 (2002)

[27] Rasmussen, C.E.: The infinite Gaussian mixture model. In: NIPS, vol. 12, pp. 554–560 (1999)