Recommender Algorithms for Form Anomaly Detection

Anne S. Parker[*] Danielle E. Gewurz [†] William J.J. Roberts [‡]

Abstract

Unsupervised anomaly detection is performed on forms assumed to be sparse and normally distributed. Maximum likelihood (ML) estimation is applied to estimate the parameter from a large collection of sparse forms. An expectation maximization algorithm from the literature that has been applied to sparse-matrix recommendation is used. Given the estimated parameter, constrained ML is applied to estimate anomalies. The constraints here are used to ensure that only anomalies in specific and predefined subspaces are detected. This formulation borrows from the literature of one-sided multivariate testing. The overall approach is tested and the results compared to a database of forms with known anomalies. The approach improves on the authors' previously developed unsupervised method for anomaly detection.

Key Words: Sparse data, expectation maximization, unsupervised, one-sided multivariate test

1. Introduction

An unsupervised anomaly detection approach that identifes field-level anomalies from sparse form data was studied by Parker *et al.* [11]. Unsupervised approaches are useful when labeled examples of anomalous and un-anomalous forms are difficult to obtain. Approaches that can contend with sparse data are useful when not all fields of the forms are populated. The Internal Revenue Service (IRS) is interested in unsupervised approaches that can handle sparse data because a) accurate labeling of anomalous form data fields requires a time-consuming, manual review process using highly experienced examiners; and b) forms are generally sparse as not all respondents complete all fields of a tax form.

[*] Internal Revenue Service Research, Applied Analytics, and Statistics, 1111 Constitution Avenue NW, Washington, DC 20224

[†] Deloitte Consulting LLP, 1919 N Lynn Street, Arlington, VA 22209

[‡] Deloitte Consulting LLP, 7900 Tysons One Place, Suite 800, Clean, VA, 22102

In [11] the sparse form data is assumed to be a Gaussian mixture model (GMM) with diagonal covariance matrices. Given a large set of forms the GMM parameter is estimated using an expectation maximization (EM) algorithm that does not require imputation of missing form field values. Anomalous fields values are identified using the probability of the observed field value, or a value more extreme, under the estimated GMM.

In this paper we address a similar problem to [11] using a Gaussian model that does not assume diagonal covariance matrices. Model training is accomplished using a well-known EM algorithm from the literature, see [12] and references therein. Anomalous form fields are identified by applying an approach by Kudo [4] and Neusch [8] developed for multivariate one-sided testing. In this approach, anomalies are represented using a field-level corrective term estimated using the maximum likelihood (ML) criterion. This can be performed in a computationally expeditious manner using quadratic programming. Numerical experiments were conducted on a training data set of over 100 million tax forms and a testing dataset consisting of $109,307$ forms that had fields that had been manually labelled as either anomalous or not. Using these data, the new approach showed improved performance as compared to the approach in [11].

The remainder of this paper as organized as follows: In Section 2 we describe the model and test statistic. In Section 3 we discuss the application of the model to real data. Section 4 concludes with some comments.

2. Model

2.1 Model Specification

Let v denote a k-dimensional random vector representing the k field values of a form. The distribution of v is Gaussian with $k \times 1$ mean vector, μ, and $k \times k$ covariance matrix, R. We write $v \sim \mathcal{N}(\mu, R)$. We seek anomalous forms, where the anomalies we are interested in have a specific definition. Define an anomalous form as a form where some of the k field values have been changed in a specific direction which is known and fixed for each field. Let z denote a k-dimensional random vector representing an anomalous form. Let θ denote a k-dimensional vector of deterministic variables representing the additive changes to v that results in an anomalous form z. Let $\delta = [\delta(1), \ldots, \delta(k)]'$ with $\delta(m) = 1$ or $\delta(m) = -1$, $m = 1, \ldots, k$. The model for anomalous forms is

$$z = v + \theta \qquad (1)$$
$$\text{subject to } D\theta \geq 0$$

where $D = \text{diag}(\delta)$ represents the diagonal matrix with diagonal elements equal to the elements of δ, and 0 represents a conforming vector of zeros. Each of the $\delta(m), m = 1, \ldots k$, sets the direction of interest for anomalies in the corresponding form field. The directions are generally specified in real applications and so δ is considered here to be a known constant vector. In the model (1) the random vector z is Gaussian with mean $\mu + \theta$ and covariance R. In anomaly detection applications z is generally observed, whereas v and θ are generally unobserved. The field-level anomaly detection problem is to estimate θ having observed an anomalous form z. Under the assumed Gaussian model, constrained maximum likelihood (ML) estimation of θ is generally a quadratic programming problem [4] [9].

A complicating aspect in form applications is that forms are often sparse. Many form fields are not applicable to some respondents and those respondents generally leave such fields blank. Let k_t denote the number of populated fields for the tth respondent. Let H_t indicate a $k_t \times k$ sub-matrix of the $k \times k$ identity matrix I where the rows of I corresponding to the indices of the unpopulated fields from the tth respondent have been deleted. Let y_t and w_t denote k_t-dimensional vectors representing the subset of populated elements of anomalous form values and true form values, respectively. Thus $y_t \sim \mathcal{N}(\mu_{y_t} + \theta_t, R_{y_t})$, where $\mu_{y_t} = H_t\mu$, $R_{y_t} = H_t R H_t'$, and θ_t is a k_t-dimensional deterministic unknown vector. The model (1) under sparse conditions is thus

$$y_t = w_t + \theta_t \qquad (2)$$
$$\text{subject to } D_t\theta_t \geq 0$$

where $D_t = \text{diag}(H_t\delta)$ represents the directions of interest for anomalies on the populated fields and v_t represents the un-anomalous form corresponding to y_t.

2.2 Anomaly Detection at Field Level

The field-level anomaly detection problem under sparse conditions is the estimation of θ in model (2) having observed an anomalous form y_t. The probability density function (pdf) of y_t in this model is denoted by $p(y_t; \mu_{y_t} + \theta_t, R_{y_t})$ and we have

$$p(y_t; \mu_{y_t} + \theta_t, R_{y_t}) = \frac{1}{\sqrt{(2\pi)^{k_t} \left|R_{y_t}\right|}} \exp\left(-\frac{1}{2}\left(y_t - \mu_{y_t} - \theta_t\right)' {R_{y_t}}^{-1} \left(y_t - \mu_{y_t} - \theta_t\right)\right) \quad (3)$$

We apply maximum likelihood (ML) estimation and thus the field-level anomaly detection problem is to find $\hat{\theta}$ such that

$$\hat{\theta}_t = \begin{array}{c} \arg\max \\ D_t\theta_t \geq 0 \end{array} p\left(y_t; \mu_{y_t} + \theta_t, R_{y_t}\right) \quad (4)$$

Similar constrained ML estimation problems arise in hypothesis testing of restricted means and have been extensively studied by Bartholomew [2], Kudo [4] and Nuesch [8], [9]. Substituting (3) into (4), taking the log and simplifying yields

$$\hat{\theta}_t = \begin{array}{c} \arg\max \\ D_t\theta_t \geq 0 \end{array} - \theta_t' {R_{y_t}}^{-1} \theta_t + 2\left(y_t - \mu_{y_t}\right)' {R_{y_t}}^{-1} \theta \quad (5)$$

Equation (5) involves the maximization of a quadratic objective function subject to linear inequality constraints. Such problems are generally referred to as quadratic programming problems. Estimation of $\hat{\theta}_t$ is explicit under certain conditions on the constraints and on the covariance R. If there were no constraints then $\hat{\theta}_t = y_t - \mu_{y_t}$. If the inequality constraint were instead an equality then an explicit optimum is known, see e.g. [3, Example 5.1]. If R is a diagonal positive definite matrix then $\hat{\theta} = \max(y_t - \mu_{y_t}, 0)$. More generally, we can apply the Karush-Kuhn-Tucker (KKT) conditions which are necessary (and in our particular case also sufficient) set of conditions for the optimum $\hat{\theta}$. These conditions, with their generally-applied labels, are given by

$$\begin{aligned} R_{y_t}^{-1}\hat{\theta} + (y_t - \mu_{y_t}) - D_t\lambda_t = 0&\text{: Stationarity} \quad (6)\\ D_t\theta_t \geq 0&\text{: Feasibility} \quad (7)\\ \lambda_t \geq 0&\text{: Nonnegativity} \quad (8)\\ \lambda_t' D\hat{\theta}_t = 0&\text{: Complementary Slackness} \quad (9) \end{aligned}$$

where λ_t is the k_t-dimensional vector of Lagrange multipliers [3]. These equations cannot be solved explicitly for $\hat{\theta}$ under general conditions. Algorithmic approaches are discussed in [4], [8], [9] and [14] which are derived from geometric representations of the problem. General purpose numerical quadratic programming routines are also readily available. These routines are particularly efficient and robust when R is positive definite, which is generally the case in the application considered here.

2.3 One-Sided Multivariate Testing

Building upon work by Bartholomew [2], Kudo [4] and Nuesch [9] developed a one-side multivariate hypothesis test that employs a maximization similar to that described above. In Kudo's [4] and Nuesch's [9] test, the distribution under the null hypothesis is zero-mean Gaussian against an alternative hypothesis having positive mean. We consider a hypothesis test where, under the null hypothesis, y_t has a non-zero mean that can be moved in any known and fixed direction under the alternative hypothesis, i.e.,

$$\begin{aligned} H_0 :&\quad y_t \sim N(\mu_{y_t}, R_{y_t}),\\ H_1 :&\quad y_t \sim N(\mu_{y_t} + \theta_t, R_{y_t}) \text{ where } D_t\theta_t \geq 0 \end{aligned} \quad (10)$$

The corresponding likelihood ratio test is given by

$$\frac{p(y_t; \mu_{y_t}, R_{y_t})}{\max\limits_{D_t\theta_t \geq 0} p(y_t; \mu_{y_t} + \theta_t, R_{y_t})} \underset{H_1}{\overset{H_0}{\gtrless}} \eta \quad (11)$$

where $0 < \eta \leq 1$ is a scalar test threshold. Substituting in the specific forms of the pdf's and simplifying yields the equivalent test

$$-2(y_t - \mu_{y_t})' R_{y_t}^{-1} \hat{\theta}_t + \hat{\theta}_t' R_{y_t}{}^{-1} \hat{\theta}_t \underset{H_1}{\overset{H_0}{\gtreqless}} \log \eta \tag{12}$$

where $\hat{\theta}_t$ is maximizer for the denominator. The test statistic can be simplified using the KKT conditions. Specifically, solving for λ in the stationarity condition (6) and then substituting the result into the complementary slackness condition (9) yields

$$\hat{\theta}_t' R_{y_t}^{-1} \hat{\theta}_t - (y_t - \mu_{y_t}) \hat{\theta}_t = 0 \tag{13}$$

Using (13) to simplify (12) yields

$$\hat{\theta}_t R_{y_t}^{-1} \hat{\theta}_t \underset{H_0}{\overset{H_1}{\gtreqless}} -\log \eta \tag{14}$$

Nuesch [8] and Kudo [4] derive the distribution of the corresponding test statistic under a null hypothesis with zero mean. The distribution of the test statistic for the non-zero-mean case developed here is not currently known.

2.4 Parameter Estimation

The model parameter is estimated from training data consisting of n sparse forms. We assume that the training data is comprised of un-anomalous forms, however we believe that an estimation approach could be developed to estimate from anomalous forms too.

We seek to identify maximum likelihood estimates for μ and R that maximize $p(w^n; \mu, R)$ where $w^n = \{w_1, \ldots, w_n\}$. There is no method to find explicit maximum likelihood estimates for both μ and R ; therefore we resort to use of the expectation-maximization algorithm and closed-form mean estimate from [6] and [7]. We provide the details for implementation here; the full derivations can be found in [12] and [7].

To initialize the model, we use the best-performing estimates from [12]. For μ, $\hat{\mu}^0 = N^{-1} \sum_{t=1}^{n} H_t' w_t$, where N is a vector of length k, with each element indicating the number of times the corresponding field is populated.

To initialize the covariance R, we first define S as

$$S = \sum_{t=1}^{n} H_t' \left(w_t - H_t \hat{\mu}^0\right) \left(w_t - H_t \hat{\mu}^0\right)' H_t \tag{15}$$

Then we use $\hat{R}^0 = N^{-1/2} S N^{-1/2}$ as a initial covariance estimate; see [12].

Next, we turn to the EM update formulas. We now require a mathematical representation of the missing data. Let J_t be a $(k - k_t) \times k$ sub-matrix of the $k \times k$ identity matrix I where the rows corresponding to the indices of the populated fields for the tth form have been deleted. Thus, $R_{x_t} = J_t R J_t'$, $R_{x_t y_t} = J_t R H_t'$, and $\mu_{x_t} = J_t \mu$.

Given R^i, the resulting EM iteration to get R^{i+1} is

$$\begin{aligned} \hat{R}^{i+1} &= \frac{1}{n} \sum_{t=1}^{n} \left(\hat{v}_t - \mu\right) \left(\hat{v}_t - \mu\right)' + & (16) \\ & J_t \left(\hat{R}^i_{x_t} - \hat{R}^i_{x_t y_t} \left(\hat{R}^i_{y_t}\right)^{-1} \left(\hat{R}^i_{x_t y_t}\right)' \right) J_t' & (17) \end{aligned}$$

where $\hat{v}_t$ and $\hat{x}_t$ are defined as follows:

$$\hat{v}_t = H_t' w_t + J_t' \hat{x}_t \tag{18}$$

$$\hat{x}_t = R_{x_t y_t} R_{y_t}^{-1} \left(w_t - \mu_{y_t}\right) + \mu_{x_t} \tag{19}$$

Once the covariance estimate has been updated, we next update the estimate for the mean. The estimate for μ^i has a closed form maximum-likelihood solution for a given R^i:

$$\hat{\mu}^i = \left(\sum_{t=1}^{n} H_t' R_{y_t}^{-1} H_t \right)^{-1} \sum_{t=1}^{n} H_t' R_{y_t}^{-1} w_t \tag{20}$$

3. Model Results

3.1 Numerical Experiments

This model and anomaly detection methodology were applied to similar testing data as described in [11]: an anonymized form database consisting of $n = 148,334,102$ entities who had each populated some, or all of, $k = 209$ real-valued form fields on an individual 1040 US tax return. This data is populated on a yearly basis by taxpayers to calculate their tax liability, and contains information about a wide variety of earnings, assets, expenses and business ownership. On average the data set was under 7% populated.

To initialize the EM, the estimates discussed in Section 2.4 were used. Likelihood increased with each iteration. Iterations ceased once the convergence criterion $\log p(y^n; \hat{\phi}^{j+1}) - \log p(y^n; \hat{\phi}^j) < n\delta$, with $\delta = .001$, was satisfied. The number of iterations required for convergence was 101. The final log-likelihood normalized by n was -37.5228.

Estimation of field-level anomalies was done using the quadratic programming routine from the cvxopt library in Python 3.7. Maximum number of iterations was set to 100.

3.2 Results

Field-level quality validation was performed using the θ discussed in Section 2.2, accounting for the directionality of the anomalies of interest as discussed. To measure the anomaly detection performance we used a similar setup to that used in [11], which, for completeness we describe again here. We tested on a collection of $109,307$ forms that had undergone two types of manual validation to identify anomalies. The first was *routine*, where field values appearing anomalous were identified using a comparatively quick review of the form and associated information. The second manual val idation was *detailed* and performed over an extended time, with many supporting documents and other types of relevant information. We considered the *detailed* result to represent the ground truth against which the performance of the model (and the routine review) was assessed. We considered two relevant error meters: a *false alarm*, i.e., an anomaly detected by the model (or routine review) that did not arise in the detailed review, and a *detection*, i.e., an anomaly detected by the model (or routine review) that did arise in the detailed review. The ROC curve across all field values appears in Fig. 1.

Also represented in this ROC curve is the Gaussian Mixture Model anomaly detection approach discussed in [11], as well as the fixed point representing the false alarms and detections obtained by the *routine* review. This plot shows that the model obtains similar performance as the manual routine review, and improves over the previous approach. Further, the performance advantage of this approach over the Gaussian Mixture Model approach widens when examining smaller, harder-to-detect anomalies that are still of interest to the Service.

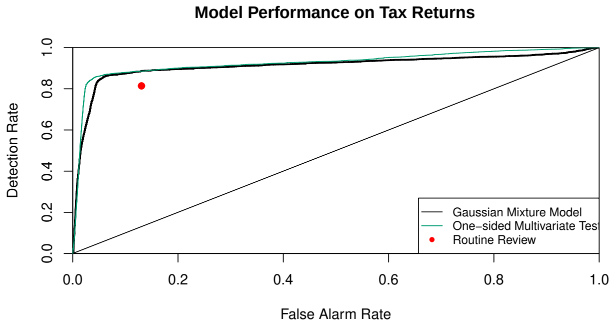

Figure 1: ROC curve showing performance of anomaly detection approach as compared with Gaussian Mixture Model approach and fixed point showing performance of routine review

4. Conclusion

We have derived a test for directional anomalies using a multivariate Gaussian model, with unstructured covariance, trained on sparse data. A constrained maximum likelihood estimate to identify anomalous values is developed. The model was applied to form quality validation. This test does not require known anomalous forms; known anomalous forms *were* used for performance measurement. We showed performance of the approach was comparable to a routine review of the model in detecting field anomalies and improves upon the approach in [11]. In operation, the results of the routine review are available to reviewers and would be expected to influence the detailed review. This suggests that direct performance projections of our results here would be conservative if our model were to replace the routine review.

References

[1] M. S. Andersen, J. Dahl, and L. Vandenberghe. CVXOPT: A Python package for convex optimization, version 1.1.5. Available at abel.ee.ucla.edu/cvxopt,

2012.

[2] D. J. Bartholomew. “A Test of Homogeneity of Means Under Restricted Alternatives," *Journal of the Royal Statistical Society, Series B*, vol. 23, no. 2, pp. 279-272, July 1961.

[3] S. Boyd, L. Vandenberghe. “Convex Optimization." Cambridge University Press, 2004.

[4] A. Kudo. “A Multivariate Analog of the One-Sided Test," *Biometrika*, vol. 50, no. 3/4, pp. 403-408, Dec. 1963.

[5] E. L. Lehmann, “Testing Statistical Hypotheses,” 2nd ed. New York: Chapman and Hall, 1994.

[6] R. J. A. Little and D. B. Rubin, “Statistical Analysis with Missing Data,” 2nd ed. Hoboken: Wiley-Interscience, 2002.

[7] D. W. McMichael, “Estimating Gaussian mixture models from data with missing features,” in *Proc. 4th Int. Symp. Sig. Proc. and its Apps., Gold Coast, Australia, Aug. 1996.* pp. 377–378.

[8] P. E. Neusch. “Multivariate tests of location for restricted alternatives." Doctoral dissertation, Swiss Federal Institute of Technology, Zurich, 1964. https://doi.org/10.3929/ethz-a-000107853.

[9] P. E. Neusch. “On the Problem of Testing Location in Multivariate Populations for Restricted Alternatives," *The Annals of Mathematical Statistics* Vol. 37, No. 1, pp. 113–119, Feb. 1966.

[10] A. S. Parker, “Recommendation system application for anomaly detection and missing value imputation,” presented at National Academy of Sciences, Big Data Day, 2018, Washington, United States. May 11, 2018.

[11] A. S. Parker, D. Gewurz and W. J. J. Roberts. “Quality and Validity Testing of Sparse Form Data using Gaussian Mixture Models," *JSM Proceedings, Social Statistics Section*, 2018.

[12] W. J. J. Roberts, “Application of a Gaussian, missing-data model to product recommendation,” *IEEE Signal Processing Letters*, vol. 17, pp. 509–512, 2010.

[13] W. J. J. Roberts, “Factor analysis parameter estimation from incomplete data,” *Computational Statistics and Data Analysis*, vol. 70, pp. 61–66, 2014.

[14] Y. Yamamoto, A. Kudo and K. Ujiie. “Computation of the Test Statistic and the Null Distribution in the Multivariate Analogue of the One-Sided Test," *Journal of the Japanese Society of Computational Statistics*, vol. 10, pp. 89-97, Jan. 1997.