# Quality and Validity Testing of Sparse Form Data using Gaussian Mixture Models

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

**Abstract**

Data quality and data validation using $p$-values obtained from a Gaussian mixture model (GMM) are studied and applied to numerical form data. Generally numerical forms completed by individuals are sparse in the sense that not all field values are populated by all individuals. Thus, estimation of the GMM parameter from sparse data is required. An expectation-maximization approach is derived here for a particular type of GMM under sparse data conditions. Given the estimated GMM parameter, the $p$-value of each populated field is calculated and used to detect anomalous fields. Performance of the approach is compared to manual data quality and validation and found to have similar performance in the detection of anomalous field values.

**Key Words:** Sparse data, expectation maximization, $p$-value

## 1. Introduction

Many government agencies use forms to collect data from respondents and are interested in the quality of that data. In particular, the Internal Revenue Service devotes extensive resources to both manual and automated detection of data errors in form submissions. There are certain very time-consuming routine manual review processes where employees identify which fields of a form have anomalous values, or issues. Replicating these manual reviews with a model-based approach frees up scarce resources and ensures consistency in the review process. Use of an unsupervised model is advantageous, as the model can be estimated after changes in tax law without waiting for new tagged examples to be generated.

*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, McLean, VA, 22102

A tax form completed by many respondents can be represented as a sparse matrix — respondents only populate fields which are relevant to them. This sparsity can complicate analysis and necessitate the application of specialized sparse data approaches. Parker [7] modeled sparse form data as Gaussian and applied parameter estimation approaches from [9]. In this paper, we extend this work by modeling sparse form data using a Gaussian mixture model (GMM). The GMM potentially allows the multi-modal nature of form data to be captured. As some forms may be completed by millions of respondents, we are interested in maximum likelihood (ML) GMM parameter estimation, as ML estimates have desirable asymptotic properties with increasing data size. Even under non-sparse conditions, no explicit ML GMM parameter estimate is known and the expectation-maximization (EM) algorithm is generally applied, see, e.g., [5]. Under sparse conditions, McMichael [6] in 1996 developed an EM algorithm that did not require imputation of unpopulated fields and estimated the covariances using steepest descent. More recently Delalleau, Courville and Bengio [1] developed an EM approach where the unpopulated fields are imputed and applied it to handwritten digit recognition. Silva and Deutsch [11] developed a similar approach and applied it to geological data.

Here we constrain the covariances to be diagonal and develop an EM approach that does not require imputation of unpopulated fields. The resulting estimation equations are particulary straightforward, allowing application to very large data sets.

Given models trained on a large body of forms, we apply a hypothesis testing framework to detect issues at the individual field level. Let $H1$ denote the hypothesis that a particular field on a particular form is anomalous. Let $H0$ denote the hypothesis that a particular field on a particular form is not anomalous. If the probability density functions (pdfs) of the forms under the two hypotheses are known, then optimum decision rule in the Neyman-Pearson sense is given by the likelihood ratio test [3, p. 32, Thm. 1]. The true pdf under $H1$ is difficult to estimate when the number of forms confirmed to be anomalous is low. IRS forms will change due to Tax Cuts and Jobs Act of 2017 and numbers of confirmed, anomalous new forms may initially be low. If the pdf for $H0$ is the only one known, then a composite

hypothesis test may be appropriate [3]. Composite hypothesis testing problems are notoriously difficult and general optimality of approaches may be difficult to prove. In practice, Bayesian approaches, see, e.g., [8] and the generalized likelihood ratio test, see, e.g., [10] are often applied. Here we perform anomaly detection without requiring a pdf for $H1$ using the probability of the observed value, or a value more extreme, under the pdf for $H0$. This is equivalent to calculating the $p$-value of the observation under the GMM. Although this approach is simple and intuitive, we do not believe it has been previously applied to form data anomaly detection using GMMs.

The above model training and anomaly detection approaches were applied on a data set consisting of over 10 million tax forms. Specialized numerical approaches were applied to minimize computations while maintaining numerical precision. Performance of the overall approach was measured using tax forms with known anomalies.

## 2. Sparse Gaussian Mixture Model

### 2.1 Model Specification

We assume that we have data from $n$ forms and each form consists of $k$ numerical fields. We assume the $k$-dimensional vector of fields is distributed according to a GMM with $r$ mixtures. Thus a $k$-dimensional vector of fields generated by the $m$th mixture has a probability density function (pdf) given by $\mathcal{N}(\mu_m, R_m)$, where $\mathcal{N}(\mu_m, R_m)$ represents the Gaussian pdf with $k \times 1$ mean vector $\mu_m$ and $k \times k$ covariance matrix $R_m$.

In general, however, not all field values are populated for all forms. For the $t$th form, $1 \leq t \leq n$, we assume that $0 < k_t \leq k$ fields are populated, and $k - k_t$ are unpopulated. Let the data from all forms be represented as $y^n = \{y_t, \ldots, y_n\}$ where the $k_t$-dimensional vector $y_t$ denotes the populated line items from the $t$th form. Let $s^n = \{s_t, \ldots, s_n\}$ where $s_t \in \{1, \ldots, r\}$ denote the sequence of mixtures corresponding to $y^n$. The conditional pdf of $y_t$ given $s_t = m$ is $\mathcal{N}(H_t\mu_m, H_t R_m H_t')$ where $H_t$ is 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 for the $t$th form have been

deleted.

Let $p(y^n;\phi)$ denote the pdf of $y^n$ where $\phi = \{\alpha_m, \mu_m, R_m\}_{m=1}^r$ is the GMM parameter consisting of the $r$ means, $r$ covariances and $r$ scalar mixture weights with $\sum_m \alpha_m = 1$. We have that

$$\begin{aligned} p(y^n;\phi) &= \prod_{t=1}^{n}\sum_{m=1}^{r} p(y_t|s_t;\phi)p(s_t;\phi) \\ &= \prod_{t=1}^{n}\sum_{i=1}^{r} \alpha_i \frac{\exp\left(-(y_t - H_t\mu_i)'(H_tR_iH_t')^{-1}(y_t - H_t\mu_i)/2\right)}{(2\pi)^{k_t/2}|H_tR_iH_t|^{1/2}} \end{aligned} \quad (1)$$

where $|\cdot|$ denotes a matrix determinant [9].

### 2.2 Parameter Estimation

We aim for the maximum likelihood (ML) estimate of the parameter $\phi$ given the data $y^n$, that is

$$\hat{\phi} = \arg\max_{\phi} p(y^n;\phi) \quad (2)$$

The parameter $\phi$ of the GMM cannot be explicitly estimated even when the data is not sparse. As is generally done in the non-sparse case, we estimate $\phi$ here using the EM algorithm to obtain a sequence of estimates $\{\hat{\phi}^j\}$ . The EM algorithm guarantees that each parameter estimate in the sequence has non-decreasing likelihood, i.e., $p(y^n;\hat{\phi}^{j+1}) \geq p(y^n;\hat{\phi}^j)$. Parameter estimates in the EM are obtained by maximizing the conditional expected value of the logarithm of the likelihood of the *complete data*, i.e.,

$$\hat{\phi}^{j+1} = \arg\max_{\phi} E\{\log p(y^n, s^n;\phi)|y^n;\hat{\phi}^j\} \quad (3)$$

We define the *complete data* to be the populated observations and the mixture sequence $\{y^n, s^n\}$ without including the un-populated fields, see, [10]. The choice of complete data within an EM formulation is often arbitrary and different choices generally lead to different algorithms. In this case, if we were to include the un-populated values as part of the complete data, the parameter estimation equations would require imputation for missing data values, see, e.g. [1, 11]. Our complete data choice avoids imputation of the un-populated values when the $\{R_m\}$ are constrained

to be diagonal, resulting in particularly straightforward estimation equations.

Substituting the definitions for the relevant pdfs into (3) yields

$$\begin{aligned}\hat{\phi}^{j+1} &= \arg\max_{\phi} \sum_{t=1}^{n} \sum_{s_t} \xi_t(s_t, \hat{\phi}^j) \times \\ &\quad (\log|H_t R_{s_t} H_t'| - (y_t - H_t\mu_{s_t})'(H_t R_{s_t} H_t')^{-1}(y_t - H_t\mu_{s_t}) + \log\alpha_{s_t}) \quad (4)\end{aligned}$$

where $\xi_t(s_t, \hat{\phi}^j) = p(s_t|y_t; \hat{\phi}^j)$ is the a posteriori mixture probability. Differentiating (4) by the individual $\mu_m$, setting the resulting equation to zero, and then solving for $\mu_m$ yields

$$\hat{\mu}_m^{j+1} = \left(\sum_{t=1}^{n} \xi_t(m, \hat{\phi}_j) H_{y_t}' R_{y_t}^{-1} H_{y_t}\right)^{-1} \sum_{t=1}^{n} \xi_t(m, \hat{\phi}_j) H_{y_t}' R_{y_t}^{-1} y_t. \quad (5)$$

This expression can be considerably simplified when the $\{R_m\}$ are constrained to be diagonal matrices. Let $z_t = H_t' y_t$ and let $z_t(i)$ denote the $i$th element of $z_t$. With diagonal $\{R_m\}$, the $i$th element of $\hat{\mu}^{j+1}$, denoted by $\hat{\mu}^{j+1}(i)$ simplifies to

$$\hat{\mu}_m^{j+1}(i) = \frac{\sum_{t=1}^{n} \xi_t(m, \hat{\phi}^j) z_t(i)}{\sum_{t=1}^{n} \xi_t(m, \hat{\phi}^j) 1_t(i)} \quad (6)$$

where $1_t(i)$ is an indicator function such that

$$1_t(i) = \begin{cases} 1 & \text{if } z_t(i) \text{ is populated} \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

Thus, for diagonal $\{R_m\}$, the mean estimates are weighted arithmetic means of the populated fields.

Constraining the covariances to be diagonal also simplifies covariance estimation. Assume that $R_m$ is diagonal with elements $\{\sigma_m^2(i), i = 1, \ldots, k\}$. Differentiating (4) with respect to each of the $\{\sigma_m^2(i)\}$, setting the result to zero, and solving yields

$$\hat{\sigma}_m^2(i)^{j+1} = \frac{\sum_{t=1}^{n} \xi_t(m, \hat{\phi}^j) 1_t(i) (z_t(i) - \mu_m(i))^2}{\sum_{t=1}^{n} \xi_t(m, \hat{\phi}^j) 1_t(i)} \quad (8)$$

Thus, the diagonal variance estimates are weighted sample variances of the populated fields.

Mixture weight $\{\alpha_m\}$ estimation is done using Lagrange multipliers to maximize (4) under the constraint $\sum_m \alpha_m = 1$, see, e.g., [4, Lemma 2, p. 1042], yielding

$$\hat{\alpha}_m^{j+1} = \frac{1}{n} \sum_t \xi_t(m, \hat{\phi}^j) \tag{9}$$

To complete the EM algorithm we need the following expression for the a posteriori mixture probability

$$\begin{aligned} \xi_t(m, \hat{\phi}^j) &= p(m|y_t; \hat{\phi}^j) \\ &= \frac{\hat{\alpha}_m^j \; p(y_t|m; \hat{\phi}^j)}{\sum_{m'=1}^r \hat{\alpha}_{m'}^j \; p(y_t|m'; \hat{\phi}^j)}. \end{aligned} \tag{10}$$

The EM algorithm is thus given by (6), (8), (9) and (10).

### 2.3 Anomaly detection

Let $Y_t$ denote $k_t$-dimensional random vector representing the form competed by the $t$th individual and let $y_t$ denote a realization of $Y_t$. Let $\phi$ denote the parameter corresponding to a form that is not anomalous. Detecting an anomaly for the $i$th field on the $t$th form is expressed as identifying which of the following hypotheses is true:

$$\begin{aligned} H0: &\quad Y_t(i) \sim p(y_t(i); \phi), \\ H1: &\quad Y_t(i) \sim p(y_t(i); \phi') \text{ where } \phi' \neq \phi. \end{aligned} \tag{11}$$

In statistical parlance, this is a classification problem for one simple and one composite hypothesis [3]. $H0$ is a *simple* hypothesis as the observations are described by a known pdf. $H1$ is a *composite* hypothesis as the observations are described by a pdf known only to be a member of a family of pdfs. A variety of approaches are available for composite hypothesis testing. One approach, if $\phi$ is assumed random with a known prior distribution, is to represent the composite hypothesis as a simple hypothesis using a Bayesian approach, see, e.g., [8]. Another approach is to apply the generalized likelihood ratio test, see, e.g., [10]. Here we perform anomaly detection using the probability of the observed value, or a value more extreme. This is

equivalent to calculating the $p$-value of the observation under the GMM. With this approach the decision is made according to

$$\Pr(|Y_t(i)| > y_t(i); \phi) \underset{H0}{\overset{H1}{\gtrless}} \eta \tag{12}$$

where $\eta$ is a threshold. If only one-sided anomalies are of interest, then the probability in the decision rule can be adjusted to be $\Pr(Y_t(i) > y_t(i))$ or $\Pr(Y_t(i) < y_t(i))$ as appropriate. Applying the specific form of the GMM pdf, (12) becomes

$$\sum_{m=1}^{r} \alpha_m \Pr(|Y_t(i)| > y_t(i) | s_t = m; \phi) \underset{H0}{\overset{H1}{\gtrless}} \eta \tag{13}$$

which can be calculated using numerical routines for the Gaussian cumulative distribution function.

## 3. Implementation and Numerical Results

### 3.1 Implementation

The techniques of section II were implemented using Python and Fortran and augmented where appropriate with explicit calls to basic linear algebra subprograms (BLAS), see e.g. [2, 9]. The use of Fortran for the most computationally burdensome parts of the EM algorithm provided an order of magnitude computational speedup over using Python alone. We applied an approach due to West [12] to improve numerical properties of the algorithm. In the parlance of West, applying the "text book" approach to $\{\sigma_m^2\}$ estimation can result in loss of numerical precision. Numerical precision can be improved by the "two pass" method at the cost of additional computation. Building on earlier work, West describes a stable weighted estimate update that reduces the risk of loss of numerical precision while requiring minimal increase in computation. Applying West's update here we have that $M_i(1) = H_t y_t, S_i(1) = 0$ and

$$\begin{aligned} M_i(t) &= M_i(t-1) + \frac{\xi_t(i)}{\sum_{\tau=1}^{t} \xi_\tau(i)} (H_t y_t - M_i(t-1)) \\ S_i(t) &= S_i(t-1) + \frac{\xi_t(i) \sum_{\tau=1}^{t-1} \xi_\tau(i)}{\sum_{\tau=1}^{t} \xi_\tau(i)} (H_t y_t - M(t-1)) \odot (H_t y_t - M(t-1)) \end{aligned}$$

with $i = 1, \ldots, r$ and $t = 2, \ldots, n$ and where $\odot$ represents element-wise multiplication. The final estimates are given by $\hat{\mu}_j = M_j(n)$ and the diagonal entries of $\hat{R}_i$ are given by $S_i(n)/\sum_{t=1}^{n} \xi_i(t)$.

### 3.2 Results

The approaches described here were tested using a sample of an anonymized form database consisting of $n = 10,000,000$ entities from tax filings from 2012-2014 who had each populated some, or all of, $k = 177$ real-valued form fields on an individual 1040 US tax return and aggregated schedules. 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. We used $r = 10$ mixtures. To initialize the EM, forms were ordered using $(\sum_{i=1}^{k_t} y_t)/k_t$ and then partitioned into $r$ equally sized groups. The sample means and variances of the $r$ groups constituted the EM algorithm's initial mean and variances estimates. With each EM iteration the likelihood increased. Iterations were ceased once the convergence criterion $\log p(y^n; \hat{\phi}^{j+1}) - \log p(y^n; \hat{\phi}^j) < n\delta$, with $\delta = .0001$, was satisfied. The number of iterations required for convergence was 29. The final log-likelihood normalized by $n$ was -163.45.

Using the resulting GMM parameter estimate, field-level quality validation was performed using (13) accounting for the directionality of the anomalies of interest as discussed above (12). The $p$-value obtained was treated as the test statistic in the binary hypothesis formation. To measure the anomaly detection performance we used $1,603,591$ forms from 2012-2014 that had undergone two types of manual validation to identify anomalies. The first manual validation is *routine*, where field values appearing anomalous were identified using a comparatively quick review of the form and associated information. The second manual validation 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. As $\eta$ in (13) is changed, the number of false alarms and detections arising from the model changes. The locus of the relative frequencies of false alarms and detections as $\eta$ is varied is called a receiver operator characteristic (ROC) curve. The ROC curve of the test (13) across all field values appears in Fig. 1. Also represented in this ROC curve is 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.

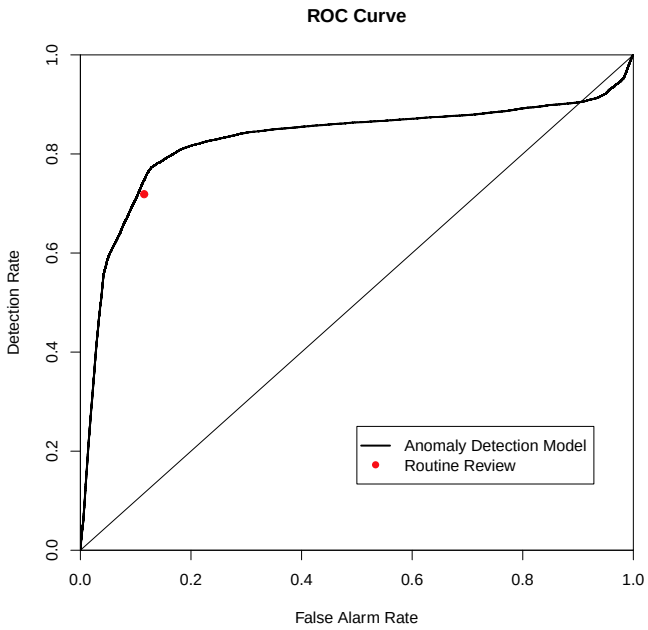


**Figure 1**: ROC curve showing performance of anomaly detection approach and fixed-point showing performance of routine review

## 4. Discussion

We have developed a simple and intuitive EM algorithm for GMM parameter estimation using sparse data that avoids imputation of unpopulated field values. Key to avoidance of imputation was the EM algorithm's complete data definition and the diagonal covariance constraint. The model was applied for form quality validation using a hypothesis testing formulation. The test statistic in the hypothesis test was the probability of the observed value, or a value more extreme. This test statis-

tic corresponds to the $p$-value of the observation. This test did not require known anomalous forms but 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. In practice, the results of the routine review are available to the detailed review, 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] O. Delalleau, A. Courville, and Y. Bengio, "Efficient EM training of Gaussian mixtures with missing data," arXiv:1209.0521, 2012.

[2] J. Dongarra, "Basic linear algebra subprograms technical forum standard," *International Journal of High Performance Applications and Supercomputing*, vol. 16, no. 1, pp. 1–111, 2002.

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

[4] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *The Bell System Technical Journal*, vol. 62, no. 4, pp. 1035–1074, Apr. 1983.

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

[6] 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.

[7] 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.

[8] W. J. J. Roberts, Y. Ephraim, and H. W. Sabrin, "Speaker classification using composite hypothesis testing and list decoding," *IEEE Trans. Speech and Audio Proc.*, vol. 13, no. 2, pp. 211–219, Mar. 2005.

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

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

[11] D. S.F. Silva and C.V. Deutsch, "Multivariate data imputation using Gaussian mixture models," in press, *Spatial Statistics*, 2016.

[12] D. H. D. West, "Updating Mean and Variance Estimates: An Improved Method," *Communications of the ACM*, vol. 22, no. 9, pp. 532–535, Sept. 1979.