Received: 19 December 2017 | Revised: 3 May 2018 | Accepted: 3 May 2018

DOI: 10.1002/qre.2339

SPECIAL ISSUE ARTICLE

WILEY

Outlier detection for multivariate categorical data

Xavier Puig | Josep Ginebra

Technical University of Catalonia,
Barcelona, Spain

Correspondence
Josep Ginebra, Technical University of
Catalonia, Barcelona, Spain.
Email: josep.ginebra@upc.edu

Funding information
Ministerio de Economia y Competitividad
of Spain, Grant/Award Number:
MTM2013-43992-R

Abstract

The detection of outlying rows in a contingency table is tackled from a Bayesian perspective, by adapting the framework adopted by Box and Tiao for normal models to multinomial models with random effects. The solution assumes a 2–component mixture model of 2 multinomial continuous mixtures for them, one for the nonoutlier rows and the second one for the outlier rows. The method starts by estimating the distributional characteristics of nonoutlier rows, and then it does cluster analysis to identify which rows belong to the outlier group and which do not. The method applies to any type of contingency table, and in particular, it could be used on the analysis of multivariate categorical control charts. Here, the use of the method is illustrated through a simulated example and by applying it to help identify heterogeneities of style among the acts in the plays of the *First Folio* edition of Shakespeare drama.

KEYWORDS

Bayesian cluster analysis, multinomial control charts, multinomial mixed model, multinomial outlier detection, textual data

1 | INTRODUCTION

Statistical inference is grounded on the assumption that data have been generated from a given statistical model, or from a mechanism that is close to a given statistical model. That assumption can fail either because the model is wrong for all or most of the observations in the data set, or just because it fails for a relatively small subset of observations. In the second case, one labels the observations that are not generated by the statistical model in place for the majority of observations, as outliers.

The literature covering the outlier detection problem from the statistical standpoint (see, eg, Barnett and Lewis[1] and Rousseeuw and Leroy[2]) deal mainly with continuous type data and are grounded mostly on the normality assumption. In practice, one often faces the existence of outliers among categorical data, often presented in terms of a contingency table. Methods for detecting outliers in contingency tables, like the ones presented in Fienberg,[3] Haberman,[4] Brown,[5] Fuchs and Kenett,[6] Simonoff,[7] Yick and Lee,[8] Kuhnt,[9] Mebane and Sekhon,[10] or Kuhnt et al,[11] deal with outlying cells rather than with outlying rows or columns of the table, and they tackle the problem mostly through the analysis of residuals based on the robust fit of multinomial regression models.

Instead, here, the rows of the contingency table are considered to be the unit of interest, and one looks for rows that depart from the distributional behavior of the majority of the rows of the table. That problem appears often in contexts like the analysis of multivariate categorical control charts, survey data, marketing data, electoral data, and stylometric data.

The solution adopted here extends the Bayesian framework for outlier detection proposed in Box and Tiao [12] for normal models, by adapting it to multinomial models with random effects. In that approach, one supposes that there exist 2 alternative models for any given observation, a basic model adequate for the majority of the observations in the data set and an alternative model adequate when the observation is an outlier. One then uses tools from Bayesian clustering to classify all the observations in the data set into either the outlier group or the nonoutlier group. The method starts by estimating the variability of the main set of observations, ideally through a subset of the observations in the sample known to be uncontaminated by the presence of outliers, and then it simultaneously identifies the outlier observations and estimates their distribution through cluster analysis.

The paper is organized as follows. In Section 2, outlier detection models are presented, first in the case where the rows of the table are unstructured and then in the more general case where they are structured. Section 3 illustrates the model usage for unstructured tables on a simulated example, while Section 4 illustrates the use of the more advanced model for structured tables to search for heterogeneities among acts in the plays of the *First Folio* edition of Shakespeare's drama.

2 | OUTLIER DETECTION MODEL FOR THE ROWS OF A TABLE

The problem starts with a $n \times J$ contingency table. In the context of the statistical analysis of literary style, for example, it is assumed that for each text, i, in a corpus of n texts, one has a vector valued categorical observation, $y_i = (y_{i1},...,y_{iJ})$, where J denotes the number of categories. The set of all the rows in the $n \times J$ table and hence of all the vectors of counts for the n acts in the corpus will be denoted by $y = (y_1,...,y_n)$. In the example of the drama by Shakespeare analyzed later on, the corpus will be 175 acts in 35 plays and y_i will be a 20-dimensional vector with the function word counts for the ith act, presented as the ith row of Table 1.

In the analysis, the ith row of the table is assumed to be multinomially distributed, Mult(N_i,θ_i), where $N_i = \sum_{j=1}^{J} y_{ij}$ is the sum of all the counts for the ith row, and where $\theta_i = (\theta_{i1},...,\theta_{iJ})$ is such that θ_{ij} is the probability of the jth category for the ith row, and hence with $\sum_{j=1}^{J} \theta_{ij} = 1$. If one can assume the row counts to be conditionally independent, then the distribution of $y = (y_1, ..., y_n)$ is

$$y|(\theta_1, \cdots, \theta_n) \sim \prod_{i=1}^{n} \text{Mult}(N_i, \theta_i). \tag{2.1}$$

Next, 2 outlier detection models are presented. The first one assumes that all the rows of the table are exchangeable, while the second model adds structure to it to cover situations where rows are nonexchangeable.

2.1 | Outlier detection model for an unstructured table

If all the rows in the table come from the same multinomial population, one might assume the vector of multinomial probabilities, $\theta_i = (\theta_{i1},...,\theta_{iJ})$, of all the n rows to be the same. A more flexible way to model that, though, is by assuming that the θ_i could be different among different rows, and yet all θ_i share the same distribution. In particular, here, it will be assumed that

$$\theta_{ij} = \frac{e^{\beta_j + \nu_{ij}}}{1 + e^{\beta_2 + \nu_{i2}} + \ ... + e^{\beta_J + \nu_{iJ}}} \quad \text{for} \quad j = 1, \ ..., J, \tag{2.2}$$

where the β_j are fixed effects and the ν_{ij} are random effects with $\beta_1 = 0$ and $\nu_{i1} = 0$ to make the parameters of this multinomial logistic model identifiable. If all rows in the table were homogeneous, without any outlying row, one could assume that the random effects, ν_{ij}, are independent and with a Normal(0,σ^2) distribution.

TABLE 1 Part of the 175 × 20 table of counts of 20 function words in the 35 plays of the *First Folio* edition of Shakespeare's drama[a]

Most Frequent Word Counts												
play	**act**	**the**	**and**	**I**	**to**	**of**	**a**	**you**	**my**	**that**	**in**	**...**
1	1	148	121	110	88	84	91	37	83	55	49	...
	2	102	102	99	71	79	96	55	52	52	38	...
	3	69	96	105	56	49	54	44	60	27	28	...
	4	47	68	51	41	30	29	29	43	20	15	...
	5	78	97	83	45	48	43	45	54	35	31	...
...	...	...	...	...	...	...	...	...	...	...	...	...
35	1	175	144	188	141	114	115	120	102	79	66	...
	2	113	97	117	73	83	57	75	34	49	28	...
	3	175	156	141	142	121	100	69	106	99	82	...
	4	142	112	106	92	67	59	42	61	47	49	...
	5	220	183	156	160	130	122	103	103	90	76	...

[a] The first play corresponds to *The Tempest* and the last one to *Cymbeline*.

Instead, if outlier rows are present, one needs to allow for the possibility that a subset of the rows have multinomial probabilities with a more dispersed distribution. Here, the θ_i for the outlying rows are assumed to be as in (2.2), with all the ν_{ij} being independent and with a normal $(0,k^2\sigma^2)$ distribution, where $k>1$.

To model the way outliers appear in a table, the distribution of the random effects of the ith row, $\nu_i=(\nu_{i2},\dots,\nu_{iJ})$, is assumed to be a 2–component mixture model, with the first component corresponding to the nonoutlying rows, and the second one to the outlying rows. Finite mixture models are often used in Bayesian analysis when heterogeneity is suspected, because they provide a flexible structure that can be used for the unsupervised classification of observations into either one of the groups. More specifically, here, we consider the ν_i to be conditionally independent and identically distributed as

$$(\nu_1,\ \dots,\nu_n)|\omega,\sigma^2,k\sim\prod_{i=1}^{n}((1-\omega)\prod_{j=2}^{J}Normal(0,\sigma^2) + \omega\prod_{j=2}^{J}Normal(0,k^2\sigma^2)), \tag{2.3}$$

where ω is a weight that determines the proportion of rows that tend to be outliers.

To allocate rows into either the outlier or the nonoutlier groups, one introduces a vector of unobserved categorical variables, $\zeta=(\zeta_1,\dots,\zeta_n)$, such that $\zeta_i=1$ when the ith row is an outlier, and hence, $\nu_i=(\nu_{i2},\dots,\nu_{iJ})$ comes from the $\prod_{j=2}^{J}Normal(0,k^2\sigma^2)$ mixture component, and such that $\zeta_i=0$ when the ith row is not an outlier, and hence, $\nu_i=(\nu_{i2},\dots,\nu_{iJ})$ comes from the $\prod_{j=2}^{J}Normal(0,\sigma^2)$ mixture component. The ζ_i are assumed to be conditionally independent with $\pi(\zeta_i=1|\omega)=\omega$ and $\pi(\zeta_i=0|\omega)=1-\omega$.

These allocation variables can be used to estimate the posterior probabilities that the ith row is an outlier, $E[\zeta_i|y]$, that can then be used to classify the rows into either being outliers or not. By imposing that $k>1$, this formulation ensures that $\zeta_i=1$ always correspond to the second component with the largest variance.

In Bayesian statistics, one needs to choose a prior distribution for the parameters of the model, which are $\omega,\beta=(\beta_2,\dots,\beta_J),\sigma^2$ and k. That prior needs to capture what one knows about the parameters before observing the data, and its choice depends on the problem at hand. Section 3 will describe the prior chosen for ω,β, and k in our simulated example. Here, σ^2 will not be modeled through a prior. Instead, its value will be calibrated based on observations known to be nonoutliers, the way described in Section 2.3.

2.2 | Outlier detection model for a structured table

Here, a more general framework is considered, in which one can depart from the assumption that all rows in the table are exchangeable in the following 2 different ways:

1. Rows can be grouped, with a different distribution for the multinomial probabilities of each group or, even more generally, the multinomial probabilities of rows could be related to a set of covariates. That will be attained by letting the fixed effects in (2.2) change from row to row.
2. The probability that a row is an outlier can change from row to row, due to a random effect. That extension will be attained by letting the distribution of the random effects in (2.2) be such that the probabilities that 2 rows are outliers becomes related and are not independent as in the previous section. That will be useful in those settings in which 2 rows that are in some sense close are more or less likely to be both outliers than 2 rows that are far apart.

As an example of the first departure, note that when looking for outliers among the acts of the plays attributed to Shakespeare, one might want to allow for different distributions for the multinomial probabilities of the acts of plays classified as Histories than the ones for the acts of plays classified as Tragedies, and the ones for the acts of Comedies. As an example of the second kind of departure, in that same example, one might want to allow for the probabilities that different acts from the same play are outliers to be related.

The first extension can be modeled by assuming that the rows are as in (2.1) with

$$\theta_{ij}=\frac{e^{\mu(i,j)+\nu_{ij}}}{1+e^{\mu(i,2)+\nu_{i2}}+\ \dots+e^{\mu(i,J)+\nu_{iJ}}}\quad\text{for}\quad j=1,\ \dots,J, \tag{2.4}$$

instead of (2.2), with the restrictions that $\mu(i,1)=0$ and $\nu_{i1}=0$ to make the logistic model with random effects identifiable. Hence, the fixed effect component, $\mu(i,j)$, here, is allowed to change with i, and it could be a function of row covariates and, in particular, it could be a function of the group to which the ith row belongs to.

The second extension can be modeled by letting the weights in the mixture distribution for the random effects of the i-th row, $\nu_i = (\nu_{i2}, ..., \nu_{iJ})$, vary from row to row,

$$(\nu_1,\ ...,\nu_n)|\omega,\sigma^2,k \sim \prod_{i=1}^{n}((1-w_i)\prod_{j=2}^{J} Normal(0,\sigma^2) + w_i \prod_{j=2}^{J} Normal(0,k^2\sigma^2)), \tag{2.5}$$

where $\omega = (\omega_1,...,\omega_n)$. This idea was proposed by Fernandez and Green[13] for Poisson mixtures for spatial data, and it was used by Puig et al[14] for multinomial cluster analysis. As a consequence of (2.5), the probability that the ith row is allocated to the outlier group, ω_i, will change from row to row, and the set of allocation variables, $\zeta = (\zeta_1, ..., \zeta_n)$, will not be identically distributed because $\pi(\zeta_i = 1|\omega) = \omega_i$.

Certain dependence among the probabilities that rows are outliers can now be incorporated by letting ω_i be such that

$$\log \frac{\omega_i}{1-\omega_i} = \gamma + \delta_i, \quad \text{for} \quad i = 1,\ ...,n, \tag{2.6}$$

where γ is a fixed effect capturing the overall amount of outliers and the δ_i's are random effects that model the dependency in the ω_i and are linked by a hierarchical structure that lets their relative contribution be determined by data. That relative contribution is usually characterized through the variance of these random components, σ_δ^2. The hierarchical structure through which the δ_i link the probabilities of being outlier, ω_i, for rows that are close will depend on the particular example. One could, for example, model that dependency through a conditional autoregressive structure mimicking the one used in disease mapping to obtain spatially smoothed estimates of Poisson means (Besag et al[15] and Mollie[16]), and the one used in stylometric analysis to take the order of texts into consideration (Puig et al[14]). Instead, in the example on the plays by Shakespeare in Section 4, that will be done through a repeated measurement type of structure that takes into account the fact that acts belonging to the same play are more likely to belong to the same cluster than acts from different plays.

The choice of prior distribution for the parameters at hand, $\beta, k, \gamma, \sigma^2$, and σ_δ^2 will depend on the problem at hand. Section 4 will describe our choice for the example on Shakespeare's plays. The posterior distribution under the models in Sections 2.1 and 2.2 can not be computed analytically. Instead, one can update the models and simulate from them through the MCMC method. In the examples that follow, that has been implemented through JAGS (see, eg, Plummer[17]). The convergence of the chains has been assessed through visual inspection of the sample traces and by monitoring diagnostic measures. For each model, 4 chains with different initial values have been run until convergence.

2.3 | Calibration of the method through the choice of σ^2

In our method, a crucial role is played by σ^2, which models the variability of the random effects that rule the multinomial probabilities in the main set of rows, which are not outliers, together with the variability of the random effects for outliers, which is $k^2\sigma^2$.

In those instances where one expects outliers to be extreme, in the sense that their distribution is far from the distribution of the nonoutliers and hence where the value of k is much larger than 1, one can use a reference prior on σ^2. In the case where the distribution of the outlier observations is close to the one of the nonoutliers though, taking the 2 groups apart by simultaneously estimating k and σ^2 becomes a lot harder due to identifiability problems. To avoid that, here, the parameter value for σ^2 will be estimated through an empirical Bayes type of calibration process that should be carefully tailored to the problem at hand, and then that value is plugged into the analysis.

In a generic case in which the table is not structured, covered in Section 2.1, the estimation of σ^2 ideally requires one to have a subset of rows that one knows that belong to the main group and therefore that are not outliers. With this set of homogeneous rows, one can estimate σ^2 by using them to update the model that assumes (2.1) and (2.2) with the random effects ν_{ij} being normal $(0, \sigma^2)$ and the fixed effects β_j being normal $(0, 100^2)$ and the prior distribution for $1/\sigma^2$ being gamma $(0.1, 0.001)$, which are reference priors. One can then use the posterior distribution for σ^2 conditioned on the counts in the set of rows known to be nonoutliers, as an estimate of σ^2. In particular, in the simulation example presented in Section 3, the $\hat{\sigma}^2$ used to calibrate the outlier detection method will be $E[\sigma^2|y]$.

In the nonideal case where one does not have a subset of rows that are known to be nonoutliers, one can repeatedly take a small number of randomly chosen subsets of rows and calculate the corresponding set of estimates of σ^2, one for each subset of rows, using the method described above. Then one can use the median of the sample of σ^2 estimates to calibrate the model.

In the case of structured tables dealt with through the model in Section 2.2, one needs to think carefully about the way one estimates σ^2. Section 4 illustrates how that can be done in a case study like the one on Shakespeare's drama.

Note that the way in which one can estimate σ^2 in a particular example will rarely be unique. Different estimates of σ^2 will lead to a different number of observations being identified as outliers, but the relative degree of outlierness of all the observations will tend to be similar across different values for $\hat{\sigma}^2$. In the following simulation exercise, a small sensitivity analysis on the choice of estimate of σ^2 is carried out to quantify the potential effect of its misspecification and so of the calibration method.

3 | OUTLIER DETECTION IN A SIMULATED TABLE

3.1 | Description of the simulated scenario and the prior

To assess the performance of the Bayesian model driven outlier detection method for rows of an unstructured table and to carry out a small sensitivity analysis on the calibration method used, one simulation scenario is designed. In it, a table with 90 rows and 3 columns is simulated in a way such that all the rows are independent and each row has a count total of $N_i = 200$.

The scenario considered here is very close to the one faced in the analysis of control charts for multivariate attribute processes. In particular, each row of the table of this example could represent a different hospital, and the counts in each row could correspond to the result of certain medical procedure on a random sample of 200 patients from that hospital, categorized as complete success, partial success, or failure. The goal of the analysis would be to identify hospitals with a performance on this procedure that deviates from the one from the majority of the 90 hospitals considered.

The first 80 rows of the table, $y_i = (y_{i1}, y_{i2}, y_{i3})$ for $i = 1, \ldots, 80$, will be considered to be the nonoutlying observations, and they will be realizations of a Dirichlet multinomial $(N_i; \tau = 100, \mu = (0.5, 0.3, 0.2))$ model, which is a mixture of multinomial $(N_i; \theta = (\theta_1, \theta_2, \theta_3))$ distributions, where the mixing distribution on the multinomial parameter is Dirichlet $(\tau = 100, \mu = (0.5, 0.3, 0.2))$. For a description of the Dirichlet multinomial $(N;\tau,\mu)$ model with the parametrization used here, see, eg, Puig and Ginebra.[18] Note that under this model, $E[\theta_i|\mu_i, \tau_i] = \mu_i$ and $E[y_i| N_i, \tau_i, \mu_i] = N_i\mu_i$, and hence, μ_i determines the mean of θ_i and of y_i. Furthermore, $V[\theta_i|\mu_i, \tau_i] = \frac{1}{1+\tau_i}\mu_i(1-\mu_i)$ and $V[y_{ij}|N_i, \tau_i, \mu_i] = N_i\frac{N_i+\tau_i}{1+\tau_i}\mu_{ij}(1-\mu_{ij})$, and hence, the larger τ_i, the smaller the overdispersion of the ith row.

The 81st to the 85th rows, y_i with $i = 81, \ldots, 85$, will be considered to be moderate outliers, and they will be realizations of a Dirichlet multinomial $(\tau = 100, \mu = (0.6, 0.3, 0.1))$. The last 5 rows in the table, y_i with $i = 86, \ldots, 90$, will be considered to be extreme outliers, and they will be realizations of a Dirichlet multinomial $(\tau = 100, \mu = (0.2, 0.5, 0.3))$.

Figure 1 presents the proportion of the counts of each one of the 3 categories for each one of the last 40 rows of the table, first ordered with the 10 outlier rows appearing randomly among these last 40 rows, and then ordered in such a way that the 10 outliers are presented at the end of the sequence. The simulation scenario was designed with only 3 categories to help visualize the existence of

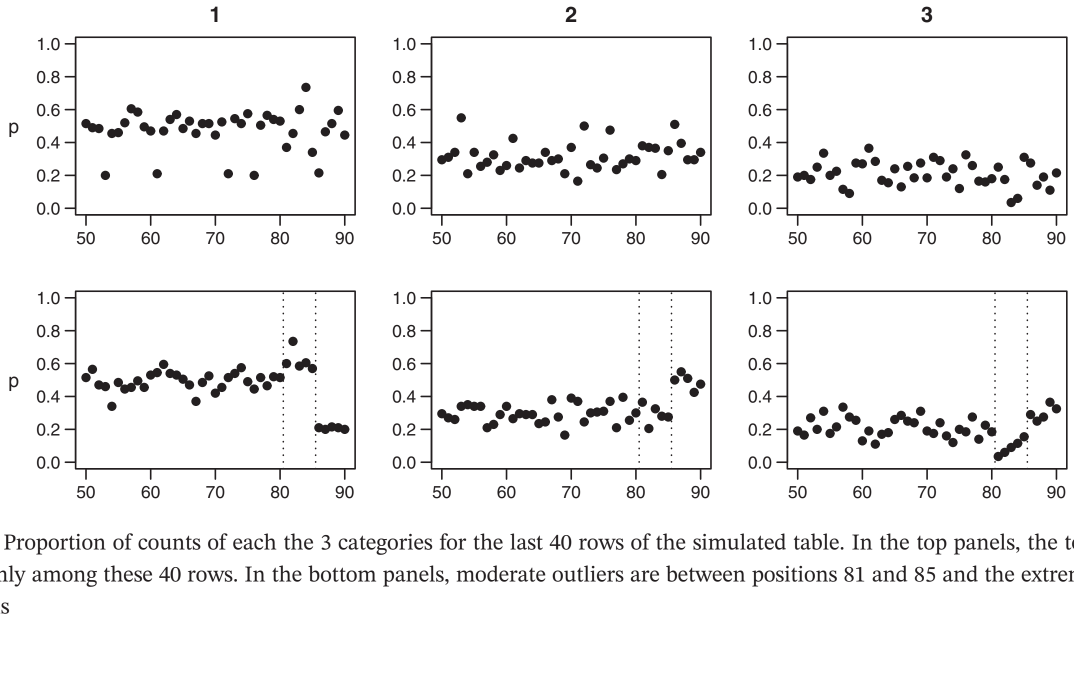

FIGURE 1 Proportion of counts of each the 3 categories for the last 40 rows of the simulated table. In the top panels, the ten outliers are placed randomly among these 40 rows. In the bottom panels, moderate outliers are between positions 81 and 85 and the extreme ones in the last 5 positions

outliers. With tables with more than 3 columns, the existence of outliers would be a lot harder to pinpoint in graphics like Figure 1.

Note that under this simulated scenario, the rows follow a multinomial mixture model different from the one assumed by our outlier detection method, and the scenario has 3 different row patterns and not just 2. By using a model for simulation that is different from the one assumed by the outlier detection method, the assessment of the method is more realistic. Note also that the scenario is quite challenging, because 1 out of every 9 rows is an outlier, and therefore, the usual outlier detection methods will fail because outliers will mask themselves.

In the Bayesian approach, one needs to choose a prior distribution for the parameters of the model, $\omega, \beta=(\beta_2, ..., \beta_J), \sigma^2$ and k. The prior for ω needs to be chosen based on the percentage of outliers that one expects to find. If one has no idea beyond the fact that the number of outlying rows is less than one-half of the total number of rows, one can choose a uniform $(0, 0.5)$, but if, for example, one expects to find less than 20% of outliers, a uniform $(0, 0.2)$ or a beta $(1, 15)$ would be better choices. In this example, ω is assumed to be uniform $(0, 0.5)$ distributed. For $\beta=(\beta_2, ..., \beta_J)$, one will assume a reference prior under which the β_j are independent and normal $(0,100^2)$ distributed, and as a prior for k, here, one assumes that it is uniform $(2.5, 25)$. Using 2.5 as the lowest possible value for k ensures that one will not consider a random effect ν_i to be outlying unless its standard deviation is 2.5 times larger than the one for nonoutliers. The larger this lowest possible value for k, the smaller the number of outlier rows that will be detected.

3.2 | Description of the results

To calibrate the method, it will first be assumed that one faces an ideal case in which one knows that the first 50 rows are homogeneous, without the presence of any outliers, and hence, that they can be used as a training set of rows to estimate σ^2. This calibration method will be labeled as method A. The value of $\hat{\sigma}^2$ used is 0.056, which is the posterior expected value of σ^2 obtained as described in Section 2.3, conditioning on the counts in the first 50 rows. That estimate is then used to calibrate the outlier detection method to tell outlier rows apart from nonoutlier rows among the set of the last 40 rows of the table.

To explore the performance of the method in a nonideal case, when one does not have any subset of rows known to be uncontaminated with outliers, a second calibration method has been implemented and labeled as method B. Under method B, one estimates σ^2 by randomly sampling 20 subsets of 5 rows and finding the corresponding set of estimates of σ^2, one for each subset of 5 rows, as explained in Section 2.3. Given that many of these subsets are bound to include outliers, we use the median of this sample of estimates of σ^2 as a robust estimate of the variability of the random effects that determine the distribution of the multinomial probabilities of the nonoutlier rows. The value of $\hat{\sigma}^2$ found in this way by sampling 20 subsets of 5 rows in our example is 0.083.

Figure 2 presents an estimate of the posterior probability that each one of the last 40 rows in this simulated scenario is an outlier, $E[\zeta_i|y]$, first, using the $\hat{\sigma}^2 = 0.056$ obtained from the ideal calibration method A and second, using the $\hat{\sigma}^2 = 0.083$ obtained from the calibration

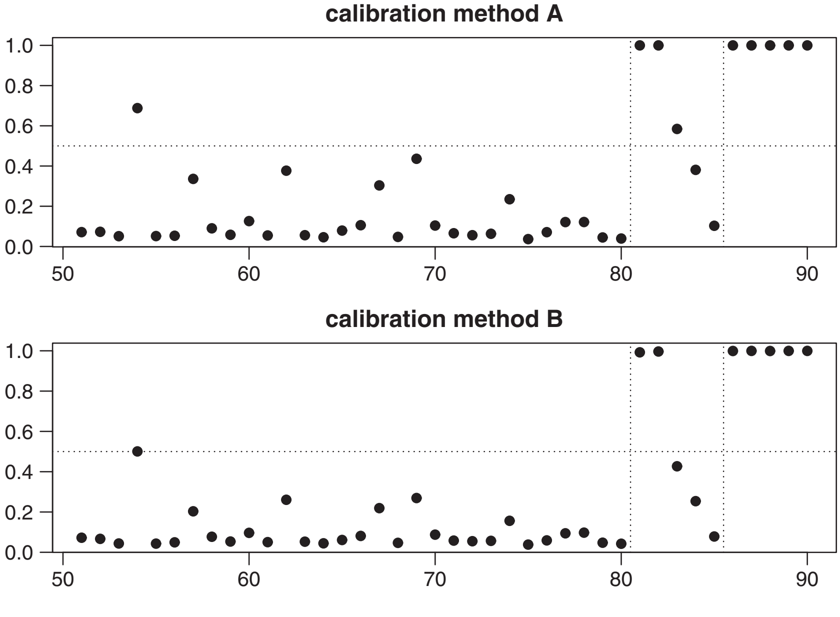

FIGURE 2 Estimate of the posterior probabilities that each one of the last 40 rows of the simulated table is an outlier, when using the model for unstructured tables in Section 2.1 with the 2 calibration methods described in Section 3.2. In reality, the first thirty rows are simulated to be nonoutliers, the next 5 rows are moderate outliers, and the last 5 rows are extreme outliers

method B. Given that under method A, one uses the fact that one knows that the first 50 rows are nonoutliers, the posterior probability that they are outliers is not presented in Figure 2. Note that all 5 extreme outliers and 3 and 2 of the moderate outliers have a posterior probability larger than .5 that they are an outlier. Only one of the 30 nonoutlier rows among the last 40 rows used as testing sample are classified as an outlier under both calibration methods. Hence, the Bayesian model proposed for the identification of outliers works adequately in this example, even though one faces a setting where 1 out of every 9 observations are outliers.

To assess how this outlier detection method fares under repeated use and to compare the performance of the 2 calibration methods considered above, the same simulation experiment described here has been repeated 500 times. That is, one has simulated 500 different tables of 90 rows with a total of 200 counts each from the model described in Section 3.1; for each one of these 500 tables, one has estimated σ^2 through the calibration methods A and B, and one has implemented the outlier detection method with these 2 calibration estimates. Table 2 presents the percentage of nonoutlier rows of moderate outlier rows and of extreme outlier rows that have been classified as outliers under both calibration methods. As one expects, the calibration method A that knows that a subset of rows are nonoutliers is better than calibration method B at identifying outliers, specially in the case of moderate outliers.

4 | OUTLIERS AMONG ACTS OF PLAYS BY SHAKESPEARE

4.1 | Description of the problem and the data

Very little is known about the life of William Shakespeare, and that has fueled a heated debate around the authorship of plays attributed to him. Even though only a minority of experts question his authorship, some claim that some or all of the plays attributed to him could be work of or joint work with Francis Bacon, Cristopher Marlowe, Ben Johnson, Sir Walter Raleigh, or Edward de Vere, among others. That debate has been going on for more than 150 years, and far too many people has contributed to it to try to summarize it here. For recent overviews of that debate, see, for example, Hope,[19] Edmondson and Wells,[20] or Shahan and Waugh.[21]

The statistical analysis of literary style has often been used to characterize the style of texts and help settle authorship-attribution problems (see, eg, Holmes[22] and Stamatatos[23]). The frequency of use of function words is one of the best tools when it comes to discriminating styles, and that is what will be used here. Examples of the use of function words can be found in Mosteller and Wallace,[24] Holmes, [25] Zhao and Zobel,[26] Giron et al,[27] Riba and Ginebra,[28] and Puig et al,[29] among many other places.

To explore possible departures from the style of Shakespeare in the plays attributed to him, the outlier detection analysis is used on the 35 plays collected in the first printing of the *First Folio* edition of the plays by Shakespeare, published posthumously in 1623. That edition includes 14 comedies, 10 histories, and 11 tragedies, and it is the only reliable version for about 20 of these plays. All plays consist of 5 acts, which will be the unit used in this study, and hence the analysis will consider a total of $n = 175$ text units.

As a stylistic characteristic, we will focus on the counts of the 20 most frequent function words in these plays, which are as follows: *the, and, I, to, of, a, you, my, that, in, is, not, it, for, with, me, your, his, this*, and *be*. Hence, data will consist of the 175×20 table with these 20 most frequent word counts, which is partially presented in Table 1.

Figure 3 presents the frequency of appearance of these 20 function words in the 175 acts, which are grouped by play and genre. Note that there is a clear difference in the use of many of these words in comedies, histories, and tragedies, and one will need to take that difference into account when searching for acts with an outlying behavior. Note also that even though all 5 acts in the same play will not be considered to be all either outliers or nonoutliers at once, by adapting the mixed multinomial cluster model in Section 2.2, one will incorporate the fact that acts from the same play are more likely to belong to the same author than acts from different plays.

TABLE 2 Percentage of outliers detected among the 3 types of rows, estimated through repeated use of the outlier detection method on tables of the same size and from the same model considered in Section 3.1

Calibration Method	Nonoutliers, %	Moderate Outliers, %	Extreme Outliers, %
A	3.0	48.3	99.6
B	0.9	28.4	94.8

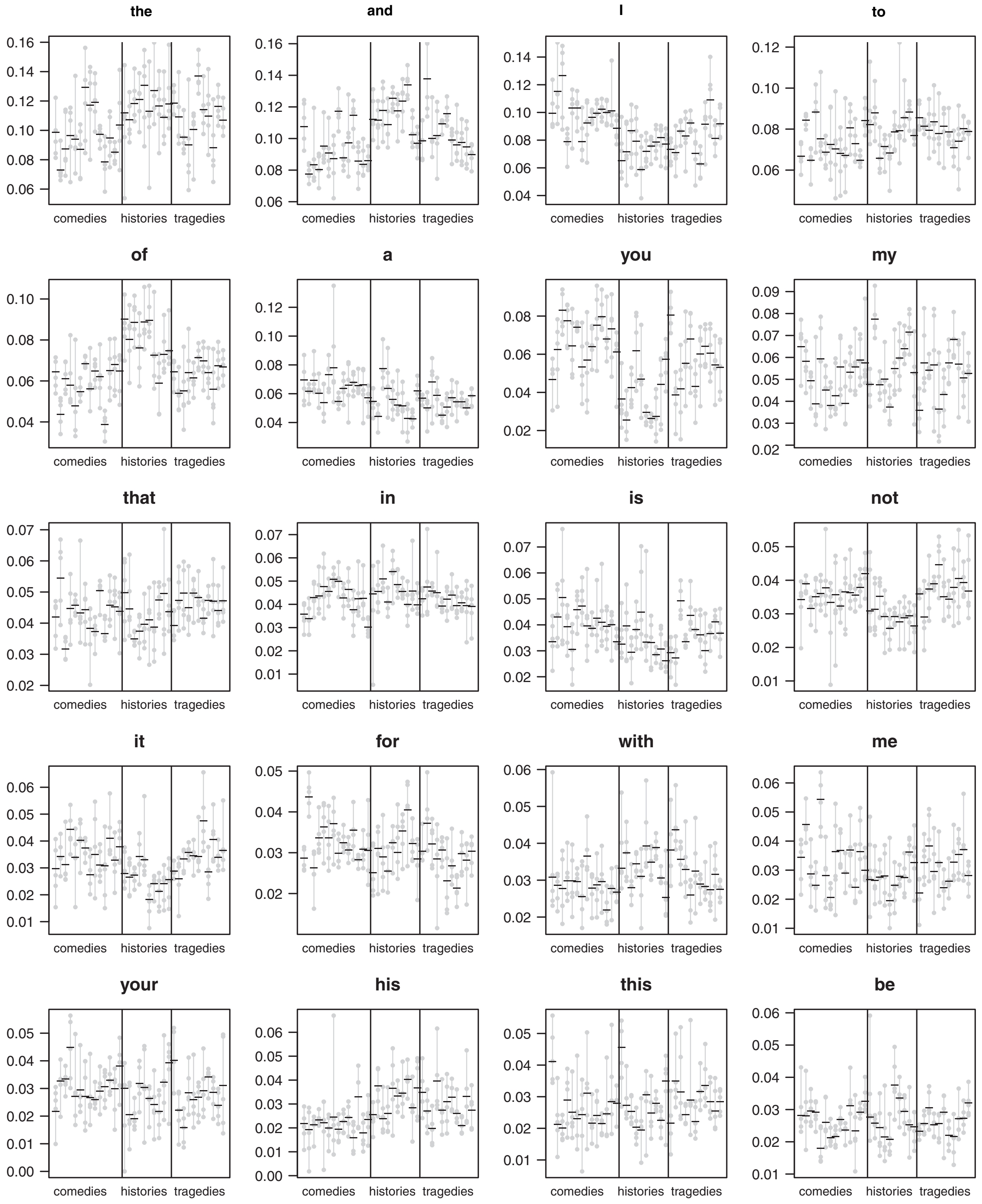

FIGURE 3 Grey dots indicate the proportion of each function word in each act, grouped by play and genre. Black lines indicate the average proportion of function word in every play

For each row (act) in Table 1, one has a vector valued categorical observation, $y_i = (y_{i1}, \ldots, y_{iJ})$, where $J = 20$ denotes the number of categories. The goal is to identify the rows of the table that depart from the style that one expects from an act from a play from the corresponding genre in that edition. If all the acts had been written by a single author and were all of the same genre, one might expect all the rows in Table 1 to come from a single distribution. If instead, the distribution of a few of these rows is very different from what is expected for an act in such a play from such genre by Shakespeare, which could indicate that these rows were written or tampered of by someone else.

4.2 | Description of the model and of the results

To adapt the outlier detection model for structured tables in Section 2.2 to this problem, one needs to specify the $\mu(i,j)$, ruling how the multinomial probabilities in different rows are related, and the distribution of ω_i, ruling the way in which the probabilities that 2 different rows are outliers are related.

In the first case, it will be assumed that $\mu(i,j) = \beta_{0j} + \beta_{1j} I_{H_i} + \beta_{2j} I_{T_i}$, where I_{H_i} and I_{T_i} are variables indicating whether the corresponding act, i, belongs to a play that is a history or a tragedy, and where β_{01},β_{11} and β_{21} are 0. That is, the multinomial probabilities of all the acts in plays of the same genre will share the same fixed effect component, $\mu(i,j)$, in (2.4).

In the second case, the log odds for ω_i will be modeled as indicated in Section 2.2 with a fixed effect, γ, and a set of random effects, δ_i. In this specific example on the plays by Shakespeare, all the 5 δ_i that correspond to the same play, p, will be considered to take the same value, $\delta_{P(i)}$, and therefore, all 5 ω_i for acts in the same given play p will take the same value, $\omega_{P(i)}$, with

$$\log \frac{\omega_{P(i)}}{1-\omega_{P(i)}} = \gamma + \delta_{P(i)}, \quad \text{for} \quad P(i) = 1, \ldots, 35, \qquad (4.1)$$

where the 35 values for $\delta_{P(i)}$ will be assumed to be normal $(0, \sigma_\delta^2)$ distributed. Hence, 2 acts in the same play will tend to be more likely both either outliers or nonoutliers together than 2 acts in different plays. Note though that the values of ζ_i for the 5 acts in the same play do not coincide, and the posterior distribution for the corresponding ζ_i will be different. As a consequence, not all acts in the same play need to be classified together in the same cluster.

Note that this model allows one to classify acts into either the outlier or the nonoutlier group through $E(\zeta_i|y)$, and it allows one to classify whole plays, p, into the same groups through $E(\omega_{P(i)})$, the way it will be illustrated in Table 3.

In this search for outlying acts in Shakespeare, the variability of the random effects, ν_{ij}, that rule the distribution of the θ_i for the nonoutlier rows, σ^2, after taking genre into consideration, needs to be estimated. Here, that will be done by assuming that 5 acts belonging to the same play are homogeneous and using the technique described in Section 2.3 on the 5 acts of each one of the 35 plays considered. That is, one updates the Bayesian model that assumes (2.1) and (2.2) with ν_{ij} being normal $(0,\sigma^2)$, based on the data on the 5 acts of each play. As a prior, one assumes that the β_{0j} are normal $(0,100^2)$ and that $1/\sigma^2$ is gamma (0.1,0.001). Figure 4 presents the posterior distribution for σ^2 given the 5 acts of a play, for each one of the 35 plays considered. The posterior expected value of each one of these distributions could be used as an estimate of σ^2. As our choice of $\hat{\sigma}^2$ to calibrate the method, here, we use an average of all the 35 posterior expected values of these distributions, which is $\hat{\sigma}^2 = 0.031$. More robust choices for $\hat{\sigma}^2$ would have been using the median of all the 35 individual estimates of σ^2, the way done in Section 3, or using an average of these estimates after discarding the largest ones, which might be contaminated with outliers. Note though that relying on these smaller estimates for $\hat{\sigma}^2$ would have lead to the identification of more outliers, but it would have not changed the relative degree of outlierness of each row, which is what is more meaningful.

If the set of 5 acts of a given play was not homogeneous because one or a few of these 5 acts are outliers, the posterior expected value of the corresponding distribution for σ^2 would tend to be larger than the rest. If one did not discard that value in the estimation of σ^2, that estimate would be larger, which would lead to fewer outliers being detected. On the other hand, by estimating σ^2 taking into account only the variability within play, one is underestimating the variability due to the pass of time. One way to estimate upper bounds for σ^2 that take into account all natural sources of variability would be to implement the idea used in Section 4.2 using all acts of all the plays in the same genre at once, instead of using 5 acts of a given play at a time. It is important to remark that different estimates of σ^2 will lead to a different number of outliers, but the relative degree of outlierness of observations will be similar.

As a prior distribution for the $\beta_{0j}, \beta_{1j}, \beta_{2j}$ determining $\mu(i,j)$, one assumes independent normal $(0,100^2)$ distributions. The prior for the parameter k determining the variability of ν_{ij} for the outlier rows will be uniform $(2.5, 25)$. Here, the smallest possible value for k is again set to be 2.5. The larger this smallest possible value for k, the harder for an act in a play to appear as an outlier, and hence the harder it is to detect false outliers. The prior on γ will be normal $(0,100^2)$, and the prior the inverse of σ_δ^2 is chosen to be gamma (0.1,0.001), which corresponds to a reference prior distribution.

Figure 5 presents the posterior probability that each one of the acts in the *First Folio* edition is an outlier, estimated through $E[\zeta_i|y]$ for $i = 1, \ldots, 175$. It is natural to classify an act as an outlier if that probability is larger than .5 and as a nonoutlier otherwise. Table 3 presents the list of 175 acts classified either as outlier or as nonoutlier based on that criteria. In a similar way, both Figure 5 and Table 3 present the posterior probabilities that each one of the plays is an outlier, estimated through $E[\omega_p|y]$ for $p = 1, \ldots, 35$.

TABLE 3 List of the 35 plays of the *First Folio* edition[a]

	Act					$E[\omega_p\|y]$
Play	**1**	**2**	**3**	**4**	**5**	
The Tempest	XXXX			XXXX		.21
The Two Gentlemen of Verona						.04
The Merry Wives of Windsor						.04
Measure for Measure						.03
The Comedy of Errors	XXXX					.17
Much Ado about Nothing						.03
Love's Labour's Lost	XXXX	XXXX	XXXX	XXXX		**.68**
A Midsummer Night's Dream		XXXX			XXXX	.27
The Merchant of Venice						.03
As You Like It						.03
The Taming of the Shrew						.04
All's Well that Ends Well						.03
Twelfth Night						.03
The Winter's Tale						.04
King John						.06
Richard II						.04
Henry IV, Part 1						.04
Henry IV, Part 2		XXXX	XXXX		XXXX	.37
Henry V	XXXX	XXXX	XXXX	XXXX	XXXX	**.80**
Henry VI, Part 1						.04
Henry VI, Part 2						.06
Henry VI, Part 3						.05
Richard III						.04
Henry VIII						.05
Coriolanus		XXXX				.19
Titus Andronicus	XXXX		XXXX			.33
Romeo and Juliet					XXXX	.25
Timon of Athens						.04
Julius Caesar						.03
Macbeth						.04
Hamlet						.03
King Lear						.03
Othello			XXXX			.16
Antony and Cleopatra						.03
Cymbeline						.03

[a] The acts in bold are classified as outliers based on estimates of their posterior probability of being so, $E[\zeta_i|y]$. Whole plays can be classified through estimates of the corresponding probability, $E[\omega_p|y]$.

According to that classification, the only play with all its 5 acts considered to be outliers is *Henry V*. Merriam[30] suggests that Shakespeare reworked a Marlowe (or Peele) original play to create *Henry V*. The only play with 4 acts considered to be so is *Love's Labour's Lost*, which has a title page stating that the play was newly corrected

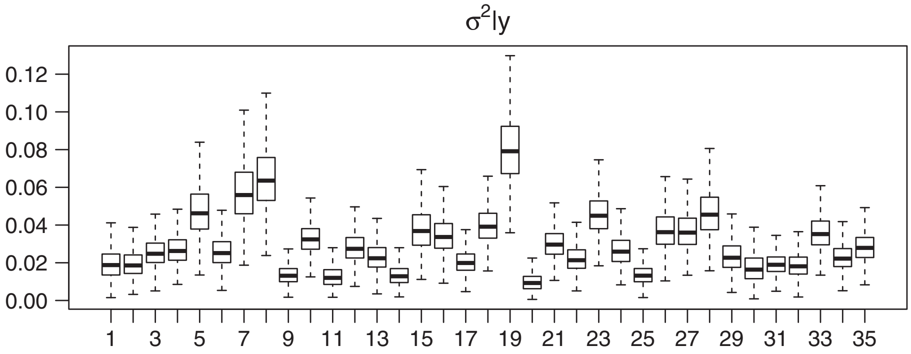

FIGURE 4 Box plots of a sample of 10 000 observations from the posterior distribution for σ^2 based on the 5 acts of each one of the plays in the *First Folio* edition. In the outlier detection implementation in Section 4, one uses the average of the 35 estimated $E[\sigma^2|y]$ as the value for $\hat{\sigma}^2$

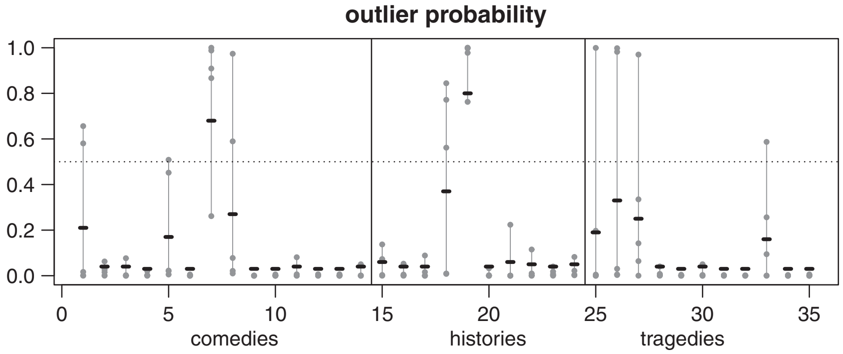

FIGURE 5 Posterior probability that each act (dots) and each play (segments) in the *First Folio* edition is an outlier, estimated through $E[\zeta_i|y]$ and $E[\omega_p|y]$, respectively. Acts are grouped by play and by genre

and augmented by Shakespeare, which has lead some scholars to suggest that it was a revision of an earlier version by someone else, maybe the same Peele suggested as a precursor author for *Henry V*. The posterior probabilities that these 2 plays are outliers, $E[\omega_p|y]$, are .80 and .68, respectively. These 2 plays are the only ones with an $E[\omega_p|y]$ larger than 0.5.

Besides these 2 plays, it is also worth remarking the fact that *Henry IV, Part 2* has 3 acts classified as outliers. One of the 3 plays with 2 acts classified to be so is *Titus Andronicus*, which is the first known printing of a Shakespearean play and is considered by many to be a collaboration between Shakespeare and at least one other dramatist.

The Derbyite theory of Shakespeare authorship defends that the true author of some of the works of William Shakespeare was William Stanley; the plays attributed to Shakespeare most often linked to Stanley are *Love's Labour's Lost, A Midsummer Night's Dream*, and *The Tempest*, which are 3 of the 6 plays with 2 or more acts identified as outliers by our method. Note though that there are several reasons that could explain why an act behaves as an outlier, besides the fact that it could have been written by another author.

5 | FINAL COMMENTS

We have presented Bayesian hierarchical models that characterize the stable pattern as well departures from that pattern in rows or a contingency table and help identify which observations follow the mainstream pattern and which ones do not. These models have been tried first on a simulated example to check that they work for unstructured tables, and then it has been used to explore the existence of heterogeneities in the drama by Shakespeare.

A critical aspect of the outlier detection method is the calibration of σ^2. In the simulation example, there was no need to calibrate, and the same results can be obtained by using a reference prior for σ^2. Instead, in the example of Shakespeare, outliers are not so easy to identify, and one needs to resort to calibrating σ^2. As it has been illustrated in Section 3, different estimates of σ^2 will lead to a different number of observations being identified as outliers, but the relative degree of outlierness of observations will be similar across different values for $\hat{\sigma}^2$. And the same applies about the choice of the smallest value allowed for k, and about the choice of the threshold value for $E[\zeta_i|y]$ and for $E[\omega_p|y]$ used to decide which acts and plays qualify as outliers. Different choices of these values lead to more or less outliers, but observations are ranked similarly from being more to being less outliers.

In the ideal case where one knows a subset of rows to be nonoutliers to start with, the computational burden of the method used on tables of sizes similar to the ones considered here is negligible. When one needs to resort to more sophisticated calibration techniques, the computational burden will largely depend on table sizes.

ACKNOWLEDGEMENTS

This work was funded in part by grant no. MTM2013-43992-R of the Ministerio de Economia y Competitividad of Spain. The authors want to thank Marti Font for providing the data used.

REFERENCES

1. Barnett V, Lewis T. *Outliers in Statistical Data*. 3rd ed. New York: Wiley; 1994.
2. Rousseeuw P, Leroy A. *Robust Regression and Outlier Detection*. New York: Wiley; 2003.
3. Fienberg SE. Preliminary graphical analysis and quasi independence for two way contingency tables. *Appl Stat*. 1969; 18:153-168.
4. Haberman SJ. The analysis of residuals in cross classified tables. *Biometrics*. 1973;29:205-220.
5. Brown MB. Identification of the sources of significance in two-way contingency tables. *Appl Stat*. 1974;23:405-413.
6. Fuchs C, Kenett R. A test for detecting outlying cells in the multinomial distribution and two-way contingency tables. *J Am Stat Assoc*. 1980;75:395-398.
7. Simonoff JS. Detecting outlying cells in two-way contingency tables via backwards stepping. *Technometrics*. 1988;30:339-345.
8. Yick JS, Lee AH. Unmasking outliers in two-way contingency tables. *Comput Stat Data Anal*. 1996;29:69-79.
9. Kuhnt S. Outlier identification procedures for contingency tables using maximum likelihood and L1 estimates. *Scand J Stat*. 2004;31:431-442.
10. Mebane WR, Sekhon JS. Robust estimation and outlier detection for overdispersed multinomial models of count data. *Am J Political Sci*. 2004;48:392-411.
11. Kuhnt S, Rapallo F, Rehage A.. Outlier detection in contingency tables based on minimal patterns. *Stat Comput*. 2014;24:481-491.
12. Box GEP, Tiao GC. A Bayesian approach to some outlier problems. *Biometrika*. 1968;55:119-129.
13. Fernandez C, Green PJ. Modelling spatially correlated data via mixtures: a Bayesian approach. *J R Stat Soc B*. 2002;64:805-826.
14. Puig X, Font M, Ginebra J. Classification of literary style that takes order into consideration. *J Quant Ling*. 2015;22:177-201.
15. Besag J, York JC, Mollie A. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math*. 1991;43:1-20.
16. Mollie A. Bayesian mapping of disease. In: Gilks WR, Richardson S, Spiegelhalter DJ, eds. *Markov Chain Monte Carlo in Practice*. Chapman Hall: New York; 1996:359-379.
17. Plummer M. *JAGS Version 4.0. 0 User Manual*. Lyon, France: International Agency for Research on Cancer; 2015.
18. Puig X, Ginebra J. A Bayesian cluster analysis of election results. *J Appl Stat*. 2014;41:73-94.
19. Hope J. *The Authorship of Shakespeare's Plays*. Cambridge: Cambridge University Press; 1994.
20. Edmondson P, Wells S. *Shakespeare Beyond Doubt: Evidence, Argument, Controversy*. Cambridge: Cambridge University Press; 2013.
21. Shahan JM, Waugh A. *Shakespeare Beyond Doubt? Exposing and Industry in Denial*. London: Llumina Press; 2013.
22. Holmes DI. The analysis of literary style. A review. *J R Stat Soc, Ser A*. 1985;148:328-341.
23. Stamatatatos E. A survey of modern authorship attribution methods. *J Am Soc Inf Sci Technol*. 2009;60:538-556.
24. Mosteller F, Wallace DL. *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*. 1st and 2nd ed. Berlin: Springer-Verlag; 1984.
25. Holmes DI. A stylometric analysis of Mormon scripture and related texts. *J R Stat Soc, Ser A*. 1992;155:91-120.
26. Zhao Y, Zobel J. Effective and scalable authorship attribution using function words. *Information Retrieval Technology*, *Lecture Notes in Computer Science*, vol. 3689. Berlin: Springer Verlag; 2005:174-189. https://doi.org/10.1007/11562382. 14.
27. Giron J, Ginebra J, Riba A. Bayesian analysis of a multinomial sequence and homogeneity of literary style. *The Am Stat*. 2005;59:19-30.
28. Riba A, Ginebra J. Change-point estimation in a multinomial sequence and homogeneity of literary style. *J Appl Stat*. 2005;32:61-74.
29. Puig X, Font M, Ginebra J. A unified approach to authorship attribution and verification. *The Am Stat*. 2016;70:232-242.
30. Merriam T. Heterogeneous authorship in early Shakespeare and the problem of Henry V. *Lit Ling Comput*. 1998;13:15-28.

Xavier Puig holds a degree in Statistics and a PhD from the Technical University of Catalonia, where he is currently an Associate Professor of Statistics. He does research on Bayesian data analysis of categorical data and on model-based cluster analysis, with applications to the analysis of election results, to the statistical analysis of literary style and in biostatistics and biomedicine.

Josep Ginebra holds an Industrial Engineering degree from UPC and a PhD in Statistics from the University of Wisconsin-Madison. Currently is a Full Professor of Statistics at the Technical University of Catalonia. He does research on statistical foundations, design of experiments and on Bayesian data analysis of discrete and categorical data, with applications to the analysis of election results and to the statistical analysis of literary style.

How to cite this article: Puig X, Ginebra J. Outlier detection for multivariate categorical data. *Qual Reliab Engng Int*. 2018;1–13. https://doi.org/10.1002/qre.2339