
Available from: http://dx.doi.org/10.2174/157341112800392634

Accessed from: http://hdl.handle.net/1959.13/1053715
A Numerical Evaluation of the Classification of Portuguese Red Wine

Eric J. Beh*, Thomas B. Farver

1School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW 2308 Australia
2School of Veterinary Medicine, University of California Davis, Davis, CA 95616, USA

*Corresponding author: eric.beh@newcastle.edu.au, Ph. +61 2 4921 5113, Fax: +61 2 4921 6898

Running Title: A Numerical Evaluation of Red Wine

Abstract: Evaluating the accuracy of classification methods for rating wines based on their physical and chemical characteristics is an important part of the wine industry. An ability to accurately predict the quality of wine based on this information provides vast opportunity for applications in other areas of science. Recently, a method was proposed that utilises data mining techniques for the prediction of the quality of wines based solely on their physicochemical properties and compared these results with those classifications obtained by experienced assessors. In this paper we explore an analytical approach to evaluating the accuracy of these classification methods using new advances in the area of statistical modelling.

Keywords: Classification methods, Confusion Matrix, Estimation methods, Ordinal log-linear models, Physicochemical characteristics, Red wine.

1. INTRODUCTION

Analytically evaluating the accuracy of classification methods that involve data derived from experimental data is an important aspect of the chemistry discipline. Such classification methods help to clearly identify the composition of compounds, or the extent to which particular compounds may be associated with each other. Recently, a study of the chemical and physical properties of red and white Portuguese wine was undertaken and a method was derived for classifying the quality of wine based on this information ([1]). In their study, [1] considered more than 6500 red and white wines and successfully demonstrated that one can predict the quality
of these wines based on 11 different characteristics. The prediction method was based on tools familiar in the
data mining areas and the quality of wine was compared with those classifications given by experienced wine
assessors.

Evaluating the accuracy of the prediction methods for determining the quality of a wine therefore becomes an
important aspect to the analyst. A graphical evaluation of the classification of the Portuguese wines was
performed in this issue ([2]) and shows that, generally, it compared extremely well with those classifications
given by the professional wine assessors. Analysts may also adopt a numerical approach to evaluating the
classification method. By considering new advances in the area of ordinal log-linear models, we shall be doing
just that in the following sections. Section 2 will describe the data that will be analysed. Such data is
interestingly referred to a confusion matrix (a special example of what is commonly called a contingency table)
and consists of ordered row categories and ordered column categories that summarise the number of wines that
were given a particular classification by the assessors and a predicted classification based on the
physicochemical data. Section 3 will briefly describe some of the fundamental qualities of ordinal log-linear
models and explore the computationally intensive estimation methods that are commonly used, and alternative
estimation methods that allow for estimation to be made easily and directly. Section 4 will demonstrate the
applicability of these modelling tools to part of the data given by [1] while Section 5 will provide some
concluding remarks.

2. MATERIALS AND METHODS

The study of [1] primarily involved determining a method of accurately predicting the classification of 4898
white wines and 1599 red wines based on 11 key chemical and physical qualities of the wines. The data from
these wines were collected from the Minho region of Portugal between May 2004 and February 2007 and
consisted of

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulphur dioxide
- Total sulphur dioxide
- Density
- pH
- Sulphates
The quality of the red and white wines based on these characteristics was measured on an 11 point scale where 0 designates a poor quality wine and 10 an excellent quality wine. Prior to the analytical approach taken by [1] to evaluate the wines, they were assessed by three professional assessors and the official observed classification of the quality of a given wine was based on the median quality of the wine from these assessors. The task was therefore to establish whether the classification method adopted by these authors, who employed support vector machine (SVM) techniques, was consistent with those given by the assessors. Therefore, one may view such a study as evaluating the accuracy of an “electronic tongue” for the wines.

The method of summary used to compare the assessor’s evaluation with the classification method used by [1] was the confusion matrix ([3]); such a matrix is sometimes referred to as a matching matrix or misclassification matrix ([4]). For the 10 point scale considered in the study, such a matrix consists of 11 rows and 11 columns and summarises the number of wines that were given a particular evaluation by the assessors and a particular predicted evaluation (not necessarily the same as the by the assessors) by analytical means. Table 1 is a partial confusion matrix for the red wines analysed; it is partial since many wines were not given a very low or a very high classification by either approach. It is size 6x5 meaning that Table 1 consists of 6 rows and 5 columns. Ideally, the confusion matrix should be perfectly square and diagonal so that if the method of prediction works perfectly, then those wines assessed by a classification of 6, say, should all be given a predicted classification of 6 by the analytical approach that considered the chemical and physical characteristics of the wine. To help better understand the meaning of the elements in Table 1, consider the number 400. It tells us that there were 400 red wines which were given a classification of 6 by both the assessors and the prediction method. The element 10 in the last row of the table shows that there were 10 red wines that were given a classification of 8 by the assessors but were given a classification of 6 by the prediction method. This suggests that either the assessors were rather liberal in their assessment of the wine or the model underestimated the predicted classification. Interestingly, while [1] point out that 1599 red wines were considered in their analysis, their confusion matrix (and therefore Table 1) consists of the cross-classification of 1600 wines.

Table 1. Confusion matrix of the red wines studied.
As described above, [1] also considered the analysis of white wine. In their graphical evaluation of the classification method (using correspondence analysis), [2] analysed both red and white wines. However, in this study we shall avoid repetition of the analytical approach and demonstrate its application to the red wines only. The interested reader is invited to consider its use for the evaluation of the 4898 white wines that were originally considered in the study.

3. MODELS OF ASSOCIATION

Suppose we consider the confusion matrix of Table 1 and refer to it by N. Denote the ij’th cell entry $n_{ij}$ (for $i = 1, \ldots, 6$ and $j = 1, \ldots, 5$) as the number of wines that were given classification $i$ by the assessors and classification $j$ using the approach of [1]. For example $n_{43} = 400$ and $n_{63} = 10$. Let $p_{ij}$ be the proportion of wines classified into the ij’th cell of Table 1 so that, say, $p_{43} = 400/1600 = 0.25$ and $p_{63} = 10/1600 = 0.00625$.

We shall also define the proportion of wines given a classification of $i$ by the assessors to be $p_{i*}$ (so that $p_{4*} = 638/1600 = 0.399$) and the proportion of wines given a predicted classification (by using the analytical approach taken by [1]) of $j$ by $p_{*j}$ (so that $p_{*5} = 693/1600 = 0.433$).

If the assessment of the wine given by the assessors and the predicted classification of the wines by more analytical means are independent then one may estimate the number of wines to fall into the (i, j)th cell of Table 1 using the log-linear model
\[
\ln m_{ij} = \mu + \alpha_i + \beta_j.
\]  

(1)

Here \( \ln \) indicates for the “natural logarithm” of \( m_{ij} \) which is the estimated number of wines in the \((i, j)\)’th cell of the table, \( \mu \) is the grand mean, and \( \alpha_i, i = 1, \ldots, 6, \) and \( \beta_j, j = 1, \ldots, 5, \) are main effects. If the analyst believes the two classification methods to be related, this may be incorporated into the model by considering instead

\[
\ln m_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}
\]

(2)

where the interaction between assessors wine classification and that obtained from the method of prediction, \( \gamma_{ij} \), is not equal to zero. This is referred to as the log-linear model of the confusion matrix (Table 1). Since the rows and columns of Table 1 consist of ordered classifications, we can modify the log-linear model to reflect this structure:

\[
\ln m_{ij} = \mu + \alpha_i + \beta_j + \phi(i - \bar{u})(j - \bar{v})
\]

(3)

Model (3) is referred to as the ordinal log-linear model and the term that numerically summarises the relationship between the row and column classifications is \( \phi \). This is called the association parameter and is positive if there is a positive association between the categories (so that an increase on the assessor’s classification will be linked to an increase in the predicted classification). A value of \( \phi \) that is negative indicates that there is a negative association between the categories (so that an increase in the assessors classification will be linked to a decrease in the predicted classification). No association will correspond to \( \phi = 0 \) and (3) will reduce to (1) in this case. The values of \( \bar{u} \) and \( \bar{v} \) are the mean row classification and column predicted classifications, respectively, such that \( \bar{u} = \sum_{i=1}^{6} i p_{i\cdot} \) and \( \bar{v} = \sum_{j=1}^{5} j p_{\cdot j} \).

The aim, therefore, for any study involving the ordinal log-linear model, (3), is to estimate the parameter \( \phi \). Estimation of \( \phi \) for the confusion matrix of Table 1 will allow for an understanding of the association between the table’s row and column classifications. There are many statistical ways in which this can be done involving a
varying degree of complexity. In the next section we shall briefly describe two very different methods of estimation.

4. ESTIMATING $\phi$

4.1 Newton-Raphson Iterative Procedure

Most commonly, iterative algorithms are used to estimate $\phi$ and the most popular is the univariate Newton-Raphson algorithm. The analyst may consider the following articles that discuss the statistical issues concerned with the implementation of this iterative estimation technique: [5 – 13].

We may begin our discussion of the univariate Newton-Raphson iterative procedure by considering the log-likelihood function (based on the observed cell counts being independent Poisson random variables), with parameter $m_{ij}$, as

$$l(m) = \sum_{i=1}^{6} \sum_{j=1}^{5} \left( n_{ij} \ln m_{ij} - m_{ij} - \ln(n_{ij}) \right)$$

where $m$ is the vector of the expected cell frequencies. There are many ways to model these values, and here we are considering

$$m_{ij} = \exp[\mu + \alpha_i + \beta_j + \phi(i - \bar{u})(j - \bar{v})].$$

Therefore, the kernel of this log-likelihood function becomes

$$l(\phi | \mu, \alpha, \beta) = \sum_{i=1}^{6} \sum_{j=1}^{5} \left( n_{ij}(\mu + \alpha_i + \beta_j + \phi(i - \bar{u})(j - \bar{v})) - \exp[\mu + \alpha_i + \beta_j + \phi(i - \bar{u})(j - \bar{v})] \right).$$

Therefore the first and second derivatives of this log-likelihood function, with respect to $\phi$, are

$$\frac{\partial l}{\partial \phi} = \sum_{i=1}^{6} \sum_{j=1}^{5} \left( i - \bar{u} \right) (j - \bar{v}) \left( n_{ij} - m_{ij} \right),$$

$$\frac{\partial^2 l}{\partial \phi^2} = -\sum_{i=1}^{6} \sum_{j=1}^{5} (i - \bar{u})^2 (j - \bar{v})^2 m_{ij}.$$
respectively. The estimate of $\phi$ is then updated iteratively by considering the univariate Newton-Raphson iterative algorithm

$$
\phi^{(t+1)} = \phi^{(t)} - \left( \frac{\partial^2 I}{\partial \phi^2} \right)^{-1} \left( \frac{\partial I}{\partial \phi} \right)_{\phi=\phi^{(t)}}.
$$

To begin the iterative cycle of this algorithm, we start with an initial value of $\phi$ that quantifies the association between the assessor’s classification and the classification based on the approach taken by [1]. We denote this initial value as $\phi_N^{(0)}$. The estimated $i$’th cell value of the confusion matrix can be estimated using this initial value of $\phi$, and (3), by considering

$$
m_{ij}^{(0)} = \exp[u + \alpha_i + \beta_j + \phi_N^{(0)} (i - \bar{u}) (j - \bar{v})].
$$

While we commence the iterative process by defining an initial value of $\phi$ in this paper, one may alternatively do so by considering an initial set of $m_{ij}^{(0)}$ values. For example, one may consider $m_{ij}^{(0)} = n_{p_i p_j}$ which coincides with complete independence between the actual classification and the estimated classification based on the study of [1].

Once the initial values of $m_{ij}$ have been specified, the estimate of $\phi$ is then updated such that

$$
\phi^{(1)}_{N} = \phi^{(0)}_{N} + \sum_{i=1}^{6} \sum_{j=1}^{5} \frac{(u_i - \bar{u})(v_j - \bar{v})(m_{ij}^{(0)} - m_{ij}^{(0)})}{\sum_{i=1}^{6} \sum_{j=1}^{5} (u_i - \bar{u})^2 (v_j - \bar{v})^2 m_{ij}^{(0)}}.
$$

This estimate of $\phi$ can be updated by estimating the $i$’th cell of the confusion matrix ($m_{ij}$) by

$$
m_{ij}^{(1)} = \exp[u + \alpha_i + \beta_j + \phi^{(0)}_{N} (i - \bar{u}) (j - \bar{v})].
$$
and then considering

\[ \phi_N^{(2)} = \phi_N^{(1)} + \frac{\sum_{i=1}^{6} \sum_{j=1}^{5} (u_i - \bar{u})(v_j - \bar{v})(n_{ij} - m_{ij}^{(1)})}{\sum_{i=1}^{6} \sum_{j=1}^{5} (u_i - \bar{u})^2 (v_j - \bar{v})^2 m_{ij}^{(1)}} \]

as the next iterative step in the estimation of \( \phi \). This process is repeated such that the \( t+1 \)'th iterative estimate of \( \phi \) can be found such that

\[ \phi_N^{(t+1)} = \phi_N^{(t)} + \frac{\sum_{i=1}^{6} \sum_{j=1}^{5} (u_i - \bar{u})(v_j - \bar{v})(n_{ij} - m_{ij}^{(t)})}{\sum_{i=1}^{6} \sum_{j=1}^{5} (u_i - \bar{u})^2 (v_j - \bar{v})^2 m_{ij}^{(t)}}. \]

The number of iterations needed before accepting \( \phi_N \) as the Newton-Raphson estimate of \( \phi \) depends on the confusion matrix being analysed and the level of precision required. In the application of this algorithm to the confusion matrix of Table 1, we shall consider a level of accuracy that is to four decimal places. This is to ensure that the estimate we obtain is as accurate as possible.

There will always be a certain amount of variation present in the estimate of \( \phi \). This is because if we select another sample of wines and they are subjected to the same classification procedures as those considered by [1], a different confusion matrix will be obtained and so the analyst will obtain a different estimate of \( \phi \). The variation may be quantified by considering the standard error

\[ \text{SE}(\phi_N) = \frac{1}{\sqrt{\sum_{i=1}^{6} \sum_{j=1}^{5} (u_i - \bar{u})^2 (v_j - \bar{v})^2 m_{ij}}} \]

where \( m_{ij} \) is the estimated value of \( n_{ij} \) based on \( \phi_N \). Ideally a small standard error needs to achieved since this indicates a reliable and precise estimate of \( \phi \).
Finding the estimate of $\phi$ in this way is an example of maximum likelihood estimation and the estimate may also be referred to as the maximum likelihood estimate (MLE) of $\phi$. Since the algorithm described in this section is only concerned with estimating a single parameter $\phi$ (and does so by treating $\mu, \alpha_i$ and $\beta_j$ as nuisance parameters), it is referred to as the univariate Newton-Raphson iterative procedure. More sophisticated multivariate algorithms may also be considered but require far more complex specifications of the algorithm than we can describe here. Such procedures involve finding the MLE of the parameters that we have treated here as being nuisance as well as finding the MLE of our parameter of interest, $\phi$. We shall not consider multivariate algorithms further in this paper.

4.2 A Direct Estimation Method

There are two key problems that can prevent iterative techniques from working effectively. Firstly, since the estimation is iterative it requires the analyst to set initial value of the parameter(s) that require estimating. Therefore the selection of a poor initial value can lead to a poor estimate. Secondly, it sometimes occurs that the algorithm will fail to converge at all, thereby not producing an estimate of the parameter of interest (which is $\phi$ in our case).

To overcome these issues, one may adopt an alternative, and more direct, approach which does not involve an iterative algorithm. One such approach was originally proposed by [14] and elaborated upon by [13]. Given the above definitions, it is sufficient to inform the reader that $\phi$ may be estimated using the original non-iterative estimate proposed by [14]:

$$\hat{\phi} = \phi_{BDNI} = \frac{1}{\sigma_i \sigma_j} \sum_{i=1}^{6} \sum_{j=1}^{5} p_{ij} (u_i - \bar{\mu}) (v_j - \bar{\nu})$$

(4)

where $\sigma_i^2 = \sum_{i=1}^{6} i^2 p_i - \bar{\mu}^2$ and $\sigma_j^2 = \sum_{j=1}^{5} j^2 p_j - \bar{\nu}^2$ is the variance of the row and column classifications, respectively. A more accurate direct estimation approach is to consider the non-iterative estimate of [13]

$$\hat{\phi} = \phi_{logNI} = \frac{1}{\sigma_i \sigma_j} \sum_{i=1}^{6} \sum_{j=1}^{5} p_{i \alpha} p_{j \beta} (u_i - \bar{\mu}) (v_j - \bar{\nu}) \ln \left( \frac{p_{ij}}{p_{i \alpha} p_{j \beta}} \right).$$

(5)
It will always be that (5) will provide a more accurate estimate of \( \phi \) than (4). This is because, in the derivation of the direct estimation methods, (4) is a first order approximation of (5). Therefore (4) will be subject to more variation than (5) and thus be generally a less reliable estimate of \( \phi \). The derivation of these results stems from considering an alternative, but related, modelling approach to that of ordinal log-linear models and involves the calculation of orthogonal polynomials. An in-depth discussion of how (4) and (5) may be derived can be found in [13].

For many analysts, these direct estimation approaches provide advantages over the traditional iterative algorithms since the latter can often require a large amount of computational time and effort. Therefore, in the following section we shall compare the accuracy of the Newton-Raphson algorithm with the direct estimation methods of (4) and (5) for analysing the confusion matrix of Table 1. The variation of the estimate obtained from these direct methods may be measured using the standard error of

\[
\text{SE}(\hat{\phi}) = \frac{1}{\sqrt{\sum_{i=1}^{6} \sum_{j=1}^{5} (u_i - \bar{u})^2 (v_j - \bar{v})^2 m_{ij}}}
\]

where \( m_{ij} \) is the estimated value of \( n_{ij} \) given \( \hat{\phi} \).

### 4.3 The Nuisance Parameters & Zero Cell Frequencies

For both estimation procedures, one may also see that the ordinal log-linear model of (3) requires the estimation of \( \mu, \alpha_i \) and \( \beta_j \) which we consider as nuisance parameters in our analysis of the confusion matrix. These may also be estimated directly by

\[
\tilde{\mu} = \ln(n) + \frac{1}{6} \sum_{i=1}^{6} \ln(p_{i*}) + \frac{1}{5} \sum_{j=1}^{5} \ln(p_{*j}),
\]

\[
\tilde{\alpha}_i = \ln(p_{i*}) - \frac{1}{6} \sum_{i=1}^{6} \ln(p_{i*}), \quad \tilde{\beta}_j = \ln(p_{*j}) - \frac{1}{5} \sum_{j=1}^{5} \ln(p_{*j}).
\]
See [14] for more details on these estimates. Such estimates are very useful if one wishes to estimate the counts of the confusion matrix, but may also be found through other means. Since our discussion is solely on evaluating the accuracy of non-iterative estimates when compared with that obtained using the Newton-Raphson univariate algorithm, we shall not consider this issue further.

The presence of zero cell values in the confusion matrix will lead to problems with evaluating the logarithm in (5); this is because ln(0) does not exist. Therefore, when estimating $\phi$ using this expression, the zero cell values will be replaced with 0.01. Although any small value (close to zero) may be substituted for the zero cell value.

4. THE ANALYSIS OF RED WINE

We shall now consider the analysis of the confusion matrix of Table 1. By implementing the univariate Newton-Raphson iterative procedure for estimating $\phi$ we obtain an estimate of $\phi_N = 1.5622$. This value was obtained after 192 iterations of the algorithm for convergence to be achieved to four decimal places. With a standard error of 0.0382, the 95% confidence interval for $\phi_N$ is $(1.5622 - 1.96 \times 0.0382, 1.5622 + 1.96 \times 0.0382) = (1.4873, 1.6371)$. Such a result suggests that there is a statistically significant positive association between the classification of the assessors and that of the prediction method adopted by [1]. This numerical result implies that the analytical method using data mining techniques produces classifications that agree well with the classifications given by the assessors. A graphical approach to analysing the association between the rows and columns of Table 1 provides a more intuitive view of the nature of this association. For example, a correspondence analysis performed on the confusion matrix, Table 1, yields the plot given by Fig. (1) and is from the results of [2] and shows that prediction method using the chemical and physical characteristics of the red wine compare very well with those of the assessors.

Fig. (1). about here

If the analyst considers (4) and (5), the non-iterative estimates of $\phi$ is $\phi_{\text{HDNI}} = 1.0787$ and $\phi_{\text{LogNI}} = 1.5199$, respectively. The standard error of $\phi$ estimated using (4) is 0.0237, while using (5) the standard error is 0.0121
thus yielding a 95% confidence interval of (1.0322, 1.1252) and (1.4962, 1.5436) for $\phi_{B\text{DNI}}$ and $\phi_{L\text{ogNI}}$, respectively. Like the Newton-Raphson estimate found above, both of these intervals provide very strong evidence that there is a statistically significant positive association between the assessor’s classifications and those obtained using the physiochemical properties of the red wines of Portugal. One interesting property of using the non-iterative estimates is that they yield a much smaller standard error than the iterative counterpart.

Traditionally, estimates obtained using iterative algorithms like the Newton-Raphson procedure are seen as providing the benchmark estimate for models of all kinds. Thus an important issue is to investigate how comparable the non-iterative estimates are with the iteratively obtained estimate. This can be achieved by calculating the Wald statistic for $\phi_{B\text{DNI}}$ and $\phi_{L\text{ogNI}}$ which, for Table 1, has a chi-squared distribution with $(6 – 1)(5 – 1) = 20$ degrees of freedom. Thus, using (4), the estimate $\phi_{B\text{DNI}} = 1.0787$ has a p-value that is less than 0.0001 thereby indicating that it is statistically different than the Newton-Raphson estimate, $\phi_{N} = 1.5622$. Conversely, using (5), the estimate $\phi_{L\text{ogNI}} = 1.5199$ has a p-value for the Wald statistic of 0.9045. Therefore, the estimate $\phi_{L\text{ogNI}}$, obtained directly (ie without resorting to any iterative process) is statistically indistinguishable when compared with the estimate of $\phi_{N}$ obtained using the Newton-Raphson procedure.

**CONCLUSION**

Progresses in the statistical discipline offer a wide range of very different and versatile tools for the analytical chemist. Some of these tools can be quite complex and demand a high level of statistical literacy and computing skill. Other skills, especially those that provide a summary of the information contained in the data, can be easily adapted and enriched. One such area includes the modelling of the association for data summarised in the form of a confusion matrix such as Table 1 or, more generally, contingency tables. Here we have presented a facet of such a branch of statistics that offers a relatively simple analytical technique without the rigour required of those tools that are computationally demanding.

The non-iterative estimation technique applied in this paper has shown that, when analysing the physical and chemical properties that determine the quality of red wine in north-east Portugal, that the classification method of [1] obtains results that are very comparable to those obtained using experienced wine assessors. Such results prove advantageous to the analyst since it allows for a formal, rigorous, and beneficial procedure for analysing...
their data. It also shows that the 11 characteristics considered in the study of [1] are reliable predictors of the quality of wine. Such statistical procedures therefore form a valuable tool for analytical chemists and the wine industry.

This paper has only considered one type of model suitable for analysing the association between the categorical variables of Table 1. Since the row and column categories are the same (since they reflect classification levels for two different classifying procedures) the best confusion matrix is one that is exactly diagonal such that $p_{ij} = 0$ when $i \neq j$. Therefore non-iterative estimation techniques of the kind given by (4) and (5) may be incorporated into models that reflect such symmetry. One may consider, for example, [15] and [16] who describe such models.

Recall that the algorithm used in this paper is the univariate Newton-Raphson method where only a single parameter ($\phi$) is estimated. One may note that the ordinal log-linear model of (3) includes fixed and main effects; while we have treated these as nuisance parameters and proposed closed form solutions when iteratively estimating $\phi$, they may also be considered parameters of the model that require estimating. A vastly more complex adaptation of the univariate Newton-Raphson algorithm may therefore be adopted to estimate these parameters (as well as $\phi$) and will generally yield very good estimates of the cell frequencies when formally modelling the counts in the confusion matrix. However the purpose of this paper has been to assess the accuracy of the closed form estimates of $\phi$ with the popular univariate Newton-Raphson algorithm. Therefore the implementation of a multivariate algorithm, as well incorporating a diagonal restriction for confusion matrices, will be considered in future research.

REFERENCES


Fig. (1). A two-dimensional plot of the classification of Table 1.