Assessing the multi-labelness of multi-label data

Laurence A. F. Park[1]✉, Yi Guo[1], and Jesse Read[2]

[1] Centre for Research in Mathematics
School of Computing, Engineering and Mathematics
Western Sydney University, Australia
{lapark,yi.guo}@westernsydney.edu.au
[2] DaSciM team, LIX Laboratory,
École Polytechnique, 91120 Palaiseau, France.
firstname.lastname@polytechnique.edu

Abstract. Before constructing a classifier, we should examine the data to gain an understanding of the relationships between the variables, to assist with the design of the classifier. Using multi-label data requires us to examine the association between labels: its multi-labelness. We cannot directly measure association between two labels, since the labels' relationships are confounded with the set of observation variables. A better approach is to fit an analytical model to a label with respect to the observations and remaining labels, but this might present false relationships due to the problem of multicollinearity between the observations and labels. In this article, we examine the utility of regularised logistic regression and a new form of split logistic regression for assessing the multi-labelness of data. We find that a split analytical model using regularisation is able to provide fewer label relationships when no relationships exist, or if the labels can be partitioned. We also find that if label relationships do exist, logistic regression with l_1 regularisation provides the better measurement of multi-labelness.

1 Introduction

Multi-label classification models allow the classification of a set of unknown binary labels conditioned on a set of known observations. A review of common multi-label classification algorithms is given in [7].

Before modelling any data, we should examine it to determine an appropriate form of model for the data. When faced with multi-label data, we must also examine the relationships between the labels to determine the *multi-labelness* of the data: if a multi-label model is appropriate and how the labels should be modelled. If we can detect that a given set of labels are independent from each other, we can include this knowledge in the model, making the fitting time faster, resulting in a less complex model.

Unfortunately, measuring high correlation between a pair of label variables does not imply that the multi-labelness of the data is high, since the correlation might be explained by a set of confounding observation variables. Therefore, to determine the set of relationships between labels, we must model each label, with respect to all other labels and all observation variables, and examine the coefficients of the model. The number of non-zero coefficients between labels of the model provide us with a measure of multi-labelness of the data.

When modelling the response of a label, with respect to the remaining labels and the observation variables, we introduce the problem of multicollinearity; there is likely to be correlation between the observation variables and the labels, so it is also likely that many subsets of observation variables and labels provide an equally good fit to the data, but our model will only provide one subset. This implies that even though a label is independent of other labels, the model may show association to other labels due to their multicollinearity with the set of observation variables, and suggest a false high multi-labelness of the data.

In this article, we investigate the use of logistic regression in a full and split form to measure the multi-labelness of the data. The contributions of this article are:

- Derivation of a split analytical model with regularisation (Section 3.2).
- Investigation of the utility of a full and split regularised model for measuring multi-labelness on synthetic data using various multi-label structures (Section 4).
- Verification of the analysis using real data (Section 5)

The article will proceed as follows: Section 2 introduces the problem and required background knowledge, Section 3 introduces measuring multi-labelness with full and split analytical models. Section 4 examines the utility of each analytical model for measuring multi-labelness on generated data. Finally, Section 5 verifies the findings using real data.

2 Background: Multi-label data and multicollinearity

The multi-label classification problem requires modelling L label set variables $\boldsymbol{y} \in \{-1, +1\}^L$ conditioned on a set of M observation variables $\boldsymbol{x} \in \mathbb{R}^M$. Typically, sample data is provided as a set of N label sets and associated observations $(\boldsymbol{y}, \boldsymbol{x})$, where the task is to construct a model f that provides good estimates of the label sets $\hat{\boldsymbol{y}}$ conditioned on the observations, such that $\hat{\boldsymbol{y}} = f(\boldsymbol{x})$, for a given metric [3, 4].

A common technique for modelling multi-label relationships is to construct a set of models that predict only one label variable y_i or a subset of labels, based on the observations and a subset of the remaining labels. The coefficients of the models $\boldsymbol{\beta}$ provide us with insight of the level of association of the label y_i to each observation variable and remaining labels. For example, single label models can be chained [6, 5], use a tree structure [2] or even retain cyclic dependencies in a network [1, 8]. In each of these cases, higher level models predict the state of a label based on the predicted states of other labels. This implies that any error in label classification will be propagated through to other labels. Therefore, when constructing these models, if we can remove model dependencies between labels and maintain accuracy, then we should do so.

Before we fit a multi-label model to data, we should examine the data and determine if there is association between the labels; we call this measuring the *multi-labelness* of the data, where the measurement of multi-labelness is the number of inter-label relationships. Why is it important to examine the multi-labelness of the data?

- If no labels are associated to each other, then the multi-label problem reduces to a set of binary classification problems.
- If there are at least two sets of labels that have no association between them, then we can split the multi-label problems into a two or more multi-label classification problems, each independent of each other.

Also, knowing which labels are correlated will assist us in designing a suitable multi-label classifier.

Problem: Confounding variables When determining the dependence of one label variable y_i to another y_j, we must note that the set of observation variables are confounding variables. Both labels y_i and y_j are dependent on the observations $\boldsymbol{x}$, so any association between the labels might actually be explained entirely by the observations $\boldsymbol{x}$. Therefore, we must take our analysis a step further and model the variance of y_i with respect to each observation $\boldsymbol{x}$ and each other label variable $\boldsymbol{y}_{-i}$. The fitted analytical model coefficients $\boldsymbol{\beta}$ will describe the level of association of y_i to $\boldsymbol{x}$ and $\boldsymbol{y}_{-i}$.

Analytical models are fitted to data to provide us with deeper insight into the generating process behind the data. For example, when using simple linear regression, we can observe the fitted model coefficients $\boldsymbol{\beta}$ to identify how each of the observed variables effects the response variable. For our data we will model a given label y_i with respect to the observations $\boldsymbol{x}$ and the remaining labels $\boldsymbol{y}_{-i}$. The coefficients of the analytical model $\boldsymbol{\beta}$ show which of the elements of $\boldsymbol{x}$ and $\boldsymbol{y}_{-i}$ are associated to y_i. If a coefficient β_i is found to be 0, we then assume that there is no association between the associated covariate and the response.

Problem: Multicollinearity Unfortunately, the correlation between labels, that we use to improve the accuracy of predictions of a multi-label model, cause problems when analysing the coefficients of the analytical models. Multicollinearity occurs when two or more dependent variables are linearly related, and therefore, the analytical model can use different linear combinations of each variable to obtain the same model accuracy. In our case, we have the response label y_i in which we want to determine its relationship to the observed variables $\boldsymbol{x}$ and the remaining labels $\boldsymbol{y}_{-i}$.

$$y_i = f(\boldsymbol{x}, \boldsymbol{y}_{-i}; \boldsymbol{\beta}_i) \tag{1}$$

If we believe that another label y_j is also dependent on $\boldsymbol{x}$, we get the relationships

$$y_i = f(\boldsymbol{x}, y_j, \boldsymbol{y}_{-(ij)}; \boldsymbol{\beta}_i), y_j = f(\boldsymbol{x}; \boldsymbol{\beta}_j) \tag{2}$$

where $\boldsymbol{y}_{-(ij)}$ is the set of labels $\boldsymbol{y}$ excluding the labels y_i and y_j. If the above relationships hold, do we then conclude that the label y_i is dependent on y_j, or do we conclude that it is not dependent on y_j but dependent on $\boldsymbol{x}$, since y_j is dependent on $\boldsymbol{x}$? The former will suggest high multi-labelness, while the latter suggests low multi-labelness. Fitting a model to data containing the above relationships will provide a fit, but will not reveal the additional information that there may be another preferred fit that is just as suitable. In fact by definition, multi-label data must contain multicollinearity between the $\boldsymbol{x}$ and $\boldsymbol{y}$, otherwise we would not be able to obtain accurate label predictions (the set of labels must be associated to the set of observations).

If multicollinearity effects all multi-label models, we must ask how it effects the fit of analytical models and the measurement of multi-labelness, and what we can do to control it. In the next sections, we will investigate the effects of different forms of regularisation on multi-label analytical models

3 Analytical models for measuring multi-labelness

Analytical models can be used to gain insight into the associations between each label, which in turn allows us measure the multi-labelness of the data. But as we showed in the previous section, multi-label data suffers from multicollinearity, therefore, there may be many combinations of observation variables and label variables that can provide a good fit to a given label y_i.

Two well known forms of regularisation may be useful in reducing the effect of multicollinearity; the l_2 and l_1 norm. Analytical models provide a set of coefficients $\boldsymbol{\beta}$ that show how much of the variance of the response is explained by each covariate. Given the choice, we would rather the model to show most of the variance to be explained by the observations $\boldsymbol{x}$, but unfortunately regularisation does not take this into account.

In this section, we present two candidates for measuring the multi-labelness of data: an analytical model with regularisation, and we introduce a split model that models each label using the observations before modelling with respect to the other labels.

3.1 Regularisation of analytical models

Analytical models (as opposed to predictive models) are fit to data to provide insight into the associations between variables. A common form of analytical model for a binary response is logistic regression.

$$\log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{\beta}_i\boldsymbol{x} \tag{3}$$

where $\boldsymbol{x}$ is the vector of observations, $\boldsymbol{\beta}_i$ is the vector of model coefficients and p_i is the probability of the response y_i being positive or negative. Once the model is fit to data, $\boldsymbol{\beta}_i$ is observed to determine which of the elements of $\boldsymbol{x}$ are associated to p_i.

A multi-label analytical model allows us to identify which of the elements of $\boldsymbol{x}$ and labels $\boldsymbol{y}_{-i}$ are associated to label y_i. Using logistic regression, we have:

$$\log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{\beta}_{xi}\boldsymbol{x} + \boldsymbol{\beta}_{yi}\boldsymbol{y}_{-i} \tag{4}$$

where $\boldsymbol{\beta}_{xi}$ are the regression coefficients associated to the observations variables, and $\boldsymbol{\beta}_{yi}$ are the coefficients associated to the set of labels excluding the ith label.

Fitting the model to data consists of identifying the coefficients $\boldsymbol{\beta}_{xi}$ and $\boldsymbol{\beta}_{yi}$ that maximise the likelihood function, or equivalently minimise the negative log likelihood function. Regularisation is used to avoid overfitting the model, by penalising the likelihood, leading to lower model variance, but introducing a bias. Common forms of regularisation are l_1 and l_2 norm regularisation, giving the loss functions:

$$\begin{aligned}
l_1: &\quad \lambda\|[\boldsymbol{\beta}_{xi}\ \boldsymbol{\beta}_{yi}]\|_1 - \mathcal{L}([\boldsymbol{\beta}_{xi}\ \boldsymbol{\beta}_{yi}]; [\boldsymbol{x}\ \boldsymbol{y}_{-i}], y_i) \\
l_2: &\quad \lambda\|[\boldsymbol{\beta}_{xi}\ \boldsymbol{\beta}_{yi}]\|_2 - \mathcal{L}([\boldsymbol{\beta}_{xi}\ \boldsymbol{\beta}_{yi}]; [\boldsymbol{x}\ \boldsymbol{y}_{-i}], y_i) \\
l_1 + l_2: &\quad \lambda(\|[\boldsymbol{\beta}_{xi}\ \boldsymbol{\beta}_{yi}]\|_1 + \|[\boldsymbol{\beta}_{xi}\ \boldsymbol{\beta}_{yi}]\|_2) - \mathcal{L}([\boldsymbol{\beta}_{xi}\ \boldsymbol{\beta}_{yi}]; [\boldsymbol{x}\ \boldsymbol{y}_{-i}], y_i)
\end{aligned}$$

where $\mathcal{L}([\boldsymbol{\beta}_{xi}\ \boldsymbol{\beta}_{yi}]; [\boldsymbol{x}\ \boldsymbol{y}_{-i}], y_i)$ is the log likelihood of the logistic regression model of label y_i with coefficients $\boldsymbol{\beta}_{xi}$ and $\boldsymbol{\beta}_{yi}$, $\|\boldsymbol{\beta}\|_1$ is the l_1 norm of $\boldsymbol{\beta}$, $\|\boldsymbol{\beta}\|_2$ is the l_2 norm, and λ is estimated using cross validation.

The l_2 norm induces bias in the coefficients $\boldsymbol{\beta}$ in an attempt to obtain a more robust set of coefficients that generalise to new data, but in the process, usually provides relationships between all variables. The l_1 norm induces bias in the coefficients to act as a variable selector, but is usually highly unstable when faced with multicollinearity. The combined $l_1 + l_2$ norm usually provides robustness and variable selection [9].

Multicollinearity in the data means that the l_1 regularisation might lead to different non-zero coefficients for a new sample. As a simple example, consider the case where all labels y_i are functions of $\boldsymbol{x}$, independent of the other y_j. Ideally, l_1 regularisation should provide zero to all coefficients of $\boldsymbol{\beta}_{yi}$, but the multicollinearity might lead to non-zero coefficient in $\boldsymbol{\beta}_{yi}$ in place of some from $\boldsymbol{\beta}_{xi}$.

3.2 Split Analytical Model

To measure the multi-labelness of data, a multi-label analytical model is used (such as logistic regression), and the label coefficients of the model are observed. We have stated that multi-label analytical models might provide superfluous inter-label relationships due to the multicollinearity of the multi-label data.

Rather than treating each observation $\boldsymbol{x}_j$ and labels $\boldsymbol{y}_{-ij}$ equally where j is the observation id, we propose that the model should first fit the variables $\boldsymbol{x}_j$ and then fit any residual variance to $\boldsymbol{y}_{-ij}$.

$$\begin{aligned}
\operatorname{logit}(p_{ij}) &= \boldsymbol{\beta}_{xi}\boldsymbol{x}_j + \epsilon_{ij} \\
\epsilon_{ij} &= \boldsymbol{\beta}_{yi}\boldsymbol{y}_{-ij} + \eta_{ij}
\end{aligned}$$

where p_{ij} is the probability of label i given $\boldsymbol{x}_j$, ϵ_{ij} is the residual of the ith label and jth observation after fitting the model using only the observations, and η_{ij} is the residual from fitting the labels to the residual from the previous model. This will force the multi-label model to not provide inter-label dependencies that could be explained with $\boldsymbol{x}_j$, due to the multicollinearity of the variables. We hypothesise that this split analytical model will provide a lower number of non-zero label coefficients $\boldsymbol{\beta}_{yi}$, since the model is forced to find associations between y_i and $\boldsymbol{x}$ first, hence providing a better measurement of multi-labelness.

The model residual cannot be measured from logistic regression in the model space, since the true label value of 0 or 1 is mapped to $-\infty$ or $+\infty$. Therefore we keep the model in the logistic space and instead supply an model offset z_{ij} for each label i and observation j.

$$\text{logit}\,(p_{ij}) = z_{ij} + \boldsymbol{\beta}_{yi}\boldsymbol{y}_{-ij} + \eta_{ij} \text{ where } z_{ij} = \boldsymbol{\beta}_{xi}\boldsymbol{x}_j$$

To generalise the model, we also add regularisation to each stage of the model, providing us with the fitting process:

$$\begin{aligned} &\text{Stage 1: } \underset{\boldsymbol{\beta}_x}{\arg\min}\left(\lambda_x \|\boldsymbol{\beta}_x\|_m - \mathcal{L}(\boldsymbol{\beta}_x|\boldsymbol{x}, y_i)\right) \\ &\text{Stage 2: } \underset{\boldsymbol{\beta}_y}{\arg\min}\left(\lambda_y \|\boldsymbol{\beta}_y\|_n - \mathcal{L}(\boldsymbol{\beta}_y|\boldsymbol{y}_{-i}, y_i, z_i)\right) \end{aligned} \tag{5}$$

where m and n are either 1 or 2 (for l_1 or l_2 regularisation), $\mathcal{L}(\boldsymbol{\beta}_y|\boldsymbol{y}_{-i}, y_i, z_i)$ is the logistic regression log likelihood function using offset z_i for each observation, and λ_x and λ_y are estimated using cross validation.

4 Analysis of Full and Split Analytical Models

A full and split analytical model were presented in the previous section. In this section we devise an experiment to deduce where each form of model is most effective at measuring multi-labelness (identifying the number of inter-label relationships) of multi-label data, with minimum superfluous relationships.

4.1 Measuring multi-labelness

Recall that multi-labelness is the number of inter-label relationships. The usual procedure for determining if an observation variable is associated to the response variable is to examine the standard error of the fitted regression coefficient, and assess if it provides evidence against the true coefficient being zero. Unfortunately, regularisation induces bias into the regression coefficient estimates $\boldsymbol{\beta}$, therefore the coefficient standard error is not as useful[3]. To determine significance we must compute the confidence interval for each coefficient. But, reporting

[3] Section 6 of `https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf`

the interval of each coefficient would ignore any association between coefficients (treating them as independent), and be misleading.

We want to determine which form of regularisation provides an analytical model with a good fit to the data, and showing the least dependence between each label to the regression label. We are not concerned with the association between the observations $\boldsymbol{x}$ and the response label; our goal is to determine which form of regularisation provides the least label interdependence.

We are also not concerned with which labels are associated to the response. We stated that we cannot show causality, and that multicollinearity exists, so are unable to determine the true associations. The best we can do is to assess which form of regression provides the least inter-dependence between labels, while providing a good fit. Therefore, we measure multi-labelness as the *number of non-zero label coefficients* of the analytical model. The lower the number, the fewer relationships have been established between labels (meaning that more association has been found to the observations $\boldsymbol{x}$) leading to lower complexity models.

4.2 Generating multi-label data

To assess the measurement of multi-labelness of different data structures from a given analytical model, we need to know the true underlying model form that generated the data. Unfortunately, we do not know the underlying model that was used to generate existing real multi-label data sets, therefore, if using them, we will not be able to determine what about the data is effecting the fit.

Therefore, we are required to simulate multi-label data using strategically designed multi-label data models, where the models exhibit the characteristics that we want to test, but remain simple in order to reduce the chance of other effects being introduced. We present ten multi-label data models, all using three observation variables and three label variables. Each data model consists of three observation variables x_1, x_2, $x_3 \in [-1, 1]$ which are independent, and all Bernoulli, with $p = 0.5$, and three response labels y_1, y_2, $y_3 \in [-1, 1]$. The models differ in the dependence of the response labels, and use the Model Factor a associated to the likelihood of the model (larger a leads to data that provides higher likelihood). The ten models are:

OneXOneY $\operatorname{logit}(y_i) = ax_i \forall i$. All y_i depend on only one x_i. A multi-label model should find no interdependence between the labels. The non-zero label coefficient count should be 0.

ManyXOneY $\operatorname{logit}(y_i) = ax_1/3 + ax_2/3 + ax_3/3 \forall i$. All y_i depend on every x_i. A multi-label model should find no interdependence between the labels. The non-zero label coefficient count should be 0.

OneXChainY $\operatorname{logit}(y_1) = ax_1$, $\operatorname{logit}(y_2) = ay_1$, $\operatorname{logit}(y_3) = ay_2$. The first label y_1 depends on one observation variable x_1, the second label depends on the first and the third depends on the second. The non-zero label coefficient count should be 2.

ManyXChainY $\operatorname{logit}(y_1) = ax_1/3+ax_2/3+ax_3/3$, $\operatorname{logit}(y_2) = ay_1$, $\operatorname{logit}(y_3) = ay_2$. The same as OneXChainY, but the first label depends on all observed variables x_i. The non-zero label coefficient count should be 2.

OneXPartitionY $\operatorname{logit}(y_1) = ax_1$, $\operatorname{logit}(y_2) = ay_1$, $\operatorname{logit}(y_3) = ax_1$ The first and third labels depend on an observed variable x_1 and the second label depends on the first label. The non-zero label coefficient count should be 1.

ManyXPartitionY $\operatorname{logit}(y_1) = ax_1/3+ax_2/3+ax_3/3$, $\operatorname{logit}(y_2) = ay_1$, $\operatorname{logit}(y_3) = ax_1/3+ax_2/3+ax_3/3$. The same as OneXPartitionY, but the first and third labels depend on all observed variables x_i. The non-zero label coefficient count should be 1.

OneXTreeY $\operatorname{logit}(y_1) = ax_1$, $\operatorname{logit}(y_2) = ay_1$, $\operatorname{logit}(y_3) = ay_1$. The first label depends on an observed variable, the second and third labels depend on the first label. The non-zero label coefficient count should be 2.

ManyXTreeY $\operatorname{logit}(y_1) = ax_1/3+ax_2/3+ax_3/3$, $\operatorname{logit}(y_2) = ay_1$, $\operatorname{logit}(y_3) = ay_1$. The same as OneXTreeY, but the first label depends on all observed variables. The non-zero label coefficient count should be 2.

OneXFanY $\operatorname{logit}(y_1) = ax_1$, $\operatorname{logit}(y_2) = ay_1$, $\operatorname{logit}(y_3) = ay_1/2 + ay_2/2$. The first label depends on an observed variable, the second label depend on the first label, and the third label depends on the second and first labels. The non-zero label coefficient count should be 3.

ManyXFanY $\operatorname{logit}(y_1) = ax_1/3 + ax_2/3 + ax_3/3$, $\operatorname{logit}(y_2) = ay_1$, $\operatorname{logit}(y_3) = ay_1/2 + ay_2/2$. The same as OneXFanY, but the first label depends on all observed variables.

The dependencies of each model are shown in Figure 1. These 10 data models contain a set of simple relationships that we would expect to find in multi-label data. We will generate data using these known models and examine how the analytical models measure their multi-labelness.

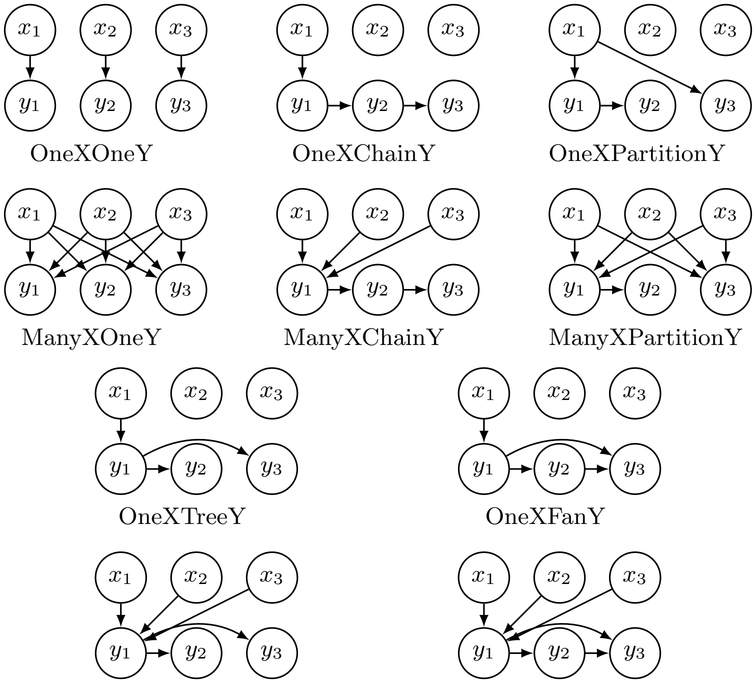
For this investigation, we generated 100 training and 100 testing data sets for each of the above ten data types using Model Factors $a = 0.1, 0.5, 1$ and 2, providing 4000 training and 4000 testing sets. The data was generated by sampling from the models using the model probabilities. We can see that as a increases, the probability of each label is likely to increase in magnitude, so the resulting data sample will have less variance.

In the following sections we will fit this generated data using logistic regression with regularisation, and examine how the regularisation effects the coefficients of the model.

4.3 Investigation: Full model with l_1 and l_2 regularisation

To begin our investigation, we examine the effect of l_1 (lasso regularisation), l_2 (ridge regularisation) and a combination of l_1 and l_2 (elastic net regularisation) regularisation.

To examine how well each analytical model is able to fit the data from each data model, we fitted the analytical model to the generated data and counted the number of non-zero coefficients associated to labels, from the fit. Since there



OneXOneY OneXChainY OneXPartitionY

ManyXOneY ManyXChainY ManyXPartitionY

OneXTreeY OneXFanY

ManyXTreeY ManyXFanY

Fig. 1. Dependencies of the 10 data models used for simulation.

are three labels in each data set, each can have association to zero, one or two other labels, providing a range of zero to six non-zero coefficients for the three fitted labels. The results are shown in Figure 2.

Figure 2 provides one plot for each data model type. Each plot contains sets of box plots for model factors 0.1, 0.5, 1 and 2, and each set of box plots contains three box plots of the non-zero label coefficient count for the three forms of regularisation over the 100 replications. As expected, l_2 regularisation provides all six of the label coefficients as non-zero. We can also see that l_1 provides either an equivalent or fewer number of non-zero label coefficients compared to $l_1 + l_2$. But we also find that each of these forms of regularisation provide non-zero label coefficients for the OneXOneY and ManyXOneY data, in which there is no interdependence on the labels. Therefore, using l_1, l_2 or a mixture will suggest that label dependencies exist, when in fact they do not.

4.4 Investigation: Split model with l_1 and l_2 regularisation

In this section, we will examine the effect of the two stage analytical model from Section 3.2 to try to force the dependence of y_i towards the observations $\boldsymbol{x}$, and then fit the remaining label variance to the labels $\boldsymbol{y}_{-i}$. The two stages of the split model from equation 5 require two norms to be set. We used the data from

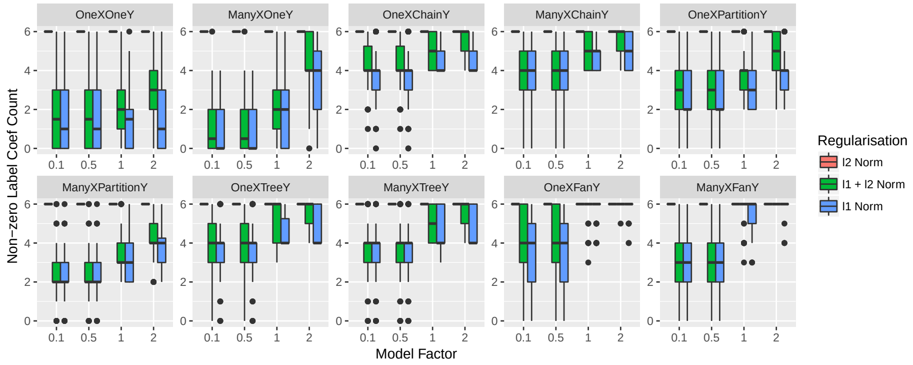


Fig. 2. The distribution of the number of non-zero label coefficients for l_1, $l_1 + l_2$ and l_2 regularisation, on each data type, using Model Factors 0.1, 0.5, 1, and 2.

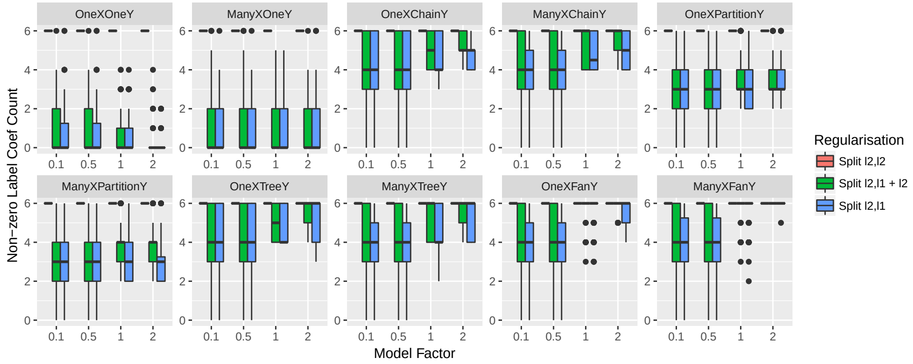


Fig. 3. The distribution of the number of non-zero label coefficients for split regularisation (l_2 for observed variables and l_1, $l_1 + l_2$ and l_2 for labels), on each data type, using Model Factors 0.1, 0.5, 1, and 2.

Section 4.2 to obtain results when each of m and n are 1 or 2. The results are shown in Figures 3, 4 and 5.

Figure 3 provides box plots for the number of non-zero label coefficients using l_2 regularisation for the $\boldsymbol{x}$ variables and a selection of l_2, l_1 and $l_1 + l_2$ regularisation for the labels $\boldsymbol{y_i}$. We find again that l_2 always provides 6 non-zero coefficients, and that l_1 provides either equivalent or fewer labels than $l_1 + l_2$. We also find that the median non-zero label coefficient count is 0 for l_1 and $l_1 + l_2$ regularisation for the OneXOneY and ManyXOneY data structures, showing that the split regularisation has had an impact in removing non-existent label inter-dependencies.

Figures 4 and 5 provide the non-zero label coefficient count when using the $l_1 + l_2$ regularisation for $\boldsymbol{x}$ and l_1 regularisation for $\boldsymbol{x}$ respectively. These results

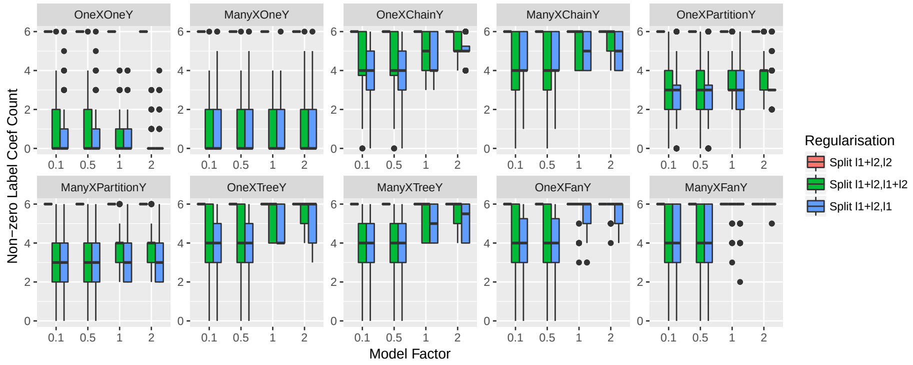


Fig. 4. The distribution of the number of non-zero label coefficients for split regularisation ($l+1+l_2$ for observed variables and l_1, l_1+l_2 and l_2 for labels), on each data type, using Model Factors 0.1, 0.5, 1, and 2.

lead us to the same conclusion, that l_1 regularisation for the labels leads to lower non-zero label coefficient counts in the analytical models.

4.5 Comparing Full and Split regression

We have provided the non-zero label coefficient count and prediction error results from using l_1 regularisation on all coefficients (Full l_1, from Section 4.3) and the split model results using l_1, l_1+l_2 and l_2 regularisation for $\boldsymbol{x}$ and l_1 regularisation for $\boldsymbol{y}_{-i}$ in Figures 6 and 7 (comparing the best forms of regularisation from the previous results).

Figure 6 shows that the split models provide a lower distribution (shifted towards zero) of non-zero label coefficients compared to the Full model for OneXOneY, ManyXManyY, and for OneXPartitionY, ManyXPartitionY when the Model Factor (a) is high. For all other data models, the Full l_1 model provides an equivalent or fewer non-zero label coefficients.

The accuracy results in Figure 7 provides the mean absolute error between the predicted label probability and the true label probability (from the model). We find that each analytical model provides equivalent accuracy, but there are a few occurrences (from the Fan, Tree and Chain data structures) of the Full l_1 model providing lower error when the Model factor (a) is 1 or 2.

The mean number of non-zero label coefficients for each regularisation method on each data type in Figure 6 are shown in Table 1. We find that the split regularisation provided significantly fewer non-zero label coefficients for the data where there was no inter-label dependencies (OneXOneY and ManyXOneY) and the partitioned labels (OneXPartY and ManyXPartY), but provided more non-zero coefficients for the Chain, Tree and Fan data. This suggests that the Split models are useful when no label relationships exists, otherwise the Full model should be used for measuring multi-labelness.

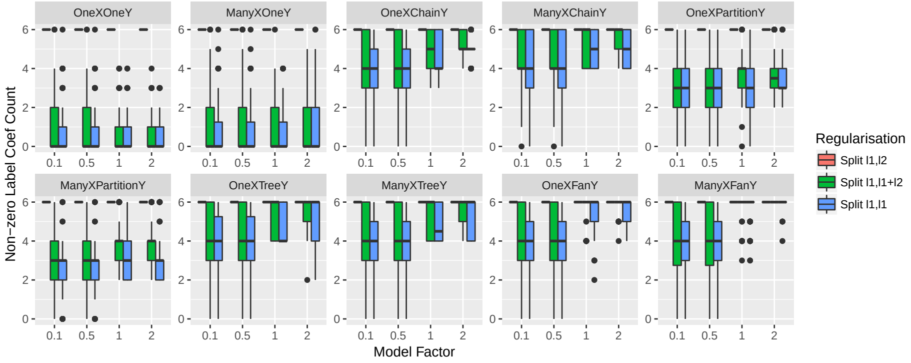


Fig. 5. The distribution of the number of non-zero label coefficients for split regularisation (l_1 for observed variables and l_1, $l_1 + l_2$ and l_2 for labels), on each data type, using Model Factors 0.1, 0.5, 1, and 2.

5 Full and Split analytical models on real data

In this section, we examine the effect of regularisation on the non-zero label coefficient proportion from five commonly used multi-label data sets. We assume that there is dependence amongst the labels in each of the data sets, due to their use in multi-label classification research.

We use the Emotions, Stare, Scene, Slashdot, and Enron multi-label data sets[4]. A 50/50 train/test split is used, and the regularisation parameters λ, λ_x and λ_y were fit using 10 fold cross validation on the training data. The models are then used to examine the effect of regularisation and the classification accuracy using the testing set. The mean non-zero label coefficient proportion for each label is reported, representing the detected number of relationships between labels. The accuracy is measured in terms of mean Hamming similarity (proportion of correctly predicted labels), Jaccard similarity (ratio of true positive count and 1 - true negative count) and Exact similarity (score 1 if all labels are correct, otherwise score 0) of the predicted label state compared to the true label state, computed over the set of test observations. Note that each model is a function of $\boldsymbol{x}$ and $\boldsymbol{y}_{-i}$, therefore, we have provided two accuracy scores for each regularisation method for each metric, providing an evaluation interval. The lower score is computed using label estimates $\hat{\boldsymbol{y}}_{-i}$ from an independent model (predicting each label based on the observations $\boldsymbol{x}$ alone, not using other label information). The upper score is computed using the true label values $\boldsymbol{y}_{-i}$. The results are presented in Table 2.

[4] All available from `http://mulan.sourceforge.net/datasets-mlc.html`, `https://sourceforge.net/projects/meka/files/Datasets/` (Slashdot), and `http://cecas.clemson.edu/~ahoover/stare/` (Stare).

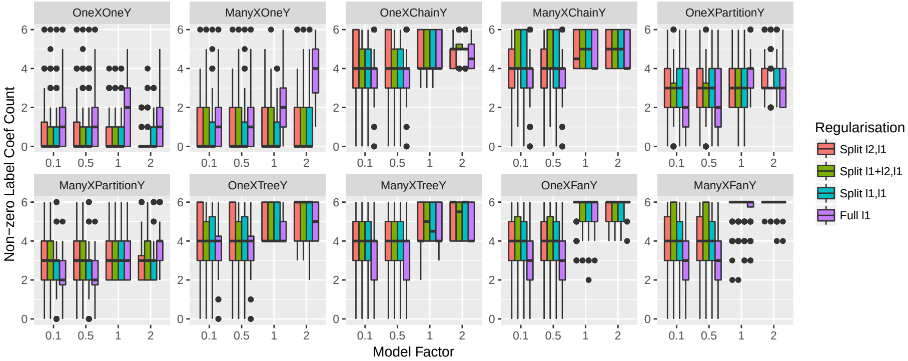


Fig. 6. The distribution of the number of non-zero label coefficients for split regularisation (l_1, $l_1 + l_2$ and l_2 for observed variables and l_1 for labels) and full l_1 regularisation, on each data type, using Model Factors 0.1, 0.5, 1, and 2.

5.1 Label interdependence

We first examine the non-zero label coefficient proportion in Table 2 to determine which form of analytical model provides the most appropriate measure of multi-labelness. We find that using Full l_1 regularisation provides the lowest proportion over all but the Emotions data set, where it is close to the minimum. This is consistent with the simulated results, assuming that each of the multi-label data sets have no independent sets of labels. Calculating the maximum likelihood score for each data set also reveals that they all are most similar to the generated data where $a = 0.1$, further reinforcing the results from the generated data.

5.2 Effect of label-interdependence reduction on accuracy

We next assess the accuracy of the model to determine if the smaller number of label relationships is due to the model making better use of the observations $\boldsymbol{x}$ (meaning that the number of non-zero label coefficients is smaller, but the accuracy is not lower), or it is simply due to a poorer use of the label set $\boldsymbol{y}_{-i}$ (meaning that the number of non-zero label coefficients is smaller, and the accuracy is lower).

The accuracy of each regularised analytical model is provided as two scores; the first when using estimates of the label state (from an independent model), and the second when using the true labels. The second score provides us with a measure of the accuracy of the analytical model, the first score gives us an indication of the effect when using computed label relationships.

Most of the split models provide significantly greater upper scores. Examining the lower scores, we find that many of the Split score are significantly worse than the Full l_1 score. This suggests that the Full l_1 provides a good base set of label relationships and that further relationships can be found using

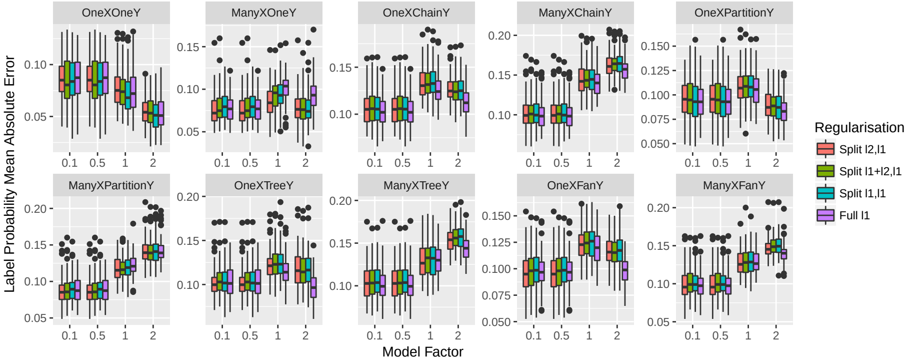


Fig. 7. The distribution of mean absolute error of the label probability and predicted probability for split regularisation (l_1, $l_1 + l_2$ and l_2 for observed variables and l_1 for labels) and full l_1 regularisation, on each data type, using Model Factors 0.1, 0.5, 1, and 2.

Split regularisation, but they are only useful when obtaining accurate label estimates. These results align with those from the simulation; where the labels are associated, Full l_1 provided the least number of non-zero label coefficients with equivalent accuracy to the other forms of regularisation.

These experiments conducted on both the generated and real data lead to the same conclusion, that a split analytical model provides a better measure of multi-labelness when the labels are all independent, or when they can be partitioned into models with high likelihood. Otherwise, using the full model with l_1 regularisation provides a better measure of multi-labelness. Analysing the results has shown that the regularisation has a major impact for the split model; it is shared for the observations and labels in the full model, but not for the split model. We will investigate this impact in future work.

6 Conclusion

Examining the relationships between labels in multi-label data before constructing a multi-label classifier, provides us with insight as to how to design the classifier. Measuring the multi-labelness of the data (the number of relationships between labels) allows us to determine if a multi-label classifier is appropriate for the data.

Multi-labelness of data cannot simply be measured using the correlation between labels, since the label relationships are confounded by the data observations. Fitting an analytical model to a label with respect to the other labels and observations can also present false label relationships due to multicollinearity between the labels and observations.

We investigated the effect of using a full model and proposed a new split analytical model to minimise the number of spurious relationships and measure

Table 1. The mean number of non-zero label coefficients for each regularisation methods and each data type. A star (*) represents a significant difference to the Full l_1 regularisation using a paired Wilcoxon test.

Reg	OneXOneY	ManyXOneY	OneXChainY	ManyXChainY	OneXPartY
Full l_1	1.55	1.84	4.24	4.31	2.98
Split l_2,l_1	0.47*	0.82*	4.59*	4.55*	2.81*
Split $l_1 + l_2$,l_1	0.62*	0.80*	4.53*	4.51*	2.86*
Split l_1,l_1	0.59*	0.76*	4.55*	4.51*	2.77*

Reg	ManyXPartY	OneXTreeY	ManyXTreeY	OneXFanY	ManyXFanY
Full l_1	3.01	4.26	4.06	4.74	4.47
Split l_2, l_1	2.66*	4.62*	4.51*	4.78	4.67*
Split $l_1 + l_2$,l_1	2.69*	4.62*	4.53*	4.84*	4.74*
Split l_1,l_1	2.72*	4.62*	4.42*	4.83	4.67*

the multi-labelness of data. We examined l_1, l_2, and combined l_1 and l_2 regularisation with each of the full and split models. It was found that split analytical models using regularisation have a greater likelihood of detecting independence of labels. But if labels are not independent from each other, a full model using l_1 regularisation provides the fewest dependencies between labels making it more suitable for measuring the multi-labelness of data.

References

1. Yuhong Guo and Suicheng Gu. Multi-label classification using conditional dependency networks. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1300, 2011.
2. Aljaž Osojnik, Panče Panov, and Sašo Džeroski. Multi-label classification via multi-target regression on data streams. *Machine Learning*, 106(6):745–770, Jun 2017.
3. Laurence A. F. Park and Jesse Read. A blended metric for multi-label optimisation and evaluation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 719–734. Springer, 2018.
4. Laurence AF Park and Simeon Simoff. Using entropy as a measure of acceptance for multi-label classification. In *International Symposium on Intelligent Data Analysis*, pages 217–228. Springer, 2015.
5. Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85(3):333, 2011.
6. L Enrique Sucar, Concha Bielza, Eduardo F Morales, Pablo Hernandez-Leal, Julio H Zaragoza, and Pedro Larrañaga. Multi-label classification with bayesian network-based chain classifiers. *Pattern Recognition Letters*, 41:14–22, 2014.
7. M. L. Zhang and Z. H. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, Aug 2014.
8. Min-Ling Zhang and Kun Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 999–1008. ACM, 2010.

9. Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Table 2. Measurement of Hamming, Jaccard and Exact accuracy, and the average number of non-zero label coefficients for five commonly used multi-label data sets. Each cell contains the score when using label estimates $\hat{y}_i$ for prediction, and the true labels y_i for prediction. An asterisk (*) shows a significant difference of each Split method compared to Full l_1.

	Accuracy			Non-zero
	Hamming	Jaccard	Exact	Coefficients
Emotions (5 labels)				
Full l2	0.765, 0.822	0.508, 0.560	0.251, 0.309	1
Full l1+l2	0.762, 0.828	0.533, 0.598	0.242, 0.358	0.866
Full l1	0.746, 0.835	0.527, 0.622	0.222, 0.373	0.832
Split l1,l1	0.764*, 0.842	0.551*, 0.611	0.248, 0.349	0.8
Split l2,l1	0.748, 0.834	0.548*, 0.600	0.227, 0.344	0.832
Split l1+l2,l1	0.757, 0.838	0.551*, 0.608	0.239, 0.358	0.832
Stare (12 labels)				
Full l2	0.920, 0.951	0.539, 0.596	0.410, 0.537	1
Full l1+l2	0.907, 0.947	0.509, 0.577	0.364, 0.520	0.538
Full l1	0.896, 0.945	0.495, 0.600	0.335, 0.531	0.461
Split l1,l1	0.864*, 0.954*	0.488*, 0.679*	0.335, 0.618*	0.872
Split l2,l1	0.860, 0.955*	0.471*, 0.676*	0.324, 0.624*	0.891
Split l1+l2,l1	0.874, 0.956*	0.497*, 0.684*	0.341, 0.635*	0.872
Scene (5 labels)				
Full l2	0.894, 0.947	0.684, 0.789	0.579, 0.740	1
Full l1+l2	0.834, 0.971	0.647, 0.904	0.528, 0.869	1
Full l1	0.774, 0.976	0.616, 0.927	0.522, 0.903	1
Split l1,l1	0.749*, 0.975	0.613*, 0.926	0.523, 0.892*	1
Split l2,l1	0.746*, 0.976	0.611*, 0.928	0.523, 0.899	1
Split l1+l2,l1	0.745*, 0.976	0.613*, 0.927	0.525, 0.897	1
Slashdot (18 labels)				
Full l2	0.956, 0.957	0.407, 0.403	0.367, 0.369	1
Full l1+l2	0.947, 0.963	0.490, 0.523	0.397, 0.472	0.462
Full l1	0.904, 0.969	0.463, 0.623	0.331, 0.570	0.424
Split l1,l1	0.857*, 0.976*	0.454, 0.736*	0.323*, 0.676*	0.886
Split l2,l1	0.907, 0.972*	0.483*, 0.652*	0.333, 0.612*	0.801
Split l1+l2,l1	0.835*, 0.976*	0.443*, 0.741*	0.324, 0.682*	0.848
Enron (47 labels)				
Full l2	0.945, 0.948	0.200, 0.229	0.001, 0.001	1
Full l1+l2	0.937, 0.957	0.318, 0.483	0.117, 0.195	0.210
Full l1	0.929, 0.957	0.275, 0.482	0.013, 0.198	0.153
Split l1,l1	0.915*, 0.956	0.243*, 0.491	0.004, 0.200	0.315
Split l2,l1	0.898*, 0.963*	0.251*, 0.590*	0.001*, 0.289*	0.326
Split l1+l2,l1	0.917*, 0.957	0.247*, 0.502*	0.003*, 0.203	0.318