The Classifier Chain Generalized Maximum Entropy Model for Multi-label Choice Problems

Supanika Leurcharusmee, Jirakom Sirisrisakulchai, Songsak Sriboonchitta and Thierry Denœux

Abstract Multi-label classification can be applied to study empirically discrete choice problems, in which each individual chooses more than one alternative. We applied the Classifier Chain (CC) method to transform the Generalized Maximum Entropy (GME) choice model from a single-label model to a multi-label model. The contribution of our CC-GME model lies in the advantages of both the GME and CC models. Specifically, the GME model can not only predict each individual's choice, but also robustly estimate model parameters that describe factors determining his or her choices. The CC model is a problem transformation method that allows the decision on each alternative to be correlated. We used Monte-Carlo simulations and occupational hazard data to compare the CC-GME model with other selected methodologies for multi-label problems using the Hamming Loss, Accuracy, Precision and Recall measures. The results confirm the robustness of GME estimates with respect to relevant parameters regardless of the true error distributions. Moreover, the CC method outperforms other methods, indicating that the incorporation of the information on dependence patterns among alternatives can improve prediction performance.

1 Introduction

The *discrete choice problem* describes how an individual chooses an alternative from $M \geq 2$ available ones. Empirically, the problem is similar to the *single-label classification problem*, in which objects are classified into M classes. However, in many situations, we observe an individual choosing more than one alternative simultaneously. This problem is then empirically equivalent to the *multi-label classification* problem, in which one object can be associated with a subset of classes. In this study, we extend existing single-label choice models to multi-label choice models.

S. Leurcharusmee (✉) · J. Sirisrisakulchai · S. Sriboonchitta
Faculty of Economics, Chiang Mai University, Chiang Mai, Thailand
e-mail: supanika.econ.cmu@gmail.com

T. Denœux
Université de Technologie de Compiègne CNRS, UNR 7253 Heudiasyc, cedex, France

V.-N. Huynh et al. (eds.), *Econometrics of Risk*, Studies in Computational Intelligence 583, DOI 10.1007/978-3-319-13449-9_13

Since the development of the random utility model, which explains individuals' decision making process, the parameters in the empirical models can be linked to those in the utility functions. The knowledge of the model parameters can contribute to behavior explanation and policy implications. Consequently, the objectives of our multi-label choice model are not only to predict a set of alternatives that each individual chooses, but also to estimate the model parameters that describe factors determining each individual's decisions.

Common methods to study discrete choice problems are the Logit and Probit models [16]. These models are limited in the sense that, being likelihood-based, they require distributional assumptions for the errors. [4, 15] introduced the Maximum Entropy (ME) model for discrete choice problems. [5] added the error components to the model and extended it to the Generalized Maximum Entropy (GME) model for multinomial choice problem to improve efficiency. The traditional discrete choice models are for single-label classification. There are a few Logit and Probit models that were developed to explain the multi-label choice problem in which each individual purchases a bundle of products. As discussed in [2], commonly used models are the Label Powerset model with multinomial Logit and Probit estimation and the multivariate Probit or Logit models [1–3]. Although both models allow each individual to choose more than one alternatives, none of them can cope with large choice sets.

Existing methodologies to analyze the multi-label classification problem in computer science follow two main approaches, referred to as *problem transformation* and *algorithm adaptation* [17]. The strategy of the former approach is to transform the multi-label problem into single-label one in order to apply traditional classification methods. Problem transformation methods include Binary Relevance, Label Powerset, Random k-labelsets, Classifier Chains, Pruned Sets, Ensemble of Classifier Chains and Ensemble of Pruned Sets [7, 13]. The algorithm adaptation approach, in contrast, tackles the multi-label problem directly. Algorithm adaptation methods include Multi-label k-Nearest Neighbors, Back-Propagation Multi-label Learning and Decision Trees [7, 13].

Since the objective of this study is to extend the single-label choice model to multi-label choice, we focus on the problem transformation approach. As discussed in [7], the problem transformation approach is generally simpler, but it has a disadvantage of not incorporating the dependence among alternatives. However, this is not true for the Classifier Chain (CC) method, which can capture the dependence pattern among alternatives. Since the choices that each individual makes are usually correlated, this study focuses on the CC method. As for the base single-label choice model, the Logit, Probit and GME models were all developed with a main objective to estimate model parameters that describe factors determining each individual's decisions. However, the GME estimates are robust to distributional assumptions. In addition, the GME model can generally estimate under-determined problems most efficiently. In other words, the GME method yields the estimated parameters with the smallest possible variances [6]. Therefore, to robustly estimate all relevant parameters and capture the dependence pattern among alternatives, we propose the CC-GME model.

For the experimental part of this study, we used Monte-Carlo simulations to compare the CC method with the Binary Relevance (BR) and Label Powerset (LP)

methods and we compare the GME method with the Logit and Probit methods. Specifically, we empirically assessed the performances CC-GME model against CC-Logit, CC-Probit, BR-GME, BR-Logit, BR-Probit, LP-GME, LP-Logit and LP-Probit models. To test the robustness of the estimations, we applied all the methods to three simulated datasets with normal, logistic and uniform errors. Moreover, we also applied all the methods to a real dataset to explain factors determining the set of occupational hazards that each individual faces. Performance measures used in this study include Hamming Loss, Accuracy, Precision and Recall [7, 13, 18]. The results show that the forecasting performances are more sensitive to the choice of the problem transformation method than to the choice of single-label estimation methods. That is, the CC model outperformed the BR and LP models with respect to all evaluation measures expect the Precision. For the parameter estimates, the GME based methods yielded smaller Mean Squared Error (MSE) than the Logit and Probit base methods.

This paper is organized as follows. The original GME model for single-label choice model is recalled in Sect. 2 and the multi-label CC-GME model is introduced in Sect. 3. In Sect. 4, the CC-GME model is evaluated using Monte-Carlo simulations. Section 5 provides an empirical example using occupational hazard data. Finally, Sect. 6 presents our conclusions and remarks.

2 The Single-Label GME Model

The concept of entropy was introduced by [14] to measure the uncertainty of a set of events. The *Shannon entropy function* is $H(p) = \sum_j p_j \log(p_j)$, where p_j is the probability of observing outcome j [8, 9]. Proposed the Maximum Entropy (ME) principle, stating that the probability distribution that best represents the data or available information is the one with the largest entropy. From the ME principle, [4, 15] developed the ME model for discrete choice problems [5]. Added error components to the model and extended it to the GME model for discrete choice problems.

Consider a problem in which each of N individuals chooses his or her most preferred choice from M alternatives. From the data, we observe dummy variables y_{ij} which equal 1 if individual i chooses alternative j and 0 otherwise. Moreover, we observe K characteristics of each individual x_{ik}, where $k = 1, \ldots, K$. The objective of the GME multinomial choice model is to predict $p_{ij} = Pr\{y_{ij} = 1|x_{ik}\}$ for all i and j, which is the probability of individual i choosing each alternative j given the set of his or her characteristics x_{ik}. That is, we want to recover p_{ij} from the observed data y_{ij} and x_{ik}.

In the GME model, the observed data y_{ij} is assumed to be decomposed into the signal component p_{ij} and error component e_{ij},

$$y_{ij} = p_{ij} + e_{ij}. \tag{1}$$

The error component is supposed to be the expected value of a discrete random variable with support $\{v_h\}$ and probabilities $\{w_{ijh}\}$: $e_{ij} = \sum_h v_h w_{ijh}$. Following [11], the error support is constructed using the three sigma rule, which states that the error support should be symmetric around zero and the bounds should be $-3\sigma_y$ and $3\sigma_y$ where σ_y is the empirical standard deviation of the dependence variable. The number of values for the error is usually fixed at 3 or 5. That is, the error support is usually set to $\{-3\sigma_y, 0, 3\sigma_y\}$ or $\{-3\sigma_y, -1.5\sigma_y, 0, 1.5\sigma_y, 3\sigma_y\}$ [6, 11]. Premultiplying (1) with x_{ik} and summing across i, we have MK stochastic moment constraints,

$$\sum_i x_{ik} y_{ij} = \sum_i x_{ik} p_{ij} + \sum_{ih} x_{ik} v_h w_{ijh}, \quad \forall j = 1, \ldots, M, \forall k = 1, \ldots, K. \tag{2}$$

From the principle of ME, p_{ij} that best represents the data must maximize the entropy function

$$\max_{p,w} H(p_{ij}, w_{ijh}) = -\sum_{ij} p_{ij} \log(p_{ij}) - \sum_{ijh} w_{ijh} \log(w_{ijh}) \tag{3}$$

subject to constraints (2) and the following normalization constraints

$$\sum_j p_{ij} = 1, \quad \forall i = 1, \ldots, N \tag{4}$$

$$\sum_h w_{ijh} = 1, \quad \forall i = 1, \ldots, N, \forall j = 1, \ldots, M. \tag{5}$$

This maximization problem can be solved using the Lagrangian method. It should be noted that we can estimate p_{ij} and w_{ijh} without making any functional form or distributional assumptions. However, to analyze marginal effects of each characteristic x_{ik} on p_{ij}, let us assume that

$$y_{ij} = p_{ij} + e_{ij} = G(x_i \beta_j) + e_{ij} \tag{6}$$

for some function G and coefficients β_j. Unlike the Logit or Probit-based models, the GME model only makes the linear assumption on $x_i \beta_j$, but it does not need to make assumption on function G. However, [5] show that the estimated Lagrange multiplier for each stochastic moment constraint λ_j is equal to $-\beta_j$ and the marginal effect can be calculated using the information from the λ_j.

3 The Multi-label CC-GME Model

Let Ω be a choice set that contains M alternatives. Let us observe a set of dummy variables y_{ij} where $y_{ij} = 1$ when individual i chooses alternative j. For the multi-label model, each individual may choose more than one alternative. In other words, it is

possible that $y_{ij} = 1$ for more than one j. Therefore, there are at most 2^M possible outcomes.

3.1 The CC Model

The multi-label CC model was introduced by [12]. The objective of the multi-label choice model is to estimate $Pr\{\underset{\sim}{y}_i = A|x_{ik}\}$ where $\underset{\sim}{y}_i$ is the set of all alternatives that individual i chooses and $A \subseteq 2^{\Omega}$. To allow the probability of choosing each alternative to be correlated, the CC method uses Bayes' rule to expand $Pr\{\underset{\sim}{y}_i|x_{ik}\}$ as follows,

$$\begin{aligned} Pr\{\underset{\sim}{y}_i|x_{ik}\} &= Pr\{y_{i1} = 1|x_{ik}\}Pr\{y_{i2} = 1|y_{i1}, x_{ik}\}\ldots \\ &\quad Pr\{y_{iM} = 1|y_{i1}, y_{i2}, \ldots, y_{1(M-1)}, x_{ik}\}, \end{aligned} \tag{7a}$$

which can be denoted as

$$Pr\{\underset{\sim}{y}_i|x_{ik}\} = \prod_{j=1}^{M} Pr\{y_{ij} = 1|\tilde{x}_{ij}\} = \prod_{j=1}^{M} G(\tilde{x}_{ij}\beta_j), \tag{7b}$$

where $\tilde{x}_{ij} = (y_{i1}, \ldots, y_{ij}, x_{i1}, \ldots, x_{iK})$ for all $j = 2, \ldots, M$ and $\tilde{x}_{i1} = (x_{i1}, \ldots, x_{iK})$. In (7a, 7b), notice that the multi-label problem is decomposed into a series of conditionally independent binary choice problems $Pr\{y_{ij} = 1|\tilde{x}_{ij}\}$ for all $j = 1, \ldots, M$. The CC method reduces the dimension of the problem significantly, as 2^{Ω} grows exponentially with the size of the choice set Ω.

Notice that different sequences of the choices y_{ij} yield different estimates and predictions. The criterion to select the sequence of the choice depends on the method used to estimate $Pr\{y_{ij} = 1|\tilde{x}_{ij}\}$. When GME is used, the criterion is to choose the sequence that maximizes the total entropy. When the Logit or Probit models are used, the criterion is to maximize the likelihood.

3.2 The CC-GME Model

To estimate the probability $Pr\{\underset{\sim}{y}_i|x_{ik}\}$ of individual i choosing a set A, we need to estimate all the components of the Bayes' decomposition in Eq. (7a, 7b). In this section, we address the problem of estimating the parameters for each of the binomial choice problems $Pr\{y_{ij} = 1|\tilde{x}_{ij}\}$ for all $j = 1, \ldots, M$ using the multinomial choice GME model with two alternatives in the choice set. In this case, the y_{ij} only can take values 0 or 1. Therefore, the two alternatives are whether individual i chooses alternative j or not.

Let $y_{ij} = \tilde{p}_{ij} + e_{ij} = G(\tilde{x}_{ij}\beta_j) + e_{ij}$, where $e_{ij} = \sum_h v_h w_{ijh}$. Let k_j be the index for elements in $\tilde{x}_{ij}$. To simultaneously estimate $\tilde{p}_{ij}$ and w_{ijh} for all j, the GME model

can be written as

$$\max_{\tilde{p},w} H(\tilde{p}_{ij}, w_{ijh}) = -\sum_{ij} \tilde{p}_{ij} \log(\tilde{p}_{ij}) - \sum_{ijh} w_{ijh} \log(w_{ijh}) \tag{8}$$

subject to

$$\sum_{i} \tilde{x}_{ijk_j} y_{ij} = \sum_{i} \tilde{x}_{ijk_j} \tilde{p}_{ij} + \sum_{ih} \tilde{x}_{ijk_j} v_h w_{ijh}, \quad \forall j = 1, \ldots, M, \quad \forall k_j = 1, \ldots, (K + j - 1) \tag{9}$$

$$\sum_{h} w_{ijh} = 1, \quad \forall i = 1, \ldots, N, \forall j = 1, \ldots, M, \tag{10}$$

where (8) is the entropy function, (9) are the stochastic-moment constraints and (10) are normalization constraints. From the maximization problem, the Lagrangian can be expressed as

$$\begin{aligned} L(\tilde{p}_{ij}, w_{ijh}) = &-\sum_{ij} \tilde{p}_{ij} \log(\tilde{p}_{ij}) - \sum_{ijh} w_{ijh} \log(w_{ijh}) \\ &+ \sum_{jk} \lambda_{jk} \left[\sum_{i} \tilde{x}_{ijk_j} y_{ij} - \sum_{i} \tilde{x}_{ijk_j} \tilde{p}_{ij} - \sum_{ih} \tilde{x}_{ijk_j} v_h w_{ijh} \right] \\ &+ \sum_{ij} \delta_{ij} \left[1 - w_{ijh} \right]. \end{aligned} \tag{11}$$

The solutions to the above Lagrangian problem are

$$\hat{p}_{ij} = \exp\left(-1 - \sum_{k} \lambda_{jk} \tilde{x}_{ijk_j} \right) \tag{12a}$$

and

$$\hat{w}_{ijh} = \frac{\exp(-\sum_k \hat{\lambda}_{jk} \tilde{x}_{ijk_j} v_h)}{\sum_h \exp(-\sum_k \hat{\lambda}_{jk} \tilde{x}_{ijk_j} v_h)}. \tag{12b}$$

3.2.1 The Concentrated CC-GME Model

Following [5], the GME model can be reduced to the *concentrated GME model*, which is the model with the minimum number of parameters that represents the original GME model. From the Lagrangian (11) and the GME solutions (12a, 12b), we

can derive the objective function for the concentrated GME model as

$$M(\lambda_{jk_j}) = \sum_{ijk_j} \lambda_{jk_j} \tilde{x}_{ijk_j} y_{ij} + \sum_{ij} \left[\exp(-1 - \sum_{k} \lambda_{jk_j} \tilde{x}_{ijk_j}) \right] + \sum_{ij} \left[\log \sum_{h} \exp(- \sum_{k_j} \lambda_{jk_j} \tilde{x}_{ijk_j} v_h) \right]. \tag{13}$$

The concentrated GME model minimizes expression (13) with respect to λ_{jk_j}. The gradient can be written as

$$\frac{\partial M}{\partial \lambda_{jk_j}} = \sum_{i} \tilde{x}_{ijk_j} y_{ij} - \sum_{i} \tilde{x}_{ijk_j} \tilde{p}_{ij} - \sum_{i} \tilde{x}_{ijk_j} v_h w_{ijh}. \tag{14}$$

Notice that the objective function of the concentrated model is no longer a function of $\tilde{p}_{ij}$ and w_{ijh}, but only a function of λ_{jk_j}. As discussed in [5], the interpretation of λ_{jk_j} from the concentrated model can be compared to that of the β_{jk_j} parameters. Specifically, it can be shown mathematically that $\beta_{jk_j} = -\lambda_{jk_j}$.

3.3 Result Analysis

The multi-label CC-GME model can capture the marginal effects of an individual characteristics on his or her decisions and the dependence pattern of the decisions on all available alternatives.

3.3.1 Marginal Effects

The marginal effects measure the effect of a change in an individual characteristic on an individual's choice decisions. For this multi-label model, the marginal effects are situated at two levels. The first level is to analyze the effect of a change in x_k on the probability that the individual will choose an alternative $j \in \Omega$. This marginal effects in this level is

$$\frac{\partial Pr\{y_j|\tilde{x}_j\}}{\partial x_k} = \beta_{jk} G'(\tilde{x}_j \beta_j). \tag{15}$$

The second level is to analyze the effect of a change in x_k on the probability that the individual will choose a set of alternatives $A \in 2^{\Omega}$. From Eq. (7a, 7b), the marginal effect of x_k on $Pr\{\underline{y}|x\}$ is

$$\frac{\partial Pr\{\underline{y}|x\}}{\partial x_k} = \sum_{j} \left[\beta_{jk} G'(\tilde{x}_j \beta_j) \prod_{q \neq j} G(\tilde{x}_q \beta_q) \right]. \tag{16}$$

3.3.2 Dependence of the Alternatives

In the multi-label model, an individual can choose multiple alternatives. The decisions of choosing each of those alternatives or not can be dependent. The dependence between an alternative j and another alternative q, where the index $q < j$, can be captured from the marginal effects of the change in y_q on $Pr\{y_j|\tilde{x}_j\}$, which is

$$\frac{\partial Pr\{y_j|\tilde{x}_j\}}{\partial y_q} = \beta_{j(K+q)} G'(\tilde{x}_j \beta_j). \tag{17}$$

3.3.3 Model Evaluations

The evaluation of multi-label choice problems requires different measures from those of single-label problems. In contrast to the single-label prediction, which can either be correct or incorrect, the multi-label prediction can be partially correct [13]. Summarized several measures to evaluate multi-label classification models. Commonly used measures include the Hamming Loss, Accuracy, Precision and Recall. The Hamming Loss measures the symmetric difference between the predicted and the true choices with respect to the size of the choice set. The other three methods measures the number of correct predicted choices. The difference is in the normalizing factors. The Accuracy measures the number of correct predicted choices with respect to the sum of all correct, incorrect and missing choices. The Precision and Recall measure the number of correct predicted choices with respect to the number of all predicted choices and the number of all true choices, respectively. The formulas for these four measures are

$$Hamming\ Loss = \sum_{i=1}^{N} \frac{|\hat{\mathbf{y}}_i \Delta \mathbf{y}_i|}{NM} \tag{18}$$

$$Accuracy = \sum_{i=1}^{N} \frac{|\hat{\mathbf{y}}_i \cap \mathbf{y}_i|}{|\hat{\mathbf{y}}_i \cup \mathbf{y}_i|} \tag{19}$$

$$Precision = \sum_{i=1}^{N} \frac{|\hat{\mathbf{y}}_i \cap \mathbf{y}_i|}{|\hat{\mathbf{y}}_i|} \tag{20}$$

$$Recall = \sum_{i=1}^{N} \frac{|\hat{\mathbf{y}}_i \cap \mathbf{y}_i|}{|\mathbf{y}_i|}. \tag{21}$$

where $|\cdot|$ is the number of elements in the set, Δ is the symmetric difference between the two sets, $\cap$ is the intersection of the two sets and $\cup$ is the union of the two sets.

4 Monte-Carlo Experiment

In this section, we used Monte-Carlo simulations to empirically evaluate our multi-label CC-GME model using three simulated datasets with normal, logistic and uniform errors. We compared the performance of the CC-GME model with some selected multi-label estimations including CC-Logit, CC-Probit, BR-GME, BR-Logit, BR-Probit, LP-GME, LP-Logit and LP-Probit models.

The BR model simplifies the multi-label model to an independent series of binary single-label choice models. For example, the BR-GME model applies the GME single-label model to estimate the probability that individual i chooses alternative j, $Pr\{y_{ij} = 1|x_{ik}\}$. The probability that individual i chooses the set of alternatives A is then $Pr\{\underline{y}_i = A|x_{ik}\} = \prod_{j=1}^{K} Pr\{y_{ij} = 1|x_{ik}\}$. The LP model transforms the multi-label problem into a single-label problem of 2^{Ω} alternatives. For example, the LP-GME model applies the GME single-label model to estimate $Pr\{\underline{y}_i = A|x_{ik}\}$ where $A \in 2^{\Omega}$.

4.1 Simulation

For simplicity, we assumed $N = 1{,}000$ individuals, $M = 3$ alternatives and $K = 2$ individual characteristics. The simulation procedures are composed of two main steps. The first step is to generate all characteristics x_i, the true parameters β_{ik}^0 and the error e_{ij}. Given the information from x_i and β_{1k}^0, we calculated the latent variable $y'_{i1} = \sum_k \tilde{x}_{i1k}\beta_{1k} + \varepsilon_{i1}$. We then generated the choice variable y_{i1} by letting $y_{i1} = 1$ when $y'_{i1} \geq 0$ and $y_{ij} = 0$ otherwise. Once we have y_{i1}, we can simulate $y_{i2}, \ldots, y_{iM}$. This first step provided us with the simulated data (y_{ij}, x_i) and true parameters β_{ik}^0.

The second step is to use the data from the first step and apply the CC-Logit, CC-Probit, BR-GME, BR-Logit, BR-Probit, LP-GME, LP-Logit and LP-Probit models. After computing the parameter estimates $\hat{\beta}_{ik}$, the predicted probability of individual i choosing choice j, $\hat{p}_{ij}$, and the corresponding predicted choices, $\hat{\underline{y}}_i$, can be obtained. Using Monte-Carlo simulation, the standard deviation of each estimated parameter and statistics can be estimated.

4.2 Results

The Monte-Carlo simulations allowed us to compare the performances of the CC-Logit, CC-Probit, BR-GME, BR-Logit, BR-Probit, LP-GME, LP-Logit and LP-Probit models.

Table 1 shows the true parameters and the estimated parameters from all the CC and the BR models. It should be noted that the LP models can also provide estimates

Table 1 True and estimated parameters for the CC and BR models

Alternative	Regressor	TRUE	Classifier chains			Binary relevance		
			GME	Logit	Probit	GME	Logit	Probit
Normal error								
y_1	x_1	0.318	0.513	0.516	0.319	0.513	0.516	0.319
			(0.071)	(0.072)	(0.044)	(0.071)	(0.072)	(0.044)
	x_2	0	0.003	0.003	0.002	0.003	0.003	0.002
			(0.068)	(0.068)	(0.042)	(0.068)	(0.068)	(0.042)
y_2	x_1	−0.223	−0.382	−0.382	−0.228	−0.331	−0.333	−0.199
			(0.073)	(0.076)	(0.045)	(0.069)	(0.071)	(0.042)
	x_2	−0.659	−1.100	−1.108	−0.665	−1.073	−1.098	−0.660
			(0.076)	(0.077)	(0.044)	(0.071)	(0.074)	(0.043)
	y_1	0.243	0.404	0.382	0.228	–	–	–
			(0.101)	(0.139)	(0.082)			
y_3	x_1	0.706	1.190	1.205	0.700	0.841	1.092	0.640
			(0.103)	(0.107)	(0.060)	(0.071)	(0.095)	(0.053)
	x_2	−0.360	−0.629	−0.633	−0.368	−0.645	−0.849	−0.498
			(0.092)	(0.100)	(0.058)	(0.068)	(0.092)	(0.052)
	y_1	0.551	0.965	0.992	0.574	–	–	–
			(0.150)	(0.177)	(0.102)			
	y_2	0.844	1.407	1.442	0.837	–	–	–
			(0.152)	(0.183)	(0.106)			
MSE			0.001	0.143	0.005	0.001	0.112	0.006
Logistic error								
y_1	x_1	0.318	0.324	0.325	0.203	0.324	0.325	0.203
			(0.070)	(0.070)	(0.043)	(0.070)	(0.070)	(0.043)
	x_2	0	0.006	0.006	0.004	0.006	0.006	0.004
			(0.060)	(0.061)	(0.038)	(0.060)	(0.061)	(0.038)
y_2	x_1	−0.223	−0.227	−0.228	−0.139	−0.209	−0.208	−0.127
			(0.068)	(0.069)	(0.042)	(0.068)	(0.069)	(0.042)
	x_2	−0.659	−0.665	−0.669	−0.409	−0.656	−0.666	−0.408
			(0.072)	(0.073)	(0.043)	(0.071)	(0.073)	(0.043)
	y_1	0.243	0.244	0.241	0.148	–	–	–
			(0.099)	(0.131)	(0.080)			
y_3	x_1	0.706	0.703	0.707	0.422	0.593	0.676	0.407
			(0.079)	(0.080)	(0.046)	(0.068)	(0.076)	(0.044)
	x_2	−0.360	−0.348	−0.348	−0.208	−0.390	−0.452	−0.272
			(0.075)	(0.079)	(0.047)	(0.067)	(0.077)	(0.046)
	y_1	0.551	0.563	0.581	0.348	–	–	–
			(0.129)	(0.146)	(0.087)			
	y_2	0.844	0.832	0.852	0.511	–	–	–
			(0.133)	(0.184)	(0.109)			

(continued)

Table 1 (continued)

Alternative	Regressor	TRUE	Classifier chains			Binary relevance		
			GME	Logit	Probit	GME	Logit	Probit
MSE			0.000	0.012	0.043	0.000	0.007	0.032
Uniform error								
y_1	x_1	0.318	1.665	1.680	1.008	1.665	1.680	1.008
			(0.095)	(0.097)	(0.054)	(0.095)	(0.097)	(0.054)
	x_2	0	−0.022	−0.023	−0.013	−0.022	−0.023	−0.013
			(0.082)	(0.082)	(0.048)	(0.082)	(0.082)	(0.048)
y_2	x_1	−0.223	−1.284	−1.351	−0.780	−0.758	−0.877	−0.505
			(0.148)	(0.173)	(0.096)	(0.103)	(0.124)	(0.069)
	x_2	−0.659	−3.812	−4.011	−2.320	−3.353	−3.769	−2.172
			(0.199)	(0.237)	(0.125)	(0.170)	(0.216)	(0.115)
	y_1	0.243	1.431	1.492	0.859	–	–	–
			(0.184)	(0.293)	(0.160)			
y_3	x_1	0.706	3.581	4.488	2.552	1.359	2.801	1.580
			(0.183)	(0.347)	(0.188)	(0.068)	(0.184)	(0.097)
	x_2	−0.360	−1.818	−2.240	−1.280	−1.163	−2.456	−1.385
			(0.182)	(0.288)	(0.155)	(0.072)	(0.168)	(0.089)
	y_1	0.551	2.811	3.506	1.997	–	–	–
			(0.293)	(0.468)	(0.253)			
	y_2	0.844	4.307	5.434	3.085	–	–	–
			(0.294)	(0.533)	(0.288)			
MSE			0.047	7.161	1.728	0.017	3.480	0.783

[1] Standard deviations in parentheses

of the parameters. However, the parameters in the LP model are not comparable to the true parameters generated in this Monte-Carlo experiment. It should be noted that the data simulation process was based on the CC model. When the errors are normally distributed, the true model is the CC-Probit model. Therefore, the Probit-based models performed better than the Logit-based models. When the errors are logistically distributed, the true model is the CC-Logit model and the Probit models performed better than the Logit models. However, regardless of the error distributions, the GME models always have the lowest MSE.

Figure 1 shows the prediction regions for an individual's decision on each alternative y_{ij} given his or her characteristics x_1 and x_2 using the CC-GME model. Each of the three lines represents the combinations of x_1 and x_2 such that $Pr\{y_{ij} = 1|\tilde{x}_{ij}\} = 0.5$. Regions (1) to (8) represent the choices $y_i = (0, 0, 0)$, $(0, 0, 1)$, $(0, 1, 0)$, $(0, 1, 1)$, $(1, 0, 0)$, $(1, 0, 1)$, $(1, 1, 0)$ and $(1, 1, 1)$, respectively. Therefore, the result shows that individuals with high value x_1 and low value of x_2 are more likely to choose all three alternatives. Individuals with lower x_1 and high x_2 are likely to choose none of the alternatives.

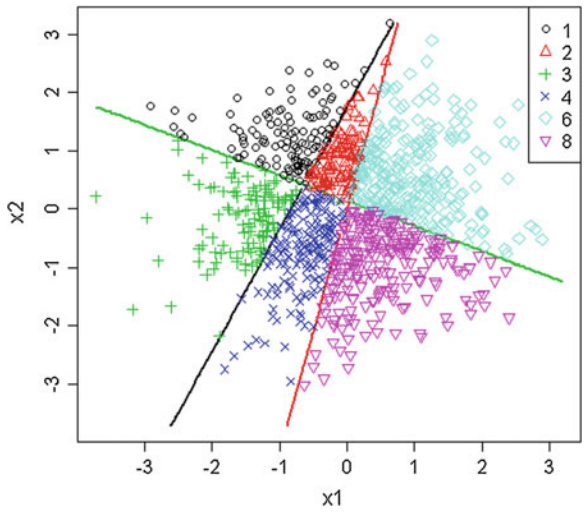

Fig. 1 Prediction regions for all possible sets of alternatives from the CC-GME estimation for the simulation with logistic errors

Table 2 reports the Hamming Loss, Accuracy, Precision and Recall statistics for all the CC, BR and LP models. The results show that the forecasting performance depends on the choice of the problem transformation methods, but not on the choice of single-label estimation methods. That is, the CC model outperformed the BR and LP models with respect to all evaluation measures, except Precision. The CC-GME, CC-Logit and CC-Probit models yielded similar results.

5 Occupational Hazards Empirical Example

Consider a problem in which an individual chooses a job with multiple job attributes. This problem can be viewed as an individual choosing a set of job attributes. In this empirical example, the job attributes are a set of occupational hazards. Therefore, each individual will choose the hazards from which he or she gains the least disutility. In this section, we apply the CC-GME model to predict a set of occupational hazards that an individual faces and the factors determining his or her choices of hazards. For the performance evaluation, we applied the five-fold cross validation method to compare the out-sample prediction performance between the CC-GME model and other models [10].

5.1 Data Description

The dataset is from *The Informal Worker Analysis and Survey Modeling for Efficient Informal Worker Management Project*, which aims at studying the structure and

Table 2 Model comparison for the simulated data

Evaluation	Classifier chains			Binary relevance			Label powerset		
Methods	GME	Logit	Probit	GME	Logit	Probit	GME	Logit	Probit
Normal error									
Hamming loss	**0.304**	**0.303***	**0.303***	0.341	**0.314**	**0.315**	0.326	0.327	0.331
	(0.008)	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)	(0.010)
Accuracy	**0.589**	**0.590***	**0.590***	0.516	**0.581**	**0.581**	0.543	0.541	0.528
	(0.011)	(0.013)	(0.013)	(0.009)	(0.013)	(0.013)	(0.013)	(0.013)	(0.016)
Precision	0.726	0.725	**0.725**	0.737	0.713	0.713	**0.736**	**0.737**	**0.745***
	(0.010)	(0.009)	(0.009)	(0.011)	(0.008)	(0.008)	(0.013)	(0.013)	(0.015)
Recall	**0.757**	**0.760***	**0.760***	0.633	**0.758**	**0.758**	0.674	0.671	0.645
	(0.009)	(0.017)	(0.017)	(0.008)	(0.017)	(0.017)	(0.017)	(0.017)	(0.025)
Logistic error									
Hamming loss	**0.364**	**0.363***	**0.364**	0.391	**0.372**	**0.372**	0.383	0.383	0.388
	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.011)	(0.011)	(0.011)
Accuracy	**0.521***	**0.521***	**0.521***	0.455	**0.516**	**0.516**	0.475	0.473	0.460
	(0.013)	(0.017)	(0.017)	(0.010)	(0.018)	(0.018)	(0.018)	(0.018)	(0.020)
Precision	**0.666**	**0.659**	**0.659**	**0.666**	**0.648**	**0.648**	**0.665**	**0.666**	**0.670***
	(0.014)	(0.010)	(0.010)	(0.014)	(0.010)	(0.010)	(0.016)	(0.015)	(0.017)
Recall	**0.713**	**0.714**	**0.714**	0.589	**0.717**	**0.718***	0.624	0.620	0.596
	(0.014)	(0.029)	(0.029)	(0.009)	(0.030)	(0.030)	(0.031)	(0.030)	(0.035)
Uniform error									
Hamming loss	**0.156**	**0.155***	**0.155***	0.216	0.174	0.174	0.184	0.184	0.185
	(0.005)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)
Accuracy	**0.768**	**0.769***	**0.769***	0.668	0.745	0.745	0.724	0.723	0.719
	(0.008)	(0.008)	(0.008)	(0.007)	(0.009)	(0.009)	(0.008)	(0.008)	(0.009)
Precision	**0.866**	**0.867**	0.867	**0.879***	0.849	0.849	**0.867**	**0.870**	**0.874**
	(0.007)	(0.005)	(0.005)	(0.008)	(0.006)	(0.006)	(0.007)	(0.007)	(0.008)
Recall	**0.871***	**0.871***	**0.871***	0.736	**0.860**	**0.859**	**0.814**	**0.811**	**0.802**
	(0.059)	(0.006)	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)	(0.009)

Standard deviations in parentheses. Statistics with * represent estimation methods that are the 'best' with respect to each evaluation metric. Statistics in bold represent estimation methods with the prediction power not statistically different from the 'best' estimation method

nature of the informal sector in Chiang Mai, Thailand in 2012. In the survey, each respondent was asked whether he or she faced each of the three types of occupational hazards, namely, (1) physical and mechanical hazards, (2) ergonomic and psychosocial hazards and (3) biological and chemical hazards. The survey also provides data for each individual's demographic, employment and financial status. Explanatory variables used in this study include (1) age, (2) number of children, (3) total income and dummy variables for (4) female, (5) high school, (6) college and (7) agricultural household.

Table 3 Model comparison for the occupational hazards data

Evaluation	Classifier chains			Binary relevance			Label powerset		
Methods	GME	Logit	Probit	GME	Logit	Probit	GME	Logit	Probit
Hamming loss	**0.263***	**0.297**	0.372	**0.284**	0.382	0.383	**0.315**	**0.325**	–
	(0.049)	(0.016)	(0.022)	(0.039)	(0.032)	(0.032)	(0.083)	(0.020)	
Accuracy	**0.702***	0.541	**0.658**	**0.675**	**0.652**	**0.652**	**0.676**	0.529	–
	(0.065)	(0.012)	(0.041)	(0.055)	(0.053)	(0.053)	(0.080)	(0.015)	
Precision	**0.753**	**0.759***	**0.758**	**0.755**	**0.748**	**0.748**	**0.701**	**0.723**	–
	(0.080)	(0.058)	(0.047)	(0.084)	(0.044)	(0.044)	(0.102)	(0.063)	
Recall	**0.914**	**0.848**	**0.844**	0.870	**0.849**	**0.850**	**0.956***	**0.883**	–
	(0.023)	(0.096)	(0.112)	(0.045)	(0.131)	(0.131)	(0.040)	(0.131)	

Standard deviations in parentheses. The LP-Probit model fails to converge. Statistics with * represent estimation methods that are the 'best' with respect to each evaluation metric. Statistics in bold represent estimation methods with the prediction power not statistically different from the 'best' estimation method

5.2 Results

For the choice of problem transforming methods, the results are similar to the simulation exercises in the sense that the CC model outperformed the BR and LP models in most measures (see Table 3). Specifically, the CC model is superior than the BR and LP models with respect to the Hamming Loss, Accuracy and Precision criteria. For the choice of single-label estimation methods, the GME model outperformed the Logit and Probit models with respect to the Hamming Loss, Accuracy and Recall measures.

6 Concluding Remarks

The empirical results obtained in this study show that the forecasting performance depends on the choice of the problem transformation methods, but not on the choice of single-label estimation methods. Specifically, the CC model outperformed the BR and LP models with respect to all evaluation measures expect the Precision. For the parameter estimates, the GME-based methods yielded smaller MSE than those of the Logit and Probit-based methods.

Although the Bayes' rule implies that $Pr\{\underline{y}_i|x_{ik}\} = \prod_{j=1}^{M} Pr\{y_{ij} = 1|\tilde{x}_{ij}\}$, it does not imply directly that $Pr\{\underline{y}_i|x_{ik}\} = \prod_{j=1}^{M} G(\tilde{x}_{ij}\beta_j)$. The CC-GME model still relies on the linearity assumption of $\tilde{x}_{ij}\beta_j$ when we set

$$Pr\{y_{ij} = 1|\tilde{x}_{ij}\} = G(\tilde{x}_{ij}\beta_j). \tag{22}$$

Therefore, other methods to incorporate the dependency among alternatives into the multi-label classification problem with weaker assumptions could potentially improve the performance.

Acknowledgments We are highly appreciated and would like to acknowledge the financial support from the Thailand Research Fund through the Royal Golden Jubilee Ph.D. Program (Grant No. PHD/0211/2556).

References

1. Baltas, G.: A model for multiple brand choice. Eur. J. Oper. Res. **154**(1), 144–149 (2004)
2. Bhat, C.R., Srinivasan, S., Sen, S.: A joint model for the perfect and imperfect substitute goods case: application to activity time-use decisions. Transp. Res. Part B: Methodol. **40**(10), 827–850 (2006)
3. Bhat, C.R., Srinivasan, S.: A multidimensional mixed ordered-response model for analyzing weekend activity participation. Transp. Res. Part B: Methodol. **39**(3), 255–278 (2005)
4. Denzau, A.T., Gibbons, P.C., Greenberg, E.: Bayesian estimation of proportions with a cross-entropy prior. Commun. Stat.-Theory Methods **18**(5), 1843–1861 (1989)
5. Golan, A., Judge, G., Perloff, J.M.: A maximum entropy approach to recovering information from multinomial response data. J. Am. Stat. Assoc. **91**(434), 841–853 (1996)
6. Golan, A.: Information and Entropy Econometrics: A Review and Synthesis. Now Publishers Inc. (2008)
7. Heath, D., Zitzelberger, A., Giraud-Carrier, C.G.: A multiple domain comparison of multi-label classification methods. In: Working Notes of the 2nd International Workshop on Learning from Multi-label Data at ICML/COLT, 21–28 (2010)
8. Jaynes, E. T.: Information theory and statistical mechanics. Phys. rev. **106**(4), 620 (1957a)
9. Jaynes, E. T.: Information theory and statistical mechanics. II. Phys. rev. **108**(2), 171 (1957b)
10. Mosteller, F., Tukey, J.W.: Data Analysis, Including Statistics. The Collected Works of John W. Tukey: Graphics pp. 1965–1985, vol. 5 (123) (1988)
11. Pukelsheim, F.: The three sigma rule. Am. Stat. **48**(2), 88–91 (1994)
12. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Mach. Learn. **85**(3), 333–359 (2011)
13. Santos, A., Canuto, A., Neto, A.F.: A comparative analysis of classification methods to multi-label tasks in different application domains. Int. J. Comput. Inform. Syst. Indust. Manag. Appl **3**, 218–227 (2011)
14. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE Mob. Comput. Commun. Rev. **5**(1), 3–55 (2001)
15. Soofi, E.S.: A generalizable formulation of conditional logit with diagnostics. J. Am. Stat. Assoc. **87**, 812–816 (1992)
16. Train, K.: Data analysis. Including Statistics Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge (2009)
17. Tsoumakas, G., Katakis, I.: Multi-label classification: an overview. Int. J. Data Warehous. Min. **3**(3), 1–13 (2007)
18. Zhang, M.L., Zhou, Z.H.: ML-KNN: a lazy learning approach to multi-label learning. Pattern Recognit. **40**(7), 2038–2048 (2007)