

# ASSESSING CONSUMER FRAUD RISK IN INSURANCE CLAIMS: AN UNSUPERVISED LEARNING TECHNIQUE USING DISCRETE AND CONTINUOUS PREDICTOR VARIABLES

Jing Ai,<sup>\*†</sup> Patrick L. Brockett,<sup>§</sup> and Linda L. Golden<sup>\*\*</sup>

---

## ABSTRACT

We present an unsupervised learning method for classifying consumer insurance claims according to their suspiciousness of fraud versus nonfraud. The predictor variables contained within a claim file that are used in this analysis can be binary, ordinal categorical, or continuous variates. They are constructed such that the ordinal position of the response to the predictor variable bears a monotonic relationship with the fraud suspicion of the claim. Thus, although no individual variable is of itself assumed to be determinative of fraud, each of the individual variables gives a “hint” or indication as to the suspiciousness of fraud for the overall claim file. The presented method statistically concatenates the totality of these “hints” to make an overall assessment of the ranking of fraud risk for the claim files without using any a priori fraud-classified or -labeled subset of data. We first present a scoring method for the predictor variables that puts all the variables (whether binary “red flag indicators,” ordinal categorical variables with different categories of possible response values, or continuous variables) onto a common  $-1$  to  $1$  scale for comparison and further use. This allows us to aggregate variables with disparate numbers of potential values. We next show how to concatenate the individual variables and obtain a measure of variable worth for fraud detection, and then how to obtain an overall holistic claim file suspicion value capable of being used to rank the claim files for determining which claims to pay and the order in which to investigate claims further for fraud. The proposed method provides three useful outputs not usually available with other unsupervised methods: (1) an ordinal measure of overall claim file fraud suspicion level, (2) a measure of the importance of each individual predictor variable in determining the overall suspicion levels of claims, and (3) a classification function capable of being applied to existing claims as well as new incoming claims. The overall claim file score is also available to be correlated with exogenous variables such as claimant demographics or high-volume physician or lawyer involvement. We illustrate that the incorporation of continuous variables in their continuous form helps classification and that the method has internal and external validity via empirical analysis of real data sets. A detailed application to automobile bodily injury fraud detection is presented.

---

\* Jing Ai, PhD, is Assistant Professor in the Department of Financial Economics and Institutions, Shidler College of Business, University of Hawaii at Manoa, Honolulu, HI 96822, jing.ai@hawaii.edu.

† Corresponding author.

§ Patrick L. Brockett, PhD, is Gus S. Wortham Chaired Professor in Risk Management and Insurance in the McCombs School of Business, Department of Information, Risk, and Operations Management, Global Fellow IC<sup>2</sup> Institute, University of Texas at Austin, Austin, TX 78712, brockett@mail.utexas.edu.

\*\* Linda L. Golden, PhD, is Marlene and Morton Meyerson Centennial Professor in Business in the McCombs School of Business, Department of Marketing, Global Fellow IC<sup>2</sup> Institute, University of Texas at Austin, Austin, TX 78712, linda.golden@mcombs.utexas.edu.

## 1. IMPORTANCE OF THE PROBLEM AND OVERVIEW OF SOLUTION

### 1.1 Problem Importance

The National Insurance Crime Bureau estimates about 10 percent of property and casualty insurance claims are fraudulent, which costs Americans \$30 billion per year with annual average household insurance premiums \$200 to \$300 higher because of the cost of fraud. When indirect costs of fraud are incorporated, this cost may rise to \$1,000 per year per family (Texas Department of Insurance 2006). More recently the Federal Bureau of Investigation (FBI) estimated these costs at more than \$40 billion per year, resulting in an increase in family insurance premiums of between \$400 and \$700 per year (Federal Bureau of Investigation 2008). The Insurance Research Council puts the cost of automobile insurance fraud alone at \$15–20 billion per year and estimates that approximately one-third of all bodily injury claims in automobile accidents have some degree of fraud in their claimed amount. If other lines of insurance are included, the total cost of insurance fraud may exceed \$120 billion annually (IRC 2008). Thus, insurance fraud is a very serious problem, and the detection of fraud is quite important economically.

Unfortunately the problem of detecting insurance fraud is very difficult. It does not lend itself to traditional supervised statistical classification methods such as logistic regression, because by the nature of insurance fraud, it is usually costly, and often impossible, to obtain a “training sample” of insurance claims with known unambiguous fraud labels (i.e., a variable indicating whether the claim is fraudulent or legitimate). This is required for estimating parameters in standard supervised classification models. Even if one could obtain these labels, they can be subject to classification errors that contaminate a supervised classification model.

Finally, those who commit fraud specifically and thoughtfully attempt to cover their tracks so as to avoid being identified, and if a supervised model for detection is developed, the fraudsters can change their behavior to reduce the effectiveness of the detection method used. Therefore, supervised learning methods (e.g., logistic regression, discriminant analysis, support vector machines [SVM], Bayesian additive regression trees [BART], neural networks), although well developed and accessible for certain types of fraud detection (such as credit card fraud), may not be applicable to insurance fraud detection. Unsupervised classification methods not requiring knowledge of fraud labels for a subset of data are suitable for these insurance claim fraud detection applications.

### 1.2 Overview of Methodology and Solution Properties

In much of the previous fraud literature involving polychotomous or continuous variables, the analysts group or “bin” the data to create discrete (usually binary) variables. This is done to ensure that the scaling is comparable across the variables. Binning, however, results in information loss.

Once the variable has been binned, one next needs to choose a method of assigning numerical values to represent the categories for subsequent numerical statistical analysis. Commonly one assigns increasing integers to the categories (raw integer scoring). A flaw with this, however, is that a raw-integer scored variable with four possible values (1, 2, 3, 4) might have an extreme value listed as a “4,” whereas another binary variable with values (1, 2) might have an extreme value listed as a “2.” Making the scale so that the extreme values are compatible has led most authors to group or bin the data so every variable has an equal number of categorical possibilities. Because many fraud detection variables are already binary (e.g., “Was a police report filed at the scene?”), in previous analysis the data were usually binned into binary variables.

For the model developed in this paper, we introduce a method for assigning a numerical score for the predictor variables by extending a technique developed in the epidemiological literature called RIDIT analysis (RIDIT stands for Relative to an Identified Distribution Integral Transformation). This scoring method was invented by Bross (1958) for the analysis of discrete ordinal data. An overview of RIDIT analysis is given in Flora (1988).

For an application to fraud detection it is important to extend Bross’s RIDIT scores to include continuous predictors as well. Many fraud predictor variables such as “time between the accident and

the filing of the claim,” “time between the accident and the filing of the police report,” “age of the claimant,” and “ambulance charges” are continuous predictor variables (Viaene et al. 2002).

Our method of scoring the discrete and continuous variables overcomes both the information loss associated with binning as well as the scaling issue (because all variables are put on a common  $[-1, 1]$  scale). Brockett and Golden (1992) provide further intuitive justification for the RIDIT extension used herein and show that it also has better statistical properties than other competitor scaling methods, including raw integer scoring and conditional mean scoring, even for use involving other standard statistical methods such as regression and logistic regression. The unified method in this paper is an extension of the technique applied in Brockett et al. (2002) to binary predictors. The extension here allows a consistent analysis incorporating more variables and more information than before.

In the model of this paper, each claim file is viewed conceptually as having a relative position on an underlying latent “fraud suspicion” dimension that underlies the dichotomous “fraud” label, and the concatenation of the ensemble of predictor variables in the claim file provides information concerning this position. The fraud predictor variables in a claim file are constructed by domain experts to be related to fraud suspicion level in a monotonic fashion, and we assume that this set of variables has been chosen for the analysis (Weisberg and Derrig 1991). The initial dichotomy of fraud classification is used in the claims settlement process for deciding whether or not to spend more resources investigating and negotiating the claim or to pay it immediately.

The developed method provides three useful outputs: (1) an ordinal measure of overall claim file fraud suspicion level capable of being ranked across claims and of being used in other statistical analysis, (2) a measure of the importance of each individual predictor variable in assessing the suspicion level of claims, and (3) a classification function that can be applied to existing or new incoming claims. Other unsupervised methods (such as cluster analysis or Kohonen’s self-organizing feature map) are usually not able to perform one or more of the above functions, making them more difficult to interpret and less useful for fraud detection.

After we present the unified method, we illustrate its application to insurance fraud detection. We provide external validation of its performance by assessing its classification accuracy on a data set that contains “fraud” labels. To further diagnose whether the fraud detection failures we uncover are due to deficiencies in our new method or deficiencies in the data set, we compare the classifications produced by this (unsupervised) method to those generated by a set of supervised learning methods that are optimized for performance on the data set. It is worth noting that these supervised methods are not really competitors for our method because they differ fundamentally in the prior information required for learning a model and in the implementation cost incurred in practice. Rather, the set of supervised learning methods (logistic regression, SVM, and BART) are used as a form of external validation of our method, just as the assessment of our method relative to the given “fraud” labels is best viewed as an external validation.

Limited unsupervised methods are currently available for insurance fraud detection. Among the few, Brockett, Xia, and Derrig (1998) use the unsupervised Kohonen’s Feature Map (Kohonen 1989) for insurance fraud classification. In this paper we examine and discuss the relative strengths of our method against two competitor unsupervised techniques: Kohonen’s self-organizing feature maps and cluster analysis.

The paper is organized as follows. The next section discusses our modification of the RIDIT scoring system of Bross (1958). Section 3 develops the unified unsupervised fraud classification method incorporating all types of predictor variables. More specifically, we develop an individual variable scoring method based on RIDITs, develop a measure of individual predictor variable importance, and create an overall claim file suspicion score using an iterative weight additive scoring method that shows its relation to principal component analysis. Section 4 demonstrates this method using a personal injury protection insurance claims data set from the Automobile Insurance Bureau in Massachusetts, where “fraud” labels are available, and assesses performance by comparison with supervised and unsupervised methods. Because this insurance fraud data set contains only binary predictor variables, in Section 5

we introduce a second data set (an income classification data set) containing both discrete and continuous predictors to demonstrate why the extension to continuous predictors is desirable, that is, the information loss and performance deterioration due to data binning. Section 6 concludes the paper.

## 2. ADAPTING BROSS'S RIDIT SCORING METHOD TO SCORE FRAUD PREDICTOR VARIABLE RESPONSES WITH A VIEW TOWARD ASSESSING THE PREDICTIVE VARIABLES' RELATIVE IMPORTANCE IN DISCERNING BETWEEN TWO CLASSES

In the context of epidemiology, Bross (1958) presented a scoring mechanism for subjectively ordinal (rank-ordered) categorical data that are not necessarily metric (numerical) in nature (e.g., “degree of paleness”). He called the scoring method (and the resulting analysis) RIDIT as an acronym for Relative to an Identified Distribution Integral Transformation to highlight that the scoring was relative to an identified probability distribution (usually some standard or control population distribution) and to signify that it was in the spirit of the usual probability integral transformation familiar to rank-ordered nonparametric statistics. In Bross's RIDIT scoring, the score (numerical value) assigned to a particular response category  $i$  of a categorical variable is defined as  $\sum_{j<i} P_j + 1/2 P_i$ , where  $\{P_i\}$  is the probability of response  $i$  based on an identified reference distribution for the responses (Bross 1958). It can be observed that these scores represent the expected relative rank of the response category  $i$  with respect to the identified or standard distribution, with tied values counted as  $1/2$ , divided by the sample size. Brockett and Levine (1977) and Brockett (1981) provide axiomatic foundations for RIDIT scoring, and Golden and Brockett (1987) show empirically that RIDIT scoring of rank-ordered categorical questionnaire data exhibits superior empirical performance relative to other standard scoring methods when used in subsequent statistical analysis. The scoring method used in our analysis generalizes this methodology to continuous as well as discrete scenarios.

The variable score we use here is an adaptation (a linear transformation) of the original RIDIT score for discrete ordinal variables, which takes the empirical distribution for the sample as the identified comparison distribution (so the relative ranking in Bross's context is with respect to the entire data set). Accordingly, in the discrete case we define the variable score  $B_{it}$  for a claim file that has a response in category  $i$  of the predictor variable  $t$  to be

$$B_{it} = \sum_{j<i} P_{jt} - \sum_{j>i} P_{jt}, \quad (2.1)$$

where  $P_{jt}$  is the *observed* proportion of responses in category  $j$  of variable  $t$  in the sample. Accordingly, the score  $B_{it}$  represents how “extreme” the response  $i$  to variable  $t$  is relative to the entire set of data. Suppressing the subscript  $t$ , if  $R_i$  is Bross's RIDIT score for category  $i$  and the identified distribution used is the sample empirical distribution, the simple relation  $B_i = 2R_i - 1$  holds.

These variable scores (2.1) possess desirable properties. First,  $B_{it}$  is bounded in  $[-1, 1]$ . Second,  $B_{it}$  is monotonically increasing in the rank of response  $i$ . Third,  $B_{it}$  is centered around zero, that is, it has mean equal to zero over the entire sample (Brockett et al. 2002).

In the unsupervised learning context of this paper's classification algorithm, it is assumed that the predictor variables have been constructed in such a fashion that the rank of response categories bears a monotonic relationship with the likelihood of being in the fraud class (i.e., without loss of generability, assume responses in lower-ranked response categories of a variable suggest higher suspicion of fraud). This is an intuitive assumption based on the domain knowledge. In essence, each predictor variable gives a “hint” as to the likelihood of the entity belonging to the fraud class, and the problem in this unsupervised learning context is how to put all these hints together to make a classification judgment and to derive related measures.

Equation (2.1) is for ordinal discrete predictor variables; however, in many applications both continuous and discrete predictor variables are desired to be incorporated in a unified unsupervised classification method.

To derive the analogue of  $B_{it}$  in the continuous case, note that from (2.1) the variable score  $B_i$  is the proportion of claim files within response categories ranked lower than  $i$  minus the proportion within higher ranked response categories (henceforth we suppress the index  $t$  when we focus on the calculation for a single variable  $t$ ). In the continuous predictor case, if  $X$  is a continuous predictor having a monotonically decreasing relationship with the unobserved fraud suspicion level (i.e., a lower value of this variable leads to higher likelihood of fraud class membership), by analogue with (2.1), we define variable score  $B(x)$  as the proportion of claim files with response value less than  $x$  minus the proportion of claim files with response value larger than  $x$ . This, we have

$$B(x) = \hat{F}(x^-) - [1 - \hat{F}(x)] = [\hat{F}(x) - \hat{P}(x)] - [1 - \hat{F}(x)] = 2\hat{F}(x) - 1 - \hat{P}(x), \quad (2.2)$$

where  $\hat{F}(x)$  is the empirical distribution of  $X$  and  $\hat{P}(x)$  is the sample proportion of response  $x$ .

The variable score  $B(x)$  in (2.2) preserves the desirable characteristics from (2.1), that is,  $B(x)$  is bounded in  $[-1, 1]$  and  $B(x)$  is monotonically increasing and centered around zero:

$$\begin{aligned} E_{\hat{p}}[B(x)] &= \sum_{k=1}^K B(x_k) \hat{P}(x_k) = \sum_{k=1}^K [2\hat{F}(x_k) - 1 - \hat{P}(x_k)] \hat{P}(x_k) \\ &= 2 \sum_{k=1}^K \left\{ \left[ \sum_{l=1}^K \hat{P}(x_l) \right] \hat{P}(x_k) \right\} - 1 - \sum_{k=1}^K [\hat{P}(x_k)]^2 \\ &= 2 \left\{ \sum_{k=1}^K [\hat{P}(x_k)]^2 + \sum_{k=2}^K \left[ \hat{P}(x_k) \sum_{l=1}^{k-1} \hat{P}(x_l) \right] \right\} - \sum_{k=1}^K [\hat{P}(x_k)]^2 - 1 \\ &= \left[ \sum_{k=1}^K \hat{P}(x_k) \right]^2 - 1 = 0, \end{aligned}$$

where variable  $X$  takes on an (increasingly ranked) value,  $x_1, \dots, x_K$ , in the sample.

### 3. A UNIFIED UNSUPERVISED CLASSIFICATION METHOD FOR DISCRETE AND CONTINUOUS PREDICTOR VARIABLES

#### 3.1 Stochastic Dominance Assumption and Variable Construction

As discussed in Section 1, an important assumption of our model is that the individual predictor variables are constructed so that the fraud class members tend to score lower on predictor variables than the nonfraud class members. In practice, this is not a difficult assumption to have hold, because the veracity of this assumption can be guaranteed by the choice of variables for the analysis (i.e., include only those that satisfy this assumption, or by rephrasing the variable description to obtain the desired ordering). Once this has been done, the individual predictor variables exhibit a first-order stochastic dominance relationship between the fraud class and the nonfraud class. Stated formally, let  $F_1(\cdot)$  be the distribution function of some variable  $t$  for class 1 (fraud class) and  $G_2(\cdot)$  be the distribution function for class 2 (nonfraud class). The distribution  $G_2(\cdot)$  first-order stochastically dominates  $F_1(\cdot)$  if and only if  $G_2(x) \leq F_1(x)$  for all  $x$ , or equivalently,  $\Delta(x) = F_1(x) - G_2(x) \geq 0$  for all  $x$  (Mas-Colell, Whinston, and Green 1995). This is equivalent to our variable construction assumption that smaller response values suggest higher potential of fraud. We formally establish this in Proposition 1.

#### Proposition 1

Let  $X$  denote the value obtained for variable  $t$  by a randomly chosen claim file from class 1, and let  $Y$  denote the value obtained for variable  $t$  by a randomly chosen claim file from class 2. Assume  $X \sim F_1$

( $\cdot$ ), and  $Y \sim G_2(\cdot)$ . If  $G_2(\cdot)$  first-order stochastically dominates  $F_1(\cdot)$ , then  $P(X < Y) \geq \frac{1}{2}$ , that is, a response (to variable  $t$ ) from class 1 is likely to be lower in value than a response from class 2.

**PROOF**

Denote  $\Delta(y) = F_1(y) - G_2(y)$ . Using Lebesgue-Stieltjes integrals, and noting that integration by u-substitution yields  $\int_{-\infty}^{\infty} G_2(y)dG_2(y) = \frac{1}{2}$ , we have

$$\begin{aligned} P(X < Y) &= \int_{-\infty}^{\infty} P(X < y)dG_2(y) = \int_{-\infty}^{\infty} F_1(y)dG_2(y) \\ &= \int_{-\infty}^{\infty} [G_2(y) + \Delta(y)]dG_2(y) \\ &= \frac{1}{2} + \int_{-\infty}^{\infty} \Delta(y)dG_2(y) \geq \frac{1}{2}. \end{aligned}$$

The inequality is established because  $\Delta(y) \geq 0$  by stochastic dominance. Note that by using the Lebesgue-Stieltjes integral, this proof applies to both discrete and continuous predictor variables.

**3.2 A Metric for Assessing the Discriminatory Power of the Predictor Variables**

The method we propose provides a discriminatory power measure  $A_t$  for each predictor variable  $t$  that assesses its ability to distinguish between the fraud and nonfraud class. If claim files in the fraud class all tend to have lower response values to variable  $t$  and claim files in the nonfraud class all tend to respond higher to this variable, variable  $t$  is considered to discriminate well, as will be reflected by a high value of  $A_t$ . We formally define  $A_t$  below.

Consider any predictor variable  $t$ . Take a (random) sample of size  $N$  consisting of  $N_1$  claim files from class 1 and  $N_2$  claim files from class 2. The expected proportion  $\theta$  of class 1 (i.e., the percentage of fraudulent claims, or the fraud rate) is  $\theta = E[N_1/N]$ , where  $N_1 \sim \text{Binomial}(N, \theta)$ . The expected population distribution is  $F(x) = \theta F_1(x) + (1 - \theta)G_2(x)$ , where  $F_1(x)$ ,  $G_2(x)$  are the variable distribution functions for class 1 and class 2, respectively. Again let  $\Delta(x) = F_1(x) - G_2(x)$ . Let  $B = B(X)$  denote the variable score corresponding to response value  $X$ . Note that  $B$  depends on  $X$  (the value of the particular response value) and on the rest of the data set  $D = (X_1, X_2, \dots, X_N)$ , that is, it is a function  $B(X; D)$ . Then the expected variable score for a member of class 1 with response value  $X$  on predictor variable  $t$  can be calculated as

$$\begin{aligned} E[B_t | \text{class 1}] &= E[B(X, D) | X \in D, X \in \text{class 1}] \\ &= \int_{-\infty}^{\infty} E_{N_1}\{E_D[B(x, D) | x \in D, N_1]\}dF_1(x) \\ &= \int_{-\infty}^{\infty} E_{N_1}\left[2\left(\frac{N_1 F_1(x)}{N} + \frac{N_2 G_2(x)}{N}\right) - 1\right]dF_1(x) \\ &= \int_{-\infty}^{\infty} [2(\theta F_1(x) + (1 - \theta)G_2(x)) - 1]dF_1(x) \\ &= 2 \int_{-\infty}^{\infty} [F_1(x) - (1 - \theta)G_2(x)]dF_1(x) - 1 \\ &= 2(\theta - 1) \int_{-\infty}^{\infty} \Delta(x)dF_1(x). \end{aligned} \tag{3.1}$$

**Definition 1**

The discriminatory power measure  $A_t$  (for variable  $t$ ) is defined by  $A_t = 2 \int_{-\infty}^{\infty} \Delta(x)dF_1(x) = 2 \int_{-\infty}^{\infty} \Delta(x)dG_2(x)$ .

Thus by (3.1), we have

$$E[B_t|\text{class 1}] = (\theta - 1)A_t < 0.$$

We show below that  $2 \int_{-\infty}^{\infty} \Delta(x)dF_1(x) = 2 \int_{-\infty}^{\infty} \Delta(x)dG_2(x)$ , by applying integration by parts and simply noting that the product of  $F_1(\cdot)$  and  $G_2(\cdot)$  is still a cumulative distribution function. Noting that  $\int_{-\infty}^{\infty} F_1(x)dF_1(x) = \frac{1}{2}$ , we have

$$\begin{aligned} 2 \int_{-\infty}^{\infty} \Delta(x)dF_1(x) &= 2 \int_{-\infty}^{\infty} [F_1(x) - G_2(x)]dF_1(x) = 1 - 2 \int_{-\infty}^{\infty} G_2(x)dF_1(x) \\ &= 1 - 2[F_1(x)G_2(x)]|_{-\infty}^{\infty} + 2 \int_{-\infty}^{\infty} F_1(x)dG_2(x) = 2 \int_{-\infty}^{\infty} F_1(x)dG_2(x) - 1 \\ &= 2 \int_{-\infty}^{\infty} F_1(x)dG_2(x) - 2 \int_{-\infty}^{\infty} G_2(x)dG_2(x) = 2 \int_{-\infty}^{\infty} \Delta(x)dG_2(x). \end{aligned}$$

A calculation similar to (3.1) yields

$$E[B_t|\text{class 2}] = 2\theta \int_{-\infty}^{\infty} \Delta(x)dG_2(x) = \theta A_t > 0.$$

Note additionally that

$$E[B_t|\text{class 2}] - E[B_t|\text{class 1}] = A_t > 0.$$

Thus, the value of  $A_t$  can be derived by calculating the conditional mean variable score for either class of claim files.

To see the interpretation of  $A_t$  as “discriminatory power” of variable  $t$ , note that the integrand  $\Delta(x)$  measures how “spread apart” the variable distributions  $F_1(x)$  and  $G_2(x)$  are from each other. No matter what the base rate  $\theta$  is, the difference between the expected scores in the nonfraudulent claim class and in the fraudulent claim class is  $A_t$ , so a predictor variable that differentiates well between the two classes will have a large positive value for  $A_t$ , whereas a poorly differentiating variable will have a small value for  $A_t$ . Also observe that if  $X$  is from class 1 and  $Y$  is from class 2, then  $P(X < Y) - \frac{1}{2}$  measures how different the class responses are, and  $A_t = 2[P(X < Y) - \frac{1}{2}]$ .

In the discrete case, substituting sums for the previous Lebesgue-Stieltjes integrals and rearranging terms, we calculate that  $A_t$  in Definition 1 is equivalent to

$$A_t = \sum_{i=1}^{k_t-1} \sum_{j>i} \{ \pi_{it}^{(1)} \pi_{jt}^{(2)} - \pi_{it}^{(2)} \pi_{jt}^{(1)} \}, \tag{3.2}$$

where for  $q = 1$  and  $2$ ,  $\pi_t^{(q)} = (\pi_{t1}^{(q)}, \dots, \pi_{tk_t}^{(q)})$  is the multinomial response probability vector of variable  $t$  for class  $q$ ,  $i = 1, \dots, k_t$  and  $j = 1, \dots, k_t$  are the indices for response categories of variable  $t$ , and  $k_t$  denotes the number of possible response categories for variable  $t$ . This is a measure developed in Brockett et al. (2002). We provide a brief proof of this equivalence below.

**PROOF**

Using Lebesgue-Stieltjes integrals and our previous notations, we rewrite the representation (3.2) of  $A_t$  for the discrete case as

$$A_t = \int_{-\infty}^{\infty} \int_x^{\infty} f_1(x)g_2(u)dudx - \int_{-\infty}^{\infty} \int_x^{\infty} g_2(x)f_1(u)dudx$$

for the continuous case where  $f_1(x)$  and  $g_2(x)$  are density functions (of variable  $t$ ) for class 1 and class 2, respectively. Then we have

$$\begin{aligned}
 A_t &= \int_{-\infty}^{\infty} \int_x^{\infty} f_1(x)g_2(u)dudx - \int_{-\infty}^{\infty} \int_x^{\infty} g_2(x)f_1(u)dudx \\
 &= \int_{-\infty}^{\infty} [1 - G_2(x)]dF_1(x) - \int_{-\infty}^{\infty} [1 - F_1(x)]dG_2(x) \\
 &= \int_{-\infty}^{\infty} [1 - G_2(x)]dF_1(x) - \left[ 1 - \int_{-\infty}^{\infty} F_1(x)dG_2(x) \right].
 \end{aligned}$$

Noting that the product of  $F_1(\cdot)$  and  $G_2(\cdot)$  is still a distribution function, we have

$$A_t = \int_{-\infty}^{\infty} [1 - G_2(x)]dF_1(x) - \left[ F_1(x)G_2(x)|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} F_1(x)dG_2(x) \right].$$

Then integrating by parts, we have

$$\begin{aligned}
 A_t &= \int_{-\infty}^{\infty} 1dF_1(x) - \int_{-\infty}^{\infty} G_2(x)dF_1(x) - \int_{-\infty}^{\infty} G_2(x)dF_1(x) \\
 &= 1 - \int_{-\infty}^{\infty} G_2(x)dF_1(x) - \int_{-\infty}^{\infty} G_2(x)dF_1(x).
 \end{aligned}$$

Noting again that  $\int_{-\infty}^{\infty} F_1(x)dF_1(x) = \frac{1}{2}$ , we have

$$\begin{aligned}
 A_t &= \int_{-\infty}^{\infty} [F_1(x) - G_2(x)]dF_1(x) + \int_{-\infty}^{\infty} [F_1(x) - G_2(x)]dF_1(x) \\
 &= 2 \int_{-\infty}^{\infty} \Delta(x)dF_1(x) = A_t.
 \end{aligned}$$

Therefore, we have established the equivalence between the  $A_t$  measure in Definition 1 and its alternative representation (3.2) for the discrete case. It follows that  $E[B_t|\text{class } 1] = (\theta - 1)A_t$  in the discrete case as well. The quantity inside the brackets in (3.2) can in fact be recognized as a  $2 \times 2$  contingency table measure of association. From representation (3.2), we can also see that  $A_t$  measures the amount of dispersion between the two classes in the latent dimension. Thus,  $A_t$  in Definition 1 extends the measure  $A_t$  utilized in Brockett et al. (2002) to general ordinal and continuous predictor variables. Next, we develop a set of variable weights which capture the importance of the predictors as reflected by  $A_t$ . However, unlike the theoretical probabilities involved in the development of  $A_t$ , these weights can be obtained in an unsupervised environment without knowing the class membership of each claim.

### 3.3 Combining Variable Weights and Variable Scores to Obtain Classification

We use an iterative scheme to motivate the determination of the weights. Essentially, we obtain weights by iteratively correlating individual variable scores with overall claim file summative scores to assess the contribution or the importance of each predictor variable. More specifically, the individual predictor variable score  $B_{it}$  for claim  $i$  on variable  $t$  is obtained as described previously (e.g., eq. [2.1] and [2.2]). These are then weighted and summed to obtain a summative suspicion score for an overall assessment of each claim. Equal weights for the predictor variables are first used to get an initial set of summative scores, and these weights are then updated to take into account each variable’s degree of “importance,” as measured by its consistency with the overall claim assessment. This updating process is executed iteratively, and the weights are successively refined until convergence takes place.

Formally in vector notation, an initial summative score matrix  $S^{(0)}$  is obtained by applying equal weights  $\hat{W}^{(0)}$  (a vector of 1’s) to the score matrix  $F$ , the  $(i, t)$ -th element of which is claim file  $i$ ’s score on variable  $t$ , or  $B_{it}$ , that is,  $S^{(0)} = F\hat{W}^{(0)}$ . Once we have the predictor variable  $t$  scores and an initial overall summative score for each claim file  $i$ , we can “correlate” the overall claim file scores

with the predictor variable  $t$  scores to assess how consistent the variable  $t$  scores are with the overall scores. If a large variable  $t$  score goes with a large overall score, this is a better predictor variable than if there is no association between the variable  $t$  score and the overall score. More specifically, the initial weight vector is updated by obtaining a new consistency weight,  $\hat{W}^{(1)} = \hat{F}^T S^{(0)} / \|\hat{F}^T S^{(0)}\|$ , where “ $\|\cdot\|$ ” represents the Euclidian distance and the denominator is a normalization factor. We can then use this set of consistency weights  $\hat{W}^{(1)}$  to weight the variables in the summation, obtaining a new vector of weighted overall claim file scores  $S^{(1)} = F\hat{W}^{(1)}$ . The individual variable scores can again be “correlated” with this new set of overall scores to obtain another new set of consistency weights, that is,  $\hat{W}^{(2)} = F^T S^{(1)} / \|F^T S^{(1)}\|$ . We continue this process of iteratively constructing new weights and new overall scores. At stage  $n$ ,  $\hat{W}^{(n)} = F^T S^{(n-1)} / \|F^T S^{(n-1)}\| = (F^T F)^{(n)} \hat{W}^{(0)} / \|(F^T F)^{(n)} \hat{W}^{(0)}\|$ .

Weights are iteratively updated until convergence is secured. Analogous to Theorem 1 in Brockett et al. (2002), being successive normalized powers of the matrix  $F^T F$ , the vector of successive weights  $\hat{W}^{(n)}$  converge to a limiting vector  $\hat{W}^{(\infty)}$ , which is the first principal component of the matrix  $F^T F$ . The matrix  $F^T F$  is the sample covariance matrix of the variable scores because the variable scores are centered with mean zero. Importantly, the final predictor variable weight  $\hat{W}_t^{(\infty)}$  turns out to be strongly related to each variable’s discriminatory power measure  $A_t$ , further justifying their use as variable weights.

Specifically, the limiting predictor variable weight vector  $\hat{W}^{(\infty)}$  is the first principal component of  $F^T F$ , which is a consistent estimate of the first principal component  $W^{(\infty)}$  of  $E[F^T F]$ ,<sup>1</sup> the  $t$ th component of which is

$$W_t^{(\infty)} = \frac{A_t}{(\mu_1 - U_{tt}) \sqrt{\sum_{s=1}^m A_s^2 / (\mu_1 - U_{ss})^2}},$$

where  $\mu_1$  is the largest eigenvalue of  $E[F^T F]$  and  $U_{tt} = N_1 \sigma_{1t}^2 + N_2 \sigma_{2t}^2$  is the “uniqueness component of variance” in a single-factor analytic model (for class  $q = 1, 2$ ,  $\sigma_{qt}^2 = \text{variance } [B_{qt}]$ , and  $B_{qt}$  is the score of a randomly selected claim from class  $q$  on variable  $t$ ).<sup>2</sup>

As shown above, although the iterative scheme provides the intuition behind the weights, we directly calculate these weights from the first principal component of  $F^T F$  without needing the iterative process. Note that in calculating the limiting weights  $\hat{W}^{(\infty)}$ , supervised information such as class membership and class distribution is not used.

Each claim’s weighted summative score, calculated by applying the limiting weights  $\hat{W}_t^{(\infty)}$  to individual predictor variable scores  $B_{it}$ , measures the claim file’s potential of being in the fraud class. Claims can be classified based on the sign of the summative scores because the individual variable scores are centered at 0, and summative scores are linear combinations of variable scores. By the assumption that lower responses to the predictor variables indicate higher potential of fraud, claim files with negative summative scores will be classified accordingly into the “fraud” class. If the fraud rate is known,<sup>3</sup> claims can be classified accordingly based on the ranking of the summative scores. Because the proposed technique uses a principal component analysis of RIDIT-based scores for predictor variables, the classification technique is called PRIDIT analysis. The ultimately determined weighted summative score for each claim file is referred to as the PRIDIT score for the claim file.

<sup>1</sup> The sample covariance matrix  $F^T F$  is a consistent unbiased estimate of  $E[F^T F]$ , so it can be shown that  $\hat{W}^{(\infty)}$  is a consistent unbiased estimate of  $W^{(\infty)}$ .

<sup>2</sup> It can be proven that the discriminatory power measure  $A_t$  is linearly related to the expectation of the Wilcoxon rank sum statistic for each class. The (nonparametric) Wilcoxon rank sum statistic has traditionally been used to measure deviation between two classes, and this has further validated  $A_t$  as a discriminatory power measure. The variable weights (essentially proportional to  $A_t$ ’s) can then be interpreted in terms of the Wilcoxon rank sum statistic and can also be used as inputs in further analyses.

<sup>3</sup> Even in the unsupervised context, we may have an estimate of the fraud rate available for use. See Ai et al. (2009).

The weights, as the first principal component, give the combination that maximizes the variance in the data. Because of the fact that the predictor variables were selected by experts to monotonically assess positioning of the claim file on an underlying latent dimension of “fraud suspicion,” the dimension of highest variance should be the underlying latent fraud suspicion. Note that if a situation occurs in which the actual relation of a predictor variable with nonfraud is negative instead of positive (because of expert error or incorrect construction of the variable), because of the correlative structure in the iterative process, it will result in this variable having a negative correlation with the overall summative score, and hence a negative value for  $A_i$ . In the summation of weighted variable scores, the negative value for  $A_i$  corrects for the negative relationship in the individual variable score  $B_i$  calculation, resulting in a correctly signed contribution of this variable to the overall PRIDIT score. Similarly, if a chosen predictor variable does not actually predict fraud, the value of  $A_i$  should be close to zero, and this variable will have little or no impact when the weighted variable score is used to calculate the summative PRIDIT score.

## 4. AN INSURANCE FRAUD DETECTION APPLICATION

A difficult question with regard to unsupervised classification methods is how do you know that the results you are getting have anything to do with the goal of performing the unsupervised learning? In particular, by what standards do you judge that an unsupervised method has “succeeded” in aiding the fraud detector, or that one unsupervised method is better than another? One method of doing this is to spend a lot of time and money to obtain expert fraud assessors to create a data set in which an ultimate fraud label is determined, and then judge the unsupervised method according to concurrence with the expert. We use this assessment method and present the results in Sections 4.2 and 4.3. We also compare the proposed PRIDIT method to competitor unsupervised methods in Section 4.5.

A next question is to what extent any performance failures of an unsupervised method are due to the method itself versus the extent to which the failures are unrelated to the unsupervised method and would be problematic for any method (e.g., data set problems such as highly overlapping data). To further assess the reasonableness of the PRIDIT classification performance we look at the accuracies that are possible with a “standard set” of supervised methods for comparison with the PRIDIT method. This provides an upper bound performance benchmark against which to measure PRIDIT accuracy for the data set and provides a type of external validation relative to more informationally demanding supervised learning methods when there is a set of known classifications. Detailed discussions are in Section 4.4.

### 4.1 Data Description

We use an insurance claims data set to examine performance of the proposed method. This data set, produced by the Automobile Insurance Bureau (AIB) in Massachusetts (Viaene et al. 2002), contains 1,399 personal injury protection (PIP) claims for automobile insurance, each of which was assessed by insurance adjusters and experts in the special investigation unit of AIB. Each claim was given a suspicion score on a 0-to-10 point scale by each expert and was assigned a code indicating whether the suspicion score was above 1, above 4, or above 7. An operational definition of fraud is used here to classify any claim with a suspicion score greater than or equal to 4 to be fraudulent (Viaene et al. 2002). This data set contains a set of binary fraud predictor variables selected by the AIB claims experts using their subject area expertise. The entire claims set is divided into two data sets A and B using a 50–50 random sample split.

### 4.2 The PRIDIT Analysis for Data Set A

To perform PRIDIT analysis, we first calculate predictor variable scores for each claim file in data set A. From equation (1), category  $i$  of a predictor variable  $t$  has variable score  $B_{it}$ . For each claim file, a weighted summative score is then calculated with the weights determined using the methodology de-

Table 1  
**PRIDIT Classification Accuracies against  
 Expert-Assessed Fraud Classification  
 (Data Set A)**

	Fraud Class	Nonfraud Class	Overall
Accuracy	77.30%	57.67%	62.86%
Sample fraud rate		26.43	

scribed in Section 3. Because the proportion of fraudulent claims is not known a priori, we adopt the classification rule based on the sign of the weighted PRIDIT summative scores, that is, claims with a negative PRIDIT score are assigned to the fraud class, and claims with a positive PRIDIT score are assigned to the nonfraud class. This is based on the variable construction that lower response values suggest higher potential of fraud.

Table 1 shows the PRIDIT classification accuracies for data set A evaluated against the expert fraud assessment. Although the overall accuracy is moderate, PRIDIT performs well in classifying the fraud class. The fraud class, which accounts for 26.43% of data set A, is the minority class so classifying all claims to the majority class would result in a 73.57% accuracy. This does not mean, however, that the PRIDIT accuracy of 62.86% is unacceptable because the goal of fraud detection is to correctly classify *both* classes, especially the minority class (fraud class). To achieve this, the PRIDIT method “defies the odds by classifying an individual in the smaller group” (Morrison 1969, p. 158).

The purpose of fraud detection algorithms is to identify cases that warrant transmission to a special investigation unit for further assessment. The vast majority of cases can be dispensed with by payment, and others dispensed with once a further investigation is undertaken. However, one does not want to overlook a potential fraud case, so fraud class accuracy is more important than overall accuracy. Cost-related benefits usually are also found with this focus on fraud class (the minority and event class) because misclassification cost is usually higher for the event class than for the nonevent class. In fraud detection, wrongly classifying a legitimate claim to the fraud class may just result in unnecessary auditing cost (and perhaps some ill will), whereas missing a fraudulent claim could lead to a substantial amount of unnecessary claim payments being made by the insurance company.

### 4.3 Evaluating PRIDIT in Data Set B

As an unsupervised method, PRIDIT analysis does not require or utilize knowledge of the known fraud classification (or any dependent variable), so it is possible to assess its performance directly using data set A by comparing the PRIDIT classification with the (ostensibly) known fraud classification, if available. However, applying the PRIDIT model obtained using data set A to classify claim files in data set B allows us to evaluate the stability of the PRIDIT scores and classification algorithm in additional samples of claims.

To obtain the PRIDIT classification for data set B, we use the individual variable scores and the variable weights calculated within data set A. The calculated variable scores are first assigned directly to each claim file in data set B according to its respective response to each predictor variable. These variable scores are then concatenated to derive a summative PRIDIT score for each claim file using the weights obtained from data set A. As was done for data set A, claim files with negative summative scores are assigned to the fraud class, and claim files with positive summative scores are assigned to the nonfraud class. Therefore, rather than conducting a new set of PRIDIT analysis for data set B, we take variable scores and weights already developed from data set A to classify claims in data set B, much in the same way as logistic regression analysis would estimate parameters from a training set A to score the data in a test set B (but without using fraud labels in either data set A or data set B).

Table 2 presents the PRIDIT classification results for data set B evaluated against expert-assessed fraud classification. The classification accuracies are generally similar to those on data set A, suggesting

Table 2  
**PRIDIT Classification Accuracies against  
 Expert-Assessed Fraud Classification  
 (Data Set B)**

	Fraud Class	Nonfraud Class	Overall
Accuracy	81.52%	60.25%	66.67%
Sample fraud rate		30.19	

that the PRIDIT method exhibits good generalizability and robustness. Thus in practice, we need not run separate PRIDIT analysis when analyzing data sets suspected to have a similar pattern (e.g., claims collected in the same time period and in the same region, such as data sets A and B in the current analysis).<sup>4</sup> We can run new PRIDIT analysis only when necessary (e.g., when structural changes in the fraud pattern occur as fraudsters learn to avoid detection).

#### **4.4 External Validations Benchmarking against Supervised Learning Methods: How Much Is Lost by Not Knowing the Classification Variable for Even a Subset?**

##### **4.4.1 Empirical Comparison Results**

In this section we perform an external validation of the unsupervised method PRIDIT using a “standard set” of supervised methods, including logistic regression, Support Vector Machines (Cristianini and Shawe-Taylor 2000), and Bayesian Additive Regression Trees (Chipman, George, and McCulloch 2006). These methods are selected because of their wide applicability and documented performance. In particular, BART is a recent predicting technique that has been shown to generate more accurate predictions than other prevalent supervised learning methods in a variety of applications (Chipman, George, and McCulloch 2006).<sup>5</sup> These supervised methods are *not* competitors for PRIDIT because supervised methods require more and different information.

The supervised methods are trained on data set A and assessed on data set B.<sup>6</sup> Each of the supervised methods is first evaluated against the “known” expert fraud assessment and then compared with the PRIDIT classification of claims presented in Tables 1 and 2. Cross-classifications between each supervised method and the PRIDIT method are presented to examine the individual claim file classification consistency. We also calculate Pearson correlation and Spearman rank correlation between the scores produced by supervised methods and by PRIDIT to assess the consistency of the predicted relative rankings underlying the fraud classification.

In the comparative analysis, a score of 0 is used as the threshold value for PRIDIT; however, for the supervised classifiers, we classify claims in both data sets A and B according to a data-driven threshold, that is, the fraud rate of the training data set A, because this information is available in the supervised

<sup>4</sup> Note that even when data set B’s structure (e.g., the fraud rate) is different from that of data set A, PRIDIT can still generate reasonably good results because of its nonparametric nature (see Bross 1958 and Golden and Brockett 1987 for more comments regarding the robustness). This was also confirmed by other experiments we have run but have not presented here.

<sup>5</sup> BART was originally designed to predict a continuous variable. The continuous predictions generated by BART in the case of a binary dependent variable can be viewed as similar to the class probability estimates produced by logistic regression. A modified version of BART specifically designed for a classification task has been proposed (e.g., Abu-Nimeh et al. 2008). It is shown, however, that the classification performance of BART is satisfactory and quite comparable to the modified version, so we still use the original BART.

<sup>6</sup> In SVM, the Gaussian (or Radial Basis Function) Kernel is selected, and parameters are tuned using 10-fold cross validation of the training set A. Default parameters are used for BART (Chipman et al. 2008).

learning context. We also consider a default threshold value of probability level 0.5 for the supervised classifiers. Classification results of both data sets along with the correlations are shown in Tables 3, 4, and 5.

Table 3 compares the classification accuracies of PRIDIT and the supervised methods evaluated against the expert fraud assessment. The performances on data set A and data set B are similar. We can see from Table 3 that PRIDIT is relatively more accurate on the fraud class, whereas supervised methods tend to have higher overall and nonfraud class accuracies. Note, however, that supervised methods require fraud labels obtained at a cost to derive a trained classification model whereas PRIDIT does not.

We also calculate and present in Table 3 the Area under the Receiver Operating Characteristic Curve (ROC Curve) measure (AUC) and the F1-measure for these methods. The ROC curve plots the false positive rates against the true positive rates under various threshold values to allow visualization of the trade-off between these two rates. Simply speaking, an AUC measure serves as an estimate of each classifier’s prediction quality (see Viaene et al. 2002 for a discussion in the context of fraud detection). The F1 measure is the harmonic mean of precision (the number of correctly classified fraudulent claims/the total number of claims classified as fraudulent) and recall (fraud class accuracy) and gives specific attention to the performance on the fraud class. None of the above measures by itself is sufficient to assess the performance of different methods because they focus on different aspects of the performance. The ensemble of accuracies, AUC measure, and F1 measure, however, helps provide a validation for the PRIDIT method.

**Table 3**  
**Accuracies, AUC, and F1 Measure for Different Classification Techniques**

	Fraud Class Accuracy	Nonfraud Class Accuracy	Overall Accuracy	AUC <sup>a</sup>	F1 <sup>b</sup>
Panel A: Data set A					
Panel A.1: Classification threshold $p = 0.5$					
PRIDIT (threshold = 0)	77.30%	57.67%	62.86%	72.95%	52.38%
LR	29.19	94.17	77.00	77.70	40.15
BART	20.00	97.86	77.29	77.18	31.76
SVM	21.08	99.42	78.71	84.00	34.36
Panel A.2: Classification threshold based on the sample fraud rate of 26.43%					
PRIDIT (threshold = 0)	77.30	57.67	62.86	72.95	52.38
LR	51.89	82.72	74.57	77.70	51.89
BART	51.89	82.72	74.57	77.18	51.89
SVM	61.08	86.02	79.43	84.00	61.08
Panel B: Data set B					
Panel B.1: Classification threshold $p = 0.5$					
PRIDIT (threshold = 0)	81.52%	60.25%	66.67%	76.69%	59.62%
LR	32.23	93.65	75.11	77.53	43.87
BART	19.43	96.31	73.10	77.36	30.37
SVM	16.59	97.13	72.82	75.39	26.93
Panel B.2: Classification threshold based on the sample fraud rate of 26.43%					
PRIDIT (threshold = 0)	81.52	60.25	66.67	76.69	59.62
LR	51.18	84.22	74.25	77.53	54.54
BART	51.18	84.22	74.25	77.36	54.54
SVM	53.08	85.04	75.39	75.39	56.57

Note: In SVM, Gaussian (or Radial Basis Function) kernel is used and parameters are tuned to be gamma = 0.01, cost = 3. All parameters take the default values in BART. In panels A.1 and B.1, a threshold value of  $p = 0.5$  is used for LR, BART, and SVM classification. In panels A.2 and B.2, data set A fraud rate (26.43%) is used for LR, BART, and SVM classification for both data sets A and B. A threshold of 0 is used for PRIDIT classification in all analyses.

<sup>a</sup> AUC refers to the Area under the Receiver Operating Characteristic Curve.

<sup>b</sup> F1 measure is a measure of classification accuracy that pays particular attention to the fraud class.

Table 4  
**Cross-Classifications between PRIDIT and Supervised Learning Methods**

	LR Nonfraud	LR Fraud	BART Nonfraud	BART Fraud	SVM Nonfraud	SVM Fraud	Total Count
Panel A: Data set A							
Panel A.1: Classification threshold $p = 0.5$							
PRIDIT nonfraud	337	2	339	0	336	3	339
PRIDIT fraud	279	82	313	48	322	39	361
Total count	616	84	652	48	658	42	700
Total agreement	59.86%		55.29%		53.57%		
Panel A.2: Classification threshold based on the sample fraud rate of 26.43%							
PRIDIT nonfraud	332	7	337	2	325	14	339
PRIDIT fraud	183	178	178	183	190	171	361
Total count	515	185	515	185	515	185	700
Total agreement	72.86%		74.29%		70.86%		
Panel B: Data set B							
Panel B.1: Classification threshold $p = 0.5$							
PRIDIT nonfraud	329	4	333	0	331	2	333
PRIDIT fraud	271	95	307	59	319	47	366
Total count	600	99	640	59	650	49	699
Total agreement	60.66%		56.08%		54.08%		
Panel B.2: Classification threshold based on the sample fraud rate of 26.43%							
PRIDIT nonfraud	323	10	329	4	322	11	333
PRIDIT fraud	191	175	185	181	192	174	366
Total count	514	185	514	185	514	185	699
Total agreement	71.24%		72.96%		70.96%		

Note: In SVM, Gaussian (or Radial Basis Function) kernel is used, and parameters are tuned to be gamma = 0.01, cost = 3. All parameters take the default values in BART. In panels A.1 and B.1, a threshold value of  $p = 0.5$  is used for LR, BART, and SVM classification. In panels A.2 and B.2, data set A fraud rate (26.43%) is used for LR, BART, and SVM classification for both data sets A and B. A threshold of 0 is used for PRIDIT classification in all analyses.

To further compare these methods, Table 4 presents PRIDIT classification against classifications produced by supervised methods. We can see that they agree on the assessment of the claims for the majority (55–75%) of the time.

Besides examining the binary fraud classifications, we also investigate the consistency of the suspicion-level assessment of the claims by calculating the Pearson correlations and Spearman rank correlations of the predicted scores produced by PRIDIT and the supervised methods. The correlations of the predicted scores may reveal more information than the comparison of the binary classifications. These correlations are presented in Table 5.

Table 5  
**Correlations between Predicted Scores by PRIDIT and Supervised Learning Methods**

	Pearson Correlation ( $p$ Value)	Spearman Rank Correlation ( $p$ Value)
Panel A: Data set A		
Between logistic regression and PRIDIT Scores	0.7543 (<0.0001)	0.7912 (<0.0001)
Between BART and PRIDIT	0.8793 (<0.0001)	0.8815 (<0.0001)
Between SVM and PRIDIT	0.5526 (<0.0001)	0.6875 (<0.0001)
Panel B: Data set B		
Between LR and PRIDIT	0.7819 (<0.0001)	0.8377 (<0.0001)
Between BART and PRIDIT	0.9040 (<0.0001)	0.9057 (<0.0001)
Between SVM and PRIDIT	0.6375 (<0.0001)	0.7351 (<0.0001)

Note: In SVM, Gaussian (or Radial Basis Function) kernel is used and parameters are tuned to be gamma = 0.01, cost = 3. All parameters take the default values in BART. The correlations do not depend on the classification threshold chosen for the models.

From Table 5 we can see that PRIDIT scores for the claims and predicted probabilities by supervised methods in both the data set A and the data set B are highly significantly correlated. This suggests that although the PRIDIT method uses only the consistency of the internal structure of the predictor variables, its predictions of the suspicion levels are similar to those of the more information-intensive supervised methods.

#### 4.4.2 Strengths and Weaknesses of Alternative Methods

The empirical results suggest that the unsupervised PRIDIT method and the set of supervised methods make consistent assessment of the suspicion levels of claims, validating the effectiveness of PRIDIT in fraud detection. Supervised learning methods generally demonstrate superior accuracies overall and on the majority (nonfraud) class, whereas the PRIDIT method has strength in predicting fraudulent claims.

The relative strengths of these methods result from their different underlying mechanisms. Supervised learning methods are optimized on a sample of audited claims for which fraud labels have been obtained (usually at an additional cost), whereas the PRIDIT method uncovers the pattern directly using the consistency of the predictor variables. If no consistent set of predictors is available, PRIDIT may not perform as well.<sup>7</sup>

Also, the PRIDIT method can be used when the applicability of the standard supervised methods is limited by various factors that may exist in the context of insurance fraud detection,<sup>8</sup> for example, when it is cost-prohibitive or impossible to obtain a training sample of audited claims. This could arise for a smaller-scale company or in an emerging or rapidly changing marketplace where expert and monetary resources are limited or temporally unstable. PRIDIT may also be useful when this data set has a severely uneven class distribution that might adversely affect supervised learning methods' performance, especially on the minority class. Finally, PRIDIT can be a valuable alternative when the fraud labels in the audited training sample contain a large amount of error because they are only a best assessment of fraud by fallible human experts (see Brockett, Xia, and Derrig 1998, which shows that there is substantial interjudge variation in automobile insurance fraud assessment). In this situation the performance of supervised learning methods may be contaminated<sup>9</sup> (Hausman, Abrevaya, and Scott-Morton 1998), whereas the PRIDIT method is relatively free of this bias because these labels are not employed.

#### 4.5 Comparison with Competitor Unsupervised Learning Methods

Two competitor unsupervised learning methods are cluster analysis and Kohonen's self-organizing map (Kohonen 1989). Cluster analysis is perhaps one of the most used unsupervised classification methods. Under this class of models, data are partitioned according to certain "similarity" or "dissimilarity" measures. In our comparisons we use the Two Step Clustering Procedure from SPSS, which first uses sequential pre-clustering and then a model-based measure ("decrease in log likelihood") to form the

<sup>7</sup> Supervised learning methods may be combined among themselves (see "ensemble learning," Dietterich 2000) or with unsupervised learning methods, such as the PRIDIT method, to achieve better performance. A detailed discussion of these methods is beyond the scope of this paper, but we have a working paper in progress to investigate the benefits of the latter combination.

<sup>8</sup> Special classes of supervised learning methods that aim at alleviating the impact of these factors, such as those dealing with imbalanced data set and/or noisy training labels, may alternatively be deployed or designed for this purpose as well.

<sup>9</sup> Although label inaccuracies are largely suspected in insurance fraud data sets (e.g., see Brockett, Xia, and Derrig 1998), we cannot examine method robustness to data errors directly using an insurance fraud data set because the (possibly incorrect) labels are already intrinsically given in such a data set. Instead, we designed an experiment using a public domain database ADULT with known accurate income classification (see Section 5.1 for data description) and a set of PRIDIT-assessed class memberships as the dependent variable values to investigate the performance of a representative supervised method, logistic regression, i.e., we are creating a "thoughtful inaccuracy" that emulates the type of "thoughtful inaccuracy" present in subjective expert assessments of fraud. We found that logistic regression is successful only in predicting the partially inaccurate PRIDIT labels but not the true labels, which suggests that supervised methods are sensitive to the quality of the labels in the training data. The results are not reported to conserve space and are available upon request.

Table 6  
**PRIDIT Classification Accuracies against Expert-Assessed Fraud Classification (Entire Data Set)**

	Fraud Class	Nonfraud Class	Overall
Accuracy	79.29%	58.52%	64.40%
Sample fraud rate		28.31	

desired number of clusters (SPSS 2001). Kohonen’s self-organizing map essentially discovers the mapping between the inputs (fraud predictor variables) and the outputs (leading to fraud suspicion levels) by iteratively updating the weight vectors for the outputs. Because both methods are unsupervised, the entire insurance fraud data set is used to make this comparison. The PRIDIT classification for the entire data set is presented in Table 6.

The classification by cluster analysis is presented in Table 7. One major concern with cluster analysis is that it is a clustering method rather than a classification method. Hence, no certified assignment of each cluster is made to each predefined class. To provide a fair comparison with the PRIDIT analysis, we assign clusters to the desired fraud and nonfraud class based on the number of positive responses to individual predictor variables in each cluster. This allows cluster analysis to use essentially the same information as PRIDIT, that is, a stochastic dominance assumption that positive responses suggest higher potential of being fraud. One additional concern is that two clusters are not guaranteed for this binary classification task. In fact, according to the “Bayesian information criteria plus distance change” criterion employed by the SPSS two-step cluster analysis (SPSS 2001), the optimal number of clusters for this data set is four, and thus we need to combine the clusters for this classification task.

More specifically, for this analysis we employ two approaches to match clusters to the two classes. First, we produce two clusters and match by assigning the cluster with more positive responses to the fraud class and the other to the nonfraud class. The classification accuracy is presented in panel A of Table 7. We can see that the resulting classification accuracies exhibit a pattern similar to those of PRIDIT with a lower overall accuracy and nonfraud class accuracy. Second, we produce the optimal number of four clusters and combine the clusters to maximize the positive responses in one of the two

Table 7  
**Cluster Analysis Classification Accuracies against Expert-Assessed Fraud Classification (Entire Data Set)**

	Fraud Class	Nonfraud Class	Overall
Panel A: Two clusters produced			
Accuracy	83.08%	48.45%	58.26%
Panel B: Four clusters produced and combined into two clusters			
Accuracy	84.34	43.77	55.25
Sample fraud rate		28.31	

*Note:* In panel A, the two clusters and the two classes are matched so that the cluster with more positive responses to predictor variables is designated as the fraud class. In panel B, the four clusters are combined to maximize the number of positive responses in one of the final two clusters, and this cluster is designated as the fraud class.

final clusters, which is then designated as the fraud class.<sup>10</sup> Panel B in Table 7 shows the classification accuracies under this approach. The fraud class accuracy is slightly increased with the nonfraud class accuracy and overall accuracy further reduced.

Table 8 shows the cross-classification between PRIDIT and cluster analysis under both matching approaches. We can see that the two unsupervised methods agree with each other on the majority of the claims.

Although PRIDIT and cluster analysis give consistent classification results for the insurance fraud data set, several difficulties hinder the use of cluster analysis in this type of classification tasks. First, as we discussed above, there is no guaranteed “correct” match between clusters and target classes (e.g., fraud and nonfraud class), especially when the optimal number of clusters does not equal the number of target classes.

Second, in cluster analysis, it is more difficult and less intuitive to determine variable importance and obtain variable weights in the numerical form,<sup>11</sup> making it difficult to interpret results and to incorporate them into further analyses. By contrast, PRIDIT assigns numerical weights to predictor variables and yields a summative PRIDIT score for each claim file based on which classifications are drawn. These numerical PRIDIT scores are easily correlated with exogenous variables and can also be used to rank the claim files to ascertain which claims should be further investigated and which should be paid promptly. Such a ranking is difficult with cluster analysis, and yet it is the most important component of the insurance fraud detection problem. Additionally, it is difficult to control the size of each cluster, which means that unlike PRIDIT we cannot make use of the fraud rate parameter even when it is known.

Another prevalent unsupervised classification method is the unsupervised Kohonen’s feature map. This method is widely applicable and has been previously used to identify fraudulent insurance claims (Brockett, Xia, and Derrig 1998). However, this method has two considerable drawbacks. First, this method produces ambiguous graph-based results. The interpretation of these graphic results usually lacks of objectivity and precision. It is also quite difficult to conduct any follow-up in-depth analysis based on these results. Second, this method is computationally intensive. As a result, although the

Table 8

**Cross-Classification between PRIDIT and Cluster Analysis (Entire Data Set)**

	Cluster Nonfraud	Cluster Fraud	Total Count
Panel A: Two clusters produced			
PRIDIT nonfraud	533	136	669
PRIDIT fraud	20	710	730
Total count	553	846	1,399
Total agreement		88.85%	
Panel B: Four clusters produced and combined into two clusters			
PRIDIT nonfraud	490	179	669
PRIDIT fraud	11	719	730
Total count	501	898	1,399
Total agreement		86.42%	

Note: In panel A, the two clusters and the two classes are matched so that the cluster with more positive responses to predictor variables is designated as the fraud class. In panel B, the four clusters are combined to maximize the number of positive responses in one of the final two clusters, and this cluster is designated as the fraud class.

<sup>10</sup> Because the fraud labels are available in this data set, one other approach is to match the clusters and the classes so as to maximize the overall classification accuracies of cluster analysis (however, notice that we are using information beyond the predictor variables in this case). For this data set, this alternative assignment with two clusters results in the same matching as presented in the text. However, the four clusters experiment results in high overall accuracy but low fraud class accuracy.

<sup>11</sup> The weights may be exogenously supplied and incorporated using a modified Mahalanobis distance type of formula for clustering (Mahalanobis 1936).

neural network method may be employed to provide some heuristic insights into the claims investigation process, it may not be a good candidate for insurance companies to establish an automated and consistent claims screening process for potential fraud cases. Interested readers are referred to Brockett, Xia, and Derrig (1998) for details of this method.

## 5. THE VALUE OF USING CONTINUOUS PREDICTOR VARIABLES AND ROBUSTNESS

Because all predictor variables could be made binary, the question arises as to why one might want to extend the method described in Brockett et al. (2002) for binary variables to ordinal discrete and continuous variables. This section shows that classification performance can be much reduced if continuous variables are made binary. The AIB insurance fraud data set used earlier contains only binary predictors, so here we use a different data set to exhibit the performance gain to be expected by incorporating continuous, ordinal discrete, and binary variables all in a single model. The ADULT database (a census income data set) from UCI Machine Learning Repository<sup>12</sup> will be used for this illustration. We also use this data set to show that the usefulness of the PRIDIT method extends beyond the insurance fraud detection context.

### 5.1 Data Description

Similar to fraud detection, the ADULT database aims at a binary classification task (an income level classification). It has a combination of continuous and categorical predictor variables as well as known income classifications and has been widely used to test different algorithms.<sup>13</sup>

This data set contains 45,222 individuals with their income classes. It has been predivided to data set C and data set D by a 2/3 (30,162 cases), 1/3 (15,060 cases) random split. There are 14 census variables that could be used as predictors of income class, six of which are continuous and eight of which are categorical. Because the PRIDIT method requires ordinal predictor variables, we select five continuous predictors and four ordinal categorical predictors to use in our analyses. The continuous predictors are *age*, *education*, *working hours per week*, *capital gain*, and *capital loss*; the categorical predictors are *sex*, *race*, *work class*, and *occupation*.<sup>14</sup>

The PRIDIT classifications of data sets C and D are presented in Table 9. Note that as for the insurance fraud data set, variable scores and variable weights are calculated within data set C and directly applied to classify data set D. The performance of PRIDIT on this data set is similar to that obtained on the insurance fraud data set.

Table 9  
**PRIDIT Classification Accuracies against True Income Classification**

	High-Income Class	Low-Income Class	Overall
Panel A: Data set C			
Accuracy Percentage of high-income individuals	71.16%	60.61% 24.89	63.24%
Panel B: Data set D			
Accuracy Percentage of high-income individuals	71.54%	60.48% 24.57	63.20%

<sup>12</sup> For details, see UCI Machine Learning Repository (2008).

<sup>13</sup> For details, see <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult/adult.names>.

<sup>14</sup> To ensure an intuitive rank order of responses for the categorical variables, we collapsed the original 14 categories and seven categories of responses for the predictors *occupation* and *work class* to four categories and five categories, respectively. We have also run a logistic regression to check the rank order of variables and found that the intuitive ordering is consistent with the logistic regression results. Detailed variable descriptions are in the data dictionary for the Current Population Survey from U.S. Census Bureau (2008).

## 5.2 Desire for a Unified Method: Effects of Discretizing Continuous Predictor Variables

In this subsection we examine the value of developing a unified PRIDIT method incorporating both continuous and ordinal categorical predictors. More specifically, we design an experiment in which continuous predictors are transformed into binary ones to investigate the potential performance loss. To single out the impact of continuous variables, only the five continuous variables are used for classification. Table 10 shows that high income class accuracies are lower as compared to using all predictors.

Based on intuitive cutoff criteria, these continuous predictor variables are turned into binary variables (the original monotonic relationship remains through this binary transformation). The specific cutoff points are listed in Table 11. Table 12 presents the PRIDIT performance using the transformed variables. Both the overall accuracy and the high-income class (event class) accuracy drop dramatically (10 percentage points and 20 percentage points, respectively) as compared to keeping the continuous

Table 10  
**PRIDIT Classification Accuracies Using Only Continuous Predictors**

	High-Income Class	Low-Income Class	Overall
Panel A: Data set C			
Accuracy Percentage of high-income individuals	66.60%	65.83% 24.89	66.02%
Panel B: Data set D			
Accuracy Percentage of high-income individuals	66.30%	64.77% 24.57	65.15%

Table 11  
**Cutoff Criteria for Transforming Continuous Predictors into Binary Predictors**

Predictor Variable	Age	Capital Gain	Capital Loss	Working Hours per Week	Education (Years)
Transformed variable Value = 1 if	<50	=0	=0	<40	<13
Transformed variable Value = 0 if	≥50	>0	>0	≥40	≥13

Table 12  
**PRIDIT Classification Accuracies Using Only Continuous Predictors Transformed into Binary Predictors**

	High-Income Class	Low-Income Class	Overall
Panel A: Data set C			
Accuracy Percentage of high-income individuals	44.14%	59.31% 24.89	55.54%
Panel B: Data set D			
Accuracy Percentage of high-income individuals	43.92%	58.68% 24.57	55.05%

predictors in their continuous form.<sup>15</sup> This experiment confirms the importance of having a classification method able to accommodate both types of predictor variables, thereby supporting the development of the unified PRIDIT method.

## 6. CONCLUSION

Brockett et al. (2002) presented an unsupervised classification method for predicting automobile insurance fraud and applied the method to a collection of binary “red flag indicator” variables. In this paper we have extended this method to incorporate all levels of ordinal predictor variables including binary, ordinal categorical variables with various different numbers of response categories, and continuous variables. We have shown that this extension improves the performance of the method. More specifically, after extending the scoring method known as RIDIT scoring from the epidemiological literature (Bross 1958) to accommodate continuous as well as ordinal categorical variables, we show how to concatenate the predictor variables so as to obtain measures of predictor variable importance, and how to obtain an overall claim file measure of fraud suspicion. The proofs of the main results are obtained in the general case using Lebesgue-Stieltjes integration, which extends as well as simplifies previous proofs.

By developing a data-based classification function that can be applied to existing and new claims, fraud examiners can continuously and in an automated manner monitor fraud levels (useful because of the huge number of claims files daily) and signal suspicious claims as they come in. The metric of overall claim file fraud suspicion level or risk allows the fraud investigator to rank claims to best allocate resources to the most suspicious ones and pay the vast majority of apparently valid claims. The measure of the importance of each individual predictor in assessing the suspicion level makes sure that only the most important variables need to be gathered and focused on. We can also investigate the association of the fraud suspicion level with other variables (e.g., geographic location of the claimant, the involvement of high claim volume attorneys and physicians) to determine additional predictor variables that might be added to claim files in the future to improve detection performance or aid in negotiating settlements. Because the method is unsupervised, it is relatively costless to update the analysis when fraud techniques used by fraudsters change over time.

The empirical performance of our method was assessed relative to other unsupervised classification methods and was externally validated by comparing the classification performance to that might have been obtained if the true fraud labels of claims had been known. We have shown statistically that exploiting the interrelation of the individual predictor variables allowed us to concatenate the “hints” about fraud contained in the individual predictor variables in such a way that very little was lost by not knowing the dependent variable value in statistical classification. With much less required information and thus lower learning cost, this method does well in identifying cases in the target class and is a valuable alternative to standard supervised models for applications such as insurance fraud detection.

## 7. ACKNOWLEDGMENTS

We wish to express our gratitude to Richard Derrig and the Automobile Insurers Bureau of Massachusetts for allowing us to use the personal injury protection claims fraud data that appears in this paper. We also thank the reviewers for their helpful comments.

---

<sup>15</sup> Discretization of continuous predictor variables also results in information and performance loss for supervised methods. We have performed a similar set of analysis for logistic regression using both a default classification threshold of  $p = 0.5$  and one suggested by the actual data (see Section 4.4.1 for discussions on threshold selection). The data binning criteria are the same as in Table 11. A similar effect of discretization on the performance as shown in Tables 10 and 12 for PRIDIT was seen for logistic regression using either threshold. Results are untabulated to conserve space and are available upon request.

## REFERENCES

- ABU-NIMEH, S., D. NAPPA, X. WANG, AND S. NAIR. 2008. Bayesian Additive Regression Trees-Based Spam Detection for Enhanced Email Privacy. In *2008 Third International Conference on Availability, Reliability and Security*, 1044–51.
- AI, J., P. L. BROCKETT, L. L. GOLDEN, AND M. GUILLEN. 2009. On the Development of a Fraud Rate Estimation Method. Working paper, University of Hawaii at Manoa.
- BROCKETT, P. L. 1981. A Note on Numerical Assignment of Scores to Ranked Categorical Data. *Journal of Mathematical Sociology* 8: 91–101.
- BROCKETT, P. L., R. A. DERRIG, L. L. GOLDEN, A. LEVINE, AND M. ALPERT. 2002. Fraud Classification Using Principal Component Analysis of RIDITs. *Journal of Risk and Insurance* 69: 341–72.
- BROCKETT, P. L., AND L. L. GOLDEN. 1992. A Comment on “Using Rank Values as an Interval Scale” by Dowling and Midgley. *Psychology and Marketing* 9: 255–61.
- BROCKETT, P. L., AND A. LEVINE. 1977. On a Characterization of RIDITs. *Annals of Statistics* 5: 1245–48.
- BROCKETT, P. L., X. XIA, AND R. A. DERRIG. 1998. Using Kohonen’s Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud. *Journal of Risk and Insurance* 65: 245–74.
- BROSS, I. 1958. How to Use RIDIT Analysis. *Biometrics* 14: 18–38.
- CHIPMAN, H., E. I. GEORGE, AND R. E. MCCULLOCH. 2006. Bayesian Ensemble Learning. *Neural Information Processing Systems*.
- CHIPMAN, H., E. GEORGE, AND R. MCCULLOCH. 2008. BART: Bayesian Additive Regression Trees. Working paper, University of Chicago. Available at: [http://www-stat.wharton.upenn.edu/~edgeroge/Research\\_papers/BART%20June%2008.pdf](http://www-stat.wharton.upenn.edu/~edgeroge/Research_papers/BART%20June%2008.pdf). Accessed September 28, 2009.
- CRISTIANINI, N., AND J. SHAWE-TAYLOR. 2000. *An Introduction to Support Vector Machines*. Cambridge: Cambridge University Press.
- DIETTERICH, T. G. 2000. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, Lecture Notes in Computer Science 1857, ed. J. Kittler and F. Roli, pp. 1–15. Berlin: Springer.
- FEDERAL BUREAU OF INVESTIGATION. 2008. Insurance Fraud. Available at: [http://www.fbi.gov/publications/fraud/insurance\\_fraud.htm](http://www.fbi.gov/publications/fraud/insurance_fraud.htm). Accessed October 12, 2008.
- FLORA, J. D., JR. 1988. RIDIT Analysis. In *Encyclopedia of Statistical Sciences*, ed. S. Kotz and N. L. Johnson, vol. 8, pp. 136–39. New York: John Wiley & Sons.
- GOLDEN, L. L., AND P. L. BROCKETT. 1987. The Effects of Alternative Scoring Techniques on the Analysis of Rank Ordered Categorical Data. *Journal of Mathematical Sociology* 12: 383–414.
- HAUSMAN, J. A., J. ABREVAYA, AND F. M. SCOTT-MORTON. 1998. Misclassification of a Dependent Variable in a Discrete-Response Setting. *Journal of Econometrics* 87: 239–69.
- INSURANCE RESEARCH COUNCIL (IRC). 2008. Available at: [http://www.ircweb.org/News/IRC\\_Fraud\\_NR.pdf](http://www.ircweb.org/News/IRC_Fraud_NR.pdf). Accessed September 28, 2009.
- KOHONEN, T. 1989. Self-Organizing Feature Maps. In *Self-Organizing and Associative Memory*. New York: Springer.
- MAHALANOBIS, P. C. 1936. On the Generalized Distance in Statistics. *Proceedings of the National Institute of Sciences of India* 12: 49–55.
- MAS-COLLEL, A., M. D. WHINSTON, AND J. R. GREEN. 1995. *Microeconomic Theory*. New York: Oxford University Press.
- MORRISON, D. G. 1969. On the Interpretation of Discriminant Analysis. *Journal of Marketing Research* 6: 156–63.
- SPSS. 2001. SPSS TwoStep Cluster Component. White paper–technical report. Available at: [http://www.spss.com/downloads/Papers.cfm?prod\\_categoryID=00012&Name=Analytical\\_Components](http://www.spss.com/downloads/Papers.cfm?prod_categoryID=00012&Name=Analytical_Components). Accessed September 28, 2009.
- TEXAS DEPARTMENT OF INSURANCE. 2006. Fraud, Facts and Frequently Asked Questions. Available at: <http://www.tdi.state.tx.us/fraud/faq.html>. Accessed October 12, 2008.
- UCI MACHINE LEARNING REPOSITORY. 2008. ADULT database. Available at: <http://www.ics.uci.edu/~mllearn/MLSummary.html>. Accessed September 28, 2009.
- U.S. CENSUS BUREAU. 2008. Data Dictionary for Current Population Survey. Available at: <http://www.census.gov/eps/>. Accessed September 28, 2009.
- VIAENE, S., R. A. DERRIG, B. BAESENS, AND G. DEDENE. 2002. A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection. *Journal of Risk and Insurance* 69: 373–421.
- WEISBERG, H. I., AND R. A. DERRIG. 1991. Fraud and Automobile Insurance: A Report on Bodily Injury Liability Claims in Massachusetts. *Journal of Insurance Regulation* 9: 497–541.

*Discussions on this paper can be submitted until April 1, 2010. The authors reserve the right to reply to any discussion. Please see the Submission Guidelines for Authors on the inside back cover for instructions on the submission of discussions.*