Active Learning Through Sequential Design, With Applications to Detection of Money Laundering

Xinwei DENG, V. Roshan JOSEPH, Agus SUDJIANTO, and C. F. Jeff WU

Money laundering is a process designed to conceal the true origin of funds that were originally derived from illegal activities. Because money laundering often involves criminal activities, financial institutions have the responsibility to detect and report it to the appropriate government agencies in a timely manner. But the huge number of transactions occurring each day make detecting money laundering difficult. The usual approach adopted by financial institutions is to extract some summary statistics from the transaction history and conduct a thorough and time-consuming investigation on those suspicious accounts. In this article we propose an active learning through sequential design method for prioritization to improve the process of money laundering detection. The method uses a combination of stochastic approximation and *D*-optimal designs to judiciously select the accounts for investigation. The sequential nature of the method helps identify the optimal prioritization criterion with minimal time and effort. A case study with real banking data demonstrates the performance of the proposed method. A simulation study shows the method's efficiency and accuracy, as well as its robustness to model assumptions.

KEY WORDS: Bayesian estimation; Optimal design; Pool-based learning; Stochastic approximation; Threshold hyperplane.

1. BACKGROUND

Money laundering is an act designed to hide the true origin of funds by sending them through a series of seemingly legitimate transactions. Its main purpose is to conceal the fact that funds were acquired as a result of some form of criminal activity. These laundered funds can in turn be used to foster further illegal activities, such as the financing of terrorist activity or trafficking of illegal drugs. Even legitimate funds that are laundered to avoid reporting them to the government, as is the case with tax evasion, have substantial costs to society. Financial institutions that have the responsibility to detect and prevent money laundering face the challenge of detecting potential suspicious activities among the millions of legitimate transactions that occur each day. Once suspicious activities are detected, the investigation process usually has to retrieve transaction data for suspicious customers, separate "inflow" and "outflow" of funds, filter relevant transaction types (e.g., wire transfer, cash), and suppress irrelevant information. Then investigators create transaction summaries for each day, week, month, or the whole period to extract such basic statistics as total amount, frequency, average, maximum, count of wire transfer, and so on. By gathering other related information (e.g., customer profiles, income) from other sources (e.g., Internet, third party), investigators use heuristics and experience to create a "story" and identify suspicious activities, including

- Who is the customer?
- What banking product does the person use (e.g., checking, credit card, investment)?
- What kind of transactions does the person conduct (e.g., Automated Clearing House [ACH], wire transfer, cash)?
- What transaction channel does the person use (e.g., ATM, Internet)?
- Where are the transactions conducted (i.e., geographic location)?
- What are the transaction amounts and frequencies?
- Any other information and past or related incidence.

Finally, investigators apply their judgment to determine the need to submit a Suspicious Activity Report (SAR) to the Financial Criminal Enforcement Network (FinCEN). An investigation can easily take 10 hours just to classify a case as either suspicious or nonsuspicious.

The challenge of detecting money laundering arises not only from the huge amount of transactions occurring each day, but also from the different kinds of businesses with money laundering activities. The behaviors of various business categories can be quite different. For example, money laundering activities in personal accounts can be completely different from those in small business accounts, and so the knowledge and experience regarding potentially suspicious money laundering activities for personal accounts cannot be applied to those for small business accounts. Even the behaviors of the same business category in different time periods appear to be different in money laundering activities.

Table 1 shows a sample of transaction data. The transaction history contains various types of information. The information structure can be very complex or can involve multiple bank accounts, financial organizations, parties, and jurisdictions. Transaction frequency and amounts can be useful information for detecting suspicious money laundering activities. For example, an account is suspicious if transactions are conducted in bursts of activity in short periods, especially in a previously dormant account. The information on different types of transactions also is an important indicator for investigating money laundering activities. The basic summary statistics can be dozens of continuous and categorical variables; thus, investigating every account

Xinwei Deng is Visiting Assistant Professor, Department of Statistics, University of Wisconsin–Madison, Madison, WI 53706 (E-mail: *xdeng@isye.gatech.edu*). V. Roshan Joseph is Associate Professor, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332. Agus Sudjianto is Senior Vice President, Bank of America, Charlotte, NC 28255. C. F. Jeff Wu is Professor, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 (E-mail: *jeffwu@isye.gatech.edu*). We are greatly thankful to the Associate Editor and the referees, whose constructive comments helped us to improve the contents and presentation of the paper. The research of Deng, Joseph, and Wu was supported in part by a grant from the U.S. Army Research Laboratory and the U.S. Army Research Office under contract number W911NF-05-1-0264.

Journal of the American Statistical Association
September 2009, Vol. 104, No. 487, Applications and Case Studies
DOI: 10.1198/jasa.2009.ap07625

Table 1. A sample of transaction data

Acct No.	D/C	PostDate	TransAmt	TransCode	Description
999999	D	1/23/2005	$1,295.00	9059	Check Check
999999	D	5/19/2004	$1,020.00	9059	Check Check
999999	D	1/23/2005	$10,000.00	9059	Check Check
999999	D	3/2/2004	$5.00	9593	Returned Item Charge Returned Item Charge
999999	D	2/24/2004	$5.00	9593	Returned Item Charge Returned Item Charge
999999	D	10/12/2004	$34.00	9203	Overdraft Charge Overdraft Charge
999999	D	7/13/2004	$60.00	9659	Check Card Purchase Dr Jm Layton And Ep Lay5194121949512823
999999	D	6/10/2004	$129.36	9905	Pos Withdrawal Costco Whse # 0001 8442627516108999910830
999999	D	6/14/2004	$51.49	9905	Pos Withdrawal Bed, Bath & Beyo 844262751650899999914310
999999	D	6/10/2004	$168.44	9905	Pos Withdrawal Costco Whse # 0001 8442627516108999910370
999999	D	7/18/2004	$34.84	9905	Pos Withdrawal Costco Whse # 0001 8442627519708999916890
999999	D	5/24/2004	$33.20	9905	Pos Withdrawal Costco Gas # 00662 8442627514408999924800
999999	D	6/22/2004	$158.65	9905	Pos Withdrawal Bed, Bath & Beyo 844262751730899999922610
999999	D	6/10/2004	$190.64	9905	Pos Withdrawal Costco Whse # 0001 8442627516108999910750
999999	C	1/14/2004	$100.00	9003	Deposit Deposit
999999	C	8/10/2004	$20.00	9003	Deposit Deposit
999999	C	5/11/2004	$10,000.00	9003	Deposit Deposit
999999	C	8/31/2004	$3,300.00	9003	Deposit Deposit 0831CA319P007160134679
999999	C	6/29/2004	$2,079.95	9003	Deposit Deposit
999999	C	10/7/2004	$2,500.00	9003	Deposit Deposit
999999	C	1/30/2005	$22.43	9699	Automatic Deposit Deposit Merchant Bankcd 267917678885
999999	C	1/30/2005	$22.43	9699	Automatic Deposit Deposit Merchant Bankcd 267917678885
999999	C	6/16/2004	$64.97	9660	Reverse Check Card Purchase The Home Depot 4715 5166010183470016
999999	C	7/21/2004	$151.61	9660	Reverse Check Card Purchase Hardware Sales 5202207788501885
999999	C	9/20/2004	$24.95	9660	Reverse Check Card Purchase Twx*Sports Illustrated 5259000879500624
999999	C	4/27/2004	$14,032.37	9039	Deposit To Close Account Deposit To Close Account
999999	C	11/30/2004	$3,243.59	9003	Deposit Deposit
999999	C	7/6/2004	$400.00	9003	Deposit Deposit
999999	C	10/6/2004	$2,981.07	9003	Deposit Deposit
999999	C	7/21/2004	$100.00	9007	Miscellaneous Deposit Transfer From Checking 22782403

for money laundering would be extremely time-consuming and cost-prohibitive.

One detection strategy to overcome this problem and improve the process of money laundering detection is segmentation and risk prioritization. First, we segment accounts into distinct clusters based on a similarity measure and business knowledge. Then, for each cluster we prioritize a group of accounts based on their likelihood of suspicious activity and severity. By incorporating statistical modeling and knowledge of investigation experience, we can extract several profile features (which are nonlinear functions of transaction data) using the transaction history of each account. This is a common method of analyzing such data in financial institutions. The profile features can be a nonlinear projection from those basic summary statistics of transaction data or a complicated augment of transactions with pooling, multiscale extraction, and smoothing. If these profile features are highly representative of the suspiciousness for the transaction history, then they can be used to set rules for prioritization of account investigation. The accounts with high priority are investigated thoroughly to determine whether or not they are suspicious. When a new account belonging to certain cluster is introduced, the corresponding prioritization rule can be used to decide whether or not the account merits a detailed investigation. This can significantly improve productivity by focusing investigations only on those cases that really matter. In this work we develop a statistical methodology to perform risk prioritization.

The rest of the article is organized as follows. We formulate the risk prioritization as a sequential design problem in Section 2. In Section 3 we review some existing methods in sequential designs and the concept of optimal designs. We propose our active learning through sequential design approach for prioritization in Section 4. In Section 5 we apply our proposed method to a real case study on detecting money laundering. In Section 6 we provide some simulation results to demonstrate the performance of our proposed. We end with a discussion and some conclusions in Section 7.

2. MATHEMATICAL FORMULATION

The problem can be formulated as follows. Let $\mathbf{x} = (x_1, \ldots, x_p)^T$ be the vector of profile features extracted from a transaction history of a group of accounts in the same cluster. Let $Y = 1$ if the account is detected as suspicious and $Y = 0$ otherwise. Then $P(Y = 1|\mathbf{x}) = F(\mathbf{x})$ gives the probability of suspiciousness at $\mathbf{x}$. When $F(\mathbf{x})$ exceeds a threshold probability α, we can investigate that account in detail. The threshold probability α can be chosen beforehand by an investigator with domain knowledge. Assume that $F(\mathbf{x})$ is an increasing function in each x_i. Define the *decision boundary* $l_\alpha(\mathbf{x})$ at level α as

$$l_\alpha(\mathbf{x}) = \{\mathbf{x} : F(\mathbf{x}) = \alpha\}. \tag{2.1}$$

The form of the decision boundary $l_\alpha(\mathbf{x})$ can be linear, nonlinear, or nonparametric in $\mathbf{x}$. Note that the profile features $\mathbf{x}$

are nonlinear functions of the transaction data and can approximately characterize the suspiciousness behavior of a transaction history. Thus a linear combination of profile features $\mathbf{x}$ as the decision boundary can be a reasonable choice and useful for business interpretation. Hereinafter, we refer to the decision boundary as the *threshold hyperplane*. Now for a new account in this cluster, if $\mathbf{x}$ falls below $l_\alpha(\mathbf{x})$, then we need not investigate that account further. But if $\mathbf{x}$ falls above $l_\alpha(\mathbf{x})$, then we must investigate the account in detail. An institution may choose a reasonable α so that only a portion of their accounts must be investigated.

Developing a procedure for efficiently finding the threshold hyperplane is important. The problem is that $F(\mathbf{x})$ is unknown, and thus $l_\alpha(\mathbf{x})$ is also unknown. Data on $\mathbf{x}$ and Y can be used to estimate $l_\alpha(\mathbf{x})$. For this purpose, a training set of the investigated accounts is needed; however, labeling the suspiciousness (1 or 0) for a large number of accounts is time-consuming and extremely expensive. Finding a way to minimize the number of investigated accounts and use these accounts to construct the threshold hyperplane would be beneficial. Thus the goal is to determine an optimal threshold hyperplane for prioritization with a minimum number of investigated accounts.

This calls for the use of active learning (Mackay 1992; Cohn, Ghahramani, and Jordan 1996; Fukumizu 2000) techniques in machine learning. Here the learner actively selects data points to be added into the training set. To minimize the number of investigated accounts and use these accounts to construct the threshold hyperplane, we need to judiciously select the accounts for investigation. Recently, active learning methods using support vector machines (SVMs) have been developed by several researchers (Campbell, Cristianini, and Smola 2000; Schohn and Cohn 2000; Tong and Koller 2001). We can apply these to the present problem.

For binary response, active learning with SVMs is mainly for two-class classification. The decision boundary in SVMs implements the Bayes rule $P(Y|\mathbf{x}) = 0.5$, which is the optimal classification rule if the underlying distribution of the data is known. Note that the decision boundary of SVMs can be considered a special case of (2.1). In money laundering detection, often the interest lies in values other than $\alpha = 0.5$. Finding the threshold hyperplane at a higher value of α, such as $\alpha = 0.75$, is important. Note that the concept of active learning in machine learning is closely related to that of sequential designs in the statistics literature. In sequential designs, the data points for investigation are selected sequentially by the users; that is, the next data point to be selected for investigation is based on information gathered from previously investigated data points. The present problem differs from classical sequential designs, however. Because the accounts are already available, we cannot arbitrarily select the setting of accounts for investigation. To address these points, we exploit the synergies between these two approaches to develop a new active learning through sequential design (ALSD) approach. Our ALSD approach provides a more flexible way to obtain the threshold hyperplane for different values of α. The sequential nature of the method helps identify the optimal threshold hyperplane with reasonable time and effort.

3. REVIEW OF SEQUENTIAL DESIGNS

The problem of estimating the threshold hyperplane is similar to that of stochastic root-finding in sequential designs. Suppose that we want to estimate the root of an unknown univariate function $E(Y|x) = F(x)$ from the data $(x_1, Y_1), \ldots, (x_n, Y_n)$. In sequential designs, the data points are chosen sequentially; that is, x_{n+1} is selected based on $x_1, x_2, \ldots, x_n$, and their corresponding response $Y_1, Y_2, \ldots, Y_n$. There are two approaches to generating sequential designs: stochastic approximation and optimal design.

In *stochastic approximation* methods, the x's are chosen such that x_n converges to the root as $n \to \infty$. Wu (1985) proposed a stochastic approximation method for binary data, known as the logit-maximum likelihood estimation (MLE) method, in which $F(x)$ is approximated by a logit function $e^{(x-\mu)/\sigma}/(1 + e^{(x-\mu)/\sigma})$. Then determination of x_{n+1} is a two-step procedure. First, maximum likelihood (ML) estimates $\hat{\mu}_n, \hat{\sigma}_n$ of μ, σ are found from $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$. Then x_{n+1} is chosen as $x_{n+1} = \hat{\mu}_n + \hat{\sigma}_n \log \frac{\alpha}{1-\alpha}$. Ying and Wu (1997) showed the almost-sure convergence of x_n to the root irrespective of the function $F(x)$. Joseph, Tian, and Wu (2007) proposed an improvement to Wu's logit-MLE method by giving more weight to data points closer to the root via a Bayesian scheme.

In the *optimal design* approach to sequential designs, first a parametric model for the unknown function is postulated, and then the x points are chosen sequentially based on some optimality criterion (Kiefer 1959; Fedorov 1972; Pukelsheim 1993). For example, Neyer (1994) proposed a sequential D-optimality–based design in which x_{n+1} is chosen so that the determinant of the estimated Fisher information is maximized. It is well known that a D-optimal criterion minimizes the volume of the confidence ellipsoid of the parameters (Silvey 1980). The root is solved from the final estimate of the function $F(x)$.

The performance of the optimal design approach is model-dependent. Performance is best when the assumed model is the true model, but deteriorates as the model deviates from the true model. One attractive property of the stochastic approximation methods, including the logit-MLE, is the robustness of their performance to model assumptions. This occurs because as n grows large, the points become clustered around the root, allowing estimation of the root irrespective of the model assumption. Understandably, the performance of the stochastic approximation method is not as good as that of the optimal design approach when the assumed model in the latter approach is valid. This point was confirmed by Young and Easterling (1994) through extensive simulations.

Our proposed ALSD approach combines the advantages of both the stochastic approximation and optimal design approaches. Our approach is expected to be robust to model assumptions like the stochastic approximation methods and also to produce performance comparable to that of the optimal design approach when the model assumptions are valid. Unlike most existing sequential design methods, our proposed approach can handle multiple independent variables, which occur in the money laundering detection example as well as other applications (e.g., junk e-mail classification).

4. METHODOLOGY

4.1 Active Learning Through Sequential Design

In pool-based active learning (Lewis and Gale 1994), there is a pool of unlabeled data. The learner has access to this pool and can request the true label for a certain amount of data in the pool. The main issue is to find a way to choose the next unlabeled data point to get the response. Our proposed ALSD approach attempts to "close in" on the region of interest efficiently while improving the estimation accuracy of $l_\alpha(\mathbf{x})$ for a given α.

For ease of exposition, we explain the methodology with two variables, $\mathbf{x} = (x_1, x_2)^T$. It can be easily extended to more than two variables. We assume that each variable has a positive relationship with the response; that is, for larger values of x_j, the probability of getting the response $Y = 1$ increases. Define a *synthetic variable* z by $z = wx_1 + (1 - w)x_2$, where w is an *unknown* weight factor in [0, 1]. By doing this, we can convert the multivariate problem into a univariate problem, allowing the existing methods for sequential designs to be easily applied.

As in the case of Wu's logit-MLE method, we model the unknown function $F(\mathbf{x})$ in (2.1) using the parametric form,

$$F(\mathbf{x}|\boldsymbol{\theta}) = \frac{e^{(z-\mu)/\sigma}}{1 + e^{(z-\mu)/\sigma}}, \tag{4.1}$$

which has three parameters, $\boldsymbol{\theta} = (\mu, \sigma, w)^T$. As noted before, here convergence is independent of the logit model if the linearity assumption in $\mathbf{x}$ is valid. By the definition given in (2.1), the threshold hyperplane $l_\alpha(\mathbf{x})$ at level α is

$$l_\alpha(\mathbf{x}) = \left\{ \mathbf{x} = (x_1, x_2)^T : \frac{z - \mu}{\sigma} = \log\left(\frac{\alpha}{1-\alpha}\right), \text{ where } z = wx_1 + (1-w)x_2 \right\}, \tag{4.2}$$

which is a linear hyperplane of $\mathbf{x}$. Suppose that we have $(\mathbf{x}_1, Y_1), (\mathbf{x}_2, Y_2), \ldots, (\mathbf{x}_n, Y_n)$ in the training set. Based on these training data, we can estimate the threshold hyperplane $l_{n,\alpha} = \{\mathbf{x} : F(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) = \alpha\}$ by

$$l_{n,\alpha}: \quad \hat{w}_n x_1 + (1 - \hat{w}_n)x_2 = \hat{\mu}_n + \hat{\sigma}_n \log\left(\frac{\alpha}{1-\alpha}\right), \tag{4.3}$$

where $\hat{\boldsymbol{\theta}}_n = (\hat{\mu}_n, \hat{\sigma}_n, \hat{w}_n)^T$ is estimated from the labeled data $(\mathbf{x}_1, Y_1), (\mathbf{x}_2, Y_2), \ldots, (\mathbf{x}_n, Y_n)$. The estimator $\hat{\boldsymbol{\theta}}_n$ is described in detail in Section 4.2. Let $\mathcal{X}$ be the pool of data. Now, using the idea in stochastic approximation, we choose the next data point from $\mathcal{X}$ as the one closest to the estimated hyperplane. Note that we must choose the closest point because none of the points in $\mathcal{X}$ may fall on the hyperplane. Thus

$$\mathbf{x}_{n+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \operatorname{dist}(\mathbf{x}, l_{n,\alpha}), \tag{4.4}$$

where $\operatorname{dist}(\mathbf{x}, l_{n,\alpha})$ is the distance from $\mathbf{x}$ to $l_{n,\alpha}$ (perpendicular distance from point to line). There can be multiple points satisfying (4.4) because $\mathbf{x} \in \mathbb{R}^2$. Moreover, as pointed out in the previous section, the stochastic approximation method produces points clustered around the true hyperplane, leading to poor estimation of some of the parameters in the model. We can overcome these problems by integrating the foregoing approach with the optimal design approach.

First, we choose k_0 points as candidates closest to the estimated threshold hyperplane $l_{n,\alpha}$. We denote these by $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_{k_0}$. Then we select the next point as the one maximizing the determinant of the Fisher information matrix among the candidates. Thus

$$\mathbf{x}_{n+1} = \arg\max_{\mathbf{x} \in \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_{k_0}\}} \det(I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \mathbf{x})). \tag{4.5}$$

The Fisher information matrix for $\boldsymbol{\theta}$ can be calculated as

$$I(\boldsymbol{\theta}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \frac{e^{g(\mathbf{x}_i)}}{(1 + e^{g(\mathbf{x}_i)})^2} \frac{\partial g(\mathbf{x}_i)}{\partial \boldsymbol{\theta}} \frac{\partial g(\mathbf{x}_i)}{\partial \boldsymbol{\theta}^T}, \tag{4.6}$$

where $g(\mathbf{x}) = (z - \mu)/\sigma$, $z = wx_1 + (1 - w)x_2$, and $\boldsymbol{\theta} = (\mu, \sigma, w)^T$. The foregoing approach inherits the advantages of both stochastic approximation and optimal design. The stochastic approximation method in (4.4) can produce reasonable estimates of μ and σ but very poor estimates of w. Because the D-optimality criterion in (4.5) ensures that the chosen points are well spread, we can get a better estimate of w. Thus through the integration of these two methods, we can expect to get good estimates of μ, σ, and w.

The number of candidate points (k_0) determines the extent of integration between the two methods; $k_0 = 1$ gives stochastic approximation, and $k_0 = N$ gives a fully D-optimal design. We study the choice of k_0 through simulations in Section 6 (see Figure 10).

The improved estimation provided by our approach can be demonstrated by considering the following version of the problem. Assume that there is at least one point in $\mathcal{X}$ lying in the hyperplane $l_{n,\alpha}$. Then the selected point $\mathbf{x}_{n+1}$ is the solution of the following optimization problem:

$$\begin{aligned} &\max_{\mathbf{x}} \det(I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \mathbf{x})) \\ &\text{s.t.} \quad \hat{w}_n x_1 + (1 - \hat{w}_n)x_2 = \hat{\mu}_n + \hat{\sigma}_n \log\left(\frac{\alpha}{1-\alpha}\right). \end{aligned} \tag{4.7}$$

As shown in the Appendix, this is equivalent to

$$\begin{aligned} &\max_{\mathbf{x}} \boldsymbol{\eta}_x^T I^{-1}(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \boldsymbol{\eta}_x \\ &\text{s.t.} \quad \hat{w}_n x_1 + (1 - \hat{w}_n)x_2 = \hat{\mu}_n + \hat{\sigma}_n \log\left(\frac{\alpha}{1-\alpha}\right), \end{aligned} \tag{4.8}$$

where $\boldsymbol{\eta}_x = (-1/\sigma, -\log(\alpha/(1-\alpha)), (x_1 - x_2)/\sigma)^T$. The objective function in (4.8) is the estimated variance of the hyperplane where the data are collected. By maximizing this variance, we are placing the next point at the location of greatest uncertainty. Note that the objective function in (4.8) is associated with $\mathbf{x}$ only through $\boldsymbol{\eta}_x$. It maximizes a quadratic form in terms of $(x_1 - x_2)$. Therefore, the point selected by (4.5) is expected to be distant from the previously selected points when projected onto the estimated threshold hyperplane $l_{n,\alpha}$. This is why our proposed approach can provide a more stable estimate of the parameter w.

4.2 Estimation

Because (4.1) is a probabilistic model, it is tempting to consider ML estimation for the parameter $\boldsymbol{\theta}$. Suppose that the labeled data are $(\mathbf{x}_1, Y_1), (\mathbf{x}_2, Y_2), \ldots, (\mathbf{x}_n, Y_n)$. It is known that the existence and uniqueness of ML estimation can be achieved only when successes and failures overlap (Silvapulle 1981; Albert and Anderson 1984; Santner and Duffy 1986); however, even when we are able to compute the ML estimator, it may suffer from low accuracy because of the small sample size, especially for nonlinear models. The use of a Bayesian approach with proper prior distributions for the parameters can overcome these problems.

We use the following priors:

$$\begin{aligned} &\mu \sim \mathrm{N}(\mu_0, \sigma_\mu^2), \qquad \sigma \sim \text{Exponential}(\sigma_0), \\ &w \sim \text{Beta}(\alpha_0, \beta_0). \end{aligned} \tag{4.9}$$

A normal prior is specified for the location parameter μ. The scale parameter σ is nonnegative, because each x_i is assumed to be positively related to the response Y. Thus an exponential prior with mean σ_0 is used as the prior for σ. Because w is a weight factor in [0, 1], a beta distribution is a reasonable prior for w.

Assuming that μ, σ, and w are independent of one another, the overall prior for $\boldsymbol{\theta}$ is the product of the priors for each of its components. Thus the posterior distribution is

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{Y}) \propto \prod_{i=1}^{n} \left(\frac{e^{(z_i-\mu)/\sigma}}{1+e^{(z_i-\mu)/\sigma}} \right)^{Y_i} \left(\frac{1}{1+e^{(z_i-\mu)/\sigma}} \right)^{1-Y_i} \\ \times e^{(\mu-\mu_0)^2/(-2\sigma_\mu^2)} \lambda_0 e^{-\lambda_0 \sigma} w^{\alpha_0-1} (1-w)^{\beta_0-1}, \end{aligned} \tag{4.10}$$

where $z_i = wx_{i1} + (1-w)x_{i2}$ and $\mathbf{x}_i = (x_{i1}, x_{i2})^T$. Finding the posterior mean of the parameters is difficult because it involves a complicated multidimensional integration. The maximum a posteriori (MAP) estimators are much easier to compute. The MAP estimators of μ, σ, and w are obtained by solving

$$\hat{\boldsymbol{\theta}}_n = (\hat{\mu}_n, \hat{\sigma}_n, \hat{w}_n)^T = \arg\max_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}|\mathbf{Y}), \tag{4.11}$$

where

$$\begin{aligned} \log f(\boldsymbol{\theta}|\mathbf{Y}) \triangleq \sum_{i=1}^{n} \frac{z_i-\mu}{\sigma} Y_i - \sum_{i=1}^{n} \log\left(1+\exp\left(\frac{z_i-\mu}{\sigma}\right)\right) \\ - \frac{(\mu-\mu_0)^2}{2\sigma_\mu^2} - \lambda_0\sigma + (\alpha_0-1)\log(w) \\ + (\beta_0-1)\log(1-w). \end{aligned}$$

Because proper prior distributions are used, the optimization in (4.11) is well defined even when $n = 1$. Thus this Bayesian approach allows us to implement a *fully* sequential procedure; that is, the proposed active learning method can begin from $n = 1$. This would not have been possible with a frequentist approach (Wu 1985), for which some initial sample is needed before the active learning method can be used.

5. CASE STUDY

We applied the proposed method to transaction data from a financial institution. The data in this example comprise 92 accounts from personal customers belonging to the same cluster. It contains the recent 2-year transaction history for each customer. By working with expert investigators, we obtained a large set of summary variables describing the transaction behaviors. Then, using lower-dimensional scoring on these summary variables, we extracted one profile feature, x_1, which measures how the customer's behavior is inconsistent with itself and inconsistent with similar customers (i.e., peer comparison). Incorporating knowledge of investigation experience, we analyzed the transaction history of each account through multilayer decomposition and accumulation, and then selected one more profile feature, x_2, which describes the velocity and amount of money flowing through the account. To maintain confidentiality, we do not disclose more details about the data used here. Based on discussions with expert investigators, we believe that these two profile features, $\mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2$, can be highly indicative of suspicious transaction history. A linear combination of these two profile features can be used to assess suspicious behaviors of these personal accounts. Larger values of profile features indicate a higher likelihood of suspiciousness. Profile features x_1 and x_2 are standardized to have mean 0 and unit variance. The standardized data are shown in Figure 1.

Before implementing our ALSD approach, we need to specify the prior for μ, σ, and w in (4.9). Here we use a heuristic procedure to do this, as follows. First, consider the prior for w. Assuming equal importance of x_1 and x_2 on the response, we would like the mean of w to be 0.5. Thus we set $\alpha_0/(\alpha_0+\beta_0) = w_0 = 0.5$, which implies $\alpha_0 = \beta_0$. To get a flat prior, we take $\alpha_0 = \beta_0 = 3/2$. Thus, $w \propto w^{1/2}(1-w)^{1/2}$. Now we consider the priors for μ and σ. We choose two extreme points (i.e., two accounts), $\mathbf{x}_l$ and $\mathbf{x}_u$, based on the lowest and highest values of z (denoted by z_l and z_u) through the mapping

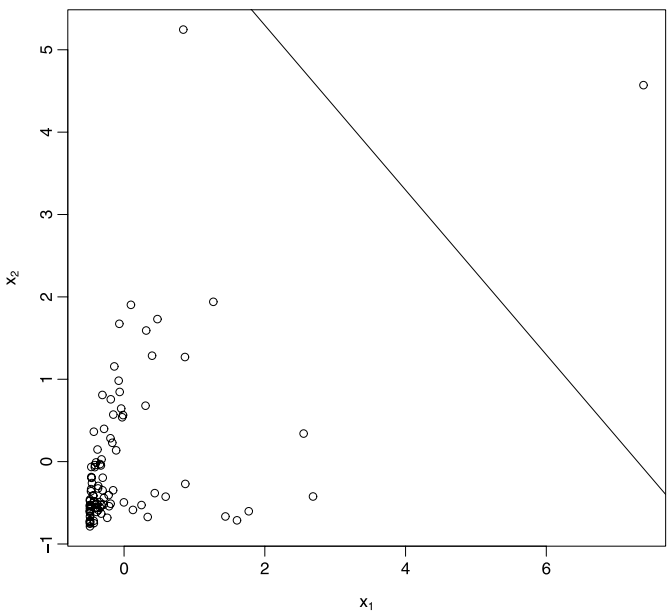

Figure 1. The standardized data. The solid line represents the initial estimated threshold hyperplane by w_0, μ_0, and σ_0.

$z = w_0 x_1 + (1 - w_0)x_2$. We assume a $\alpha_l = 5\%$ suspicious level for $\mathbf{x}_l$ and a $\alpha_u = 95\%$ suspicious level for $\mathbf{x}_u$. Plugging these values into the model (4.1), we obtain

$$z_l = \mu + \sigma \log \frac{\alpha_l}{1 - \alpha_l},$$

$$z_u = \mu + \sigma \log \frac{\alpha_u}{1 - \alpha_u}.$$

Then we obtain μ_0 and σ_0 by solving the foregoing equations as

$$\mu_0 = \left(z_l \log \frac{\alpha_u}{1 - \alpha_u} - z_u \log \frac{\alpha_l}{1 - \alpha_l} \right) \Big/ \left(\log \frac{\alpha_u}{1 - \alpha_u} - \log \frac{\alpha_l}{1 - \alpha_l} \right) = \frac{z_l + z_u}{2}, \quad (5.1)$$

$$\sigma_0 = (z_u - z_l) \Big/ \left(\log \frac{\alpha_u}{1 - \alpha_u} - \log \frac{\alpha_l}{1 - \alpha_l} \right). \quad (5.2)$$

We take σ_μ^2 as the sample variance of z_i, $i = 1, \ldots, n$, where $z_i = w_0 x_{i1} + (1 - w_0)x_{i2}$. This completes the prior specification for the three parameters.

Now we can implement our ALSD method. Suppose that our objective is to find the threshold hyperplane with $\alpha = 0.75$. The initial estimated hyperplane based on only the prior is shown in Figure 1. The points are then selected one at a time using the procedure described in the previous section. In this example, we took $k_0 = 15$ in (4.5). The performance of the proposed method for the first 20 points is shown in Figures 2 and 3.

Figure 2 shows a series of threshold hyperplanes estimated using our proposed approach. Data points marked with "×" were selected, and the response was 1. Data points marked with "+" were selected, and the response was 0. At the beginning, there were significant changes in the threshold hyperplane. After about 10–15 points, it began to converge, as shown in the bottom left panel of the figure. The final estimated threshold hyperplane (i.e., after 20 points) is shown in Figure 3. The points above this hyperplane should be given higher priority and be investigated thoroughly. There are only a few remaining accounts that require thorough investigation, clearly demonstrating the

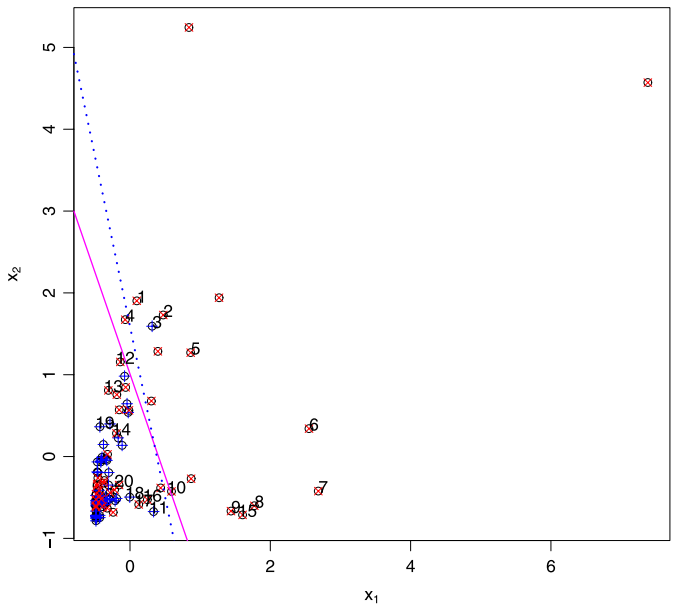

Figure 3. Comparison with the estimate based on full information. The solid line represents the estimated threshold hyperplane after 20 points are sequentially selected. The dashed line represents the estimated threshold hyperplane when all data are labeled.)

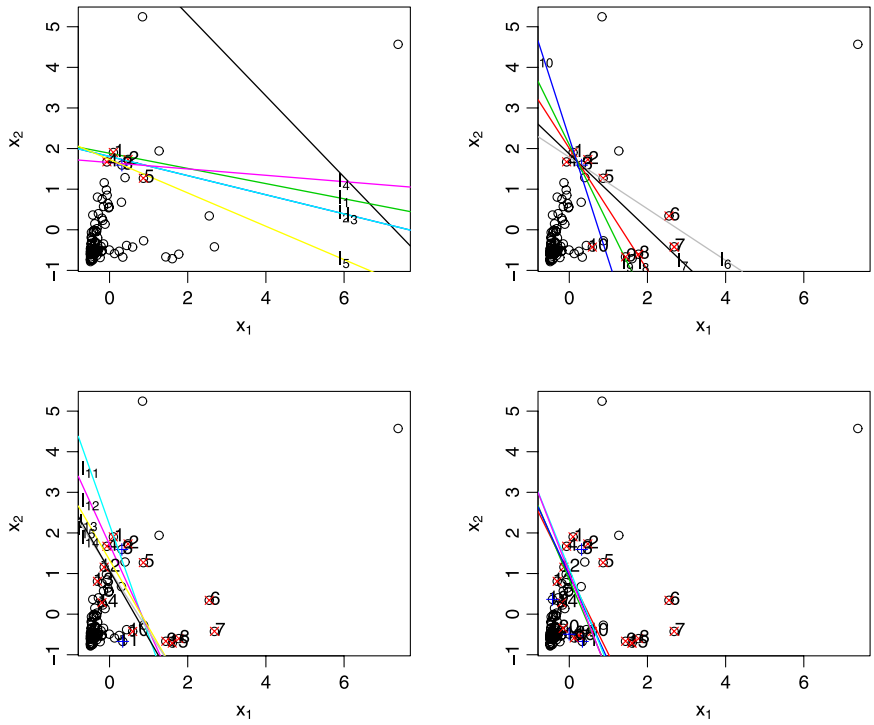

Figure 2. Our ALSD approach. Line l_5 represents the estimated threshold hyperplane at iteration 5, and so on. The four panels, from top left to bottom right, represent the first five to the last five (i.e., iterations 15–20) estimated threshold hyperplanes.

efficiency of our proposed method. For a new personal account in this cluster, the corresponding profile features $\mathbf{x}_{new}$ can be calculated based on the transaction history. If $\mathbf{x}_{new}$ falls above the estimated threshold hyperplane, then a thorough investigation of this account is performed; otherwise, a detailed investigation is not performed.

To assess the accuracy of the proposed method, we requested that the investigators at this financial institution carefully investigate all 92 accounts. Based on the information for all of these accounts, we estimated the threshold hyperplane, shown as a dashed line in Figure 3. This estimated threshold hyperplane is very close to that estimated by the active learning method (shown as a solid line). Thus our proposed method can identify the true hyperplane using only about 22% ($\approx 20/92$) of the data, providing significant savings for the financial institution.

To check the efficiency of the proposed method, we also compared the proposed method with two naive methods, one method that randomly selects the next data point for getting the response and a sequential space-filling procedure using a maxmin distance criterion (Johnson, Moore, and Ylvisaker 1990). The aim is to select the next data point $\mathbf{x}_{n+1}$ that maximizes the minimum distance from the chosen data points, that is,

$$\mathbf{x}_{n+1} = \arg\max_{\mathbf{x}\in\mathcal{X}} \min_{\mathbf{x}_i\in\mathcal{D}} \text{dist}(\mathbf{x}, \mathbf{x}_i), \tag{5.3}$$

where $\mathcal{D}$ is the training set containing n chosen data points $\mathbf{x}_i$, $i = 1, \ldots, n$, and $\text{dist}(\mathbf{x}, \mathbf{x}_i)$ is the distance between two data points $\mathbf{x}$ and $\mathbf{x}_i$.

To gauge the performance of the three methods, we assessed the closeness of the estimated threshold hyperplane $l_{n,\alpha}$ and the optimal threshold hyperplane $l_\alpha(\mathbf{x})$ when all data are labeled. The adopted measure is

$$\text{dist}(l_{n,\alpha}, l_\alpha(\mathbf{x})) \triangleq \sum_{\mathbf{t}_i\in\mathbf{T}} d_i^2, \tag{5.4}$$

where $\mathbf{T} = \{\mathbf{t}_i\}$ is a set of points lying evenly on the optimal threshold hyperplane $l_\alpha(\mathbf{x})$, ranging from -0.5 to 1, on the coordinate of x_1. Here d_i is the distance between $\mathbf{t}_i$ and the estimated hyperplane $l_{n,\alpha}$. Based on (5.4), a distance-based performance measure is defined as

$$Dist\_PM = \frac{1}{M}\sum_{j=1}^{M} \text{dist}_j(l_{n,\alpha}, l_\alpha(\mathbf{x})), \tag{5.5}$$

where M is the number of simulations and dist_j represents $\text{dist}(l_{n,\alpha}, l_\alpha(\mathbf{x}))$ for the jth simulation.

We also measured the misclassification error of the three methods. Misclassification error is estimated by $(\alpha FP + (1-\alpha)FN)/N$, where FP is the number of false-positive results (i.e., a nonsuspicious account assigned the suspicious label 1), FN is the number of false-negative results (i.e., a suspicious account assigned the nonsuspicious label 0), and N is the total number of accounts. As *Dist_PM* increases, the estimated hyperplane deviates more from the true hyperplane, increasing the misclassification error. Thus these two measures agree with each other if the linearity assumption in the model holds. Otherwise, the misclassification error should be used, because it is a more direct and relevant measure for gauging the performance.

Figure 4 shows the learning curves for the three methods in terms of *Dist_PM* and misclassification error. Each naive method was implemented for 100 simulations. Clearly, our proposed ALSD method is much more efficient than the two naive methods. The threshold hyperplane estimated by our proposed method moves toward the optimal threshold hyperplane more quickly and consistently. It converges in about 15 steps, and its misclassification error reaches 0.086, which is the error rate of the hyperplane estimated with the *full* data. Our proposed method also has a much smaller misclassification error than the two naive methods in each iteration.

To check the linearity assumption in (4.1), we use dispersion as a measure of goodness of fit. The dispersion parameter is

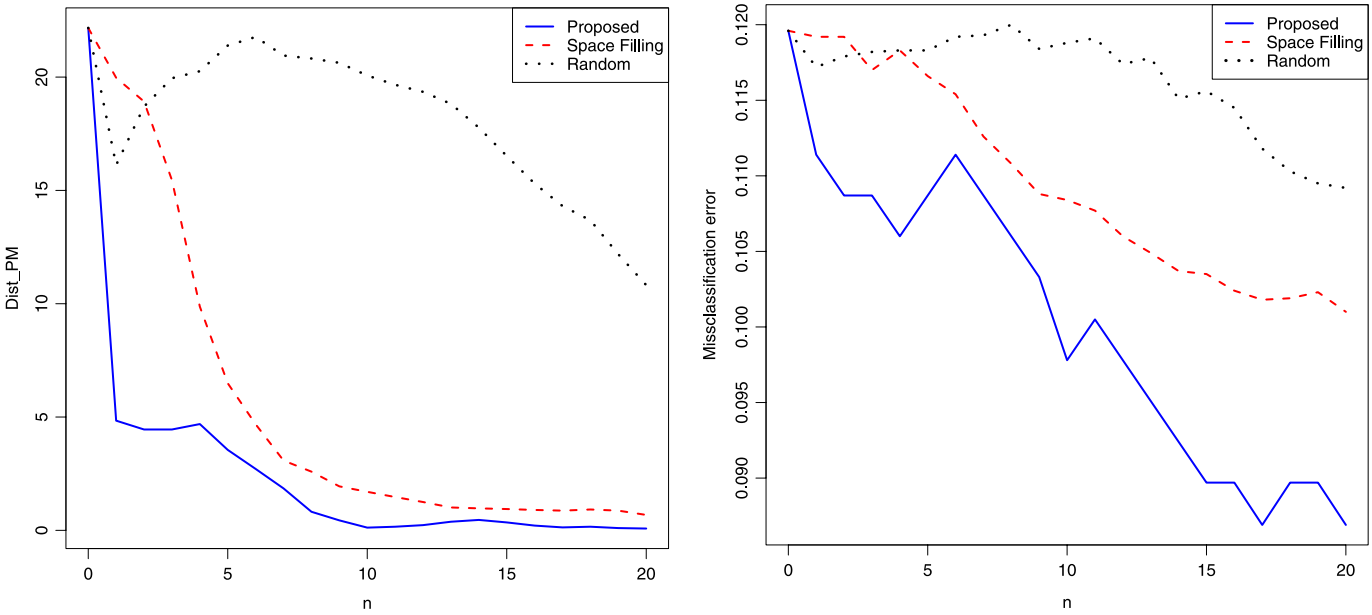

Figure 4. Learning curves of the proposed method compared with two naive methods.

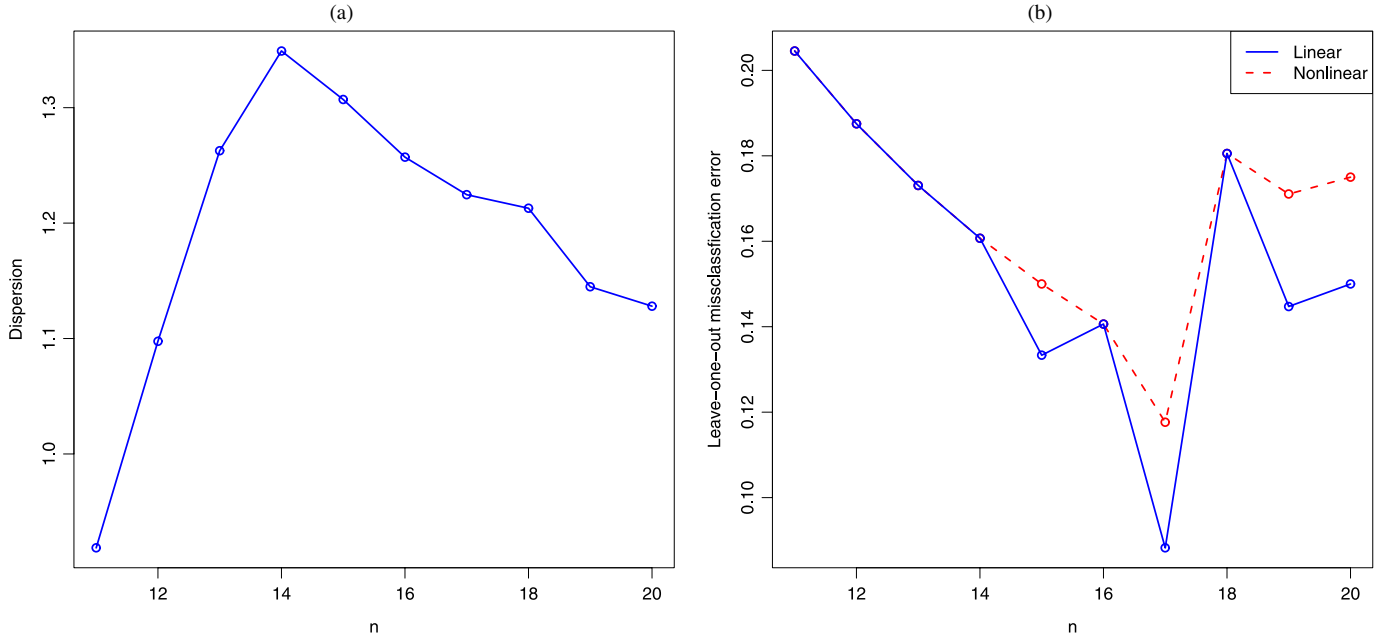

Figure 5. Diagnostics for our proposed ALSD method: estimates of dispersion parameter (a) and leave-one-out misclassification error (b).

estimated by $\hat{\phi} = X^2/(n-p)$, where n is the number of observations; p is the number of parameters in the model; X^2 is the Pearson statistic, defined as $X^2 = \sum_{i=1}^{n}(y_i - \hat{p}_i)^2/(\hat{p}_i(1-\hat{p}_i))$, and $\hat{p}_i$ is the estimated probability of $y_i = 1$ from our proposed method. If the logit model is appropriate, then ϕ should be 1. Figure 5(a) shows that $\hat{\phi}$'s are around 1 in the active learning procedure.

They are also much smaller than the 95% critical values (1.94 for $n = 11$ and 1.62 for $n = 20$) in the frequentist approach. The figure shows that the linearity assumption is adequate in the current study.

To assess the prediction error due to model lack of fit, we computed the leave-one-out misclassification error, given in Figure 5(b) under "linear." The prediction errors were reasonably low. To further improve the prediction accuracy, we could use a nonlinear hypersurface. One possible nonlinear model is $z = wx_1^{\alpha_1} + (1-w)x_2^{\alpha_2}$ with $\alpha_1, \alpha_2 \geq 0$, where α_1 and α_2 are estimated from the data. This is a reasonable model because it maintains the monotonicity of the profile features. Figure 5(b) plots the leave-one-out misclassification error of the model under "nonlinear." It shows that using this particular nonlinear model does not further reduce the misclassification error. In fact, there is a slight increase for $14 \leq n \leq 20$. Thus the linear model seems to be adequate for this problem. Both models estimated from the full data are shown in Figure 6, which clearly shows that the linear hyperplane is a good approximation to the nonlinear hypersurface.

6. SIMULATIONS

6.1 Numerical Examples

As stated earlier, the proposed method is expected to be flexible and robust to model assumptions. We conducted some simulations to study its performance. The simulated data were based on different models of $F(\mathbf{x})$. Four models were used in the study:

$$\text{Logistic distribution:} \quad F(\mathbf{x}) = \frac{\exp((z-\mu)/\sigma)}{1+\exp((z-\mu)/\sigma)},$$

$$\text{Uniform distribution:} \quad F(\mathbf{x}) = (z-\mu)/\sigma - (-2)/2 - (-2),$$

$$\text{Normal distribution:} \quad F(\mathbf{x}) = \Phi\left(\frac{z-\mu}{\sigma}\right),$$

$$\text{Cauchy distribution:} \quad F(\mathbf{x}) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{z-\mu}{\sigma}\right),$$

where $z = wx_1 + (1-w)x_2$ and Φ is the standard normal distribution function. The true values of the parameters were set

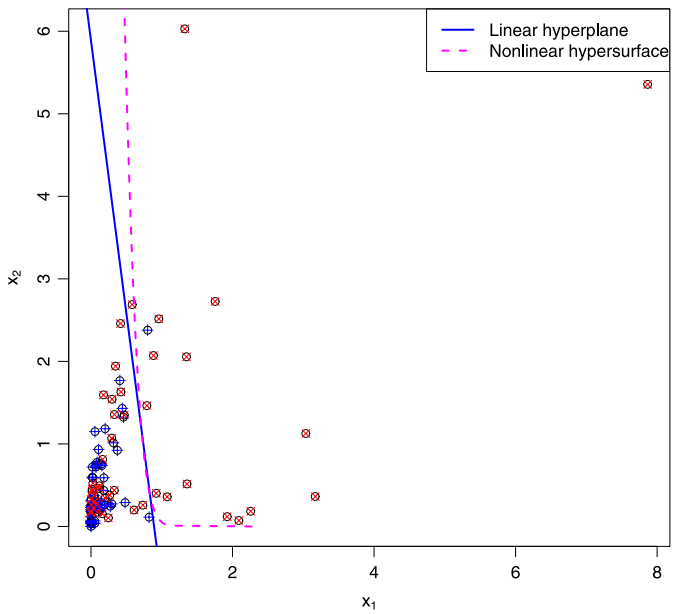

Figure 6. Comparison of the linear hyperplane and the nonlinear hypersurface estimated from the full data.

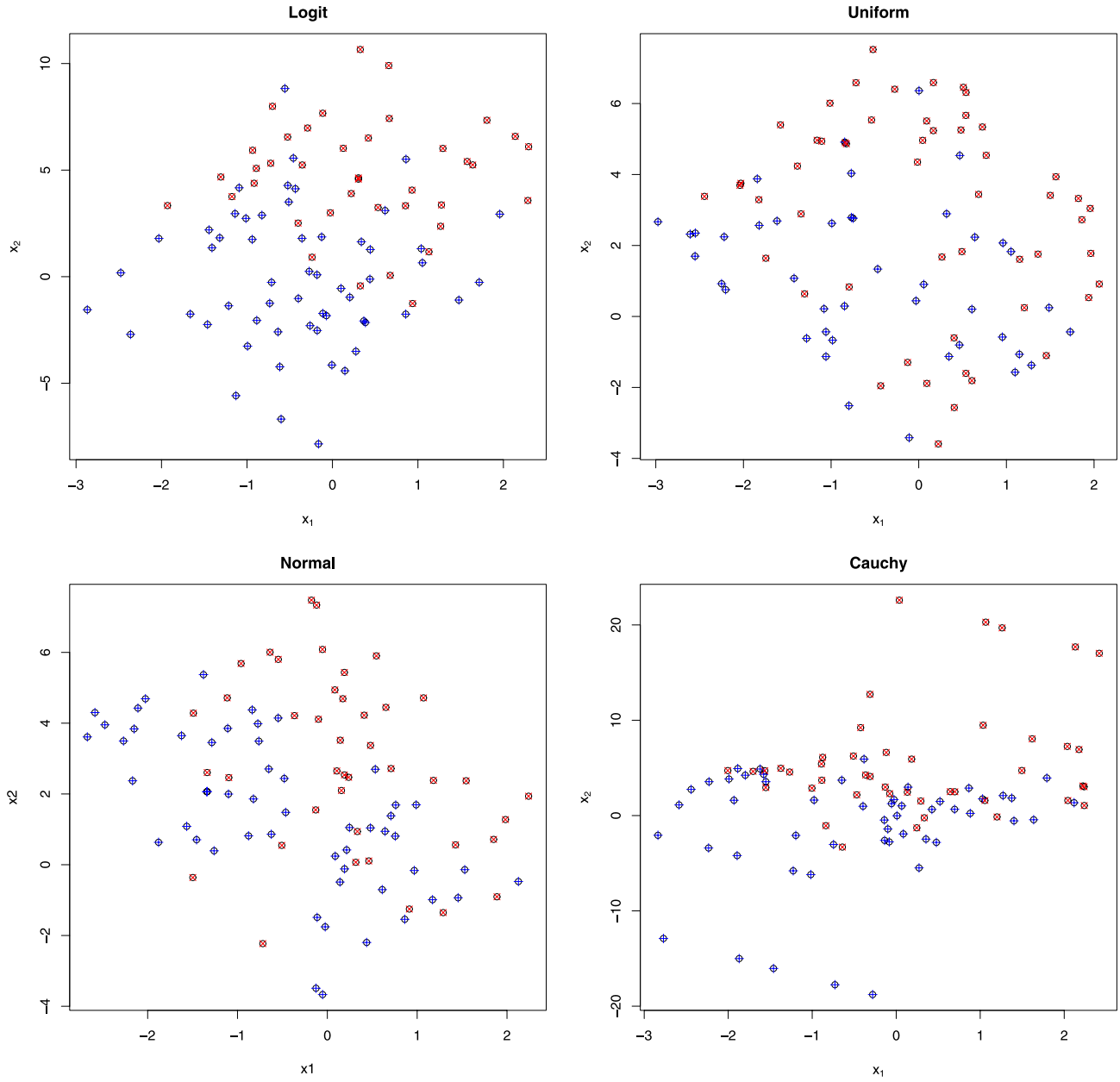

Figure 7. Illustrations of simulated data.

as $\mu = 0.5$, $\sigma = 1$, and $w = 0.7$. The response outcome at each point was generated according to $F(\mathbf{x})$. Figure 7 shows the simulated data from the four models. The range of x_1 remainined in $(-3, 3)$ for all of the plots in the figure.

In this simulation, we chose $\alpha = 0.5$ and $\alpha = 0.8$ for illustration. We used the same performance measure as in (5.5). Let $k_0 = 15$ in (4.5). The hyperparameters were specified following the heuristic procedure outlined in the previous section. We performed 100 simulations and sequentially selected $n = 30$ points in each simulation. To calculate the *Dist_PM* in (5.5), we used $\mathbf{T} = \{t_i\}$ in (5.4) points lying evenly on the true threshold hyperplane of each model. To make $\mathbf{T} = \{t_i\}$ more concentrated in the data region of the true threshold hyperplane, we adjusted the spread of $\mathbf{t}_i$ according to the value of α. When $\alpha = 0.5$, the $\mathbf{t}_i$ were evenly spaced on the true threshold hyperplane ranging from -1.5 to 1.5 on the coordinate of x_1, as shown in Figure 7. For $\alpha = 0.8$, the $\mathbf{t}_i$ were evenly spaced on the true threshold hyperplane ranging from -0.5 to 2.5 on the coordinate of x_1.

Because the random method (i.e., select the next data point randomly) outlined in Section 5 is very naive and performed poorly, we compared the proposed method with the sequential space-filling procedure used in Section 5. Note that in our problem, the accuracy of the estimated threshold hyperplane is crucial for risk prioritization. We used the *Dist_PM* to evaluate the performance of each simulation. The performance of these two methods for two α values is illustrated in Figures 8 and 9.

Clearly, our proposed ALSD method performs much better than the sequential space-filling procedure. Comparing Figures 8 and 9 shows that the methods perform better when $\alpha = 0.5$. It is well known that the estimation of extreme quantiles is much more difficult than with $\alpha = 0.5$ (see, e.g., Joseph 2004).

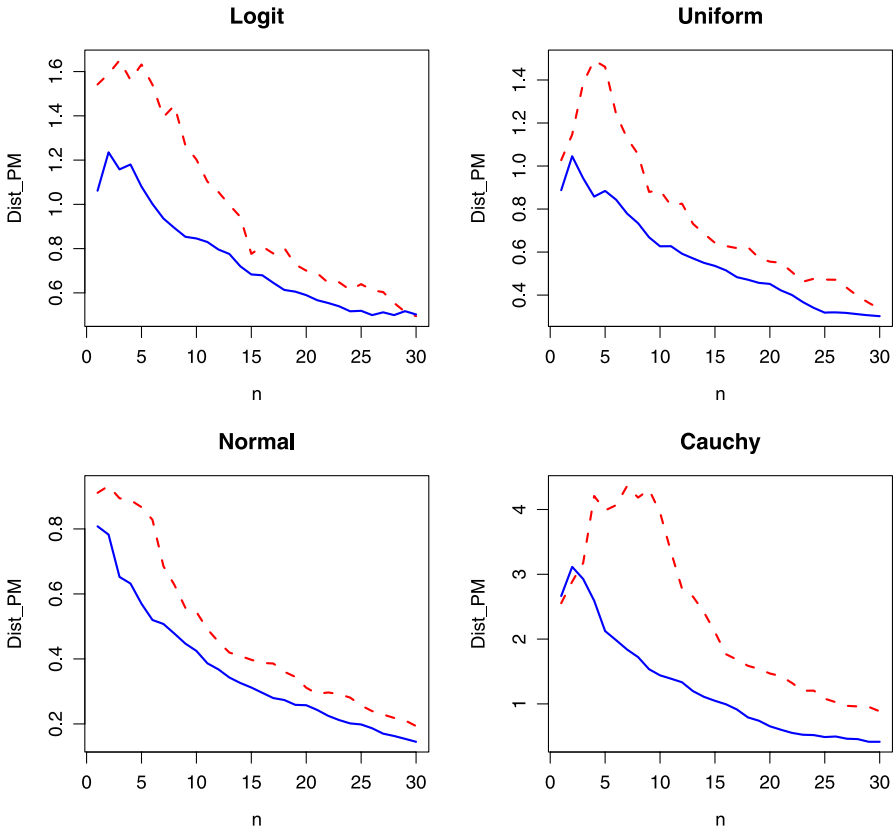

Figure 8. *Dist_PM* for four models with $\alpha = 0.5$. The solid line represents our proposed ALSD method; the dashed line, the sequential space-filling procedure.

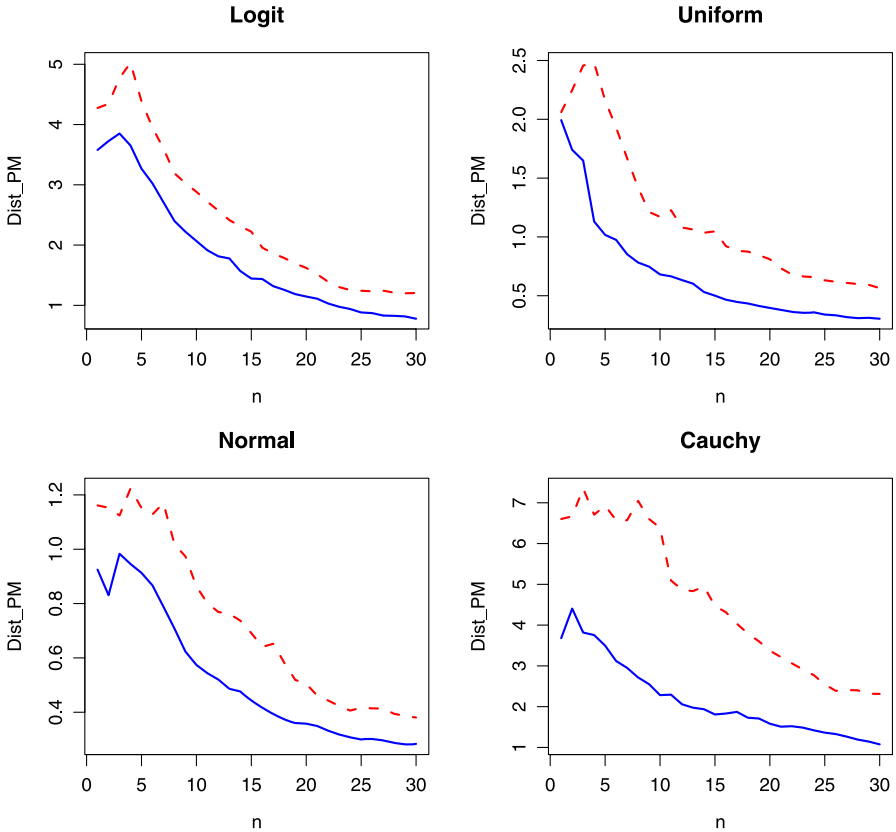

Figure 9. *Dist_PM* for four models with $\alpha = 0.8$. The solid line represents our proposed ALSD method; the dashed line, the sequential space-filling procedure.

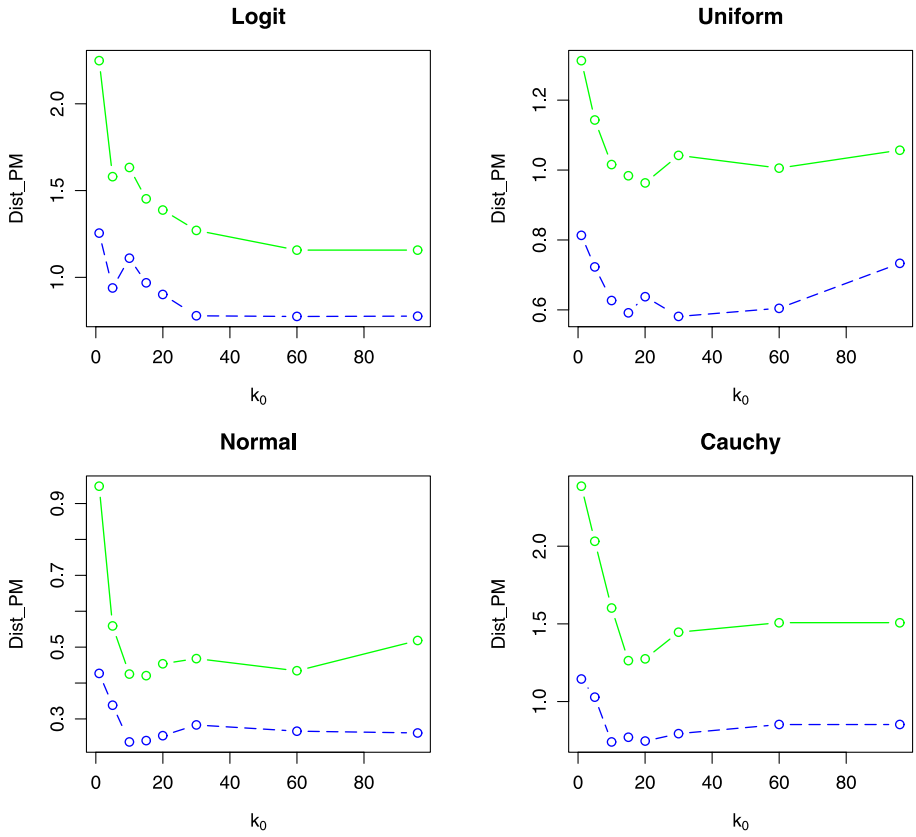

Figure 10. Performance with different k_0. Solid line, $n = 10$; dashed line, $n = 20$.

The figures also clearly show that our proposed method is quite robust to model assumptions.

In our proposed ALSD approach in (4.5), k_0 candidate points closest to the estimated hyperplane are selected. Here k_0 is considered a tuning parameter, but its optimal value has not yet been addressed. We conducted an additional experiment regarding the choice of k_0. Setting $\alpha = 0.6$, we used our proposed ALSD method for different k_0 (i.e., $k_0 = 1$, 5, 10, 15 and $k_0 = N$, where N is the total number of data points in the data set). Here $k_0 = 1$ means active learning using stochastic approximation, whereas $k_0 = N$ means active learning using a fully D-optimal–based sequential design. We generated 100 simulations for each k_0 and each model. The hyperparameters were chosen as outlined in Section 4. Figure 10 shows the simulation results.

As shown in Figure 10, except for the logistic distribution, the *Dist_PM* decreased up to some value of k_0 and then increased thereafter. This agrees with our initial intuition that choosing a large value of k_0 may not be good if the assumed model is not correct. Our method assumes the logistic model. Thus, when the model was changed to uniform, normal, or Cauchy, the method did not do well with a large k_0. As expected, the performance did not deteriorate with k_0 when the true model was logistic. It also is clear that $k_0 = 1$ was a bad choice, because the *Dist_PM* was the largest in all cases. Thus using a purely stochastic approximation method for active learning was not good in this particular problem. The best value of k_0 is not clear; the simulation results suggest choosing k_0 to be 20%–50% of N.

6.2 Comparison With Support Vector Machines

Active learning using SVMs for classification has been proposed with several variations (e.g., Campbell, Cristianini, and Smola 2000; Schohn and Cohn 2000; Tong and Koller 2001). The basic idea is to label points that lie closest to the SVM's dividing hyperplane. It is known that the hyperplane in SVM converges to the Bayes rule $P(Y = 1|\mathbf{x}) = \alpha$, where $\alpha = 0.5$. Our proposed ALSD method also can converge to the threshold hyperplane when $\alpha = 0.5$. Active learning with SVMs requires an initial sample of data points. To provide a fair comparison, we used eight points as the initial sample, chosen based on the stratified random sampling. We implemented the method as follows. With the initial guess on the parameters μ_0, σ_0, and w_0, we got $z = w_0 x_1 + (1 - w_0)x_2$. We then divided the range of z into four strata as $(-\infty, \mu_0 - 1.6\sigma_0)$, $[\mu_0 - 1.6\sigma_0, \mu_0)$, $[\mu_0, \mu_0 + 1.6\sigma_0)$, and $[\mu_0 + 1.6\sigma_0, +\infty)$. Because each point $\mathbf{x}$ can be mapped into the z value, we randomly choose two $\mathbf{x}'$s in each stratum based on the corresponding z value. The choice of the constant ± 1.6 was based on the asymptotic optimality of the estimators under the logistic distribution (see, e.g., Neyer 1994). We chose the hyperparameters as before. We generated 100 simulations for comparison.

From Figure 11, we can see that our proposed ALSD method has much smaller *Dist_PM* values compared with the active learning with SVM approach. Moreover, our ALSD approach is quite stable, whereas the SVM approach is quite unstable for small n. The SVM is not robust, because adding one more point into the training set can cause significant changes in the SVM's dividing hyperplane. Thanks to the use of the Bayesian

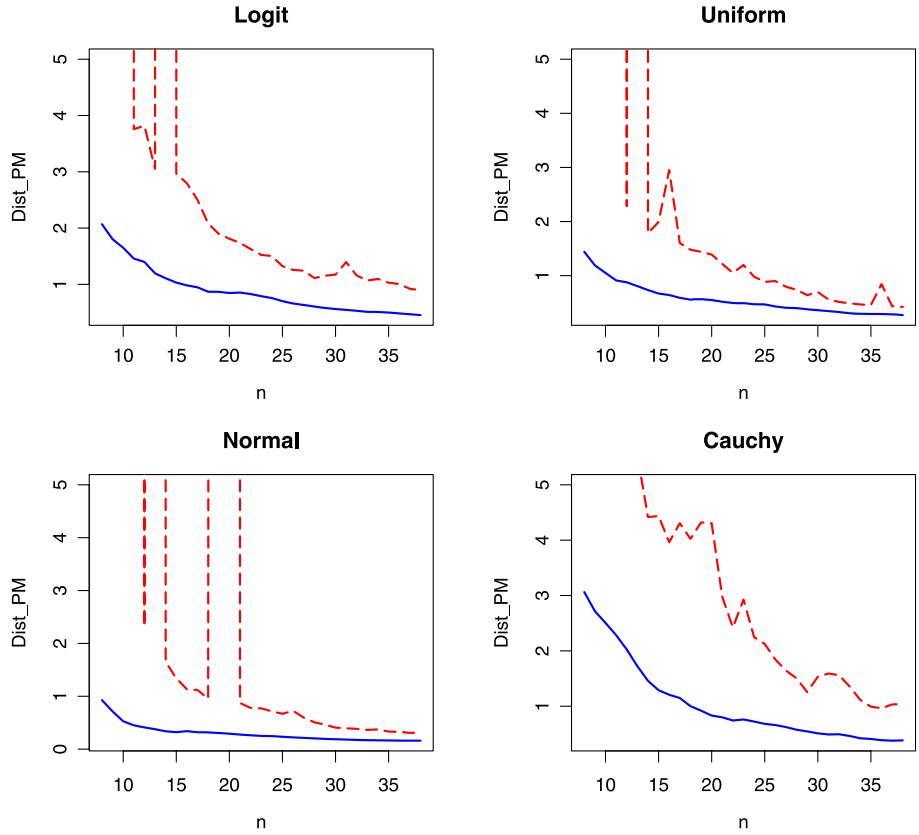

Figure 11. Comparison of our ALSD method and active learning with SVM. The solid line represent our proposed ALSD method; the dashed line, active learning with SVM.

approach, the estimation in the proposed active learning method is stable.

Our proposed ALSD method seems to converge within 20 steps, whereas the active learning with SVM approach needs at least 10 more steps to achieve a similar performance. The improvement is even more pronounced with heavy-tailed distributions, such as Cauchy. Thus, in this particular problem, our proposed ALSD method outperformed active learning with SVM in all respects, including accuracy, stability, and robustness.

7. DISCUSSION AND CONCLUSION

In this article we have proposed an ALSD approach and reported its application to a real-world problem in money laundering detection. Because of the large numbers of transactions and various business categories of investigational resources, finding an efficient way to determine the threshold hyperplane for prioritization is crucial. Our proposed method can efficiently and accurately estimate the threshold hyperplane, and its performance is robust to model assumptions. It can help investigators focus their efforts on those accounts of most importance and thus significantly improve money laundering detection.

Our proposed ALSD method uses a combination of stochastic approximation and optimal design methods. From the sequential design perspective, we have shown that our proposed method works better than either the stochastic approximation or optimal design approaches. Through simulations, we also have shown that our proposed method outperforms active learning methods using SVMs. Regarding the choice of k_0 [i.e., the number of candidate points in (4.5)], the results of our simulation study suggest choosing k_0 to be 20%–50% of N.

We have explored the use of our proposed method for two profile features, $\mathbf{x} = (x_1, x_2)^T$, with the threshold hyperplane is linear in $\mathbf{x}$. When the linearity assumption is reasonable, it can be easily extended to higher dimensions. In multivariate situations with $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$, we can define a synthetic variable z as a convex combination of the profile features (i.e., $z = \sum_{i=1}^{p} w_i x_i$), where $w_i \geq 0$ and $\sum_{i=1}^{p} w_i = 1$. We then can apply the active learning criterion (4.5) to select the next data point. Regarding the choice of the priors, we use the normal prior for the location parameter μ, the exponential prior for the scale parameter σ, and the Dirichlet prior for the weight parameters $w = (w_1, w_2, \ldots, w_p)^T$. For this to work, we need to assume monotonic effects for each of the profile features. This assumption seems reasonable in problems that we have encountered so far. If the threshold hyperplane has a nonlinear form in the profile features $\mathbf{x}$, then the linearity assumption in the model may lead to lack of fit and poor prediction accuracy. This can be alleviated by using a nonlinear model as described in Section 5, that is, by taking $z = \sum_{i=1}^{p} w_i x_i^{\alpha_i}$, where $\alpha_i \geq 0$ for all $i = 1, \ldots, p$. Another strategy for incorporating the nonlinearity is to consider the so-called "kernel trick" (Schölkopf and Smola 2002) on the synthetic variable z for the logit model in (4.1); that is, z can be expressed as an inner product in the reproducing kernel Hilbert space (Wahba 1990). Generalizing the active learning criterion (4.5) for the nonlinear threshold surface is an interesting topic for future research.

Our proposed ALSD method is flexible in estimating the threshold hyperplane for different values of α. In contrast, the standard SVM is appropriate mainly for classification problems with $\alpha = 0.5$. Lin, Lee, and Wahba (2002) proposed a modified SVM to account for α different from 0.5; however, an evaluation of this active learning with modified SVM approach is not available in the literature.

Although our proposed ALSD method was motivated by the problem of detecting money laundering, the method's sequential nature can be linked to other applications, such as sensitivity experiments (Neyer 1994), bioassay experiments (McLeish and Tosh 1990), contour estimation in computer experiments (Ranjan, Bingham, and Michailidis 2007), and identification of lead compounds in drug discovery (Abt et al. 2001). Abt et al. (2001) used a two-stage sequential approach to minimize the number of physical tests and select a set of good candidate compounds. Their work shares two common features with ours: large numbers of independent variables and high cost to measure the potencies of chemical compounds. These features are common in other fields; for example, in signal processing or image recognition, observations often are available but are not labeled or investigated to determine the response. Each observation can be a functional curve or can consist of many data points in a high-dimensional space. Because of the complexity of the observations, obtaining the responses is time-consuming and costly. By transforming the data into several uncorrelated monotonic profile features, our proposed ALSD method can efficiently exploit the level of interest of the response.

APPENDIX: EQUIVALENCE BETWEEN (4.7) AND (4.8)

From (4.6), $I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \mathbf{x}) = I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) + \kappa_x \boldsymbol{\eta}_x \boldsymbol{\eta}_x^T$, where $\kappa_x = e^{g(\mathbf{x})}/(1 + e^{g(\mathbf{x})})^2$ and $\boldsymbol{\eta}_x = \frac{\partial g(\mathbf{x})}{\partial \boldsymbol{\theta}}$. Under mild regularity conditions, the Fisher information matrix $I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ is positive semidefinite and nonsingular. Thus, applying the identity $\det(A + c\mathbf{x}\mathbf{x}^T) = \det(A)(1 + c\mathbf{x}^T A^{-1}\mathbf{x})$, we obtain

$$\begin{aligned}
&\det(I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \mathbf{x})) \\
&\quad = \det\bigl(I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) + \kappa_x \boldsymbol{\eta}_x \boldsymbol{\eta}_x^T\bigr) \\
&\quad = \det(I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)) \\
&\qquad \times \bigl(1 + \kappa_x \boldsymbol{\eta}_x^T I^{-1}(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)\boldsymbol{\eta}_x\bigr).
\end{aligned}$$

Thus $\min_{\mathbf{x}} \det(I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \mathbf{x}))$ is the same as $\min_{\mathbf{x}} \kappa_x \boldsymbol{\eta}_x^T \times I^{-1}(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)\boldsymbol{\eta}_x$. Now under the constraint in (4.7), $\kappa_x = \alpha(1 - \alpha)$ is a constant. Thus we get (4.8). Note that $\boldsymbol{\eta}_x = (-1/\sigma, -\log(\alpha/(1 - \alpha))/\sigma, (x_1 - x_2)/\sigma)^T$ under the constraint in (4.7).

[Received November 2007. Revised December 2008.]

REFERENCES

Abt, M., Lim, Y., Sacks, J., Xie, M., and Young, S. (2001), "A Sequential Approach for Identifying Lead Compounds in Large Chemical Databases," *Statistical Science*, 16, 154–168.

Albert, A., and Anderson, J. A. (1984), "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models," *Biometrika*, 71, 1–10.

Campbell, C., Cristianini, N., and Smola, A. (2000), "Query Learning With Large Margin Classifiers," in *Proceedings of 17th International Conference on Machine Learning*, pp. 111–118.

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996), "Active Learning With Statistical Models," *Journal of Artificial Intelligence Research*, 4, 129–145.

Fedorov, V. V. (1972), *Theory of Optimal Experiments*, New York: Academic Press.

Fukumizu, K. (2000), "Statistical Active Learning in Multilayer Perceptrons," *IEEE Transactions on Neural Networks*, 11 (1), 17–26.

Johnson, M., Moore, L., and Ylvisaker, D. (1990), "Minimax and Maximin Distance Design," *Journal of Statistical Planning and Inference*, 26, 131–148.

Joseph, V. R. (2004), "Efficient Robbins–Monro Procedure for Binary Data," *Biometrika*, 91, 461–470.

Joseph, V. R., Tian, Y., and Wu, C. F. J. (2007), "Adaptive Designs for Stochastic Root-Finding," *Statistica Sinica*, 17, 1549–1565.

Kiefer, J. (1959), "Optimum Experimental Designs," *Journal of the Royal Statistical Society*, Ser. B, 21, 272–304.

Lewis, D., and Gale, W. (1994), "A Sequential Algorithm for Training Text Classifiers," in *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, New York: Springer-Verlag, pp. 3–12.

Lin, Y., Lee, Y., and Wahba, G. (2002), "Support Vector Machines for Classification in Nonstandard Situations," *Machine Learning*, 46, 191–202.

MacKay, D. J. C. (1992), "Information-Based Objective Functions for Active Data Selection," *Neural Computation*, 4 (4), 590–604.

McLeish, D. L., and Tosh, D. (1990), "Sequential Designs in Bioassay," *Biometrics*, 46, 103–116.

Neyer, B. T. (1994), "D-Optimality-Based Sensitivity Test," *Technometrics*, 36, 61–70.

Pukelsheim, F. (1993), *Optimal Design of Experiments*, New York: Wiley.

Ranjan, P., Bingham, D., and Michailidis, G. (2007), "Sequential Experiment Design for Contour Estimation From Complex Computer Codes," *Technometrics*, 50 (4), 527–541.

Santner, T. J., and Duffy, D. E. (1986), "A Note on A. Albert and J. A. Andersons Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models," *Biometrika*, 73, 755–758.

Schohn, G., and Cohn, D. (2000), "Less Is More: Active Learning With Support Vector Machines," in *Proceedings of the Seventeenth International Conference on Machine Learning*.

Schölkopf, B., and Smola, A. (2002), *Learning With Kernels*, Cambridge: MIT Press.

Silvapulle, M. J. (1981), "On the Existence of Maximum Likelihood Estimators of the Binomial Response Model," *Journal of the Royal Statistical Society*, Ser. B, 43, 310–313.

Silvey, S. D. (1980), *Optimal Design*, London: Chapman & Hall.

Tong, S., and Koller, D. (2001), "Support Vector Machine Active Learning With Applications to Text Classification," *Journal of Machine Learning Research*, 2, 45–66.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadephia: SIAM.

Wu, C. F. J. (1985), "Efficient Sequential Designs With Binary Data," *Journal of the American Statistical Association*, 80, 974–984.

Ying, Z., and Wu, C. F. J. (1997), "An Asymptotic Theory of Sequential Designs Based on Maximum Likelihood Recursions," *Statistica Sinica*, 7, 75–91.

Young, L. J., and Easterling, R. G. (1994), "Estimation of Extreme Quantiles Based on Sensitivity Tests: A Comparative Study," *Technometrics*, 36, 48–60.