Published in final edited form as:
J Am Stat Assoc. 2013 June 1; 108(502): 553–565. doi:10.1080/01621459.2013.775949.

Auxiliary marker-assisted classification in the absence of class identifiers

Yuanjia Wang [Assistant Professor],
Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032

Huaihou Chen [post-doctoral fellow],
New York University

Donglin Zeng [Professor],
Department of Biostatistics, University of North Carolina at Chapel Hill

Christine Mauro [graduate student],
Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032

Naihua Duan [Professor], and
Department of Psychiatry, Columbia University

M. Katherine Shear [Professor]
School of Social Work and College of Physicians and Surgeons, Columbia University

Yuanjia Wang: yuanjia.wang@columbia.edu; Huaihou Chen: Huaihou.Chen@nyumc.org; Donglin Zeng: dzeng@email.unc.edu; Christine Mauro: cmm2212@columbia.edu; Naihua Duan: naihua.duan@columbia.edu; M. Katherine Shear: ks2394@columbia.edu

Abstract

Constructing classification rules for accurate diagnosis of a disorder is an important goal in medical practice. In many clinical applications, there is no clinically significant anatomical or physiological deviation exists to identify the gold standard disease status to inform development of classification algorithms. Despite absence of perfect disease class identifiers, there are usually one or more disease-informative auxiliary markers along with feature variables comprising known symptoms. Existing statistical learning approaches do not effectively draw information from auxiliary prognostic markers. We propose a large margin classification method, with particular emphasis on the support vector machine (SVM), assisted by available informative markers in order to classify disease without knowing a subject's true disease status. We view this task as statistical learning in the presence of missing data, and introduce a pseudo-EM algorithm to the classification. A major distinction with a regular EM algorithm is that we do not model the distribution of missing data given the observed feature variables either parametrically or semiparametrically. We also propose a sparse variable selection method embedded in the pseudo-EM algorithm. Theoretical examination shows that the proposed classification rule is Fisher consistent, and that under a linear rule, the proposed selection has an oracle variable selection property and the estimated coefficients are asymptotically normal. We apply the methods to build decision rules for including subjects in clinical trials of a new psychiatric disorder and present four applications to data available at the UCI Machine Learning Repository.

Keywords

Large margin classification; Support vector machine; Statistical learning; Classification rules; Missing data; Diagnostic and Statistical Manual of Mental Disorders

1 Introduction

Statistical learning has been a powerful tool for classification problems. An effective statistical learning algorithm provides a decision rule for classification based on feature variables which can minimize the misclassification rate evaluated against a gold standard. Traditional statistical learning approaches for classification are either regression-model based (e.g., multiple logistic regression) or density-based (e..g, linear discriminant analysis) which rely on various model assumptions that may not be satisfied especially for high-dimensional feature data. In contrast, over the past decade, the large margin classification (e.g., Shen et al. 2003; Wang, Shen, Pan 2009) in particular the support vector machine (SVM, Vapnik 1995), has been proven to be a successful model-free statistical learning technique for classification and prediction problems, and is widely used in many applications including cancer epidemiology, gene expression studies, personalized medicine, and image classification (e.g., Moguerza and Muoz 2006; Klöppel et al. 2008; Orru et al. 2012).

The success of the large margin classification including the SVM in applications is supported by its optimal theoretical properties: Lin (2002) showed that the SVM directly estimates the Bayes rule which minimizes the expected missclassification rate without estimating the class probabilities. There is a large body of literature on SVM and its derivatives such as generalizing the standard SVM to the multi-category case (Liu et al. 2005), improving robustness to outliers (Wu and Liu 2007), and using SVM to maximize the area under a receiver operating characteristic curve (Wang et al. 2011). With high dimensional feature variables, usually only a small proportion of variables is expected to be truly associated with an outcome. Therefore the feature space contains a large number of noise variables. In this case, the standard SVM may not perform well, and alternative approaches combines the L_1-penalty or the non-convex penalty (SCAD) with SVM (e.g., Zhu et al. 2003; Zhang et al. 2006) to provide sparse solutions.

Applying the above methods to classify patients into disease classes (diseased versus non-diseased, i.e., disease status) using subject-specific feature information requires knowing all class labels. However, in clinical practices, perfect diagnosis or classification of a subject's disease status is often difficult. In cases where the clinically significant anatomical and/or physiological deviation from the normal structure or function of any body part is not known, it can be impossible to determine exactly what criteria should be used to identify a disease status. Historically, expert opinion has been used for diagnosis. However, these opinions may be biased or objective, and greater emphasis has been placed on incorporating empirical evidence and using data-driven approaches to inform clinical decision making (Kraemer, Shrout and Rubio-Stipec 2007).

Although gold standard disease status is often unknown, one or more auxiliary markers informative of the latent disease status may be available, and these markers have different distributions in diseased and non-diseased subjects. For example, for some psychiatric disorders, especially new disorders such as complicated grief (Shear et al. 2005), different diagnostic criteria sets have been proposed in the clinical literature, but there is no definitive measure for this condition (Prigerson et al. 2009; Shear et al. 2011). Despite this fact, feature variables such as symptom rating scales are available and appear to be effective screening tools for complicated grief. In addition, informative markers for disease severity (e.g., Work Social Assessment Scales, Mundt et al. 2002) are often measured in clinical studies to assess disease-related impairment in functioning. Another example where gold standard class labels might be unavailable is in cancer research. The research goal is to build a prediction model for tumor subtype from feature variables such as subject's genotypes, in a situation in which accurate but invasive biomarkers (e.g., tumor histology or physiology) are available,

but do not completely identify subtypes. It is desirable to build a model for tumor subtypes from the less invasive tests (genotype profile) instead of an invasive diagnostic test (tumor histology) obtained from biopsy.

There is a large body of literature proposed for unsupervised learning where all the class labels are unknown (e.g., Hinton et al. 1999; Xu et al. 2004). In particular, Nigam et al. (1998) proposed parametric methods to perform text mining through mixture models and an EM algorithm. Xu et al. (2004) proposed a maximum margin clustering using SVM where all possible combination of the class labels enters in an exhaustive search and the one with the minimum loss is selected. Xu and Schuurmans (2005) generalized this earlier work to semi-supervised learning and multiclass problems. Due to heavy computational burden, the exhaustive search quickly becomes infeasible with increasing sample size (exponential increase). Zhang et al. (2007) proposed an approximation to the exact solution to improve computational speed and made methods in Xu et al. (2004) practical. When there are class labels on a subgroup of subjects, Rigollet (2007) derived error bounds for semisupervised classification and used a density based approach to achieve derived rates of convergence. Wang, Shen and Pan (2009) proposed semi-supervised learning through an EM algorithm (Dempester, Laird, and Rubin 1977) where conditional probabilities of missing class labels given feature variables are computed through an iterative algorithm (Wang et al. 2008). Culp (2011) proposed semi-supervised learning that combines feature-based data and graph-based data for classification.

In our problem, none of the subjects have a known disease status, or class label, but all of them have informative markers which can viewed as partial representations of the class labels. Therefore none of the current semi-supervised large margin classification techniques are directly applicable. In addition, existing unsupervised learning algorithms are not appropriate because they do not efficiently extract information from the informative markers to infer the unobserved disease status. Using informative markers directly as a surrogate outcome for the disease status may lose information and lead to classification rules only partially predicting the true disease status. Furthermore, when there are more than one disease markers, it is not clear which marker to choose or what combination to use as a surrogate for disease status.

To tackle these challenges, in this work, we propose a large margin classification-based learning approach, implemented with the SVM, to construct classification rules without observing a subject's true disease status but effectively incorporating information from available informative markers. Specifically, we view the lack of gold standard disease status as a missing data problem, and introduce the EM algorithm for missing data to the classification: we propose a novel pseudo-EM algorithm based on loss function as a pseudo-likelihood. One major distinction from the traditional EM algorithm is that we no longer model the distribution of missing data given the observed feature variables either parametrically or semiparametrically. Instead, we leave this distribution completely unspecified and treat the loss function of the classification as a natural surrogate. To select important feature variables, we use a penalty function that encourages sparse fit to perform simultaneous feature selection and classification in the pseudo-EM algorithm. The proposed general framework is implemented with SVM. We study theoretical properties of the proposed methodology and show that the derived decision function is Fisher consistent. Additionally, we show that when using an appropriate sparse penalty in the EM algorithm, this method possesses an oracle variable selection property as if the true coefficients of the decision rule were known when the decision function is linear. In this case, the estimated coefficients of the non-null variables are asymptotic normal. Extensive simulation studies are conducted to compare the proposed approach with parametric alternatives and sensitivity analyses are performed to examine the effect of violations of some assumptions. Finally,

practical utility of the proposed method is demonstrated through constructing decision rules for selecting patients for clinical trials of complicated grief (Shear et al. 2005) and four other data sets available at the UCI Machine Learning Repository.

2 Methodology

2.1 Model framework and assumptions

Let D_i denote the true disease status (class label) of the ith participant coded as $D_i = 1$ for diseased and $D_i = -1$ for non-diseased. Let $\boldsymbol{X}_i$ denote a vector of a potentially large number of feature variables, which are to be selected for classifying a subject into a diseased or non-diseased group. In many applications, D_i's are not observed; instead, some disease-informative markers are available and we denote the vector of these markers by $\boldsymbol{Z}_i$ for subject i. The observed data from n i.i.d. subjects are $(\boldsymbol{X}_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$. The informative markers $\boldsymbol{Z}_i$ are collected on the observed sample, but are not available on future subjects for whom predictions are performed using feature variables $\boldsymbol{X}_i$.

We assume that $\boldsymbol{Z}_i$ and $\boldsymbol{X}_i$ are conditionally independent given a subject's disease status D_i; that is, $\boldsymbol{Z}_i$ is non-informative of the feature variables within diseased or non-diseased group. Although this is a common assumption in the literature (e.g., Chung et al. 2006), our proposed methods can be easily generalized to allow $\boldsymbol{Z}_i$ to depend on $\boldsymbol{X}_i$ in each group (for example, see Section 6). Furthermore, we assume that given the true class label D_i, markers $\boldsymbol{Z}_i$ follow a multivariate normal distribution:

$$\boldsymbol{Z}_i | D_i = d \sim MVN(\mu_d, \textstyle\sum_d), \quad d = 1, -1.$$

To avoid non-identifiability of group labels, we assume that it is known in advance a particular marker component has population means that differ between groups, and for this marker it is known whether the diseased group has a larger population mean or a smaller mean. This assumption is usually met in practice since substantive knowledge usually informs whether disease subjects have a larger mean or a smaller mean of an informative marker. In our Grief study data example (see Section 5.1), Z_i is a marker measuring subjects' functioning impairment with a larger population mean value (more impairment) associated with diseased group. Note that this assumption does not imply that subjects with the higher marker value are diseased due to random variability in marker values. In other words, the informative marker alone does not fully identify disease groups. Moreover, because of the marker variability, using a single marker may not be as informative as using all the markers which may contain additional information about the disease status. Finally, the Gaussian distribution assumption is not essential and it can be replaced by any parametric distributions to model this mixture population.

2.2 Large margin classification via a pseudo-EM algorithm

When disease labels D_i's are observed, large margin classification can be used to identify classification rules based on the feature variables, $\boldsymbol{X}_i$. Specifically, it is equivalent to minimizing a margin loss subject to a penalty, that is,

$$\arg\min_{g \in \mathcal{H}_K} \left[\sum_i L\{D_i g(\boldsymbol{X}_i)\} + \frac{\lambda_n}{2} \|g\|^2 \right], \quad (1)$$

where $L(\cdot)$ is a margin loss defined by functional margin $dg(\boldsymbol{x})$, $\mathcal{H}_K$ is a reproducible kernel Hilbert space (RKHS) with the kernel function $K(\boldsymbol{x}, \boldsymbol{y})$, $\|g\|$ is the norm of g in the RKHS,

and λ_n is a tuning parameter depending on the sample size. Here, $g(\boldsymbol{x})$ is a general decision function to be estimated. Note that in many applications $g(\boldsymbol{x})$ is often assumed to be a linear function, that is, $g(\boldsymbol{x}) = b + \boldsymbol{\beta}^T\boldsymbol{x}$, and $\|g\| = \|\boldsymbol{\beta}\|$. Examples of margin loss functions include the hinge loss $L(z) = (1 - z)_+$ for the SVM and its variations, $(1-z)_+^q$ with $q > 1$ (Lin 2002), the ψ-loss with $L(z) = 1 - \text{sign}(z)$ if $z \geq 1$ or $z < 0$ and $2(1 - z)$ otherwise (Shen et al. 2003), and the logistic loss $\log\{1 + \exp(z)\}$ (Zhu and Hastie, 2005).

When D_i's are not available, we treat D_i's as missing data and adopt a pseudo-EM algorithm for estimation. However, we do not attempt to specify a model for D_i given $\boldsymbol{X}_i$ as required in the usual EM context; instead, we construct nonparametric pseudo-probabilities and aim to minimize the original loss function in (1) in the M-step of the algorithm. The intuition of the proposed method is as following: since D_i are missing but $\boldsymbol{Z}_i$ are partial manifestation of D_i, one would naturally consider to infer disease status using available information from $\boldsymbol{Z}_i$ and use feature variables X_i to form classification rules. Taking a uni-dimensional Z_i as example, a subject with a large value of Z_i has a high probability of being diseased; at the same time, if the subject is classified as diseased by X with little loss, then one should not reclassify this subject as non-diseased with high probability. Iterating through an EM-type algorithm reflects how to appropriately combine information from Z_i and X_i for classification.

To be more specific, we first use the assumed Gaussian mixture model to estimate μ_d and Σ_d by maximizing the marginal log-likelihood,

$$\sum_i \log\left[\sum_d \Pr(D_i=d) f(\boldsymbol{Z}_i|D_i=d;\mu_d, \textstyle\sum_d)\right],$$

where $f(\boldsymbol{Z}_i|D_i = d; \mu_d, \Sigma_d)$ is the multivariate normal density function. The estimation can easily be implemented by many existing numerical routines (e.g, McLachlan and Pee 2000). We denote these estimators as $\hat{\mu}_d$ and $\hat{\Sigma}_d$. Next, we impose the pseudo-likelihood function for D_i given $\boldsymbol{X}_i$ as proportional to the exponent of the negative of the objective function in (1). Therefore, under the conditional independence assumption, the pseudo-log-likelihood for the complete data $(\boldsymbol{Z}_i, D_i, \boldsymbol{X}_i)$ for $i = 1, \ldots, n$ is (up to some constant):

$$\sum_i \log\left[f(\boldsymbol{Z}_i|D_i;\widehat{\mu}_{D_i}, \widehat{\textstyle\sum}_{D_i})\right] - \left[\sum_i L\{D_i g(X_i)\} + \frac{\lambda_n}{2}\|g\|^2\right].$$

The proposed pseudo-EM algorithm is then based on maximizing the conditional expectation of this pseudo-log-likelihood given the observed data.

We carry out the following E- and M-step at the mth iteration.

(E-step)—In this step, we compute the posterior pseudo-log-likelihood of complete data $(\boldsymbol{Z}_i, D_i, \boldsymbol{X}_i)$ given the observed data. The posterior probability of $D_i = 1$ given the observed data is

$$w_i^{(m)} = \frac{f(\boldsymbol{Z}_i;\widehat{\mu}_1, \widehat{\textstyle\sum}_1)\Pr^{(m-1)}(D_i=1|\boldsymbol{X}_i)}{\sum_{d\in\{-1,1\}} f(\boldsymbol{Z}_i;\widehat{\mu}_d, \widehat{\textstyle\sum}_d)\Pr^{(m-1)}(D_i=d|\boldsymbol{X}_i)},$$

where $\Pr^{(m-1)}(D_i = d|\boldsymbol{X}_i)$ denotes the conditional probability of D_i given $\boldsymbol{X}_i$. For the usual EM algorithm, a probability model would be assumed for $\Pr(D_i|\boldsymbol{X}_i)$ (e.g., logistic regression). We propose a nonparametric pseudo-probability model of D_i given $\boldsymbol{X}_i$ based on the loss function without introducing extra modeling assumptions. To be specific, we construct

$$\Pr^{(m-1)}(D_i=d|\boldsymbol{X}_i)=\frac{p[d, g^{(m-1)}(\boldsymbol{X}_i)]}{p[1, g^{(m-1)}(\boldsymbol{X}_i)]+p[-1, g^{(m-1)}(\boldsymbol{X}_i)]}, \quad d=1,-1,$$

where a pseudo-density function is

$$p[d, g(\boldsymbol{x})]=\exp[-L\{dg(x)\}], \quad (2)$$

and $g^{(m-1)}(\boldsymbol{x})$ is the decision function obtained in the $(m-1)$th iteration.

(M-step)—We update the decision function $g(\cdot)$ by minimizing the negative conditional pseudo-log-likelihood given the observed data, or equivalently, the conditional expectation of (1) given the observed data. This is equivalent to solving

$$\min_{g\in\mathscr{H}_K}\left(\sum_i\left[w_i^{(m)}L\{g(\boldsymbol{X}_i)\}+(1-w_i^{(m)})L\{-g(\boldsymbol{X}_i)\}\right]+\frac{\lambda_n}{2}\|g\|^2\right). \quad (3)$$

In the special case when $L(z)$ is the hinge loss, this minimization problem can be carried out as a weighted version of the usual SVM. For example, if $g(\boldsymbol{x})$ is assumed to be a linear function, i.e., $g(\boldsymbol{x}) = b + \boldsymbol{\beta}^T\boldsymbol{x}$, and $\|g\| = \|\boldsymbol{\beta}\|$, the minimization problem is

$$\min_{b,\beta}\sum_{i=1}^n\left\{1-u_i^{(m)}(b+\beta^T\boldsymbol{X}_i)\right\}_+ +\frac{\lambda_n}{2}\|\beta\|^2, \quad (4)$$

where $u_i^{(m)}$ is the posterior expectation of D_i given the observed data defined as

$$u_i^{(m)}=\frac{f(\boldsymbol{Z}_i;\widehat{\mu}_1,\widehat{\sum}_1)\Pr^{(m-1)}(D_i=1|\boldsymbol{X}_i)}{\sum_{d\in\{-1,1\}}f(\boldsymbol{Z}_i;\widehat{\mu}_d,\widehat{\sum}_d)\Pr^{(m-1)}(D_i=d|\boldsymbol{X}_i)}-\frac{f(\boldsymbol{Z}_i;\widehat{\mu}_{-1},\widehat{\sum}_{-1})\Pr^{(m-1)}(D_i=-1|\boldsymbol{X}_i)}{\sum_{d\in\{-1,1\}}f(\boldsymbol{Z}_i;\widehat{\mu}_d,\widehat{\sum}_d)\Pr^{(m-1)}(D_i=d|\boldsymbol{X}_i)}.$$

It follows that the quadratic optimization is equivalent to

$$\min_{b,\beta}\left(\sum_i\xi_i+\frac{\lambda_n}{2}\sum_{k=1}^q\beta_k^2\right)$$
$$\text{subject to} \quad u_i^{(m)}(b+\beta^T\boldsymbol{X}_i)\geq 1-\xi_i, \quad \xi_i\geq 0, \quad i=1,\cdots,n,$$

where $q = \dim(\boldsymbol{\beta})$ and ξ_i are slack variables allowing for overlaps between classes. The dual form of this optimization problem is

$$\max\left(\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j u_i^{(m)}u_j^{(m)}\boldsymbol{X}_i^T\boldsymbol{X}_j\right),$$

$$0\le\alpha_i\le C_n,\quad i=1,\ldots,n;\quad \sum_{i=1}^{n}\alpha_i\mathrm{sign}(u_i^{(m)})=0.$$

In fact, the computation can be easily implemented using any existing SVM package (e.g., Becker et al. 2009): one can label the ith case as one if $u_i^{(m)}\ge 0$, as negative one if $u_i^{(m)}<0$, and rescale the feature variables for each subject to be $|u_i^{(m)}|\boldsymbol{X}_i$. A similar dual problem is used to solve for the general non-linear decision function (3) through reproducible kernels.

We iterate the M-step and E-step until the change in the loss function is smaller than a pre-specified threshold to obtain $\hat{b}$ and $\hat{\beta}$. Note that the distribution of $\boldsymbol{Z}$ given D enters in the computation of a subject's probability of being diseased (E-step) and enters the subsequent computation of classifying disease status using $\boldsymbol{X}$ (M-step). Here $\boldsymbol{Z}$ and $\boldsymbol{X}$ are associated through D. The methods attempt to combine information in $\boldsymbol{Z}$ and $\boldsymbol{X}$ to recover class labels and construct classification rules.

2.3 Sparse large margin classification in EM

When the dimension of $\boldsymbol{X}_i$ is large, it is expected that only a few feature variables may be informative for classification. Additionally, in practice it may be desirable to use less number of feature variables for disease classification especially when the feature variables are costly to obtain. In this case, a sparse penalty term, $p_{\lambda_n}(|\cdot|)$, has been suggested to replace the L_2-norm in (1) to yield a regularized SVM when the hinge loss function is used (Zhu et al. 2003; Zhang et al. 2006; Becker et al. 2009). Commonly used sparse penalty terms include L_1-norm or SCAD penalty. However, since L_1-norm or SCAD tend to choose only one or a few of the correlated variables, an improved regularization method with elastic net (Enet) penalty was proposed in Zou and Hastie (2005) for regression models. Enet was adapted to the SVM in Wang, Zhu and Zou (2008). Specifically, a large margin classifier with a linear classification rule and elastic net penalty is equivalent to the following optimization problem when class labels are observed:

$$\min_{b,\beta}\left\{\sum_i L\{D_i(b+\beta^T\boldsymbol{X}_i)\}+\lambda_{1n}\|\beta\|_1+\frac{\lambda_{2n}}{2}\|\beta\|^2\right\},\quad (5)$$

where $\lambda_{1n}\ge 0$ and $\lambda_{2n}\ge 0$ are the two tuning parameters depending on the sample size. The parameter λ_{1n} controls the number of feature variables selected, while λ_{2n} controls a grouping effect.

The use of a sparse penalty can be easily incorporated into the proposed large margin classification-EM algorithm, where we replace the L_2-norm in the M-step by one of the above penalty functions. For example, with correlated feature variables we use the elastic net penalty in the implementation of the the M-step by solving

$$\min_{b,\beta}\left[\sum_i\left\{w_i^{(m)}L(b+\beta^T\boldsymbol{X}_i)+(1-w_i^{(m)})L(-b-\beta^T\boldsymbol{X}_i)\right\}+\lambda_{1n}\|\beta\|_1+\frac{\lambda_{2n}}{2}\|\beta\|^2\right].\quad (6)$$

3 Theoretical Properties

First, we justify the proposed method by showing the convergence of the algorithm in the following theorem. A proof is given in the appendix.

Theorem 1—Let $p[d, g(\cdot)] = \exp[-L\{dg(\cdot)\} - p_{\lambda_n}(g)]$, where $p_{\lambda_n}(g)$ is a penalty function, and define

$$\mathcal{Q}_n(g) \equiv \sum_{i=1}^{n} \log\left\{\sum_d \tilde{p}[d, g(\boldsymbol{X}_i)] f(\boldsymbol{Z}_i | D_i = d; \widehat{\mu}_d, \widehat{\sum}_d)\right\}.$$

Then $\mathcal{Q}_n(g)$ is non-decreasing after each iteration in the EM algorithm. Furthermore, the value of $\mathcal{Q}_n$ does not increase if and only if the decision rule based on $g(\cdot)$ does not change after an iteration.

From Theorem 1, at the final convergence of the EM algorithm, we expect the estimator for $g(\cdot)$, denoted by $\hat{g}(\cdot)$, to be a local minimum for

$$\sum_{i=1}^{n} \log\left\{\sum_d \tilde{p}[d, g(\boldsymbol{X}_i)] f(\boldsymbol{Z}_i | D_i = d; \widehat{\mu}_d, \widehat{\sum}_d)\right\}.$$

To guarantee that we can have a global minimum, we suggest to start from a wide range of initial values in the EM algorithm then choose the one yielding the smallest value.

By standard assumptions for Gaussian mixture models (McLachlan and Peel 2000), $\hat{\mu}_d$ and $\hat{\Sigma}_d$ are consistent estimators for μ_{d0} and Σ_{d0}, the true values of μ_d and Σ_d, respectively. In addition to these assumptions, we assume that

- **(c.1)** $\hat{g}(x)$ belongs to a compact set $\mathcal{G}$ in some normed space with norm $\|\cdot\|$;
- **(c.2)** $\sup_{g \in \mathcal{G}} p_{\lambda_n}(g) = o(n)$;
- **(c.3)** $[\log\{\sum_d p[d, g(\boldsymbol{X})] f(\boldsymbol{Z}|D = d; \mu_d, \Sigma_d)\}: g \in \mathcal{G}, \|\mu_d - \mu_{d0}\| < \delta_0, \|\Sigma_d - \Sigma_{d0}\| < \delta_0]$ is a P-Donsker class for some constant $\delta_0 > 0$, where $p[d, g(\boldsymbol{x})]$ is defined in (2);
- **(c.4)** $\sup_{\|g - g^*\| \geq \varepsilon} \mathcal{Q}(g) < \mathcal{Q}(g^*)$, where

$$\mathcal{Q}(g) = E\left(\log\left\{\sum_d p[d, g(\boldsymbol{X})] f(\boldsymbol{Z}|D = d; \mu_{0d}, \sum_{0d})\right\}\right)$$

and $g^*(\cdot)$ maximizes $\mathcal{Q}(g)$ on $\mathcal{G}$ (it exists due to the compactness).

Conditions (c.1) through (c.3) hold naturally when $g(\cdot)$ is a linear decision rule and the corresponding β's are bounded component-wise and $L(x)$ is the hinge loss. Then under (c.1) to (c.3), we immediately obtain $\sup_{g \in \mathcal{G}} |n^{-1} \mathcal{Q}_n(g) - \mathcal{Q}(g)| = o_p(1)$. Note that $\mathcal{Q}(g)$ is equivalent to the Kullback-Leibler information from a mixture of density functions, then condition (c.4) is the uniqueness assumption on the maximum of the mixture distribution. This condition can be satisfied in many cases. For example, if $\boldsymbol{X}$ is assumed to be multivariate normal and $g(\boldsymbol{x})$ is linear, then our subsequent Theorem 3, which requires

condition (c.4), states that any maximum of $\mathcal{Q}(g)$ should possess the same sign as $g_0(\boldsymbol{x})=b_0+\beta_0^T\boldsymbol{x}$. Thus, any maximum of $\mathcal{Q}(g)$ can be shown to be proportional to the true linear score up to some positive constant. Consequently, the uniqueness of the maximum for $\mathcal{Q}(g)$ is equivalent to the uniqueness of the maximum of the one-dimensional function $\mathcal{Q}(\theta g_0)$ for $\theta > 0$, which can be easily verified. Particularly, condition (c.4) ensures that $\hat{g}$ must converge to g^* but nothing else. The proof of this result follows from Pollard (1990) and Theorem 5.7 in van der Vaart (1996).

The next theorem proves that the decision rule based on $g^*(\cdot)$ coincides with the oracle Bayes rule where all subjects' disease labels are known if the latter belongs to $\mathcal{C}$.

Theorem 2—Let $g_0(\boldsymbol{x})$ be $p_1(\boldsymbol{x}) - p_{-1}(\boldsymbol{x})$ where $p_d(\boldsymbol{x}) = \Pr(D = d|\boldsymbol{X} = \boldsymbol{x})$. If $g_0(\boldsymbol{x}) \in \mathcal{C}$, then $\text{sign}[g^*(\boldsymbol{x})] = \text{sign}[g_0(\boldsymbol{x})]$.

The proof of Theorem 2 is given in the appendix. Theorem 2 shows that the proposed learning algorithm is Fisher consistent.

Next, assume that the true decision rule $g_0(\boldsymbol{x})$ is a linear function, $b_0+\beta_0^T\boldsymbol{x}$. Then along with the above consistency results in Theorem 2, we can further show that the estimator for the true coefficients β_0, denoted by $\hat{\beta}$, is asymptotically normal and the sparse classification method possesses the oracle variable selection property if an appropriate sparse penalty term is used.

Theorem 3—In addition to (c.1)–(c.4), we assume that $\boldsymbol{X}$ has continuously differentiable density with respect to some dominating measure and that $(1, \boldsymbol{X}^T)^T(1, \boldsymbol{X}^T)$ is full rank with positive probability. If $p_{\lambda_n}(|\beta|)=\sum_{k=1}^{q} p_{\lambda_n k}(|\beta_k|)$ where β_k is the kth component of β and it satisfies: for non-zero θ, $\lim n^{1/2} p'_{\lambda_n k}(|\theta|)=0$ and $\lim p''_{\lambda_n k}(|\theta|)=0$, and for any $M > 0$, $\lim \sqrt{n}\inf_{|\theta|\le Mn^{-1/2}} p'_{\lambda_n k}(|\theta|)=\infty$, then

- **a.** $\hat{\beta}$ is consistent;
- **b.** with probability tending to 1, $\text{sign}(\hat{\beta}_k) = \text{sign}(\beta_{0k})$, where $\hat{\beta}_k$ and β_{0k} are the kth component of $\hat{\beta}$ and β_0, respectively;
- **c.** $\sqrt{n}(\hat{\beta}^{1}-\beta_0^1)$ converges in distribution to a normal distribution with mean zero where $\hat{\beta}^1$ and β_0^1 denote the components of $\hat{\beta}$ and β_0 corresponding to non-zero β_{0k}'s.

The proof of Theorem 3 utilizes the local quadratic approximation of $\mathcal{Q}(\beta^T\boldsymbol{x})$ at β_0 and the M-estimation theory. We provide the proof in the appendix.

4 Simulation Studies

4.1 Simulation design

For all simulation experiments, we generated binary disease labels D_i with a success probability of 0.5. Given a subject's disease label $D_i = d$, we generated a disease-informative marker $Z_i \sim N(\mu_d, \sigma_d^2)$ and q-dimensional feature variables $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{iq})^T \sim MVN(\theta_d, \Sigma_d)$ independently of Z_i. We fitted a linear decision boundary, $g(\boldsymbol{x}) = b+\beta^T\boldsymbol{x}$, and compared the proposed SVM with pseudo-EM algorithm (SVM-EM) to: (1) an oracle procedure where we used the gold standard disease labels as the outcomes of SVM with various penalty functions (SVM-Oracle); (2) a two-step approach where in the first step, subjects were classified into diseased and non-diseased group by a clustering analysis based on a Gaussian mixture model, and in the second step, the obtained disease labels were used as outcomes in

a regularized SVM (SVM-2stage). We evaluated performance of various methods by miss-classification rate, area under the ROC curve (AUC), and sparsity of the fitted decision rule since variable selection is performed. We used generalized approximate cross-validation (GACV, Wahba et al. 2000) to select the tuning parameters when variable selection is performed. In all simulations, the initial values of class labels were obtained from fitting a Gaussian mixture model. We also tried a wide range of initial values and the results were very similar.

We let $q = 10$, and let the sample size $n = 300$ or 500. For each set of simulations, 500 runs were conducted. It is well known that using the same data to fit a model and evaluate its performance may lead to over-optimism. To compute an honest miss-classification rate and AUC, we simulated 100 independent validation sets. We examined three common types of feature variables: continuous, binary and multinomial. The mean of the disease-informative marker Z_i was $(\mu_1, \mu_{-1}) = (1.5, 0)$ for continuous feature variable cases and $(\mu_1, \mu_{-1}) = (2, 0)$ for binary and multinomial variable cases. The variances were $(\sigma_1^2, \sigma_{-1}^2)=(1, 1)$.

We considered four simulation settings. The first three settings simulated independent continuous, binary and multinomial feature variables, respectively, and the fourth setting simulated correlated continuous variables. The SCAD penalty was used in the first three settings and both SCAD and Enet penalties (Wang et al. 2008) are used in the fourth setting to compare performance of different penalty functions when feature variables are correlated. For the continuous feature variables, the mean vector for non-diseased subjects was $\theta_{-1} = (0, 0, \cdots, 0)^T$ and for diseased was $\theta_1 = (0, 2, 0, 2, 0, 0, 2, 0, 0, 0)^T$. In the binary variable cases, we first generated ten continuous variables from the same multivariate normal distribution, and dichotomized them as one if $X_{ik} > 0$ and 0 if $X_{ik} \leq 0$ for $k = 1, \cdots, 10$. In the multinomial variable cases, we imitated the real data example and generated the variables from a multinomial distribution taking values 0, 1, 2, 3, or 4. The probability vector for non-diseased subjects was (0.2, 0.2, 0.2, 0.2, 0.2) and (0, 0, 0.1, 0.2, 0.7) for diseased.

4.2 Simulation results

Table 1 reports the AUC, miss-classification rate, and variable selection properties under the four simulation settings. Since SVM-Oracle uses the true disease labels, it provides the lowest miss-classification rate and the largest AUC. The miss-classification rate of proposed SVM-EM is only slightly higher than SVM-Oracle: the difference is less than 1.5% for the continuous and multinomial feature variable cases and less than 2% for the binary variable cases. The reduction in missclassification rate from our approach compared to SVM-2stage can be 6% (5.4% versus 11.3%). The difference in the miss-classification rate between different approaches decreases as the sample size increases to 500. The SVM-EM provides an AUC almost identical to SVM-Oracle for all three types of feature variables. The AUC of SVM-EM is higher than SVM-2stage for the continuous feature variable cases. The improvement is larger for the binary variable cases, where the difference is 5.3% for $n = 300$ and 4.7% for $n = 500$. In addition, a larger sample size is required for SVM-EM with binary variables to achieve similar performance as continuous variable, which is expected since less information is conferred by binary feature variables. The performance of SVM-EM with multinomial variables is in between continuous and binary cases in terms of miss-classification rate and AUC.

In the fourth setting of Table 1, we report results under the assumption of weak correlation between the feature variables where they have an AR-1 structure and an autocorrelation parameter of $\rho = 0.2$. The miss-classification rates of the three approaches with SCAD penalty are slightly higher than the corresponding quantities where the feature variables are independent, and the AUCs are slightly lower. Here we also implemented the Enet penalty

due to its desirable grouping effect with correlated variables. The Enet SVM-EM has higher AUC and lower miss-classification rate than SCAD SVM-EM, which is expected (last block of Table 1, setting IV).

In Table 1, we also report variable selection properties. The column indexed as "C" is the mean number of variables with non-zero coefficients correctly estimated to be nonzero; "IC" is the mean number of variables with zero as coefficients incorrectly estimated as non-zero in the model; and "Correct-fit" is the proportion of models that correctly select the exact subset of the non-null variables. The correct-fit percentage of SVM-Oracle is the highest, which reflects information gained from observing the gold standard group labels. The percentage of correct-fit of SVM-2stage is much lower than SVM-EM in many cases. For example, the correct-fit percentage is 75.2% for SVM-2stage versus 87.8% for SVM-EM with $n = 500$ under the multinomial case. The performances of SVM-EM and SVM-2stage improve substantially when the sample size is increased to 500.

When the feature variables are correlated (the fourth setting in Table 1), we see that the performance has a similar trend as those with independent variables in setting I. In terms of percentage of correct-fit, the improvement of using elastic penalty is considerable compared to SCAD for both SVM-EM and SVM-2stage. Specifically, the improvement in correct-fit percentage is about 37.4% for SVM-EM and about 20% for SVM-2stage with $n = 300$, and 34.4% for SVM-EM and 22.6% for SVM-2stage with $n = 500$. As for the computational speed, on average the computing time for $n = 300$ is 1.2 minutes for each repetition and 1.5 minutes for $n = 500$ with a Dell Workstation with 2.67GHz CPU and 4G memory. The median number of EM iterations is 16 for $n = 300$ and 12 for $n = 500$.

4.3 Sensitivity analyses

We present four sets of sensitivity analyses. In (A), we study sensitivity to the normality assumption in (1); in (B), we study sensitivity to the conditional independence assumption of Z_i and $\boldsymbol{X}_i$ given D_i; in (C), we examine sensitivity to the pseudo-class-probabilities in (2); and in (D), we examine semi-supervised learning using different proportions of missing disease labels. For analysis (A), we generated Z_i from a Laplace distribution with mean zero and variance two for the non-diseased subjects and mean two and variance two for the diseased subjects. Other simulation parameters remain the same as the independent continuous feature variable case. For (B), we generated Z_i and $\boldsymbol{X}_i$ jointly from a multivariate normal distribution given a subject's group status, where Z_i was correlated with X_{ik} with $\rho = 0.1$ and X_{ik}'s were mutually independent. The other settings are the same as the first scenario in Table 1. For (C), we generated data the same way as in the third scenario in Table 1. For (D), the setting is the same as the first scenario in Table 1.

Tables 2 summarizes the sensitivity analyses results. From setting (A), we see that although the informative marker was generated from a Laplace distribution, the miss-classification rates and the AUCs are comparable to those where normality is satisfied. The estimated coefficients show ignorable biases for all three approaches at both sample sizes (results not shown here). In terms of variable selection properties, the performance of SVM-EM is also similar to the case where the distribution of Z_i is indeed normal.

From setting (B) in Table 2, we see that although Z_i is correlated with $\boldsymbol{X}_i$ given D_i, the relative performance of the three approaches shows similar patterns as the independent case. The miss-classification rates and AUCs exhibit minimal changes. The variable selection performance (e.g., percent of correct-fit) is affected by about 10% to 20%.

In analysis (C), we investigated an alternative approach of computing pseudo-class-probabilities:

$$\Pr^{(m)}[D_i = d_i^{(m)} | \boldsymbol{X}_i] = \frac{1}{1 + [1 - d_i^{(m)} g(\boldsymbol{X}_i)]_+}. \quad (7)$$

This expression implies that when the slack variables $\xi_i = [1 - d_i^{(m)} g(\boldsymbol{X}_i)]_+ = 0$, which suggests no misclassification error, we predict a subject's disease status with high confidence (a probability of one). When $0 < \xi_i < 1$, we predict a subject's disease status with a probability less than one but greater than 1/2 (random guess). If $\xi_i > 1$, then misclassification occurs, and we predict the disease status with probability less than 1/2, which is worse than a random guess. The results show that using (7) achieves similar miss-classification rate, AUC and model sparsity as using (2). This suggests that SVM-EM is not sensitive to the estimation of class probabilities as long as the ranking is preserved through a monotone transformation of the slack variables.

In analysis (D), we compared the proposed methods with semi-supervised learning where the true class labels are known on a subset of subjects. In the implementation of semi-supervised learning, we randomly selected a subset (10%, 20%, or 30%) of subjects to reveal their true disease labels, while using the estimated conditional probabilities $\Pr^{(m-1)}(D_i | \boldsymbol{X}_i)$ by the method in Wang, Shen, and Liu (2008) to recover disease status of the rest of the subjects. We see that the informative marker guided SVM-EM without using any disease labels (setting I in Table 1) has better performance than semi-supervised learning when 10% of the true labels are revealed, and SVM-EM has about the same AUC and missclassification rate as semi-supervised learning when 20% of the disease labels are available (setting D in Table 2). Note that with 30% of the labels revealed, the missclassification rate and AUC of the semi-supervised learning is similar to SVM-oracle. In other words, in this setting using more than 30% of the disease labels does not appear to improve prediction when the informative marker is available on all subjects. In terms of variable selection properties, performance of SVM-EM is in between revealing 20% and 30% of the disease labels in a semi-supervised learning.

In summary, the proposed method is not sensitive to the normality assumption, the conditional independence of $\boldsymbol{X}_i$ and Z_i given D_i under a weak dependence structure, or the computation of the pseudo-class-probabilities. The informative marker recovers information from disease labels to perform classification when the labels are completely missing.

5 Real Data Examples

5.1 Application to complicated grief studies

An application of the proposed methods is to contribute to the recent efforts on constructing a classification algorithm for a new psychiatric disorder, Complicated Grief (CG; Shear et al. 2005). CG refers to a form of grief in which the natural healing process is derailed and debilitating symptoms persist (Shear et al. 2005). CG (renamed as persistent complex bereavement disorder) is currently being proposed for inclusion to the *Diagnostic Statistical Manual of Mental Disorders*, with specific criteria still to be determined. Crucial to its ultimate usefulness is the ability to accurately diagnose patients suffering from this disorder. Several criteria sets for CG have been proposed (e.g., Prigerson et al. 2009; Shear et al. 2011a), but there is no gold standard for disease status. However, there are several disease-specific symptom rating scales that have been shown to distinguish patients with CG from other bereaved patients and bereaved people in the general population. Most of the existing studies use some version of the Inventory of Complicated Grief (ICG), a self-report questionnaire (Prigerson et al. 1995) to assess disease symptoms. In addition, there are also

other measures that assess grief-related impairment in functioning, such as the Work and Social Adjustment Scale (WSAS), a self-report that measures work and social impairment attributable to CG (Mundt et al. 2002).

The goal of this analysis is to construct a classification rule to screen CG patients for clinical trials by combining ICG variables, using WSAS as our disease informative marker. WSAS is an indicator of disease severity and a prognostic marker of future functioning/disease severity outcomes (e.g., correlation between baseline WSAS and end of study ICG was 0.74 in Shear et al. 2005). Here the feature variables are disease-specific symptoms as measured by the ICG. WSAS is not treated as part of the feature variables since it measures functioning impairment caused by complicated grief. By accounting for WSAS, we identify subjects with CG symptoms who may have differing future outcomes.

We use independent data sets for training the classifier (training set) and evaluating its performance (validation set). Subjects included in the training set were recruited in a randomized controlled treatment study of individuals with CG (Pittsburgh study; Shear et al. 2005) comparing a specific CG Psychotherapy compared to Interpersonal Psychotherapy. Subjects included in the validation set are being recruited for two ongoing treatment studies for CG: Optimizing Treatment for Complicated Grief (also called Healing Emotions After Loss: HEAL, Duan et al. 2011) and Complicated Grief for Older Adults (CGTOA, Shear et al. 2011b). HEAL is a multi-site study to compare response to antidepressant medication administered with and without CGT among bereaved individuals. CGTOA is a CGT treatment study in an older population (at least 50 years of age).

There were 175 subjects (67% women) with a baseline visit in the Pittsburgh study included in the training set. Here our feature variables to be selected and combined include 19 variables with integer values ranging from 0 to 4 in the ICG questionnaire and the disease-informative marker is the continuous functioning measure, WSAS. We applied the proposed SVM-EM with a linear decision rule and SCAD or Enet penalty. We compared the proposed method with a standard approach treating WSAS as the outcome and fitting a penalized regression with Enet penalty (standard Enet). We used generalized cross validation to choose the tuning parameter for SVM-EM, and ten-fold cross validation for the standard penalized regression approach.

The selected variables and their coefficients from the three analyses are summarized in Tables 3 and 4. SVM-EM with SCAD selected 5 variables, SVM-EM with Enet selected 8, and standard Enet selected 4. A previous factor analysis limited to people clinically diagnosed with CG shows that the ICG variables can be grouped into six domains (Simons et al. 2011). We see that there are three variables that are selected by all three approaches, and they appear to have strong effects across the analyses. Standard Enet missed the domain “Shock and disbelief”, while the two SVM-EM approaches selected one variable from this domain. We note a clustering effect of Enet SVM-EM which is consistent with the motivation for such a penalty function: it tends to select multiple correlated variables from the same domain. For example, it chose four variables from the first domain, while SVM-EM with a SCAD penalty only selected two. None of the approaches chose any variables from the domain “Hallucinations of the deceased”.

Next, we evaluated performances of the three approaches using the independent validation data collected in HEAL and CGTOA. There were 196 subjects with a baseline visit in these studies, including 109 from HEAL and 87 from CGTOA. Among those, 77% were female. We computed the correlation of the fitted predictive scores (i.e., $\hat{\beta}^T \boldsymbol{X}_i$) with various other clinical measures known to be associated with CG. The measures being evaluated include the Structured Clinician Interview of Complicated Grief (SCI-CG), Total Impact of Event

Scale (IES-T), Impact of Event Scale – Avoidance (IES-A), and Impact of Event Scale – Intrusion (IES-I). We see from Table 4 that the two SVM-EM approaches have higher correlation than the standard Enet on all four measures. For example, Enet SVM-EM yields predictive scores with a correlation of 0.47 with SCI-CG, compared to standard Enet with a correlation of just 0.34. As for the two SVM-EM approaches with different penalty functions, SVM-EM with Enet penalty provides a higher correlation for SCI-CG, while it yields a higher correlation on the three IES measures with SCAD penalty.

We present the decision boundary of Enet SVM-EM in Figure 1. We show the decision boundary as a function of two summary indices: the first index sums over selected variables in “Domain 1” and “Domain 2” in Table 3 and the second index sums over the selected variables in the remaining domains. We color each subject in the validation set using a median split of the total ICG scores: the red dots indicate subjects with ICG≥42 (more symptoms), and the black triangles indicate subjects with ICG < 42 (less symptoms). We can see that using a median split to classify patients’ symptom severity does not always agree with the fitted decision rule which incorporates the information from WSAS. The disagreement rate is about 15%.

In summary, these analyses suggest superior performance of SVM-EM due to accounting for a mixture of diseased and non-diseased subjects, borrowing information from a disease-informative marker and using an appropriate penalty function to account for correlated variables. Future studies may be designed to collect data on the natural history of bereaved subjects’ normal grief recovery process as well as CG patients’ symptomatology and functioning process. This information can be used as markers informative of disease progression to assist deriving classification rules to better distinguish diseased and non-diseased subjects.

5.2 Application to UCI data

In addition, we analyzed four benchmark data examples provided to the UCI Machine Learning Repository (Blake and Merz 1998) including Wisconsin breast cancer (WBC), Pima Indians diabetes (PIMA), HEART and Spam email (SPAM). WBC data contained subjects with benign or malignant breast tumors. There are 9 biomarkers computed from a digitized image of a fine needle aspirate of a breast mass, and they describe characteristics of the cell nuclei present in the image. PIMA data include females at least 21 years old of Pima Indian heritage among whom, 268 tested positive for diabetes, 500 tested negative, and there are 8 biological attributes available. HEART data contain 270 subjects and 13 biological attributes associated with presence and absence of heart disease. SPAM data classifies spam emails using 57 attributes and 4601 instances which are frequencies of a particular words or characters.

Although the class labels for all these data are available, to use these data as examples to illustrate our methods, we do not use any of the class labels, but combine 20% to 30% of the attributes as informative markers, and implemented SVM-EM using the remaining attributes as feature variables. We also implemented SVM-oracle where the class labels on all training cases were used. For all these analyses, we randomly partitioned data into a training set and a testing set. For the SPAM data, 1000 cases were randomly selected as training and the rest as testing. For the other three data sets, 200 cases were randomly selected as the training set. The testing set is used to compute the AUC and missclassification rate of the decision function fitted from the training set by SVM-EM or SVM-oracle.

We summarize the results for each data set in Table 5. We see that when the marker is informative and has a strong correlation with the disease status, SVM-EM can almost fully recover information contained in disease labels. For example, for the WBC data where the

correlation between the informative markers and the true disease status is 0.867, SVM-EM, the missclassificaiotn rate and AUC of SVM-EM is almost identical to the case where the class labels are all known (SVM-Oracle). For the SPAM data where the correlation is moderate, the performance of our approach is very close to SVM-Oracle. For the HEART data when the correlation is weak (0.378), in the absence of any class label, SVM-EM still recovers information from the distribution of the informative marker and its missclassification rate and AUC are only 7% and 6% different from the ideal case of knowing all class labels. Lastly, we compare our approach with a none-SVM based approach, K-means clustering. We found that SVM-EM has notably lower missclassification rate than K-means clustering for SPAM data (i.e., SVM-EM: 16.1% versus K-means: 41.2%) and PIMA data (i.e., 36.5% versus 40.4%) and it has comparable performance for WBC (i.e., 4.5% versus 4.8%) and HEART (i.e., 26.7% versus 23.4%) data. In summary, the UCI data analysis shows that SVM-EM achieves similar classification accuracy as SVM-Oracle with a moderate or large correlation between disease status and the informative marker, and it still recovers partial information when the correlation is weak.

6 Discussion

We present large margin classification in the setting of missing class labels but with presence of informative markers. Such scenario occur frequently in clinical research on disease diagnosis/classification and biomarker studies, especially for disorders where no clinically significant anatomical or physiological deviation exists to be the gold standard disease status. Our theoretical examination shows that the proposed method is Fisher consistent and has an oracle variable selection property under some general conditions in Johnson, Lin and Zeng (2008). Simulations show that SVM-EM has a competitive AUC and missclassification rate compared to SVM-Oracle where the gold standard class labels are observed. An intuitive explanation is that SVM-EM recovers information available in the class-informative markers to inform discriminating two classes of subjects. The missing class labels affects the ability to choose the correct model more than the AUC or missclassification rate. This suggests that the unobserved gold standard outcomes have a greater influence on variable selection than on classification or prediction performance. Although the proposed methods are presented through SVM, they can be easily generalized to other loss-function-based learning algorithms such as binomial deviance or squared-loss-based classifiers.

We constructed pseudo-class-probabilities to guarantee the classification rule is a Bayes rule as if the class labels were observed. However, as long as a monotonicity constraint is preserved, the classification performance of SVM-EM is not sensitive to the specific form of the pseudo-class-probabilities. It may be worthwhile to bear in mind that these pseudo-probabilities do not estimate the true class probabilities. A more computationally involved approach in Wang, Shen, and Liu (2008) may be used to construct nonparametric estimation of class probabilities using the final class labels and outputs from SVM-EM. Our proposed algorithm implemented on a Dell PC with 2.67GHz CPU and 4Gz memory is more than 10 times faster than the maximum margin clustering in Xu et al. (2004), because Xu et al. (2004) used an exhaustive search over a 2^n dimensional space of all possible combinations of class labels (an average of 17 minutes on data with sample size of about 350; See Table V in Zhang et al. 2007; In the table Xu et al. method is referred as MMC).

Here we assumed conditional independence of the auxiliary markers and feature variables given the true class labels. It is possible to relax this condition by including the feature variables in the Gaussian mixture model, i.e., replace $f(\boldsymbol{Z}|D)$ with $f(\boldsymbol{Z}|D, \boldsymbol{X})$. When a nonparametric approach is desirable, one may consider a joint learning algorithm of $\boldsymbol{Z}$ on $(D, \boldsymbol{X})$ simultaneously with D on $\boldsymbol{X}$.

Other applications of the proposed approach include clinical studies where accurately labeling a subject is far more expensive or intrusive (e.g., biopsy to test tumor type) than measuring an informative marker (e..g, obtaining genotypes associated with tumor type). In this case, it is desirable to classify feature patients using the cheaper or less intrusive feature variables. For example, it may be desirable to derive disease screening rules from a self-report questionnaire that is easy to administer while drawing information from a more expensive clinician administered interview, despite neither being a gold standard disease diagnosis. A practical note is that it is easy to include interactions of feature variables in the proposed framework. In most clinical applications, interaction between symptoms is usually not considered when constructing the diagnostic criteria set of a disorder. Furthermore, the proposed method has the potential to provide a channel to introduce subject's disease prognostic factors and treatment response markers into the diagnosis criteria set of a disorder.

In some applications the gold standard outcome may be more than two groups (for example, a measure of the confidence of disease diagnosis ranging between 0 and 4). It is possible to extend the SVM-EM to the multi-class problems by accounting for missing class labels in loss functions for multi-category SVMs (Liu et al. 2005). Another extension is to use support vector regression for dimensional disease diagnosis (Kraemer, Shrout and Rubio-Stipec 2007) which can be treated as continuous outcomes.

It is easy to extend the the proposed method such as the SVM-EM to handle the case when some class labels are observed and others are not, thereby easily incorporating expert's opinions in the current learning algorithm. For example, when there are gold standard disease diagnoses provided by clinical experts on a sub-sample of the subjects, the observed D_j on these subjects are readily incorporated into the pseudo-EM algorithm framework. Let $\mathcal{J}$ index the set of all subjects with observed class labels, and let $\mathcal{I}$ index all subjects without class labels. Then the objective function in the M-step is

$$\min_{g\in\mathcal{H}_K}\left(\sum_{j\in\mathcal{J}}L\{D_j g(\boldsymbol{X}_j)\}+\sum_{i\in\mathcal{I}}\left[w_i^{(m)}L\{D_i g(\boldsymbol{X}_i)\}+[1-w_i^{(m)}L\{D_i g(\boldsymbol{X}_i)\}]\right]+p_{\lambda_n}(|g|)\right),$$

and in the E-step we only update the posterior probability for subjects with missing disease labels. This leads to a semi-supervised learning approach assisted by auxiliary markers. Another related scenario is that in some studies, an imperfect reference measure of class labels may be collected on all subjects. We can incorporate this reference diagnosis as part of the vector of $\boldsymbol{Z}_i$ to improve prediction and classification.

In clinical applications it is often the case that false positives and false negatives have very different consequences, and therefore should not be treated equally. The method can be modified to incorporate the different costs associated with these two different kinds of errors. Specifically, Lin, Yee and Wahba (2002) argued that in the case of disease classification, the expected cost of the future misclassification, rather than the expected misclassification rate, should be used to measure the performance of a classifier. Based on this, a weighted large margin classifier solves

$$\arg\min_{g\in\mathcal{H}_K}\left(\sum_{i=1}^{n}\Omega(D_i)L\{D_i g(\boldsymbol{X}_i)\}+p_{\lambda_n}(|g|)\right),$$

where $\Omega(D_j)$ is a weight function that represents the costs for false positives and false negatives. The current method can be modified to incorporate the different costs of false positives and false negatives. Furthermore, weights can be used to adjust for sampling design such as in a case-control study when population prevalence is available.

Lastly, evaluating validity of a classification rule without a gold standard class label is an important statistical problem in its own right. There is a body of literature in diagnostic medicine on evaluating accuracy of diagnostic tests in the absence of a gold standard (see for example, Rutjes et al. 2007), which is not elaborated on here, but will be considered in future work.

Acknowledgments

We thank the Editor, Associate Editor and two anonymous reviewers for their constructive comments which have led to significant improvement of the quality and presentation of this work. This work is supported by NIH grants NS073671-01, MH60783 and MH70741.

References

Becker N, Werft W, Toedt G, Lichter P, Benner A. Penalized SVM: a R-package for feature selection SVM classification. Bioinformatics. 2009; 25:1711–1712. [PubMed: 19398451]

Chung H, Flaherty BP, Schafer JL. Latent class logistic regression: application to marijuana use and attitudes among high school seniors. J R Statist Soc A. 2006; 169:723–743.

Culp M. On Propagated Scoring for Semisupervised Additive Models. Journal of the American Statistical Association. 2011; 106:493, 248–259.

Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B. 1977; 39 (1):138.

Duan, N.; Lebowitz, B.; Reynolds, C.; Simon, N.; Wang, Y.; Zisook, S.; Shear, K. Factorial Clinical Trials for Hybrid (Explanatory and Pragmatic) Research Studies: Design of Optimizing Treatment for Complicated Grief. Poster presentation at the annual meeting of the American College of Neuropsychopharmacology; Waikoloa, HI. 2011 Dec.

Hinton, Geoffrey; Sejnowski, Terrence J. Unsupervised Learning: Foundations of Neural Computation. MIT Press; 1999.

Johnson B, Lin D, Zeng D. Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models. Journal of the American Statistical Association. 2008; 103(482):672–680. [PubMed: 20376193]

Klöppel S, Draganski B, Golding CV, Chu C, Nagy Z, Cook PA, Hicks SL, Kennard C, Alexander DC, Parker GJ, Tabrizi SJ, Frackowiak RS. White matter connections reflect changes in voluntary-guided saccades in pre-symptomatic Huntington's disease. Brain. 2008; 131:196–204. [PubMed: 18056161]

Kraemer HC, Shrout PE, Rubio-Stipec M. Developing the diagnostic and statistical manual V: what will "statistical" mean in DSM-V? Soc Psychiatry Psychiatr Epidemiol. 2007; 42(4):259–267. [PubMed: 17334899]

Lin Y. Support vector machine and the Bayes rule in classification. Data Mining and Knowledge Discovery. 2002; 6:259–275.

Lin Y, Lee Y, Wahba G. Support vector machines for classification in non-standard situations. Machine Learning. 2002; 46:191–202.

Liu Y, Shen X, Doss H. Multicategory ψ-learning and support vector machine: computational tools. Journal of Computational and Graphical Statistics. 2005; 14(1):219–236.

McLachlan, G.; Pee, D. Finite mixture models. New York: Wiley; 2000.

Moguerza J, Munoz A. Support Vector Machines with Applications. Statistical Science. 2006; 21(3): 322–336.

Mundt J, Marks I, Shear M, Greist J. The work and social adjustment scale: a simple measure of impairment in functioning. The British Journal of Psychiatry. 2002; 180:461–464. [PubMed: 11983645]

Nigam K, McCallum A, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. Mach Learn. 1998; 39:103–134.

Orru G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. Neurosci Biobehav Rev. 2012; 36(4):1140–1152. [PubMed: 22305994]

Park C, Kim K, Myung R, Kood J. Oracle properties of SCAD-penalized support vector machine. Journal of Statistical Planning & Inference. 2012; 142:2257–2270.

Pollard, D. Empirical Processes: Theory and Applications. NSF-CMBS Regional Conference Series in Probability and Statistics; Hayward, CA: Institute of Mathematical Statistics; 1990.

Prigerson H, Maciejewski P, Reynolds C, Bierhals A, Newsom J, Fasiczka A, Miller M. Inventory of complicated grief: A scale to measure maladaptive symptoms of loss. Psychiatry Resarch. 1995; 59:65–79.

Rocha, GV.; Wang, X.; Yu, B. Asymptotic distribution and sparsistency for l1-penalized parametric M-estimators with applications to linear SVM and logistic regression. 2009. http://arxiv.org/abs/0908.1940v1

Rigollet P. Generalization error bounds in semi-supervised classification under the cluster assumption. Journal of Machine Learning Research. 2007; 8:1369–1392.

Rutjes A, Reitsma J, Coomarasamy A, Khan K, Bossuyt P. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess. 2007; 11(50)

Shear K, Frank E, Houck PR, Reynolds CF 3rd. Treatment of complicated grief: a randomized controlled trial. Journal of the American Medical Association. 2005; 293:2601–2608. [PubMed: 15928281]

Shear MK, Simon N, Wall M, Zisook S, Neimeyer R, Duan N, Reynolds C, Lebowitz B, Sung S, Ghesquiere A, Gorscak B, Clayton P, Ito M, Nakajima S, Konishi T, Melhem N, Meert K, Schiff M, O'Connor MF, First M, Sareen J, Bolton J, Skritskaya N, Mancini AD, Keshaviah A. Complicated grief and related bereavement issues for DSM-5. Depress Anxiety. 2011a; 28(2): 103–17. [PubMed: 21284063]

Shear, K.; Skritskaya, N.; Duan, N.; Mauro, C.; Wang, Y.; Lebowitz, B.; Reynolds, C.; Simon, N.; Zisook, S.; Glickman, K.; Guesquiere, A.; Worthington, J.; LeBlanc, N.; Young, IT. Suicide, depression and complicated grief. Poster presentation at the annual meeting of the American College of Neuropsychopharmacology; Waikoloa, HI. 2011b Dec.

Shen X, Tseng GC, Zhang X, Wong W. On psi-learning. Journal of the American Statistical Association. 2003; 98:724–734.

Simon N, Wall MM, Keshaviah A, Dryman M, LeBlanc N, Shear K. Informing the symptom profile for Complicated Grief. Depression and Anxiety. 2011; 28(2):118–126. [PubMed: 21284064]

van der Vaart, AW. Asymptotic Statistics. Cambridge University Press; Cambridge: 1996.

Vapnik, V. The Nature of Statistical Learning Theory. Springer-Verlag; New York: 1995.

Wahba, G.; Lin, Y.; Zhang, H. GACV for support vector machines, or, another way to look at margin-like quantities. In: Smola, AJ.; Bartlett, P.; Scholkopf, B.; Schurmans, D., editors. Advances in Large Margin Classifiers. Cambridge, Massachusetts: MIT Press; 2000. p. 297-309.

Wang L, Zhu J, Zou H. Hybrid huberized support vector machines for microarray classification and gene selection. Bioinformatics. 2008; 24 (3):412–419. [PubMed: 18175770]

Wang Y, Chen H, Schwartz T, Duan N, Parcesepe A, Lewis-Fernandez R. Prediction based structured variable selection through penalized support vector machine. Biometrics. 2011; 67:896–905. [PubMed: 21175555]

Wang J, Shen X, Liu Y. Probability estimation for large margin classifiers. Biometrika. 2008; 95:149–167.

Wang J, Shen X, Pan W. On Efficient Large Margin Semisupervised Learning: Method and Theory. Journal of the Machine Learning Research. 2009; 10:719–742.

Wu Y, Liu Y. Robust truncated-hinge-loss support vector machines. Journal of the American Statistical Association. 2007; 102:479, 974–983.

Xu, L.; Schuurmans, D. Unsupervised and semi-supervised multi-class support vector machines. AAAI-05, The Twentieth National Conference on Artificial Intelligence; 2005.

Xu, L.; Neufeld, J.; Larson, B.; Schuurmans, D. Advances in Neural Information Processing Systems. Vol. 17. Cambridge, MA: MIT Press; 2005. Maximum margin clustering.

Zhang H, Ahn J, Lin X, Park C. Gene selection using support vector machine with non-convex penalty. Bioinformatics. 2006; 22:88–95. [PubMed: 16249260]

Zhu, J.; Rosset, S.; Hastie, T.; Tibshirani, R. Advances in Neural Information Processing Systems. 2003. 1-norm support vector machines.

Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B. 2005; 67(2):301–320.

APPENDIX

Proof of Theorem 1

The proof follows similar arguments as the traditional EM algorithm which shows that the objective function increases at each iteration. We denote $\tilde{E}$ as the conditional expectation given the observed data based on the working distribution $p(d, g(\boldsymbol{x}))f(\boldsymbol{z}|D = d; \mu_d, \Sigma_d)$. Let $g^{(m)}(\cdot)$ denote the estimate for $g(\cdot)$ at the mth iteration. Then we have

$$\begin{aligned}&\tilde{E}\left[\sum_{i=1}^{n}\log\tilde{p}(D_i;g^{(m)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i;\widehat{\mu}_{D_i},\widehat{\Sigma}_{D_i})\middle|\boldsymbol{Z}_i,\boldsymbol{X}_i;g^{(m-1)},\widehat{\mu}_d,\widehat{\Sigma}_d,d\in\{-1,1\}\right]\\ &\geq\tilde{E}\left[\sum_{i=1}^{n}\log\tilde{p}(D_i;g^{(m-1)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i;\widehat{\mu}_{D_i},\widehat{\Sigma}_{D_i})\middle|\boldsymbol{Z}_i,\boldsymbol{X}_i;g^{(m-1)},\widehat{\mu}_d,\widehat{\Sigma}_d,d\in\{-1,1\}\right].\end{aligned}$$

Thus,

$$\begin{aligned}&\sum_{i=1}^{n}\log\left[\sum_d\tilde{p}(d;g^{(m)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i=d;\widehat{\mu}_d,\widehat{\Sigma}_d)\right]\\ &+\sum_{i=1}^{n}\tilde{E}\left[\log\frac{\tilde{p}(D_i;g^{(m)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i;\widehat{\mu}_{D_i},\widehat{\Sigma}_{D_i})}{\sum_d\tilde{p}(d;g^{(m)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i=d;\widehat{\mu}_d,\widehat{\Sigma}_d)}\middle|\boldsymbol{Z}_i,\boldsymbol{X}_i;g^{(m-1)},\widehat{\mu}_d,\widehat{\Sigma}_d,d\in\{-1,1\}\right]\\ &\geq\sum_{i=1}^{n}\log\left[\sum_d\tilde{p}(d;g^{(m-1)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i=d;\widehat{\mu}_d,\widehat{\Sigma}_d)\right]\\ &+\sum_{i=1}^{n}\tilde{E}\left[\log\frac{\tilde{p}(D_i;g^{(m-1)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i;\widehat{\mu}_{D_i},\widehat{\Sigma}_{D_i})}{\sum_d\tilde{p}(d;g^{(m-1)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i=d;\widehat{\mu}_d,\widehat{\Sigma}_d)}\middle|\boldsymbol{Z}_i,\boldsymbol{X}_i;g^{(m-1)},\widehat{\mu}_d,\widehat{\Sigma}_d,d\in\{-1,1\}\right].\end{aligned}$$

By the property of the Kullback-Leibler information,

$$\begin{aligned}&\tilde{E}\left[\log\frac{\tilde{p}(D_i;g^{(m-1)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i;\widehat{\mu}_{D_i},\widehat{\Sigma}_{D_i})}{\sum_d\tilde{p}(d;g^{(m-1)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i=d;\widehat{\mu}_d\widehat{\Sigma}_d)}\middle|\boldsymbol{Z}_i,\boldsymbol{X}_i;g^{(m-1)},\widehat{\mu}_d,\widehat{\Sigma}_d,d\in\{-1,1\}\right]\\ &\geq\tilde{E}\left[\log\frac{\tilde{p}(D_i;g^{(m)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i;\widehat{\mu}_{D_i},\widehat{\Sigma}_{D_i})}{\sum_d\tilde{p}(d;g^{(m)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i=d;\widehat{\mu}_d\widehat{\Sigma}_d)}\middle|\boldsymbol{Z}_i,\boldsymbol{X}_i;g^{(m-1)},\widehat{\mu}_d,\widehat{\Sigma}_d,d\in\{-1,1\}\right],\end{aligned}$$

we obtain

$$\sum_{i=1}^{n}\log\left[\sum_{d}\tilde{p}(d;g^{(m)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i=d;\widehat{\mu}_d,\widehat{\sum}_d)\right]\geq\sum_{i=1}^{n}\log\left[\sum_{d}\tilde{p}(d;g^{(m-1)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i=d;\widehat{\mu}_d,\widehat{\sum}_d)\right].$$

That is, $\mathcal{Q}(g^{(m)},\widehat{\mu}_d,\widehat{\Sigma}_d,\ d\in\{-1,1\})\geq\mathcal{Q}(g^{(m-1)},\widehat{\mu}_d,\widehat{\Sigma}_d,\ d\in\{-1,1\})$. Furthermore, the equality holds if and only if

$$\tilde{p}(D_i;g^{(m-1)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i;\widehat{\mu}_{D_i},\widehat{\sum}_{D_i})=\tilde{p}(D_i;g^{(m)}(\boldsymbol{X}_i))f(\boldsymbol{Z}_i|D_i;\widehat{\mu}_{D_i},\widehat{\sum}_{D_i}).$$

We can easily show that the latter holds if and only if $\text{sign}(g^{(m)})=\text{sign}(g^{(m-1)})$.

Proof of Theorem 2

Since $g^*(\boldsymbol{X})$ maximizes

$$E\left\{\log\left[\sum_{d}p(d,g(\boldsymbol{X}))f(\boldsymbol{Z}|d;\mu_{0d},\sum_{0d})\right]\right\},$$

for each $\boldsymbol{x}$ in $\boldsymbol{X}$'s support, $y=p(1,g^*(\boldsymbol{x}))/[p(1,g^*(\boldsymbol{x}))+p(-1,g^*(\boldsymbol{x}))]$ maximizes

$$\int_{\boldsymbol{z}}\left[\sum_{d}p_d(\boldsymbol{x})f(\boldsymbol{z}|d;\mu_{0d},\sum_{0d})\right]\log\left\{\sum_{d}[I(d=1)y+I(d=-1)(1-y)]f(\boldsymbol{z}|d;\mu_{0d},\sum_{0d})\right\}d\boldsymbol{z},$$

where $p_d(\boldsymbol{x})$ is the true probability of $D=d$ given $\boldsymbol{X}=\boldsymbol{x}$.

Differentiating the above function, we obtain that y solves

$$\int_{\boldsymbol{z}}\left[\sum_{d}p_d(\boldsymbol{x})f(\boldsymbol{z}|d;\mu_{0d},\sum_{0d})\right]\frac{f(z|d=1;\mu_{01},\sum_{01})-f(\boldsymbol{z}|d=-1;\mu_{0,-1},\sum_{0,-1})}{\sum_d[I(d=1)y+I(d=-1)(1-y)]f(\boldsymbol{z}|d;\mu_{0d},\sum_{0d})}d\boldsymbol{z}=0.$$

On the other hand, we know

$$\int_{\boldsymbol{z}}\left[\sum_{d}p_d(\boldsymbol{x})f(\boldsymbol{z}|d;\mu_{0d},\sum_{0d})\right]\frac{f(\boldsymbol{z}|d=1;\mu_{01},\sum_{01})-f(\boldsymbol{z}|d=-1;\mu_{-1},\sum_{-1})}{\sum_d p_d(\boldsymbol{x})f(\boldsymbol{z}|d;\mu_{0d},\sum_{0d})}d\boldsymbol{z}=0.$$

Take the difference of the above equations and it yields

$$\int_z \frac{[f(z|d=1;\mu_{01},\sum_{01})-f(z|d=-1;\mu_{0,-1},\sum_{0,-1})]^2(y-p_1(x))dz}{\sum_d[I(d=1)y+I(d=-1)(1-y)]f(z|d;\mu_{0d},\sum_{0d})}=0$$

so $y = p_1(x)$. Therefore, $p_1(x) > 1/2$ if and only if $p(1, g^*(x)) > p(-1, g^*(x))$ so if only if $g^*(x) > 0$. In other words, $\text{sign}[g^*(x)] = \text{sign}[p_1(x) - p_{-1}(x)] = \text{sign}[g_0(x)]$.

Proof of Theorem 3

The proof of Theorem 3 follows similar arguments as Rocha et al. (2009) and Park et al. (2012). Thus, we only sketch the proof for the choice of the hinge loss in the following proof. In conditions (c.1)–(c.4), take the norm as the usual Euclidean norm for (b, β) and $p_{\lambda_n}(g) = p_{\lambda_n}(|\beta|)$. Clearly, g^* in condition (c.4) should be g_0. For notational convenience, we absorb the constant term into X, so β includes the intercept coefficient. Furthermore, we let $\mathbf{P}_n$ denote the empirical distribution and $\mathbf{P}$ be the true expectation. We use $\Gamma(\beta, \mu, \Sigma)$ to denote $\log\{\sum_d \exp[-L(d\beta^T X)]f(Z|D = d; \mu_d, \Sigma_d)\}$.

We first establish the consistency of $\hat{\beta}$ in (a). From assumption (c.1), for any subsequence, we can choose a further subsequence such that $\hat{\beta} \to \beta^+$. For this chosen subsequence, from the fact that $\mathcal{Q}_n(\hat{g}) \geq \mathcal{Q}_n(g_0)$ where $\hat{g}(x) = \hat{\beta}^T x$ and $g_0(x) = \beta_0^T x$, we have

$$\mathbf{P}_n\Gamma(\widehat{\beta},\widehat{\mu},\widehat{\sum})-p_{\lambda_n}(|\widehat{\beta}|) \geq \mathbf{P}_n\Gamma(\beta_0,\widehat{\mu},\widehat{\sum})-p_{\lambda_n}(|\beta_0|).$$

Under conditions (c.1) and (c.2), it gives

$$\mathbf{P}_n\Gamma(\widehat{\beta},\widehat{\mu},\widehat{\sum}) \geq \mathbf{P}_n\Gamma(\beta_0,\widehat{\mu},\widehat{\sum})+o(1).$$

Furthermore, since $\hat{\mu}_d \to_p \mu_0$ and $\Sigma_d \to_p \Sigma_0$, according to condition (c.3), we take the limits of both sides, and from the Glivenko-Cantelli theorem, it gives $\mathbf{P}\Gamma(\beta^+, \mu_0, \Sigma_0) \geq \mathbf{P}\Gamma(\beta_0, \mu_0, \Sigma_0)$. That is, $\mathcal{Q}(g^+) \geq \mathcal{Q}(g_0)$ where $g^+(x) = \{\beta^+\}^T x$. Therefore, $g^+ = g_0$ by condition (c.4). Since this holds for any subsequence, we conclude that $\hat{\beta} \to_p \beta_0$. Next, we establish the convergence rate of $\hat{\beta}$ as follows. For any $\beta = \beta_0 + un^{-1/2}$ where $|u| = M$ for a large M, we have

$$\begin{aligned} n^{-1}\mathcal{Q}_n(\beta^T x)-n^{-1}\mathcal{Q}_n(\beta_0^T x) \leq{}& \mathbf{P}_n\Gamma(\beta_0+un^{-1/2},\widehat{\mu},\widehat{\sum})-\mathbf{P}_n\Gamma(\beta_0,\widehat{\mu},\widehat{\sum}) \\ &-\sum_{\beta_{k0}\neq 0}[p_{\lambda_n k}(|\beta_{k0}+u_k n^{-1/2}|)-p_{\lambda_n k}(|\beta_{k0}|)]. \end{aligned}$$

Since

$$\begin{aligned} \mathbf{P}_n\Gamma(\beta_0+un^{-1/2},\widehat{\mu},\widehat{\sum})-\mathbf{P}_n\Gamma(\beta_0,\widehat{\mu},\widehat{\sum}) &\leq O_p(n^{-1})-\mathbf{P}\Gamma(\beta_0,\mu_0,\sum_0)+\mathbf{P}\Gamma(\beta_0+un^{-1/2},\mu_0,\sum_0) \\ &\leq O_p(n^{-1})|u|-c|u|^2/n \end{aligned}$$

for some positive constant c, where the last step uses the condition (c.4), and

$$\sum_{\beta_{k0}\neq 0}[p_{\lambda_n k}(|\beta_{k0}+\boldsymbol{u}_k n^{-1/2}|)-p_{\lambda_n k}(|\beta_{k0}|)=O(|\boldsymbol{u}|n^{-1})]$$

by the property of $p_{\lambda_n k}$, we obtain

$$n^{-1}\mathcal{Q}_n(\beta^T\boldsymbol{x})-n^{-1}\mathcal{Q}_n(\beta_0^T\boldsymbol{x})\leq O_p(n^{-1})|\boldsymbol{u}|-c|\boldsymbol{u}|^2/n.$$

Thus, in probability, $\mathcal{Q}_n(\beta^T\boldsymbol{x})<\mathcal{Q}_n(\beta_0^T\boldsymbol{x})$ when M is chosen large enough. This shows that $\hat{\beta}$ is $\sqrt{n}$-consistent.

To prove the oracle property in (*b*), it suffices to show that if $\beta_{0k}=0$ for the kth component, then $\hat{\beta}_k=0$ with probability tending to 1. To do that, consider the probability set $\{|\hat{\beta}-\beta_0|\leq Mn^{-1/2}\}$ for a large M. This set has probability tending to 1 if n then M goes to infinity. We now prove (b) by contradiction. On this set, if $\hat{\beta}_k\neq 0$, then from $n^{-1}\mathcal{Q}_n(\widehat{\beta}^T\boldsymbol{x})\geq n^{-1}\mathcal{Q}_n(\widehat{\beta}_{-k}^T\boldsymbol{x})$, where $\hat{\beta}_{-k}$ is equal to $\hat{\beta}$ except that its kth component is 0, we have

$$\mathbf{P}_n\Gamma(\beta,\widehat{\mu},\widehat{\sum})-\mathbf{P}_n\Gamma(\widehat{\beta}_{-k},\widehat{\mu},\widehat{\sum})\geq p_{\lambda_n,k}(|\widehat{\beta}_k|).$$

The left-hand side of the above equation is equal to

$$n^{-1/2}\mathbf{G}_n\Gamma(\widehat{\beta},\widehat{\mu},\widehat{\sum})-n^{-1/2}\mathbf{G}_n\Gamma(\widehat{\beta}_{-k},\widehat{\mu},\widehat{\sum})+\mathbf{P}\Gamma(\widehat{\beta},\widehat{\mu},\widehat{\sum})-\mathbf{P}\Gamma(\widehat{\beta}_{-k},\widehat{\mu},\widehat{\sum}).$$

After the Taylor expansion at (β_0,μ_0,Σ_0) and the use of the stochastic differentiability of $\mathbf{G}_n$, this term is equal to $O_p(n^{-1/2}|\hat{\beta}_k|)+O\{|\hat{\beta}_k|(|\hat{\mu}-\mu_0|+|\hat{\Sigma}-\Sigma_0|)\}=O_p(n^{-1/2})|\hat{\beta}_k|$. Thus, it follows $O_p(n^{-1/2})\geq p_{\lambda_n k}(|\widehat{\beta}_k|)/|\widehat{\beta}_k|\geq\inf_{|\theta|\leq Mn^{-1/2}}p'_{\lambda_n k}(|\theta|)$, and we obtain a contradiction. Therefore, with probability tending to 1, $\hat{\beta}_k=0$ if $\beta_{k0}=0$. We have proved (b).

To prove (c), we now let $\widehat{h}=\sqrt{n}(\widehat{\beta}^1-\beta_0^1)$, and fix $\hat{\beta}^2=0$. We define $\Gamma_1(\beta_1,\mu,\Sigma)=\Gamma[(\beta_1,0),\mu,\Sigma]$. Then

$$\begin{aligned}\widehat{h}&=\operatorname{argmax}\left[\mathbf{P}_n\Gamma_1(\beta_0^1+n^{-1/2}\boldsymbol{h},\widehat{\mu},\widehat{\sum})-p_{\lambda_n}(|\beta_0^1+n^{-1/2}\boldsymbol{h}|)\right]\\&=\operatorname{argmax}\left\{\mathbf{G}_n\left[\Gamma_1(\beta_0^1+n^{-1/2}\boldsymbol{h},\widehat{\mu},\widehat{\sum})-\Gamma_1(\beta_0^1,\widehat{\mu},\widehat{\sum})\right]+\sqrt{n}\mathbf{P}\left[\Gamma_1(\beta_0^1+n^{-1/2}\boldsymbol{h},\widehat{\mu},\widehat{\sum})-\Gamma_1(\beta_0^1,\widehat{\mu},\widehat{\sum})\right]-\sqrt{n}\left[p_{\lambda_n}(|\beta_0^1+n^{-1/2}\boldsymbol{h}|)-p_{\lambda_n}(|\beta_0^1|)\right]\right\}.\end{aligned}$$

On the right-hand side, the first term converges uniformly in $\boldsymbol{h}\in\boldsymbol{K}$ for some compact set $\boldsymbol{K}$ to a Gaussian process $\boldsymbol{W}^T\boldsymbol{h}$ where $\boldsymbol{W}$ is a normal random variable with mean zero. By the Taylor expansion and the differentiability of $\mathbf{P}$ $[\Gamma_1(\beta,\hat{\mu},\hat{\Sigma})]$ in β, the second term is asymptotically equivalent to

$$\nabla_\beta \mathbf{P}\left[\Gamma_1(\beta_0^1, \widehat{\mu}, \widehat{\sum})\right]^T \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^T \nabla_{\beta\beta}\mathbf{P}\left[\Gamma_1(\beta_0^1, \mu_0, \sum_0)\right]\boldsymbol{h}.$$

The third term vanishes by the assumption of $p_{\lambda_n}(\cdot)$. Therefore, from the Argmax theorem in van der Vaart and Wellner (1993), $\hat{\boldsymbol{h}}$ is asymptotically equivalent to the argument that maximizes

$$\boldsymbol{W}^T\boldsymbol{h} + \nabla_\beta \mathbf{P}\left[\Gamma_1(\beta_0^1, \widehat{\mu}, \widehat{\sum})\right]\boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^T \nabla_{\beta\beta}\mathbf{P}\left[\Gamma_1(\beta_0^1, \mu_0, \sum_0)\right]\boldsymbol{h}.$$

The latter maximizer is equal to

$$\begin{aligned} &-\nabla_{\beta\beta}\mathbf{P}\left[\Gamma_1(\beta_0^1, \mu_0, \sum_0)\right]^{-1}\left\{\boldsymbol{W} + \nabla_\beta \mathbf{P}\left[\Gamma_1(\beta_0^1, \widehat{\mu}, \widehat{\sum})\right]\right\} \\ =&-\nabla_{\beta\beta}\mathbf{P}\left[\Gamma_1(\beta_0^1, \mu_0, \sum_0)\right]^{-1}\left\{\boldsymbol{W} + \nabla_{\beta\mu}\mathbf{P}\left[\Gamma_1(\beta_0^1, \mu_0, \sum_0)\right](\widehat{\mu} - \mu_0)\right. \\ &\left. + \nabla_{\beta\sum}\mathbf{P}\left[\Gamma_1(\beta_0^1, \mu_0, \sum_0)\right](\widehat{\sum} - \sum_0) + o_p(n^{-1/2})\right\}. \end{aligned}$$

The asymptotic normality for $\widehat{\boldsymbol{h}} = \sqrt{n}(\widehat{\beta}^1 - \beta_0^1)$ thus follows from the asymptotic normality of $(\hat{\mu}, \hat{\Sigma})$. We have proved (c) in Theorem 3.

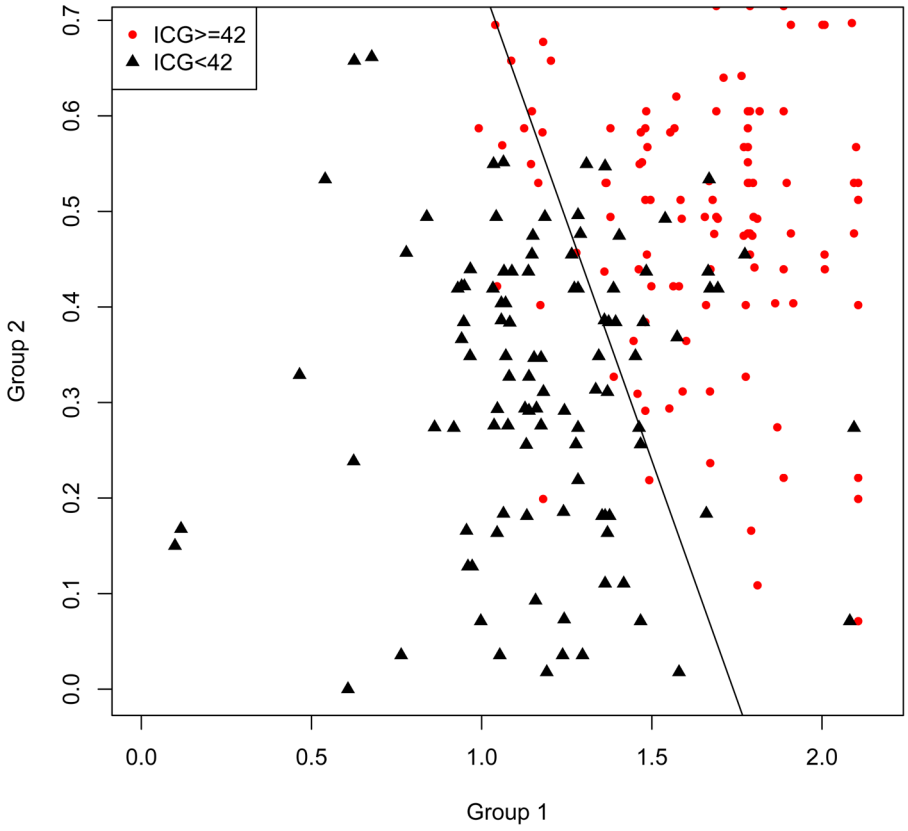

Figure 1.
Fitted decision boundary and agreement with total ICG ≥ 42*.
*: Group 1 includes variables selected from the first two domains in Table 5, and group 2 includes variables selected in the remaining domains.

Table 1

Summary of prediction and feature selection performance from simulation studies

Setting	Method	Miss	AUC	C	IC	CF%	Miss	AUC	C	IC	CF%
		***n* = 300**					***n* = 500**				
I	SVM-Oracle	0.045	0.991	2.956	0.018	93.8%	0.043	0.993	2.990	0.002	98.8%
	SVM-EM	0.059	0.987	2.882	0.442	63.0%	0.050	0.991	2.928	0.054	88.4%
	SVM-2stage	0.084	0.985	2.890	0.476	56.8%	0.075	0.987	2.948	0.046	90.4%
II	SVM-Oracle	0.098	0.852	2.974	0.000	97.4%	0.094	0.864	2.994	0.000	99.4%
	SVM-EM	0.117	0.848	2.830	0.740	52.2%	0.104	0.861	2.950	0.410	72.2%
	SVM-2stage	0.144	0.795	2.534	0.126	53.4%	0.130	0.814	2.618	0.010	62.2%
III	SVM-Oracle	0.103	0.953	2.978	0.000	97.8%	0.102	0.955	2.998	0.000	99.8%
	SVM-EM	0.108	0.952	2.976	0.456	68.4%	0.104	0.955	3.000	0.142	87.8%
	SVM-2stage	0.128	0.945	2.908	1.090	48.8%	0.115	0.953	2.988	0.472	75.2%
IV	SVM-Oracle	0.048	0.990	2.938	0.198	76.0%	0.045	0.992	2.990	0.162	84.2%
	SVM-EM	0.066	0.985	2.716	0.436	45.2%	0.054	0.989	2.900	0.316	65.4%
	SVM-2stage	0.093	0.982	2.842	0.512	52.0%	0.076	0.986	2.938	0.292	67.4%
	SVM-Oracle*	0.045	0.992	3.000	0.150	85.4%	0.044	0.992	3.000	0.066	99.8%
	SVM-EM*	0.057	0.990	2.996	0.200	82.6%	0.054	0.991	2.998	0.000	93.6%
	SVM-2stage*	0.134	0.987	2.966	0.308	71.8%	0.113	0.990	2.988	0.096	90.0%

Notes. “Miss: misclassification rate in validation data; “AUC”: the area under the ROC curve in validation data; “C”: average number of non-noise features selected (the true number is 3); “IC”: average number of noise features selected (the true number is 0); “CF%”: the proportion of selecting exactly the true model (proportion of correct fit). All four settings are:

I: X are continuous and independent features;

II: X are binary and independent features;

III: X are multinomial and independent features;

IV: X are continuous and correlated features;

*: Elastic net penalty is used instead of SCAD penalty in each of the method.

Table 2

Results from sensitivity analysis

Setting	Method	Miss	AUC	C	IC	CF%	Miss	AUC	C	IC	CF%
		$n = 300$					$n = 500$				
A	SVM-Oracle	0.044	0.992	2.970	0.012	95.8%	0.043	0.992	2.990	0.002	98.8%
	SVM-EM	0.069	0.985	2.908	1.016	41.8%	0.059	0.989	2.866	0.082	80.0%
	SVM-2stage	0.096	0.985	2.896	1.118	39.6%	0.083	0.987	2.936	0.028	91.4%
B	SVM-Oracle	0.045	0.992	2.972	0.012	96.2%	0.044	0.992	2.984	0.002	98.2%
	SVM-EM	0.080	0.981	2.682	0.656	36.8%	0.061	0.988	2.864	0.390	64.0%
	SVM-2stage	0.091	0.981	2.870	0.830	40.8%	0.074	0.987	2.946	0.526	56.2%
C	SVM-Oracle	0.098	0.852	2.974	0.000	97.4%	0.097	0.857	2.994	0.000	99.4%
	SVM-EM	0.124	0.838	2.788	0.386	59.4%	0.114	0.844	2.864	0.156	74.4%
	SVM-2stage	0.144	0.795	2.534	0.126	53.4%	0.132	0.809	2.612	0.002	61.8%
D	Semi-10%	0.075	0.981	2.896	1.338	27.0%	0.059	0.988	2.878	0.454	57.4%
	Semi-20%	0.057	0.988	2.942	0.572	57.0%	0.049	0.991	2.922	0.118	83.0%
	Semi-30%	0.053	0.990	2.942	0.286	75.2%	0.045	0.992	2.958	0.046	91.4%

Notes see Table 1. The four scenarios are:

A: Misspecify the distribution of Z_i as Laplace;

B: Misspecify $\boldsymbol{X}$ and Z as conditionally independent given D_i but $\boldsymbol{X}$ and Z are truly correlated given D;

C: Use the alternative pseudo-class probabilities (7)

D: Randomly reveal 10%, 20% or 30% of disease labels.

Table 3

Selected variables and their effects in the CG study

Analysis	SVM-EM1*	SVM-EM2**	Standard method†
Domain 1: Yearning and preoccupation with the deceased			
"I think about this person so much that it's hard for me to do the things I normally do"	1.08	0.22	0.95
"I feel that life is empty without the person who died"	0	0.09	0
"I feel that it is unfair that I should live when this person died"	0	0.10	0.13
"I feel lonely a great deal of the time ever since he/she died"	0.57	0.12	1.37
Domain 2: Anger and bitterness			
"I feel bitter over this persons death"	0	0.01	0
Domain 3: Shock and disbelief			
"Disbelief over what happened"	0.62	0.02	0
Domain 4: Estrangement from others			
"Lost the ability to care about other people"	1.03	0.13	0.17
"Envious of others who have not lost someone close"	0	0.04	0
Domain 5: Hallucinations of the deceased			
No variable selected			
Domain 6: Behavior change, including avoidance or proximity seeking			
"I feel drawn to places and things associated with the person who died"	0.58	0	0

*SVM-EM with SCAD penalty

**SVM-EM with Enet penalty

†Standard method using WSAS as the gold standard outcome with Enet penalty

Table 4

Correlation between the classification scores and external measures using independent data sets in CG studies.

Measure	SVM-EM1*	SVM-EM2**	Standard method†
SCI-CG	0.42	0.47	0.34
IES-T	0.26	0.19	0.18
IES-I	0.22	0.20	0.20
IES-A	0.24	0.15	0.13

NOTES: See Table 3.

Table 5

Analysis results of four UCI Machine Learning Repository data sets

		WBC	SPAM	PIMA	HEART
Number of feature variables Correlation†		7	47	7	9
		0.867	0.695	0.443	0.378
Miss	SVM-EM	0.045	0.161	0.365	0.267
	SVM-Oracle	0.043	0.130	0.325	0.197
AUC	SVM-EM	0.986	0.889	0.643	0.806
	SVM-Oracle	0.988	0.931	0.703	0.866

†Pearson correlation between the informative marker and disease status