Journal of Machine Learning Research 23 (2022) 1-25 Submitted 3/20; Revised 1/22; Published 3/22

Prior Adaptive Semi-supervised Learning with Application to EHR Phenotyping

Yichi Zhang* YICHIZHANG@URI.EDU
Department of Computer Science and Statistics
University of Rhode Island

Molei Liu* MOLEI_LIU@G.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health

Matey Neykov MNEYKOV@STAT.CMU.EDU
Department of Statistics and Data Science
Carnegie Mellon University

Tianxi Cai TCAI.HSPH@GMAIL.COM
Department of Biostatistics
Harvard T.H. Chan School of Public Health

Editor: Erik Sudderth

Abstract

Electronic Health Record (EHR) data, a rich source for biomedical research, have been successfully used to gain novel insight into a wide range of diseases. Despite its potential, EHR is currently underutilized for discovery research due to its major limitation in the lack of precise phenotype information. To overcome such difficulties, recent efforts have been devoted to developing supervised algorithms to accurately predict phenotypes based on relatively small training datasets with gold standard labels extracted via chart review. However, supervised methods typically require a sizable training set to yield generalizable algorithms, especially when the number of candidate features, p, is large. In this paper, we propose a semi-supervised (SS) EHR phenotyping method that borrows information from both a small, labeled dataset (where both the label Y and the feature set $\boldsymbol{X}$ are observed) and a much larger, weakly-labeled dataset in which the feature set $\boldsymbol{X}$ is accompanied only by a surrogate label S that is available to all patients. Under a *working* prior assumption that S is related to $\boldsymbol{X}$ only through Y and allowing it to hold *approximately*, we propose a prior adaptive semi-supervised (PASS) estimator that incorporates the prior knowledge by shrinking the estimator towards a direction derived under the prior. We derive asymptotic theory for the proposed estimator and justify its efficiency and robustness to prior information of poor quality. We also demonstrate its superiority over existing estimators under various scenarios via simulation studies and on three real-world EHR phenotyping studies at a large tertiary hospital.

1. Zhang and Liu contributed equally to this work.

©2022 Yichi Zhang*, Molei Liu*, Matey Neykov, and Tianxi Cai.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v23/20-290.html.

Keywords: High dimensional sparse regression, regularization, single index model, semi-supervised learning, electronic health records.

1. Introduction

Electronic Health Records (EHRs) provide a large and rich data source for biomedical research aiming to further our understanding of disease progression and treatment response. EHR data has been successfully used to gain novel insights into a wide range of diseases, with examples including diabetes (Brownstein et al., 2010), rheumatoid arthritis (Liao et al., 2014), inflammatory bowl disease (Ananthakrishnan et al., 2014), and autism (Doshi-Velez et al., 2014). EHR is also a powerful discovery tool for identifying novel associations between genomic markers and multiple phenotypes through analyses such as phenome-wide association studies (Denny et al., 2010; Kohane, 2011; Wilke et al., 2011; Cai et al., 2018).

Despite its potential, ensuring unbiased and powerful biomedical studies using EHR is challenging because EHR was primarily designed for patient care, billing, and record keeping. Extracting precise phenotype information for an individual patient requires manual medical chart reviews, an expensive process that is not scalable for research studies. To overcome such difficulties, recent efforts including those from Informatics for Integrating Biology and the Bedside (i2b2) (Liao et al., 2015; Yu et al., 2015, e.g.) and the Electronic Medical Records and Genomics (eMERGE) network (Newton et al., 2013; Gottesman et al., 2013) have been devoted to developing phenotyping algorithms to predict disease status using relatively small training datasets with gold standard labels extracted via chart review.

Various approaches to EHR phenotyping have been proposed. Supervised machine learning methods have been shown to achieve robust performance across disease phenotypes and EHR systems (Carroll et al., 2012; Liao et al., 2015). However, supervised methods typically require a sizable training set to yield generalizable algorithms especially when the candidate features, denoted by $\boldsymbol{X}$, is of high dimensionality p. One approach to overcome the high dimensionality is to consider unsupervised methods. Unfortunately, standard unsupervised methods such as clustering are likely to fail when the dimension of $\boldsymbol{X}$ is large, but a majority of the features are unrelated to the phenotype of interest and predictive of some other underlying subgroups. Recently, unsupervised methods based on "silver standard labels" have been proposed. These methods leverage a surrogate outcome S that is highly predictive of the true phenotype status Y, such as the count of International Classification of Diseases (ICD) billing codes of the disease, to train the phenotyping algorithm against the features $\boldsymbol{X}$. Specifically, Halpern et al. (2016) and Zhang et al. (2020) utilized anchor variables with high positive predictive value as the surrogate S to estimate $Y \mid \boldsymbol{X}$ under the conditional independence assumption $S \perp\!\!\!\perp \boldsymbol{X} \mid Y$. Agarwal et al. (2016) trained penalized logistic regression on $S \sim \boldsymbol{X}$ for phenotyping of Y against $\boldsymbol{X}$. Chakrabortty et al. (2017) provided theoretical justification for this strategy. They showed that a regularized estimator constructed from an unlabeled subset consisting of those with extreme values of S can be used to infer the direction of $\boldsymbol{\beta}$ under single index models $S \sim f(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{X}, \epsilon)$ and $Y \sim g(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X})$. Their method relies on the similarity between the directions of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to make efficient estimation. However, it is not robust to poor surrogacy resulted from violation of such assumptions. Furthermore, their method cannot be directly used to predict Y using both S and $\boldsymbol{X}$ or accurately recover the scale of $\Pr(Y = 1 \mid S, \boldsymbol{X})$.

A number of semi-supervised or weakly supervised deep learning procedures have also been proposed recently and shown to attain better performance than the supervised counterparts. For example, Ratner et al. (2017) proposed a weakly supervised approach that trains a deep model with imperfect labels generated from user–specified label functions from sources such as patterns, heuristics, and external knowledge bases. Wang and Poon (2018) developed a framework for weak supervision from multiple sources by composing probabilistic logic with deep learning. McDermott et al. (2018) designed a semi-supervised cycle Wasserstein regression generative adversarial networks (CWR-GAN) approach using adversarial signals to learn from unlabelled samples and improve prediction performance in scarcity of gold-standard labels. However, it remains unclear when and how the surrogate features, along with the unlabeled dataset can improve the prediction performance of these deep models, due to their complex architectures.

In this paper, we propose an semi-supervised (SS) method for estimating $Y \mid \mathbf{W} = (S, \boldsymbol{X}^\intercal)^\intercal$ that borrows information from both a small labeled dataset with n realizations of $(Y, \mathbf{W}^\intercal)^\intercal$ and a much larger unlabeled dataset with N observations on $\mathbf{W}$, under a high dimensional setting with $N \gg p \gg n$. We consider a logistic phenotype model for $Y \mid S, \boldsymbol{X}$, a single index model (SIM) for $S \mid \boldsymbol{X}$, as well as a *working* prior assumption that S is independent of $\boldsymbol{X}$ given Y. We obtain the estimator through regularization with penalty functions reflecting the prior knowledge. When the prior assumption holds exactly, we show that the unlabeled dataset can naturally be used to assist in the estimation of the phenotype model. Allowing the prior assumption to hold approximately or to be highly violated, our prior adaptive semi-supervised (PASS) estimator adaptively incorporate the prior knowledge by shrinking the estimator towards a direction derived under the prior.

The proposed PASS estimator is similar to the prior LASSO (pLASSO) procedure of Jiang et al. (2016) in that both approaches aim to incorporate prior information into the ℓ_1 penalized estimator in a high-dimensional setting. Nevertheless, the differences are substantial and clear. Jiang et al. (2016) assumed that the prior information was summarized into prediction values and contributed to the likelihood term. In contrast, we use prior information to guide the shrinkage and put them into the penalty term. In this sense, PASS and pLASSO complement each other to some extent. However, as shown in both theory and simulations, putting prior information into the likelihood term tends to lead to the "take it or leave it" phenomenon: the usefulness of the prior information is determined based on the overall effect of all predictors. On the other hand, by putting prior information into the penalty term, the PASS approach provides more flexible control: it is able to scrutinize the individual effect of each predictor. This gained flexibility can result in improved theoretical and numerical performances.

The rest of this paper is organized as follows. We discuss the motivation, an important special scenario and the general methodology in Section 2. We analyze the theoretical properties of the proposed approach in Section 3, and access its finite sample performance via simulation studies in Section 4. Furthermore, we illustrate the practical value of the proposed approach on three real EHR datasets in Section 5. Finally, we conclude this paper with some discussions and extensions in Section 6. All technical proofs and additional numerical results are given in the Supplementary Materials.

2. Methodology

2.1 Setup

We assume that the underlying data consists of N independent and identically distributed (i.i.d.) observations $\{(Y_i, S_i, \boldsymbol{X}_i^\intercal)^\intercal = (Y_i, \mathbf{W}_i^\intercal)^\intercal, i = 1, \ldots, N\}$, where Y_i is a binary indicator of the disease status of the ith patient, S_i is a scalar surrogate variable that is reasonably predictive of Y_i chosen via domain knowledge, and $\boldsymbol{X}_i$ is a p-dimensional feature vector. Examples of S_i includes the total count of ICD codes or mentions for the disease of interest in clinical notes extracted via natural language processing (NLP). Candidate features $\boldsymbol{X}$ may include the ICD9 code counts for competing diagnosis, lab results, as well as NLP mentions of relevant signs/symptoms, medications and procedures. We may also include various transformations of original features in $\boldsymbol{X}$ to account for non-linear effects. While $\{\mathbf{W}_i, i = 1, ..., N\}$ is fully observed, Y_i is only observed for a random subset of n patients. Hence the observed data are $\mathscr{L} \cup \mathscr{U}$, where without loss of generality, the first n observations are assumed fully observed as $\mathscr{L} = \{(Y_i, \mathbf{W}_i^\intercal)^\intercal, i = 1, \ldots, n\}$, and the rest constitute the unlabeled dataset as $\mathscr{U} = \{\mathbf{W}_i, i = n+1, \ldots, N\}$.

Throughout, for a d-dimensional vector $\boldsymbol{u}$, the ℓ_q-norm of $\boldsymbol{v}$ is $\|\boldsymbol{v}\|_q = (\sum_{j=1}^d |v_j|^q)^{1/q}$. The ℓ_∞-norm of $\boldsymbol{v}$ is $\|\boldsymbol{v}\|_\infty = \max_{1 \le j \le d} |v_j|$. The support of $\boldsymbol{v}$ is $\mathrm{supp}(\boldsymbol{v}) = \{j : v_j \neq 0\}$. If $\mathcal{J}$ is a subset of $\{1, \ldots, p\}$, then $\boldsymbol{v}_{\mathcal{J}}$ denotes a d-dimensional vector whose jth element is $v_j 1_{j \in \mathcal{J}}$, and 1_B is the indicator function for set B. The independence between random variables/vectors $\boldsymbol{U}$ and $\boldsymbol{V}$ is written as $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{V}$. We also denote the negative log-likelihood function associated with the logistic model with $\ell(y, \eta) = -y\eta + \log(1 + e^\eta)$.

2.2 Model Assumptions

To predict Y using $\mathbf{W} = (S, \boldsymbol{X}^\intercal)^\intercal$, we assume

$$\Pr(Y = 1 \mid \mathbf{W}) = \sigma(\zeta_0 + S\gamma_0 + \boldsymbol{X}^\intercal \boldsymbol{\beta}_0) = \sigma(\boldsymbol{\vartheta}_0^\intercal \vec{\mathbf{W}}) \quad \text{with} \quad \boldsymbol{\vartheta}_0 = (\zeta_0, \gamma_0, \boldsymbol{\beta}_0^\intercal)^\intercal, \qquad (\mathcal{M}_Y)$$

where for any vector $\mathbf{w}$, $\vec{\mathbf{w}} = (1, \mathbf{w}^\intercal)^\intercal$, and $\sigma(t) = e^t/(1 + e^t)$. To leverage the data in $\mathscr{U}$, we further assume a single index model (SIM) for $S \mid \boldsymbol{X}$, i.e. there exists $\boldsymbol{\alpha}_0 \in \mathbb{R}^p$ such that

$$S = f(\boldsymbol{X}^\intercal \boldsymbol{\alpha}_0, \epsilon), \text{ with some } \epsilon \perp\!\!\!\perp \boldsymbol{X} \text{ and } f \text{ satisfying } \mathbb{E}\{f^2(\boldsymbol{X}^\intercal \boldsymbol{\alpha}_0, \epsilon)\} < \infty, \qquad (\mathcal{M}_S)$$

where $\boldsymbol{X}^\intercal \boldsymbol{\alpha}_0$ is a single linear combination of the features $\boldsymbol{X}$ and f is an unknown link function. Here ζ_0, γ_0, $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ are parameters to be estimated where only the direction of $\boldsymbol{\alpha}_0$ is identifiable and its norm does not affect our construction introduced below. If $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ are similar in certain ways, one would expect that the unlabeled dataset $\mathscr{U}$ may be used to improve upon the standard supervised estimator for $\boldsymbol{\beta}_0$ using $\mathscr{L}$ alone. For example, if S is a noisy representation of Y with random measurement error, then it is reasonable and common in the EHR literature (Hong et al., 2019; Zhang et al., 2020, e.g.) to assume

$$\boldsymbol{X} \perp\!\!\!\perp S \mid Y. \qquad (\mathcal{C}^{\text{prior}})$$

Note that a similar conditional independence assumption to $(\mathcal{C}^{\text{prior}})$ was imposed between the input and the pretext target given the label, in the context of self-supervised learning to demonstrate its advantage (Lee et al., 2020). Under $(\mathcal{C}^{\text{prior}})$, we have Proposition 1 with proof given in Supplementary Materials.

Proposition 1. *Under ($\mathcal{M}_Y$), ($\mathcal{M}_S$), ($\mathcal{C}^{prior}$), and assuming $\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\intercal)$ is positive–definite, and it holds that: (C1) for any two vectors $\boldsymbol{a}_1, \boldsymbol{a}_2$, $\mathbb{E}(\boldsymbol{X}^\intercal\boldsymbol{a}_2 \mid \boldsymbol{X}^\intercal\boldsymbol{a}_1)$ is linear in $\boldsymbol{X}^\intercal\boldsymbol{a}_1$, there exist scalars $k_1, k_2 \in \mathbb{R}$ such that $\boldsymbol{\alpha}_0 = k_1\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}^* = k_2\boldsymbol{\beta}_0$ where*

$$(\tau^*, \boldsymbol{\alpha}^*) = \arg\min_{\tau,\boldsymbol{\alpha}} \mathbb{E}(S - \tau - \boldsymbol{X}^\intercal\boldsymbol{\alpha})^2.$$

Remark 1. Condition (C1) holds for elliptical distributions including multivariate normal. By Diaconis and Freedman (1984) and Hall and Li (1993), this assumption tends to hold for non-elliptical design when the dimensionality is high. Specifically, one can show that under mild regularity conditions, for two projection vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ uniformly randomly drawn from $\mathbb{S}^{p-1} = \{\boldsymbol{v} \in \mathbb{R}^{p-1} : \|\boldsymbol{v}\|_2 = 1\}$, the pair $(\boldsymbol{X}^\intercal\boldsymbol{a}_2, \boldsymbol{X}^\intercal\boldsymbol{a}_1)$ weakly converges to a bivariate normal distribution with high probability, and thus $\mathbb{E}(\boldsymbol{X}^\intercal\boldsymbol{a}_2 \mid \boldsymbol{X}^\intercal\boldsymbol{a}_1)$ is at least approximately linear in $\boldsymbol{X}^\intercal\boldsymbol{a}_1$; see Theorem 1.1 of Diaconis and Freedman (1984) and equation (1.9) of Hall and Li (1993).

Proposition 1 hinges on the main result of Li and Duan (1989) that when the features $\boldsymbol{X}$ satisfy (C1), the direction of the coefficients of a SIM could be estimated using least squares regression for the response against $\boldsymbol{X}$. It suggests that $\mathscr{U}$ can greatly improve the estimation of $\boldsymbol{\beta}_0$ under ($\mathcal{C}^{\text{prior}}$) because the phenotype model ($\mathcal{M}_Y$) may be rewritten as $\operatorname{logit}\Pr(Y = 1|\mathbf{W}) = \zeta + S\gamma + \rho\boldsymbol{X}^\intercal\boldsymbol{\alpha}$ for some ρ. Under this model, a simple SS estimator for ζ, γ and $\boldsymbol{\beta}$ in ($\mathcal{M}_Y$) can be obtained as $\widehat{\zeta}$, $\widehat{\gamma}$ and $\widehat{\rho}\widehat{\boldsymbol{\alpha}}$, where

$$(\widehat{\zeta}, \widehat{\gamma}, \widehat{\rho})^\intercal = \arg\min_{\zeta,\gamma,\rho} \sum_{i=1}^n \ell(Y_i, \zeta + \gamma S_i + \rho\boldsymbol{X}_i^\intercal\widehat{\boldsymbol{\alpha}}), \quad (\widehat{\tau}, \widehat{\boldsymbol{\alpha}}^\intercal)^\intercal = \arg\min_{\tau,\boldsymbol{\alpha}} \sum_{i=1}^N (S_i - \tau - \boldsymbol{X}\boldsymbol{\alpha})^2.$$

By doing so, the direction of the high dimensional vector $\boldsymbol{\beta}$ is estimated based on the entire $\mathscr{L} \cup \mathscr{U}$, and only the parameters $(\zeta, \gamma, \rho)^\intercal$ are estimated using the small labeled dataset $\mathscr{L}$. Hereafter we shall refer to this SS estimator derived under ($\mathcal{C}^{\text{prior}}$) as SS^{prior}.

Nevertheless, SS^{prior} is only valid when ($\mathcal{C}^{\text{prior}}$) and (C1) holds exactly. Our goal is to develop a more robust SS estimator under ($\mathcal{M}_Y$) and ($\mathcal{M}_S$) that can efficiently exploit $\mathscr{U}$ when ($\mathcal{C}^{\text{prior}}$) and (C1) may only hold approximately. In this more general setting, a desirable SS estimator should improve upon the standard supervised estimator when the directions of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ are similar in their magnitude and/or support. In addition, it should perform similarly to the supervised estimator when the two directions are not close. We shall now detail our PASS estimation procedure which automatically adapts to different cases as reflected in the observed data.

2.3 Prior Adaptive Semi-Supervised (PASS) Estimator

With $\mathscr{L}$ only, a supervised estimator for $\boldsymbol{\beta}$ can be obtained via the standard ℓ_1-penalized regression:

$$\breve{\boldsymbol{\vartheta}} = (\breve{\zeta}, \breve{\gamma}, \breve{\boldsymbol{\beta}}^\intercal)^\intercal = \arg\min_{\boldsymbol{\vartheta}} \frac{1}{n}\sum_{i=1}^n \ell(Y_i, \boldsymbol{\vartheta}^\intercal\vec{\mathbf{W}}_i) + \lambda\|\boldsymbol{\beta}\|_1. \tag{1}$$

With properly chosen λ, the consistency and rate of convergence for $\breve{\boldsymbol{\vartheta}}$ has been established (van de Geer, 2008). To improve the estimation of $\boldsymbol{\beta}$ through leveraging $\mathscr{U}$, we note that

when ($\mathcal{C}^{\text{prior}}$) holds approximately, the magnitude of $\boldsymbol{\beta}_0 - \rho\boldsymbol{\alpha}_0$ is small for some ρ, and the support of $\boldsymbol{\beta}_0 - \rho\boldsymbol{\alpha}_0$ is of small size as well.

To incorporate such prior belief on the relationship between $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$, we construct the penalty term

$$\min_{\rho}\{\lambda_1\|(\boldsymbol{\beta} - \rho\boldsymbol{\alpha}_0)_{\mathcal{A}_0}\|_1 + \lambda_2\|(\boldsymbol{\beta} - \rho\boldsymbol{\alpha}_0)_{\mathcal{A}_0^c}\|_1\},$$

where $\mathcal{A}_0 = \text{supp}(\boldsymbol{\alpha}_0)$, and $\lambda_1, \lambda_2 > 0$ are tuning parameters. Since $(\boldsymbol{\alpha}_0)_{\mathcal{A}_0^c} = \mathbf{0}$, the penalty term is equivalent to

$$\lambda_1\{\min_{\rho}\|(\boldsymbol{\beta} - \rho\boldsymbol{\alpha}_0)_{\mathcal{A}_0}\|_1\} + \lambda_2\|\boldsymbol{\beta}_{\mathcal{A}_0^c}\|_1. \tag{2}$$

The first term in the penalty measures how far $\boldsymbol{\beta}$ is from the closest vector along the $\boldsymbol{\alpha}_0$ direction, and hence encourages smaller magnitude of $\boldsymbol{\beta} - \rho\boldsymbol{\alpha}_0$. The second term shrinks $\boldsymbol{\beta}_{\mathcal{A}_0^c}$ towards $\mathbf{0}$, which reflects our prior that predictors irrelevant to S are likely to be irrelevant to Y as well. The tuning parameters λ_1, λ_2 control the strength of the belief imposed. When they are sufficiently large, $\boldsymbol{\beta}$ will be forced to be a multiple of $\boldsymbol{\alpha}_0$ and thus it ends up with the same estimator as in the case where ($\mathcal{C}^{\text{prior}}$) holds.

Since we have $N \gg p$ samples to estimate $\boldsymbol{\alpha}_0$, we use the adaptive LASSO (ALASSO) penalized least square estimator $\widehat{\boldsymbol{\alpha}}$ (Zou, 2006; Zou and Zhang, 2009), where

$$\widehat{\tau}, \widehat{\boldsymbol{\alpha}} = \arg\min_{\tau,\boldsymbol{\alpha}} \frac{1}{N}\sum_{i=1}^{N}\left(S_i - \tau - \boldsymbol{X}_i^{\intercal}\boldsymbol{\alpha}\right)^2 + \mu\sum_{j=1}^{p}\widehat{\omega}_j|\alpha_j|,$$

where $\widehat{\omega}_j = |\widehat{\alpha}_{\text{init},j}|^{-\nu}$ for some constant $\nu > 0$, $\widehat{\boldsymbol{\alpha}}_{\text{init}} = (\widehat{\alpha}_{\text{init},1}, ..., \widehat{\alpha}_{\text{init},p})^{\intercal}$,

$$\widehat{\tau}_{\text{init}}, \widehat{\boldsymbol{\alpha}}_{\text{init}} = \arg\min_{\tau,\boldsymbol{\alpha}} \frac{1}{N}\sum_{i=1}^{N}(S_i - \tau - \boldsymbol{X}_i^{\intercal}\boldsymbol{\alpha})^2 + \mu_{\text{init}}\|\boldsymbol{\alpha}\|_1,$$

μ_{init} and μ are tuning parameters that can be chosen via the cross-validation or Bayesian information criterion (BIC). Here, $\widehat{\boldsymbol{\alpha}}$ is actually an estimator of $\boldsymbol{\alpha}^*$, which has the same direction as $\boldsymbol{\alpha}_0$ under the conditions in Proposition 1.

Appending the penalty term (2) to the likelihood and replacing $\boldsymbol{\alpha}_0$ with its estimate $\widehat{\boldsymbol{\alpha}}$, we propose to estimate $\boldsymbol{\vartheta}_0 = (\zeta_0, \gamma_0, \boldsymbol{\beta}_0^{\intercal})^{\intercal}$ by

$$\widehat{\boldsymbol{\vartheta}} = (\widehat{\zeta}, \widehat{\gamma}, \widehat{\boldsymbol{\beta}}^{\intercal})^{\intercal} = \arg\min_{\boldsymbol{\vartheta}} \frac{1}{n}\sum_{i=1}^{n}\ell(Y_i, \boldsymbol{\vartheta}\vec{\mathbf{W}}_i) + \lambda_1\{\min_{\rho}\|(\boldsymbol{\beta} - \rho\widehat{\boldsymbol{\alpha}})_{\widehat{\mathcal{A}}}\|_1\} + \lambda_2\|\boldsymbol{\beta}_{\widehat{\mathcal{A}}^c}\|_1,$$

where $\widehat{\mathcal{A}} = \text{supp}(\widehat{\boldsymbol{\alpha}})$. The estimators can be equivalently obtained as

$$\widehat{\rho}, \widehat{\boldsymbol{\vartheta}} = \arg\min_{\rho,\boldsymbol{\vartheta}} \frac{1}{n}\sum_{i=1}^{n}\ell(Y_i, \boldsymbol{\vartheta}\vec{\mathbf{W}}_i) + \lambda_1\|(\boldsymbol{\beta} - \rho\widehat{\boldsymbol{\alpha}})_{\widehat{\mathcal{A}}}\|_1 + \lambda_2\|\boldsymbol{\beta}_{\widehat{\mathcal{A}}^c}\|_1 \tag{3}$$

The impact of the tuning parameters λ_1, λ_2 can be understood from a bias-variance tradeoff viewpoint. When λ_j's are large, $\widehat{\boldsymbol{\beta}}$ tends to be a multiple of $\widehat{\boldsymbol{\alpha}}$ and thus is an estimator with high bias and low variance. In contrast, when λ_j's are small, the likelihood term based on the labeled dataset $\mathscr{L}$ is the dominant part, and hence $\widehat{\boldsymbol{\beta}}$ will have low bias and high variance. By varying the values of λ_j's, we are able to obtain a continuum connecting these two extremes. In practice, λ_1 and λ_2 can be chosen via standard data-driven approaches such as cross-validation.

2.4 Computation Details

The minimization in (3) can be solved with standard software for LASSO estimation. Let $\boldsymbol{\delta} = \boldsymbol{\beta} - \rho\widehat{\boldsymbol{\alpha}}$. We can re-parametrize the expression above in terms of ρ, ζ, γ, and $\boldsymbol{\delta}$ as

$$\widehat{\zeta}, \widehat{\gamma}, \widehat{\rho}, \widehat{\boldsymbol{\delta}} = \underset{\zeta,\gamma,\rho,\boldsymbol{\delta}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \zeta + S_i\gamma + \rho \boldsymbol{X}_i^\intercal \widehat{\boldsymbol{\alpha}} + \boldsymbol{X}_i^\intercal \boldsymbol{\delta}) + \lambda_1(\|\boldsymbol{\delta}_{\widehat{\mathcal{A}}}\|_1 + \kappa\|\boldsymbol{\delta}_{\mathcal{P}\setminus\widehat{\mathcal{A}}}\|_1),$$

where $\mathcal{P} = \{1, \ldots, p\}$ and $\kappa = \lambda_2/\lambda_1$. This is a typical LASSO problem with covariates $(1, S_i, \boldsymbol{X}_i^\intercal\widehat{\boldsymbol{\alpha}}, \boldsymbol{X}_i^\intercal)^\intercal$, parameters $(\zeta, \gamma, \rho, \boldsymbol{\delta})^\intercal$, and a weighted ℓ_1 penalty on the parameters. Hence it can be solved by essentially any algorithm for ALASSO fitting. In this paper, we use the `R` package `glmnet` (Friedman et al., 2010) to compute $\widehat{\zeta}$, $\widehat{\gamma}$, $\widehat{\rho}$, and $\widehat{\boldsymbol{\delta}}$, and construct the final estimator for $\boldsymbol{\vartheta}_0$ as $\widehat{\boldsymbol{\vartheta}} = (\widehat{\zeta}, \widehat{\gamma}, \widehat{\boldsymbol{\beta}}^\intercal)^\intercal$ with $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\delta}} + \widehat{\rho}\widehat{\boldsymbol{\alpha}}$.

3. Theoretical Properties

In this section, we present non-asymptotic risk bounds for the PASS estimator. We also make theoretical comparisons with the supervised LASSO estimator to shed light on when PASS outperforms the LASSO and where such improvement comes from.

3.1 Notations

A random variable V is sub-Gaussian(τ^2) if $\mathbb{E}\{\exp(\lambda|V|)\} \leq 2\exp(\lambda^2\tau^2/2)$ holds for all $\lambda > 0$. Throughout, we define

$$\boldsymbol{U} = (\boldsymbol{X}^\intercal, 1)^\intercal,\ \boldsymbol{K} = \mathbb{E}\left(\boldsymbol{U}\boldsymbol{U}^\intercal\right),\ \boldsymbol{\xi} = (\boldsymbol{\alpha}^\intercal, \tau)^\intercal, \boldsymbol{Z}_{\boldsymbol{\alpha}} = (\boldsymbol{X}^\intercal, \boldsymbol{X}^\intercal\boldsymbol{\alpha}, S, 1)^\intercal, \boldsymbol{G} = \mathbb{E}\left(\boldsymbol{Z}_{\boldsymbol{\alpha}^*}\boldsymbol{Z}_{\boldsymbol{\alpha}^*}^\intercal\right),$$
$$\boldsymbol{\theta} = (\boldsymbol{\delta}^\intercal, \rho, \gamma, \zeta)^\intercal,\ \boldsymbol{H} = \mathbb{E}\left[\sigma(\boldsymbol{Z}_{\boldsymbol{\alpha}^*}^\intercal\boldsymbol{\theta}_0)\{1 - \sigma(\boldsymbol{Z}_{\boldsymbol{\alpha}^*}^\intercal\boldsymbol{\theta}_0)\}\boldsymbol{Z}_{\boldsymbol{\alpha}^*}\boldsymbol{Z}_{\boldsymbol{\alpha}^*}^\intercal\right],$$

where $\boldsymbol{\alpha}^*$ is given by $(\boldsymbol{\alpha}^{*\intercal}, \tau^*)^\intercal = \boldsymbol{\xi}^* = \arg\min_{\boldsymbol{\xi}} \mathbb{E}(S - \boldsymbol{U}^\intercal\boldsymbol{\xi})^2$, and $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} : \boldsymbol{\delta} + \rho\boldsymbol{\alpha}^* = \boldsymbol{\beta}_0, \zeta = \zeta_0, \gamma = \gamma_0\}$. Denote by $\mathcal{B}_0 = \text{supp}(\boldsymbol{\beta}_0)$, $\mathcal{A}^* = \text{supp}(\boldsymbol{\alpha}^*)$ and $q^* = |\mathcal{A}^*|$. We assume $\|\boldsymbol{\alpha}^*\|_2 = 1$ without loss of generality since $\boldsymbol{\alpha}^*$ is used to recover only the direction of $\boldsymbol{\beta}_0$ in SIM and one can change ρ correspondingly to make any $\boldsymbol{\beta} = \boldsymbol{\delta} + \rho\boldsymbol{\alpha}^*$ invariant to $\|\boldsymbol{\alpha}^*\|_2$. Note that under $(\mathcal{M}_Y)$, any $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$ minimizes $\mathbb{E}\{\ell(Y, \boldsymbol{Z}_{\boldsymbol{\alpha}^*}^\intercal\boldsymbol{\theta})\}$, and due to perfect multicollinearity in $\boldsymbol{Z}_{\boldsymbol{\alpha}^*}$, $\boldsymbol{\theta}_0$ is not unique. However, any $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$ corresponds to the unique $\boldsymbol{\beta}_0 = \boldsymbol{\delta}_0 + \rho_0\boldsymbol{\alpha}^*$ and thus $\boldsymbol{Z}_{\boldsymbol{\alpha}^*}^\intercal\boldsymbol{\theta}_0 = \zeta_0 + S\gamma_0 + \boldsymbol{X}^\intercal\boldsymbol{\beta}_0 = \boldsymbol{\vartheta}_0^\intercal\vec{\mathbf{W}}$ is well-defined. Moreover, any quantity depending on $\boldsymbol{\theta}_0$ through $\boldsymbol{Z}_{\boldsymbol{\alpha}^*}^\intercal\boldsymbol{\theta}_0$ is well-defined. Since the main results in this section depend on $\boldsymbol{\theta}_0$ solely through $\boldsymbol{Z}_{\boldsymbol{\alpha}^*}^\intercal\boldsymbol{\theta}_0$, we will use $\boldsymbol{\theta}_0$ to represent any $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ for simplicity.

For $\boldsymbol{\theta} = (\boldsymbol{\delta}^\intercal, \rho, \gamma, \zeta)^\intercal$, define $\Omega(\boldsymbol{\theta}) = \lambda_0(|\rho| + |\gamma| + |\zeta|) + \lambda_1\|\boldsymbol{\delta}_{\mathcal{A}^*}\|_1 + \lambda_2\|\boldsymbol{\delta}_{\mathcal{P}\setminus\mathcal{A}^*}\|_1$, $\Delta_{\boldsymbol{\alpha}} = 2\mu_{\text{init}}q^*/\varphi^2$ and $\Pi(\boldsymbol{\theta}) = |\rho|$, where φ is a constant defined by Assumption (A4) in Section S2.1 of the Supplement Material, and $\lambda_0 = 36B\{\log(6/\epsilon)/n\}^{1/2}$. To introduce the oracle $\boldsymbol{\theta}^*$, we define the oracle risk function as:

$$\begin{aligned}\mathscr{E}(\boldsymbol{\theta}, \mathcal{S}_+, \mathcal{S}_-) = &\,\mathbb{E}\,\ell(Y, \boldsymbol{Z}_{\boldsymbol{\alpha}^*}^\intercal\boldsymbol{\theta}) - \mathbb{E}\,\ell(Y, \boldsymbol{Z}_{\boldsymbol{\alpha}^*}^\intercal\boldsymbol{\theta}_0) \\ &+ 256\frac{\kappa(\mathcal{S}_+)^2|\mathcal{S}_+|}{\varpi\psi(\mathcal{S}_+)} + 8\lambda_1\|\boldsymbol{\theta}_{\mathcal{S}_-\cap\mathcal{A}^*}\|_1 + 8\lambda_2\|\boldsymbol{\theta}_{\mathcal{S}_-\cap(\mathcal{P}\setminus\mathcal{A}^*)}\|_1 + 8\lambda_1\Delta_{\boldsymbol{\alpha}}\Pi(\boldsymbol{\theta}),\end{aligned} \tag{4}$$

where

$$\psi(\mathcal{S}_+) = \inf_{\boldsymbol{v}:\Omega(\boldsymbol{v}_{\mathcal{S}_-})\le 3\Omega(\boldsymbol{v}_{\mathcal{S}_+})} \frac{\boldsymbol{v}^\intercal \boldsymbol{G}\boldsymbol{v}}{\boldsymbol{v}_{\mathcal{S}_+}^\intercal \boldsymbol{v}_{\mathcal{S}_+}},$$

$$\kappa(\mathcal{S}_+) = \begin{cases} \lambda_0, & \text{if } \mathcal{S}_+ \cap \mathcal{A}^* = \emptyset \text{ and } \mathcal{S}_+ \cap (\mathcal{P} \setminus \mathcal{A}^*) = \emptyset \\ \lambda_2, & \text{if } \mathcal{S}_+ \cap \mathcal{A}^* = \emptyset \text{ and } \mathcal{S}_+ \cap (\mathcal{P} \setminus \mathcal{A}^*) \neq \emptyset \\ +\infty, & \text{if } \mathcal{S}_+ \cap \mathcal{A}^* \neq \emptyset \end{cases}$$

Define $\boldsymbol{\theta}^* = (\boldsymbol{\delta}^{*\intercal}, \rho^*, \gamma^*, \zeta^*)^\intercal$, $\mathcal{S}_+^*$ and $\mathcal{S}_-^*$ as the solution to

$$\underset{\{\boldsymbol{\theta},\mathcal{S}_+,\mathcal{S}_-\}:\mathcal{S}_+\cap\mathcal{S}_-=\emptyset,\ \mathcal{S}_+\cup\mathcal{S}_-=\mathrm{supp}(\boldsymbol{\theta})\cup\overline{\mathcal{P}},\ \mathcal{S}_+\supseteq\overline{\mathcal{P}},\ \text{and } \|\boldsymbol{G}^{1/2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2\le\eta,}{\arg\min} \mathscr{E}(\boldsymbol{\theta}, \mathcal{S}_+, \mathcal{S}_-)$$

where $\overline{\mathcal{P}} = \{p+1, p+2, p+3\}$, and η is a constant as defined by Assumption (A3) in Section S2.1 of the Supplement. Let $\mathcal{S}^* = \mathcal{S}_+^* \cup \mathcal{S}_-^* = \mathrm{supp}(\boldsymbol{\theta}^*) \cup \overline{\mathcal{P}}$, $\kappa^* = \kappa(\mathcal{S}_+^*)$, and $\boldsymbol{\beta}^* = \boldsymbol{\delta}^* + \rho^*\boldsymbol{\alpha}^*$. Intuitively, one may view $\mathcal{S}_+$ as the union set of unpenalized predictors and the predictors with large coefficients but not recovered by $\mathcal{A}^*$. While $\mathcal{S}_-$ can be viewed as the union set of predictors with small nonzero coefficients and the predictors recovered by $\mathcal{A}^*$. Partitioning the support of $\boldsymbol{\theta}$ into $\mathcal{S}_+$ and $\mathcal{S}_-$ is inspired by Bühlmann and Van De Geer (2011, Section 6.2.4), which leads to a refined bound.

3.2 Main result

We first establish the risk bounds for the PASS estimator in the following theorem. Its proof can be found in Section S2 of the Supplementary Materials.

Theorem 1. *For any $\epsilon > 0$, if the assumptions (A1)–(A8) (introduced in Section S2.1 of the Supplementary Materials) hold, the following inequalities hold simultaneously with probability at least $1 - 10\epsilon$:*

$$\begin{aligned} &\textit{Excess risk:} && \mathbb{E}\,\ell(Y, \boldsymbol{Z}_{\widehat{\boldsymbol{\alpha}}}^\intercal \widehat{\boldsymbol{\theta}}) - \mathbb{E}\,\ell(Y, \boldsymbol{Z}_{\boldsymbol{\alpha}^*}^\intercal \boldsymbol{\theta}_0) \le \Xi, \\ &\textit{Linear prediction error:} && \mathbb{E}(\boldsymbol{Z}_{\widehat{\boldsymbol{\alpha}}}^\intercal \widehat{\boldsymbol{\theta}} - \boldsymbol{Z}_{\boldsymbol{\alpha}^*}^\intercal \boldsymbol{\theta}_0)^2 \le \Xi/\varpi, \\ &\textit{Probability prediction error:} && \mathbb{E}\{\sigma(\boldsymbol{Z}_{\widehat{\boldsymbol{\alpha}}}^\intercal \widehat{\boldsymbol{\theta}}) - \sigma(\boldsymbol{Z}_{\boldsymbol{\alpha}^*}^\intercal \boldsymbol{\theta}_0)\}^2 \le \Xi/\varpi, \end{aligned}$$

where ϖ is a positive constant defined in (A2), $\Xi = 64\mathscr{E}(\boldsymbol{\theta}^*, \mathcal{S}_+^*, \mathcal{S}_-^*)$, and $\mathscr{E}$ is an oracle risk function as defined in equation (4).

Remark 2. *As detailed in Section S2.1 of the Supplement Material, Assumptions (A1)–(A8) are imposed on tail behaviour of the regression residuals, regularity of the design matrix, minimum signal strength of $\boldsymbol{\alpha}^*$, sample sizes and rates of the tuning parameters. These assumptions are commonly used conditions in the theoretical literature of LASSO, such as the sub-Gaussian variable condition and the restricted eigenvalue condition; see e.g. van de Geer and Bühlmann (2009); Bickel et al. (2009); Bühlmann and Van De Geer (2011).*

Remark 3. *The last term of the risk bound $\mathscr{E}(\boldsymbol{\theta}^*, \mathcal{S}_+^*, \mathcal{S}_-^*)$ is of order $O(\lambda_1 \Delta_{\boldsymbol{\alpha}} |\rho^*|)$, which reflects the estimation error in $\widehat{\boldsymbol{\alpha}}$. Following Lemma S8 in the Supplement, one can*

show that $\Delta_{\boldsymbol{\alpha}} = O_p(N^{-1/2}|\mathcal{A}^*|)$. *All the other terms in* Ξ *describe the estimation error in* $\widehat{\boldsymbol{\theta}}$ *as if* $\widehat{\boldsymbol{\alpha}}$ *is replaced with* $\boldsymbol{\alpha}^*$. *When* N *is sufficiently large,* $O(\lambda_1 \Delta_{\boldsymbol{\alpha}}|\rho^*|)$ *is typically negligible relative to other terms. Specifically, if* $N \gg n|\mathcal{A}^*|^2 \log(p)$, $O(\lambda_1 \Delta_{\boldsymbol{\alpha}}|\rho^*|) = O(\{Nn\}^{-1/2} \log(p)^{1/2}|\mathcal{A}^*|) = o(n^{-1})$. *In general, as long as* $N \gg \max(n, p)$ *and* $\boldsymbol{\alpha}^*$ *is not much denser than* $\boldsymbol{\beta}_0$ *as in the typical EHR application cases,* $O(\lambda_1 \Delta_{\boldsymbol{\alpha}}|\rho^*|)$ *is dominated by the risk of the supervised LASSO estimator and even the supervised oracle estimator obtained under the knowledge of* $\mathrm{supp}(\boldsymbol{\beta}_0)$.

To gain a better understanding of how the key quantity Ξ in Theorem 1 changes with respect to the similarity between the prior information $\boldsymbol{\alpha}^*$ and the target $\boldsymbol{\beta}_0$, we shall discuss several specific cases in Section 3.3, based on the risk bound derived in Theorem 1.

3.3 Specific Cases

Following Remark 3, we focus our discussions on the settings where N is sufficiently large such that the last term of the risk bound is negligible. We consider three different scenarios as illustrated in Figure 1: (Case 1) $\boldsymbol{\alpha}^*$ recovers both the support and direction of $\boldsymbol{\beta}_0$; (Case 2) $\boldsymbol{\alpha}^*$ almost recovers the support of $\boldsymbol{\beta}_0$ but has a substantially different direction from $\boldsymbol{\beta}_0$; (Case 3) $\boldsymbol{\alpha}^*$ fails to recover the support of $\boldsymbol{\beta}_0$ (let alone its direction) and provides poor information. These three cases depict perfect, good, and poor qualities of the prior information $\boldsymbol{\alpha}^*$ in recovering the support and direction of $\boldsymbol{\beta}_0$. Next, we rigorously characterize the three cases by properly specifying the parameters ρ, $\boldsymbol{\delta}$, $\mathcal{S}_+$, and $\mathcal{S}_-$, and derive the convergence rate of Ξ, the risk bound of the PASS estimator, based on Theorem 1.

Case 1. *Let* $\overline{\rho} = \min_\rho \|\boldsymbol{\beta}_0 - \rho\boldsymbol{\alpha}^*\|_1$, $\overline{\boldsymbol{\delta}} = \boldsymbol{\beta}_0 - \overline{\rho}\boldsymbol{\alpha}^*$, $\overline{\boldsymbol{\theta}} = (\overline{\boldsymbol{\delta}}^\intercal, \overline{\rho}, \gamma_0, \zeta_0)^\intercal$, $\overline{\mathcal{S}}_+ = \overline{\mathcal{P}}$ *and* $\overline{\mathcal{S}}_- = \mathrm{supp}(\boldsymbol{\delta}_0)$. *If* $\boldsymbol{\alpha}^*$ *successfully recovers the support and direction of* $\boldsymbol{\beta}_0$ *(see the left panel in Figure 1),* $\overline{\mathcal{S}}_- \approx \emptyset$ *and* $\|\boldsymbol{\delta}_0\|_1 \approx 0$. *Since* $\|\boldsymbol{G}^{1/2}(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|_2 = 0$ *and* $\overline{\mathcal{S}}_+ \cap \mathcal{A}^* = \emptyset$, *we have* $\Xi = O\{\mathscr{E}(\overline{\boldsymbol{\theta}}, \overline{\mathcal{S}}_+, \overline{\mathcal{S}}_-)\}$ *by the definition of* $\boldsymbol{\theta}^*$. *Hence by Theorem 1, the excess risk of* $\widehat{\boldsymbol{\theta}}$

$$\Xi = O_p(\lambda_0^2 + \lambda_1 \|\overline{\boldsymbol{\delta}}_{\overline{\mathcal{S}}_- \cap \mathcal{A}^*}\|_1 + \lambda_2 \|\overline{\boldsymbol{\delta}}_{\overline{\mathcal{S}}_- \cap \mathcal{A}^{*c}}\|_1) \approx O_p(\lambda_0^2) = O_p(n^{-1}),$$

recalling that $\lambda_0 = O(n^{-1/2})$.

As a standard result (Negahban et al., 2009), the rate of the excess risk of the supervised LASSO estimator is either $O_p\{n^{-1}\log(p)|\mathcal{B}_0|\}$ or $O\{n^{-1/2}\log(p)^{1/2}\|\boldsymbol{\beta}_0\|_1\}$. These two rate bounds are established under different sparsity norms of $\boldsymbol{\beta}_0$, and generally comparable, e.g. when order of average magnitude of the non-zero entries in $\boldsymbol{\beta}_0$ is $n^{-1/2}\log(p)^{1/2}$. In comparison with them, $O_p(n^{-1})$, the risk rate of PASS in Case 1, is much more refined. Further, $O_p(n^{-1})$ is actually the rate of the estimator of a low (fixed) dimensional logistic regression. Thus, if $\boldsymbol{\beta}_0$ is very close to a multiple of $\boldsymbol{\alpha}^*$, PASS could outperform the vanilla LASSO and be comparable with a low dimensional regression in terms of the convergence rate. This big gain is owing to the use of N unlabeled dataset to obtain the direction of $\boldsymbol{\beta}_0$, and thus reduce the high dimensional regression to a low dimensional one where only the intercept and the scalar of $\boldsymbol{\beta}_0$ need to be estimated.

Case 2. *Consider the same choice of* $\overline{\boldsymbol{\theta}}$, $\overline{\mathcal{S}}_+$ *and* $\overline{\mathcal{S}}_-$ *as in Case 1. If* $\boldsymbol{\alpha}^*$ *recovers the support but not the direction of* $\boldsymbol{\beta}_0$ *(see the middle of Figure 1), we will only have* $\|\overline{\boldsymbol{\delta}}_{\overline{\mathcal{S}}_- \cap \mathcal{A}^{*c}}\|_1 \approx 0$

but not $\|\boldsymbol{\delta}_0\|_1 \approx 0$. *Then by Theorem 1, the excess risk of PASS is*

$$\Xi = O_p(\lambda_0^2 + \lambda_1\|\overline{\boldsymbol{\delta}}_{\overline{\mathcal{S}}_-\cap\mathcal{A}^*}\|_1 + \lambda_2\|\overline{\boldsymbol{\delta}}_{\overline{\mathcal{S}}_-\cap\mathcal{A}^{*c}}\|_1) \approx O_p\{n^{-1/2}\log(q^*)^{1/2}\|\overline{\boldsymbol{\delta}}_{\overline{\mathcal{S}}_-\cap\mathcal{A}^*}\|_1\},$$

recalling that $\lambda_1 = O\{n^{-1/2}\log(q^*)^{1/2}\}$.

In Case 2, the convergence rate of the excess risk of PASS is still better than that of the supervised LASSO estimator when $q^* \ll p$:

$$O\{n^{-1/2}\log(q^*)^{1/2}\|\overline{\boldsymbol{\delta}}_{\overline{\mathcal{S}}_-\cap\mathcal{A}^*}\|_1\} \ll O\{n^{-1/2}\log(p)^{1/2}\|\boldsymbol{\beta}_0\|_1\},$$

by $\|\overline{\boldsymbol{\delta}}_{\overline{\mathcal{S}}_-\cap\mathcal{A}^*}\|_1 \leq \min_\rho\|\boldsymbol{\beta}_0 - \rho\boldsymbol{\alpha}^*\|_1 \leq \|\boldsymbol{\beta}_0\|_1$. Namely, if $\boldsymbol{\alpha}^*$ might not recover the direction of $\boldsymbol{\beta}_0$ very well but the prior information $\mathcal{A}^* = \text{supp}(\boldsymbol{\alpha}^*)$ is sparse and covers $\text{supp}(\boldsymbol{\beta}_0)$ successfully, which is reflected as $\overline{\mathcal{S}}_+ = \overline{\mathcal{P}}$, the PASS estimator still benefits from the prior information. This is because recovering the support of $\boldsymbol{\beta}_0$ reduces the dimensionality of the empirical errors needed to be controlled from p to $q^* = |\mathcal{A}^*|$. In this case, it is also interesting to compare the proposed PASS estimator with the prior LASSO (pLASSO) procedure of Jiang et al. (2016). When $\text{supp}(\boldsymbol{\alpha}^*)$ and $\text{supp}(\boldsymbol{\beta}_0)$ are close but the directions of $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}_0$ are quite different, the pLASSO procedure is unable to utilize this information and will only result in the same convergence rate as supervised LASSO, as shown to be essentially slower than that of PASS.

Case 3. *Let* $\overline{\rho} = 0$, $\overline{\boldsymbol{\delta}} = \boldsymbol{\beta}_0$, $\overline{\boldsymbol{\theta}} = (\overline{\boldsymbol{\delta}}^\intercal, \overline{\rho}, \gamma_0, \zeta_0)^\intercal$, $\overline{\mathcal{S}}_+ = \overline{\mathcal{P}} \cup (\mathcal{B}_0 \setminus \mathcal{A}^*)$ *and* $\overline{\mathcal{S}}_- = \mathcal{B}_0 \setminus \overline{\mathcal{S}}_+$. *If* $\boldsymbol{\alpha}^*$ *fails to recover the support of* $\boldsymbol{\beta}_0$, *i.e.* $\mathcal{A}^* \cap \mathcal{B}_0 \approx \emptyset$ *and* $\|\boldsymbol{\beta}_{0,\mathcal{A}^*\cap\mathcal{B}_0}\|_1 \approx 0$, *we have* $\|\overline{\boldsymbol{\delta}}_{\overline{\mathcal{S}}_-}\|_1 \leq \|\overline{\boldsymbol{\delta}}_{\mathcal{A}^*}\|_1 \approx \|\boldsymbol{\beta}_{0,\mathcal{A}^*\cap\mathcal{B}_0}\|_1 \approx 0$. *Then again using Theorem 1,*

$$\Xi = O(\lambda_2^2|\overline{\mathcal{S}}_+| + \lambda_1\|\overline{\boldsymbol{\delta}}_{\overline{\mathcal{S}}_-\cap\mathcal{A}^*}\|_1 + \lambda_2\|\overline{\boldsymbol{\delta}}_{\overline{\mathcal{S}}_-\cap\mathcal{A}^{*c}}\|_1) \approx O_p\{n^{-1}\log(p)|\mathcal{B}_0|\},$$

recalling that $\lambda_2 = O\{n^{-1/2}\log(p)^{1/2}\}$.

In Case 3, the excess risk of the PASS is of the same order as that of supervised LASSO. Therefore the PASS approach is robust against low-quality prior information that recovers neither the direction nor the support of $\boldsymbol{\beta}_0$. This benefit is a result of using a data-adaptive parameter ρ to control the influence of the prior information on the estimator.

4. Simulation Studies

4.1 Main setups

We conducted extensive simulation studies to examine the finite-sample performance of the PASS estimator and to compare it with existing approaches. We first considered the case where the logistic model for $Y \mid S, \boldsymbol{X}$ is correctly specified, $S \mid \boldsymbol{X}$ follows an SIM, and $\boldsymbol{X}$ is near elliptical, but the similarity between $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ varies. Since EHR features are often zero inflated and skewed count variables, we generated $\boldsymbol{X}_{500\times 1}$ from

$$\boldsymbol{X}_i = h(\boldsymbol{Z}_i), \quad \boldsymbol{Z}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}), \quad h(t) = \log(1 + [e^t]),$$

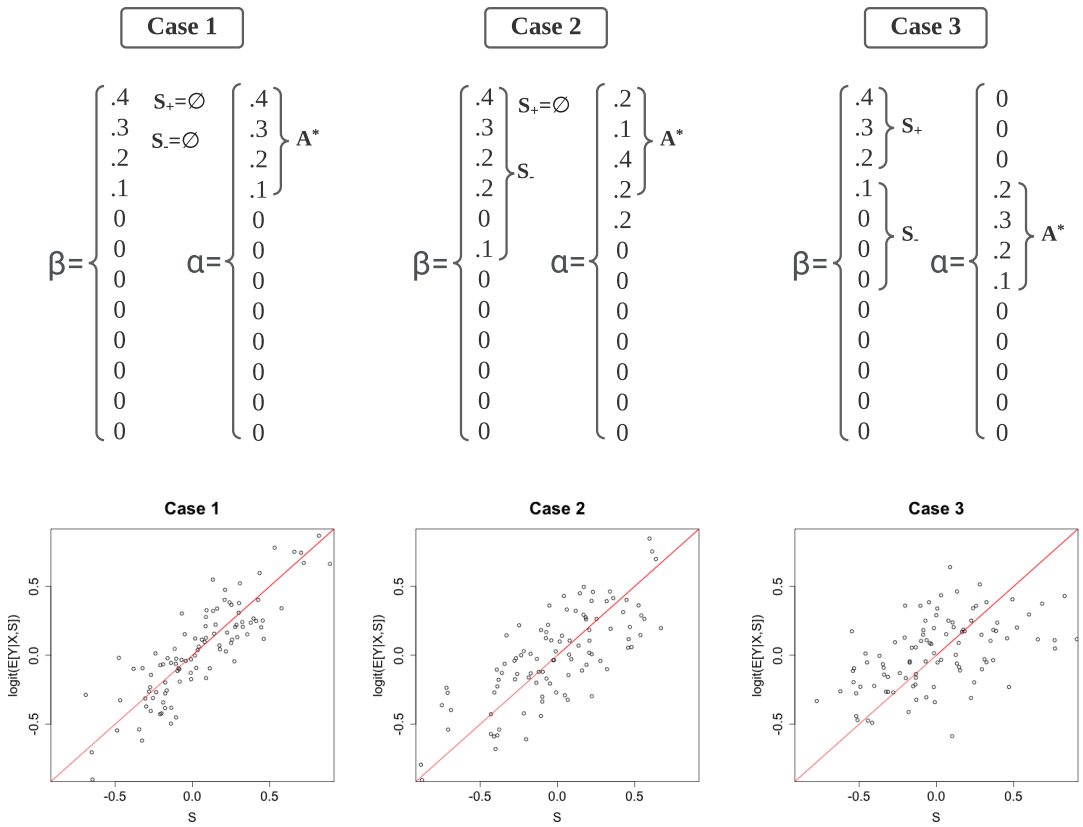


Figure 1: Examples of the coefficients $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}^*$ in the three cases of Section 3.3. Labels S_+, S_-, and A^* in the diagrams represent $\overline{\mathcal{S}}_+ \setminus \overline{\mathcal{P}}$, $\mathcal{S}_-$, and $\mathcal{A}^*$ as chosen and defined in Cases 1–3. $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}^*$ are aligned for comparison of their directions and supports. Presented below are scatter plots for $\sigma^{-1}\{\Pr(Y = 1 \mid S, \boldsymbol{X})\}$ against S of the simulated samples generated under Cases 1–3.
Case 1 (presented in the left panel): $\boldsymbol{\alpha}^*$ recovers both the support and direction of $\boldsymbol{\beta}_0$. $\sigma^{-1}\{\Pr(Y = 1 \mid S, \boldsymbol{X})\}$ shows strong collinearity with S. PASS largely outperforms supervised LASSO and has the same convergence rate as the low dimensional regression.
Case 2 (middle): $\boldsymbol{\alpha}^*$ (nearly) recovers the support but not the direction $\boldsymbol{\beta}_0$. $\sigma^{-1}\{\Pr(Y = 1 \mid S, \boldsymbol{X})\}$ shows moderate collinearity with S. PASS still outperforms both supervised LASSO and pLASSO in terms of convergence rate.
Case 3 (right): $\boldsymbol{\alpha}^*$ fails to recover the support of $\boldsymbol{\beta}_0$. $\sigma^{-1}\{\Pr(Y = 1 \mid S, \boldsymbol{X})\}$ shows weak collinearity with S. PASS is of the same convergence rate as supervised LASSO.

where $[u]$ denotes the integer nearest to u, $\boldsymbol{\Sigma_Z} = (\sigma_{i,j})_{i,j=1}^p$ and $\sigma_{i,j} = 4(0.5)^{|i-j|}$. Here $[e^{Z_{ij}}]$ mimics the skewed raw EHR feature, which is typically transformed via $t \to \log(1+t)$ prior to model fitting. We then generated the surrogate S from a SIM of $\boldsymbol{X}$:

$$S_i = h(1 + \boldsymbol{X}_i^\intercal \boldsymbol{\alpha}_0 + \epsilon_i), \quad \text{with} \quad \epsilon_i \sim N(0, 2^2).$$

Following the model assumption ($\mathcal{M}_Y$), the disease status Y_i was generated from

$$\sigma^{-1}\{\Pr(Y_i = 1 \mid \mathbf{W}_i)\} = -4 + 0.5S_i + \boldsymbol{X}_i^\intercal \boldsymbol{\beta}_0.$$

To mimic different qualities of the prior information one could encounter in practice, we design six scenarios with different similarities between the true $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$:

$$\begin{aligned}
&\text{I}: && \boldsymbol{\alpha}_0 = (\boldsymbol{a}_1^\intercal, \boldsymbol{a}_2^\intercal, \mathbf{0}_{p-10}^\intercal)^\intercal, && \boldsymbol{\beta}_0 = 1.5(\boldsymbol{a}_1^\intercal, \boldsymbol{a}_2^\intercal, \mathbf{0}_{p-10}^\intercal)^\intercal;\\
&\text{II}: && \boldsymbol{\alpha}_0 = (\boldsymbol{a}_1^\intercal, \boldsymbol{a}_2^\intercal, \mathbf{0}_{p-10}^\intercal)^\intercal, && \boldsymbol{\beta}_0 = 1.5(\boldsymbol{a}_1^\intercal + \boldsymbol{d}_1^\intercal, \boldsymbol{a}_2^\intercal + \boldsymbol{d}_2^\intercal, \mathbf{0}_{p-10}^\intercal)^\intercal;\\
&\text{III}: && \boldsymbol{\alpha}_0 = (\boldsymbol{a}_1^\intercal, \boldsymbol{a}_2^\intercal, \boldsymbol{a}_2^\intercal, \boldsymbol{a}_2^\intercal, \mathbf{0}_{p-20}^\intercal)^\intercal, && \boldsymbol{\beta}_0 = 1.5(\boldsymbol{a}_1^\intercal + \boldsymbol{d}_1^\intercal, \boldsymbol{a}_2^\intercal + \boldsymbol{d}_2^\intercal, \mathbf{0}_{p-10}^\intercal)^\intercal;\\
&\text{IV}: && \boldsymbol{\alpha}_0 = (\boldsymbol{a}_1^\intercal, \mathbf{0}_{p-5}^\intercal)^\intercal, && \boldsymbol{\beta}_0 = 1.5(\boldsymbol{a}_1^\intercal + \boldsymbol{d}_1^\intercal, \boldsymbol{a}_2^\intercal + \boldsymbol{d}_2^\intercal, \mathbf{0}_{p-10}^\intercal)^\intercal;\\
&\text{V}: && \boldsymbol{\alpha}_0 = (\boldsymbol{a}_1^\intercal, \boldsymbol{a}_2^\intercal, \mathbf{0}_{p-10}^\intercal)^\intercal, && \boldsymbol{\beta}_0 = 1.5(\boldsymbol{a}_2^\intercal, \boldsymbol{a}_1^\intercal, \mathbf{0}_{p-10}^\intercal)^\intercal;\\
&\text{VI}: && \boldsymbol{\alpha}_0 = (\boldsymbol{a}_1^\intercal, \boldsymbol{a}_2^\intercal, \mathbf{0}_{p-10}^\intercal)^\intercal, && \boldsymbol{\beta}_0 = 1.5(\boldsymbol{a}_2^\intercal, \mathbf{0}_5, \boldsymbol{a}_1^\intercal, \mathbf{0}_{p-15}^\intercal)^\intercal.
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{a}_1 &= (0.5, 1, -0.8, 0.6, 0.2)^\intercal, & \boldsymbol{d}_1 &= (-0.05, -0.5, 1.4, 0.5, -0.6)^\intercal,\\
\boldsymbol{a}_2 &= (0.1, -0.2, -0.2, 0.2, 0.7)^\intercal, & \boldsymbol{d}_2 &= (0.02, 0.05, 0.02, -0.02, -0.05)^\intercal.
\end{aligned}$$

Our specifications of $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ are motivated by the three key specific cases introduced in Section 3.3 and illustrated in Figure 1. Scenario I is the ideal case where $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ have identical direction. In Scenario II, most of the components of $\boldsymbol{\beta}_0$ differ slightly from a scalar multiple of $\boldsymbol{\alpha}_0$, while a few components differ substantially. Scenarios I and II are designed to examine the performance of PASS estimator when the prior information is highly or somewhat reliable. In Scenario III, $\boldsymbol{\alpha}_0$ is denser than $\boldsymbol{\beta}_0$ and contains quite a few weak signals. On the contrary, in Scenario IV $\boldsymbol{\beta}_0$ is denser than $\boldsymbol{\alpha}_0$. In Scenario V, the magnitude of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ are quite different, whereas they still share the same support. Scenarios III, IV and V are designed to examine the performance of PASS estimator with respect to different degrees of accuracy of the support information. In Scenario VI, both the magnitude and the support of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ differs substantially, which means the unlabeled dataset provides little information. This scenario allows us to see whether the PASS estimator is robust against unreliable prior information. See Figure 2 for a visualization of $\boldsymbol{\beta}_0$ and $\rho\boldsymbol{\alpha}_0$ across different scenarios.

We compare PASS to following existing methods: (1) supervised LASSO penalized logistic regression with n training samples (LASSO$_n$); (2) supervised ALASSO penalized logistic regression with n training samples, denoted by ALASSO$_n$; (3) the SS^{prior} estimator as described in section 2.2; and (4) two variants of pLASSO estimators as proposed in Jiang et al. (2016). For pLASSO, we fit a penalized logistic model with an LASSO penalty imposed on predictors outside supp($\widehat{\boldsymbol{\alpha}}$), as in equation (8) of Jiang et al. (2016), and then

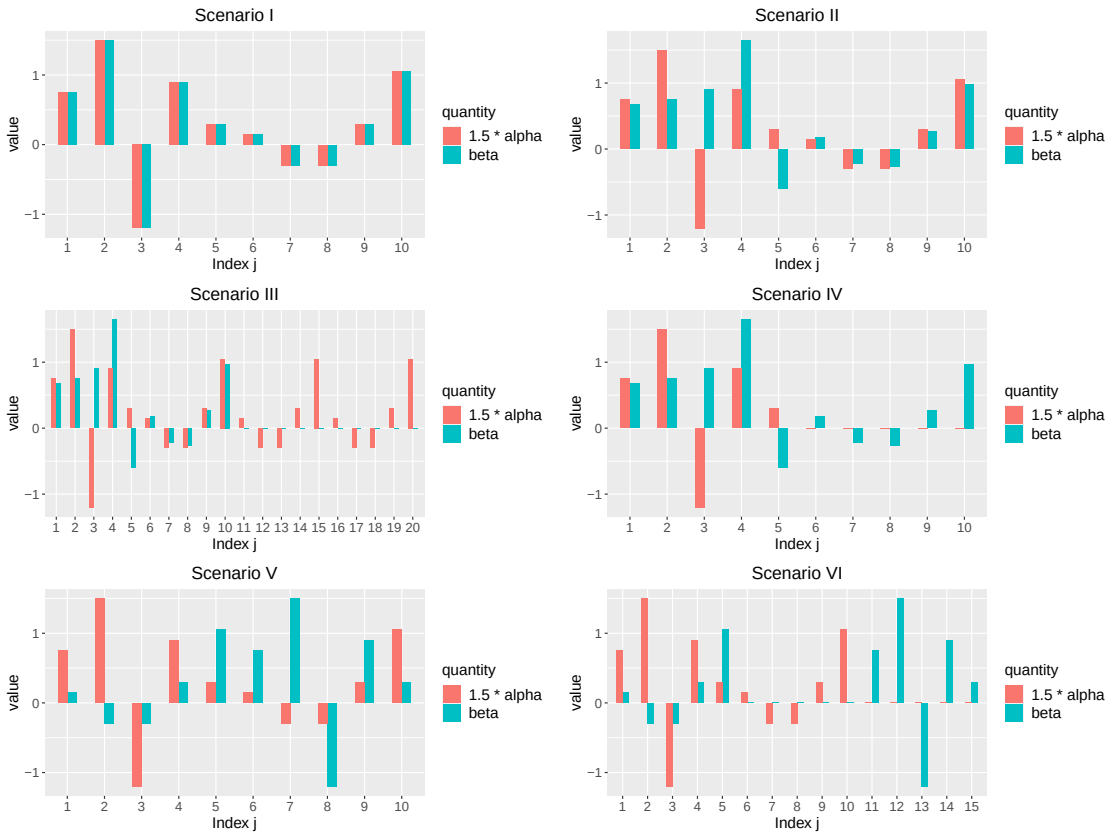


Figure 2: Supports and values of the coefficients $\boldsymbol{\beta}_0$ and $1.5\boldsymbol{\alpha}_0$ under Scenarios I–VI introduced in Section 4.1. Only those indices j satisfying $\beta_{0,j} \neq 0$ or $\alpha_{0,j} \neq 0$ are shown in the plots.

use the predicted probability from that model as Y_i^{p} in equation (7) of Jiang et al. (2016), denoted by pLASSO^1; (2) use the predicted probability given by the SS^{prior} approach as Y_i^{p} in equation (7) of their paper, denoted by pLASSO^2.

Throughout, we let $N = 10000$ and let $\nu = 1$ in the ALASSO weights. We use Bayesian information criterion (BIC) to select μ_{init} and μ in the estimation of $\boldsymbol{\alpha}$ due to large N, and use 10-fold cross-validation to select λ_1, λ_2 for the estimation of $\boldsymbol{\beta}$, so that the phenotype model is tuned towards prediction performance. We quantify the average prediction performance of the estimated linear score, $\widetilde{\boldsymbol{\vartheta}}^{\mathsf{T}}\vec{\mathbf{W}}$, with $\widetilde{\boldsymbol{\vartheta}}$ obtained via different methods in an independent test dataset with size 10000. For each choice of $\widetilde{\boldsymbol{\vartheta}}^{\mathsf{T}}\vec{\mathbf{W}}$, we consider the area under the receiver operating characteristic curve (AUC) for classifying Y, the excess risk (ER) as defined in Section 3, and the mean squared error of the predicted probabilities (MSE-P) which is the mean squared differences between the predicted probability and the true probability. We summarize results based on 1000 simulated datasets for each configuration.

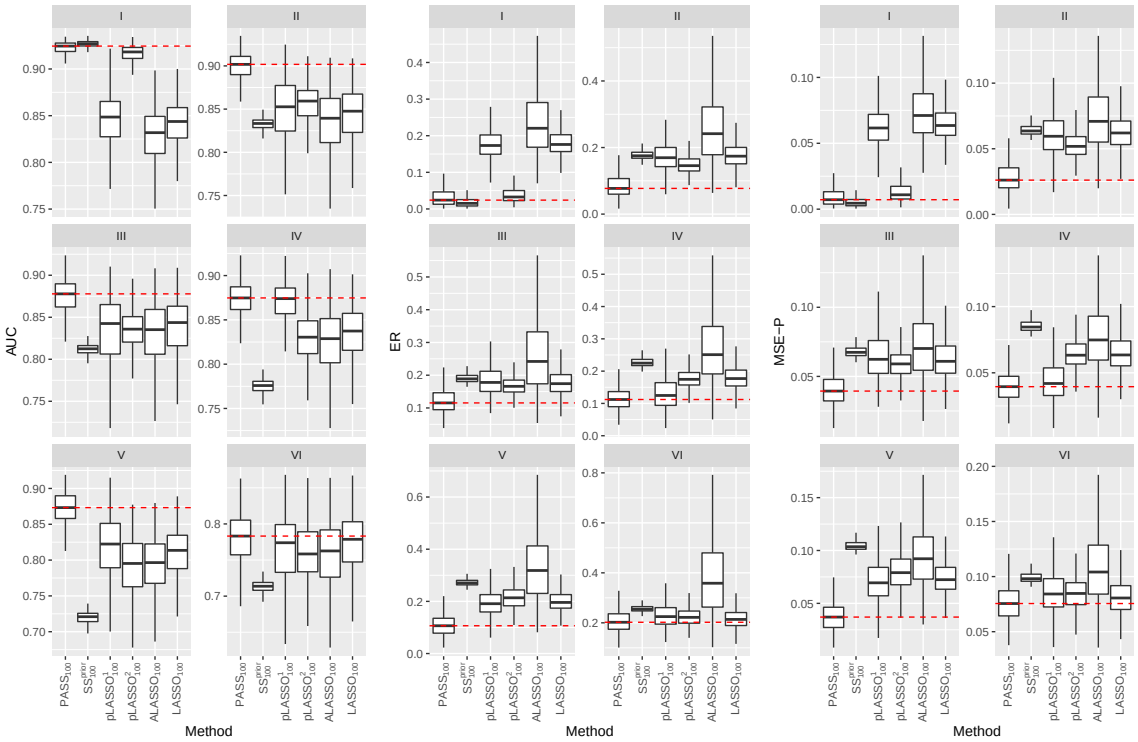


Figure 3: AUC (left), ER (middle) and MSE-P (right) evaluated on the test set for simulation studies under Scenarios I–VI. Outliers are not drawn. Mean performance of the PASS approach are marked using dashed lines for ease of comparison. The size of the labeled dataset is fixed at $n = 100$.

In Figures 3, we compare prediction measures for estimators obtained with $n = 100$. In Scenario I where the directions of $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ coincide, the SS^{prior} approach performs the best as expected, yet the proposed PASS method attained very similar accuracy followed by pLASSO[2] which performed only slightly worse. When the directions of $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ are somewhat different as in Scenario II, the SS^{prior} and the pLASSO estimators deteriorated quickly. In contrast, the PASS estimator maintains high accuracy and outperforms all competing estimators substantially. We observe qualitatively similar patterns for Scenarios III and IV under which $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ have somewhat different support. No matter whether $\boldsymbol{\alpha}_0$ is denser than $\boldsymbol{\beta}_0$ as in Scenario IV, or $\boldsymbol{\beta}_0$ is denser than $\boldsymbol{\alpha}_0$ as in Scenario V, the PASS method consistently outperforms the supervised estimators. Additionally, the performances of the SS^{prior} and pLASSO approaches are not quite satisfactory. In Scenario V, $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ have the same support but are quite different in terms of magnitude. The proposed method managed to utilize the same-support information, whereas the pLASSO approaches failed to do so. Finally, the goal of Scenario VI is to examine the robustness of the methods when $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ differs a lot, possibly due to the use of an inappropriate surrogate. The PASS estimator performs similarly to the supervised estimators, indicating that our procedure is indeed adaptive to how well the data supports the prior assumption. Across all scenarios, the ALASSO approach performs slightly worse than LASSO, possibly due to the presence of some small nonzero coefficients in $\boldsymbol{\beta}_0$.

In Figure 4, we present the AUC, ER and MSE-P of the PASS estimator trained with $n = 100$ and the supervised LASSO estimator with varying label size. In Scenario I where the prior assumption holds exactly, PASS$_{100}$, the PASS approach with 100 labeled samples, even outperforms LASSO$_{400}$, the LASSO approach with 400 labeled samples. When the prior assumption holds approximately as in Scenarios II through V, PASS$_{100}$ consistently outperforms LASSO$_{150}$, and achieves similar performance as LASSO$_{200}$, which requires twice as many labels. Finally, in Scenario VI where the prior information is highly inaccurate, the PASS method maintains comparable performance against LASSO$_{100}$.

4.2 Efficiency and Robustness Evaluations under Mis-specifications

We conducted simulation studies under three additional scenarios to further investigate the efficiency and robustness of PASS when the model assumptions and elliptical design assumptions are violated. We again set $p = 500$ and generate $\boldsymbol{X}_i = 2\Phi(\boldsymbol{Z}_i) - 1$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal, $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\intercal \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}'_{\boldsymbol{Z}})$, $\boldsymbol{\Sigma}'_{\boldsymbol{Z}} = (\sigma'_{i,j})_{i,j=1}^p$, $\sigma'_{i,j} = (0.5)^{|i-j|}$ if $i = j$ or both i and j are ≤ 20 or both i and j are > 20 and $\sigma'_{i,j} = 0$ otherwise. We make $\boldsymbol{\Sigma}'_{\boldsymbol{Z}}$ block-diagonal for the convenience of obtaining the population solution of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ through the best logistic or least square approximation under model mis-specifications. In real EHR studies, a paradigm of data generation is that the features $\boldsymbol{X}$, e.g. some genetic variants, precedes the disease status Y, and Y precedes some clinical surrogate S, e.g. the count of ICD codes associated with the disease. To mimic this scenario, we generated Y_i and S_i from the following models:

$$
\begin{aligned}
Y_i &= I\{(0.8, 1, -1, 0.8, 0.4, \boldsymbol{0}_{p-5}^\intercal)\boldsymbol{X}_i + \epsilon_{yi} \geq 0\}, \quad && \epsilon_{yi} \sim N(0,1), \\
S_i &= \mu Y_i + \boldsymbol{\eta}_1^\intercal \boldsymbol{X}_i + Y_i \boldsymbol{\eta}_2^\intercal \boldsymbol{X}_i + \epsilon_{si}, && \epsilon_{si} \sim N(0,1).
\end{aligned}
$$

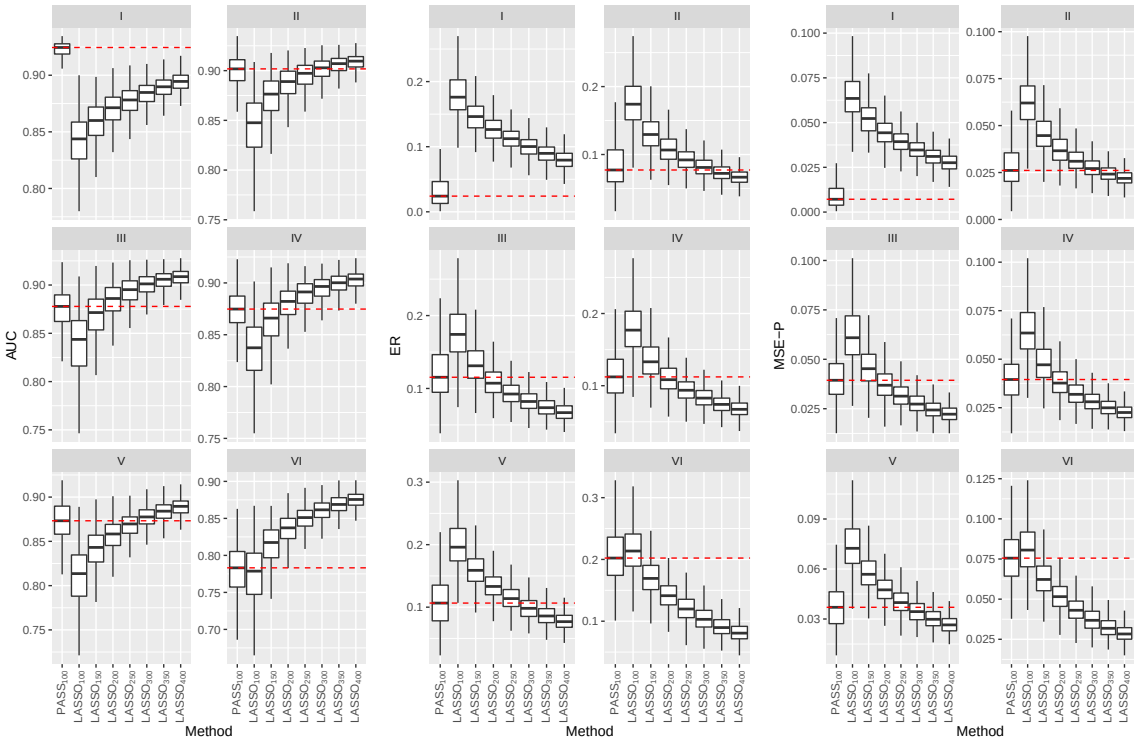


Figure 4: AUC (left), ER (middle) and MSE-P (right) evaluated on the test set for simulation studies under Scenarios I–VI. Outliers are not drawn. Mean performance of the PASS approach are marked using red dash lines for ease of comparison. The size of the labeled dataset is $n = 100$ for PASS, while it varies for LASSO, as indicated in the subscripts.

Assumptions ($\mathcal{C}^{\text{prior}}$) and ($\mathcal{M}_S$) hold when $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2 = \mathbf{0}$, and would be severely violated when $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are large. We design three scenarios with $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ representing different degrees of violation on the surrogate assumptions:

i: $\mu = 1$, and $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2 = \mathbf{0}$;

ii: $\mu = 1.5$, $\boldsymbol{\eta}_1 = (\boldsymbol{a}_3^\intercal, \mathbf{0}_{p-5}^\intercal)^\intercal$, and $\boldsymbol{\eta}_2 = (\boldsymbol{d}_3^\intercal, \mathbf{0}_{p-5}^\intercal)^\intercal$;

iii: $\mu = 2$, $\boldsymbol{\eta}_1 = (\boldsymbol{a}_3^\intercal, \boldsymbol{a}_3^\intercal, \boldsymbol{a}_3^\intercal, \mathbf{0}_{p-15}^\intercal)^\intercal$, and $\boldsymbol{\eta}_2 = (\boldsymbol{d}_3^\intercal, \boldsymbol{d}_3^\intercal, \boldsymbol{d}_3^\intercal, \mathbf{0}_{p-15}^\intercal)^\intercal$,

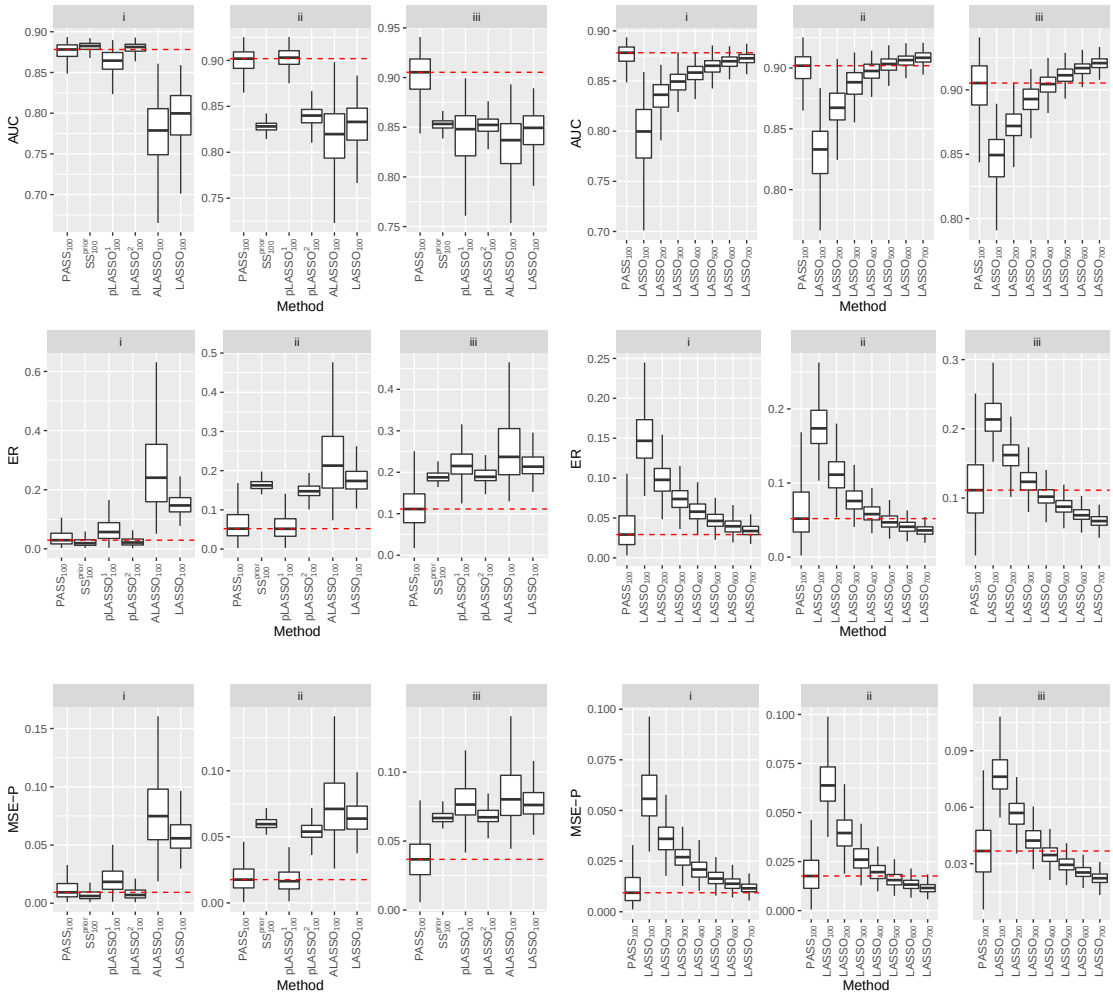
where $\boldsymbol{a}_3 = (0.6, -0.4, 0.4, 0.5, -0.5)^\intercal$ and $\boldsymbol{d}_3 = (0.3, 0.4, 0.6, -0.5, -0.5)^\intercal$. Here μ depict the marginal effect of Y_i on S_i, and are set to make the AUC of target models at a similar level across the three scenarios. Across all scenarios, $\Pr(Y_i = 1 \mid S_i, \boldsymbol{X}_i)$ is no longer a parametric logistic model, i.e. ($\mathcal{M}_Y$) is misspecified. Our goal is to estimate the limiting coefficients $\zeta_0, \gamma_0, \boldsymbol{\beta}_0$ defined as the minimizor of $\mathbb{E}\,\ell(Y_i, \zeta + \gamma S_i + \boldsymbol{X}_i^\intercal \boldsymbol{\beta})$. Benchmark methods, and their implementation, tuning, and evaluation procedures are the same as in Section 4.1, except that we implement supervised LASSO with n ranging from 100 to 700.

In Figure 5, we present AUC, ER and MSE-P of the methods under Scenarios i–iii. In Scenario i, PASS has similar performances as the semi-supervised benchmarks SS^{prior}and pLASSO, and all the semi-supervised estimators significantly outperform the two supervised estimators since ($\mathcal{C}^{\text{prior}}$) holds and $\boldsymbol{\alpha}^*$ basically recovers the direction of $\boldsymbol{\beta}_0$ well. Among the semi-supervised estimators, the SS^{prior}and pLASSO^2 estimators have a slight advantage with a smaller variation than expected since both heavily rely on the prior information which is of high quality in this setting. In Scenario ii, the key assumption ($\mathcal{C}^{\text{prior}}$) is violated, which drastically impacts the performance of SS^{prior}and pLASSO^2. On the other hand, PASS and pLASSO^1 still effectively leverage the imperfect information from $\boldsymbol{\alpha}^*$ to approximately recover the support of $\boldsymbol{\beta}_0$, and thus outperform SS^{prior}, pLASSO^2, and the supervised methods. In Scenario iii, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ become denser than those in Scenario ii. This can make the recovery of $\text{supp}(\boldsymbol{\beta}_0)$ using $\text{supp}(\boldsymbol{\alpha}^*)$ less accurate, and interestingly, PASS outperforms all methods including pLASSO^1 that also leverages $\text{supp}(\boldsymbol{\alpha}^*)$. In all the three scenarios, PASS significantly outperforms supervised LASSO using the same number or even 2-3 more times the number of samples, which displays a large gain of using the unlabelled dataset to assist the regression. Finally, the results demonstrate that our method can still efficiently leverage the prior information from S in estimating the target parameters when $S \mid Y, \boldsymbol{X}$ highly depends on $\boldsymbol{X}$ so $\mathcal{C}^{\text{prior}}$ is violated, ($\mathcal{M}_Y$) is misspecified, and the design is non-elliptical.

5. Application to EHR Phenotying

We examine the performance of PASS along with other approaches in three real world EHR phenotyping studies with the goal of developing classification models for the diseases of interest. All studies are performed at a large tertiary hospital system with EHR spanning over multiple decades. Each study has n_0 labeled observations for algorithm training and validation. We consider three choices of training size n no more than $n_0/2$ in all examples. First, we randomly split the labelled samples into four folds of equal sizes. Then we pick each fold as the validation set, sample n training labels from the other three folds for 20 times, train and validate the algorithms, and finally average the evaluation metrics and their

Figure 5: Evaluation metrics on the test set for simulation studies under Scenarios i–iii introduced in Section 4.2. Outliers are not drawn. Mean performance of the PASS approach are marked using red dash lines for ease of comparison. On the left panel, we present the evaluation metrics of all methods for comparison when $n = 100$. On the right panel, we compare the performance of PASS when $n = 100$ with supervised LASSO obtained using labelled samples with various n (from 100 to 700).



standard errors over the validation results on the four folds. We replicate this procedure 10 times and report the average performance.

Data Example 1 (CAD Phenotyping)**.** The goal of this study is to identify patients with coronary artery disease (CAD) based on their EHR features. The study cohort consists of $N = 4164$ patients, out of which a random subset of $n_0 = 181$ patients have their true CAD status annotated via chart review by domain experts. We use the sum of the counts for the CAD ICD code and NLP mention of CAD as the surrogate. There are $p = 585$ additional EHR features consisting of the total count of all ICD codes as a healthcare utilization measure, 10 ICD codes related to CAD, and 574 NLP variables. For the size of training labels, we consider $n = 50, 70, 90$. This de-identified dataset has been analyzed in previous studies (Zhang et al., 2019, e.g.) and is publicly available online: `https://celehs.github.io/PheCAP/articles/example2.html`.

Data Example 2 (RA Phenotyping)**.** Similar to the CAD phenotyping study, the goal is to identify patients with rheumatoid arthritis (RA) based on their EHR features. There are $N = 46114$ patients in total and out of which, $n_0 = 435$ patients have their RA status annotated. Again, we choose the sum of the ICD code and NLP mention of RA as the surrogate. The $p = 924$ additional EHR features consist of the healthcare utilization and 923 NLP variables potentially predictive of RA. For the size of training labels, we consider $n = 50, 125, 200$.

Data Example 3 (Depression Phenotyping)**.** The goal is to identify patients with depression based on their codified EHR features. There are $N = 9474$ patients in total and $n_0 = 236$ labeled observations. The surrogate is chosen as the counts of depression ICD code. There are $p = 231$ additional EHR features, including the healthcare utilization and 230 codified EHR features on depression related medication prescriptions, laboratory tests and ICD codes. For the size of training labels, we consider $50, 85, 120$.

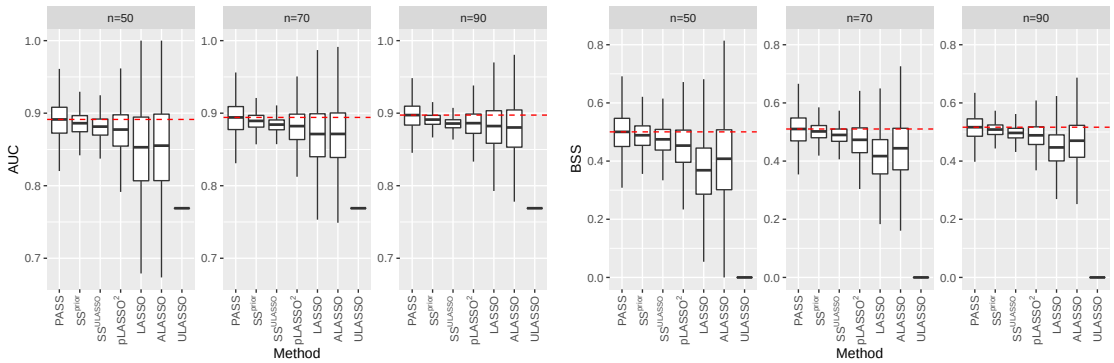
In the three data examples, N is significantly larger than p with $N/\max(p, n)$ being approximately 7 for CAD, 50 for RA, and 41 for Depression. In all the three studies, we apply $x \to \log(1 + x)$ transformation for all count variables. Also, since patients with higher healthcare utilization tend to have higher counts of most features, we orthogonalize all features against the healthcare utilization before regression fitting. Since $\boldsymbol{\vartheta}_0$ is unknown in applications, we quantify the performance of an estimator $\widetilde{\boldsymbol{\vartheta}}$ based on the AUC and Brier skill score (BSS) of $\sigma(\widetilde{\boldsymbol{\vartheta}}^\mathsf{T}\mathbf{W})$ for predicting Y, where the BSS is defined as $1 - \widehat{\mathbb{E}}_v[\{Y - \sigma(\widetilde{\boldsymbol{\vartheta}}^\mathsf{T}\mathbf{W})\}^2]/\widehat{\mathbb{E}}_v[\{Y - \widehat{\mathbb{E}}_v(Y)\}^2]$, and $\widehat{\mathbb{E}}_v$ denotes the empirical expectation on the validation sample. The BSS is essentially a binary version of the R-square.

For comparison, we included PASS, SS^{prior}, pLASSO^2, supervised LASSO and ALASSO on the three data examples to estimate the phenotyping model ($\mathcal{M}_Y$). We exclude pLASSO^1 since it requires fitting of an unpenalized regression on $\text{supp}(\widehat{\boldsymbol{\alpha}})$, which is infeasible when $|\text{supp}(\widehat{\boldsymbol{\alpha}})| > n$. In addition, we compare to the unsupervised LASSO (ULASSO) approach of Chakrabortty et al. (2017), which estimates direction of the logistic coefficients for $Y \sim \sigma(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X})$ by regressing $I(S > c_u)$ against $\boldsymbol{X}$ on the subset whose S is either greater than c_u or smaller than c_l, for some pre-specified c_u and c_l typically chosen such that $\Pr(S > c_u)$ and $\Pr(S < c_l)$ are small. Since the ULASSO approach only provides an estimate $\widetilde{\boldsymbol{\beta}}$ to optimize the prediction of $\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}$ for $Y \mid \boldsymbol{X}$ without using S explicitly as an additional predictor, we

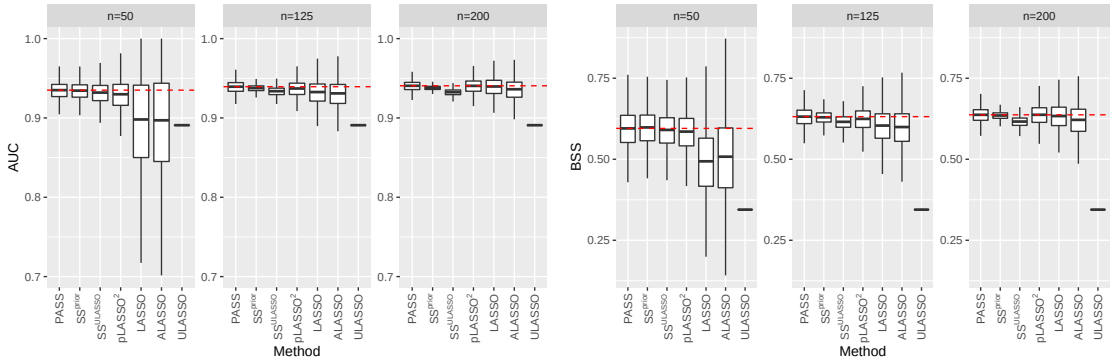
also derive a semi-supervised variant of ULASSO, denoted by $\text{SS}^{\text{ULASSO}}$, by regressing the labeled Y against $\widetilde{\boldsymbol{\beta}}^{\mathsf{T}}\boldsymbol{X}$ and S as for SS^{prior}.

Figure 6: Out of sample AUC and BSS on the data examples 1–3, with various sizes of labelled training samples denoted as n. Median performance of PASS are marked using red dash line for ease of comparison.

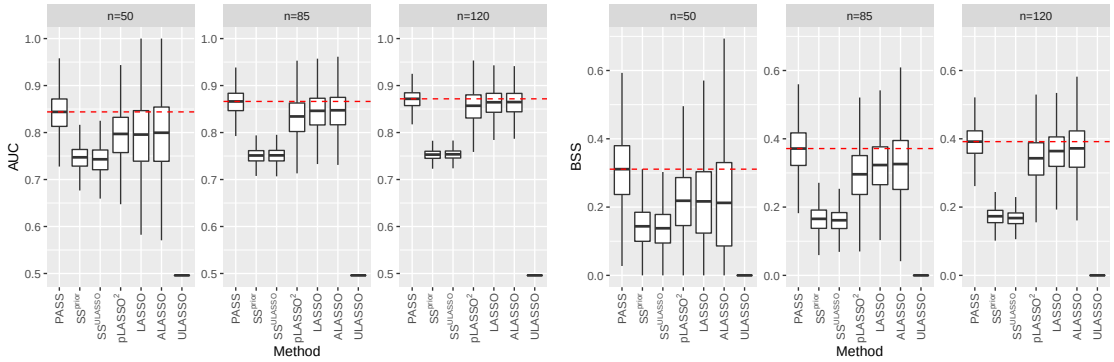
(a) CAD



(b) RA



(c) Depression



As shown in Figure 6, PASS significantly outperforms the supervised LASSO and ALASSO when $n = 50$ in all three examples. As the label size n increases, their performances get closer. Compared with the semi-supervised benchmarks, PASS has slightly or moderately better performance on the CAD and RA studies. For Depression, PASS substantially outperforms them, especially $\text{SS}^{\text{prior}}$and $\text{SS}^{\text{ULASSO}}$. For example, when $n = 50$, PASS attained average AUC in classifying depression about 0.1 higher than that of $\text{SS}^{\text{prior}}$and $\text{SS}^{\text{ULASSO}}$and 0.05 higher than pLASSO. The gap becomes smaller when n increases as expected. Interestingly, the supervised estimators outperform pLASSO, SS^{prior}, and $\text{SS}^{\text{ULASSO}}$on the Depression dataset as well but has similar or worse performance than these semi-supervised approaches on the other two examples. This could in part be attributed to the relatively poor quality of the surrogate information, which makes existing semi-supervised approaches fail. In contrast, PASS could utilize such prior information more effectively and robustly, and still preserves better performance than the supervised estimators. Thus, we can conclude that incorporating prior information from the unlabeled dataset could improve and stabilize the prediction performance of phenotyping models in EHR applications, and PASS is more robust and efficient in leveraging the prior information compared with existing semi-supervised methods. In addition, ULASSO shows much worse performance than the other supervised and semi-supervised methods in all examples. This illustrates the importance of collecting labels and including the surrogate in the regression models for EHR phenotyping.

6. Discussion

In this paper, we propose PASS, a high dimensional sparse estimator adaptively incorporating the prior knowledge from surrogate under a semi-supervised scenario commonly found in application fields like EHR analysis. Compared to the supervised approaches, the proposed PASS approach can substantially reduce the required number of labeled samples when the model assumptions ($\mathcal{M}_S$) and ($\mathcal{C}^{\text{prior}}$) and the elliptical design assumption (C1) hold exactly or approximately, and thus the prior information $\boldsymbol{\alpha}^*$ is trustworthy. Compared to existing pLASSO and SS^{prior} approaches that also incorporates prior information, the PASS approach is robust against unreliable prior information $\boldsymbol{\alpha}^*$, which might be the case when the surrogate model assumptions are violated or the design $\boldsymbol{X}$ is highly non-elliptically distributed.

One of the main challenges in our theoretical analysis comes from the colinearity of covariates $(1, S_i, \boldsymbol{X}_i^\intercal \widehat{\boldsymbol{\alpha}}, \boldsymbol{X}_i^\intercal)^\intercal$ due to the enrollment of ρ to leverage the prior information in $\widehat{\boldsymbol{\alpha}}$. We overcome this by properly constructing the oracle coefficients $\boldsymbol{\theta}^*$ and the restricted eigenvalue assumption (A6). The formulation of our problem falls into the missing data framework with missing completely at random. However, the missing probability approaches 1 as $N \to \infty$. This together with the high dimensionality of $\boldsymbol{X}$ makes the theoretical justifications more challenging than those used in the standard missing data literature. Without prior assumptions of $\boldsymbol{\beta}_0 - \rho\boldsymbol{\alpha}_0$ being sparse in certain sense, the unlabeled dataset cannot directly contribute to the estimation of $\boldsymbol{\beta}_0$. Our proposed PASS procedure hinges on the sparsity of $\boldsymbol{\beta}_0 - \rho\boldsymbol{\alpha}_0$ to leverage the unlabeled dataset.

We have restricted the discussion to a single surrogate variable for simplicity. However, the proposed method can be easily extended to multiple surrogates. Specifically, consider

K surrogates, denoted by $S^{[1]}, \ldots, S^{[K]}$. Let $\widehat{\boldsymbol{\alpha}}^{[k]}$ be the ALASSO estimator regressing $S_i^{[k]}$ against $\boldsymbol{X}_i$, $\widehat{\mathcal{A}} = \cup_{k=1}^K \mathrm{supp}(\widehat{\boldsymbol{\alpha}}_k)$, $\boldsymbol{S}_i = (S_i^{[1]}, \ldots, S_i^{[K]})^\mathsf{T}$ and $\boldsymbol{\rho} = (\rho^{[1]}, \ldots, \rho^{[K]})^\mathsf{T}$. We can obtain an estimator for the model parameters as

$$\widehat{\zeta}, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\rho}}, \widehat{\boldsymbol{\beta}} = \underset{\zeta, \boldsymbol{\gamma}, \boldsymbol{\rho}, \boldsymbol{\beta}}{\arg\min}\, n^{-1} \sum_{i=1}^n \ell(Y_i, \zeta + \boldsymbol{S}_i^\mathsf{T}\boldsymbol{\gamma} + \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}) + \lambda_1 \|(\boldsymbol{\beta} - \textstyle\sum_k \rho_k \widehat{\boldsymbol{\alpha}}_k)_{\widehat{\mathcal{A}}}\|_1 + \lambda_2 \|\boldsymbol{\beta}_{\widehat{\mathcal{A}}^c}\|_1.$$

Theoretical justification and finite sample performance of $\widehat{\boldsymbol{\beta}}$ under this setting warrant further research. In our numerical studies, we only focus on fully simulated datasets and real examples. We are further interested in investigating the performance of our approach through semi-synthetic experiments with various setups for the surrogate variables. In addition, it may be interesting to extend the semi-supervised PASS estimator under a high dimensional sparse parametric regression to semi-parametric settings such as the sparse additive model (Ravikumar et al., 2009) and the sparse varying coefficient model (Noh and Park, 2010). Under semi-parametric models, one could still leverage prior information through shrinking the coefficients to "$\rho\widehat{\boldsymbol{\alpha}}$" with some sparse penalty function, to gain statistical efficiency. Studying the specific forms and theoretical properties of such approaches via a semi-supervised framework warrants future research.

R codes for implementing PASS and the benchmark methods, and replicating the simulation results can be found at `https://github.com/moleibobliu/PASS`.

References

Vibhu Agarwal, Tanya Podchiyska, Juan M Banda, Veena Goel, Tiffany I Leung, Evan P Minty, Timothy E Sweeney, Elsie Gyang, and Nigam H Shah. Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association*, 23(6):1166–1173, 2016.

Ashwin N Ananthakrishnan, Su-Chun Cheng, Tianxi Cai, Andrew Cagan, Vivian S Gainer, Peter Szolovits, Stanley Y Shaw, Susanne Churchill, Elizabeth W Karlson, Shawn N Murphy, et al. Association between reduced plasma 25-hydroxy vitamin d and increased risk of cancer in patients with inflammatory bowel diseases. *Clinical Gastroenterology and Hepatology*, 12(5):821–827, 2014.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 08 2009. doi: 10.1214/08-AOS620. URL `http://dx.doi.org/10.1214/08-AOS620`.

John S Brownstein, Shawn N Murphy, Allison B Goldfine, Richard W Grant, Margarita Sordo, Vivian Gainer, Judith A Colecchi, Anil Dubey, David M Nathan, John P Glaser, et al. Rapid identification of myocardial infarction risk associated with diabetes medications using electronic medical records. *Diabetes care*, 33(3):526–531, 2010.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

Tianxi Cai, Yichi Zhang, Yuk-Lam Ho, Nicholas Link, Jiehuan Sun, Jie Huang, Tianrun A. Cai, Scott Damrauer, Yuri Ahuja, Jacqueline Honerlaw, Jie Huang, Lauren Costa, Petra

Schubert, Chuan Hong, David Gagnon, Yan V. Sun, J. Michael Gaziano, Peter Wilson, Kelly Cho, Philip Tsao, Christopher J. O'Donnell, Katherine P. Liao, and for the VA Million Veteran Program. Association of Interleukin 6 Receptor Variant With Cardiovascular Disease Effects of Interleukin 6 Receptor Blocking Therapy: A Phenome-Wide Association Study. *JAMA Cardiology*, 3(9):849–857, 09 2018. ISSN 2380-6583. doi: 10.1001/jamacardio.2018.2287. URL `https://doi.org/10.1001/jamacardio.2018.2287`.

Robert J Carroll, Will K Thompson, Anne E Eyler, Arthur M Mandelin, Tianxi Cai, Raquel M Zink, Jennifer A Pacheco, Chad S Boomershine, Thomas A Lasko, Hua Xu, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*, 19(e1):e162–e169, 2012.

Abhishek Chakrabortty, Matey Neykov, Raymond Carroll, and Tianxi Cai. Surrogate aided unsupervised recovery of sparse signals in single index models for binary outcomes. *arXiv preprint 1701.05230*, 2017.

Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210, 2010.

Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12(3):793–815, 09 1984. doi: 10.1214/aos/1176346703. URL `http://dx.doi.org/10.1214/aos/1176346703`.

Finale Doshi-Velez, Yaorong Ge, and Isaac Kohane. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*, 133(1):e54–e63, 2014.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. ISSN 1548-7660. doi: 10.18637/jss.v033.i01. URL `https://www.jstatsoft.org/index.php/jss/article/view/v033i01`.

Omri Gottesman, Helena Kuivaniemi, Gerard Tromp, W Andrew Faucett, Rongling Li, Teri A Manolio, Saskia C Sanderson, Joseph Kannry, Randi Zinberg, Melissa A Basford, et al. The electronic medical records and genomics (emerge) network: past, present, and future. *Genetics in Medicine*, 15(10):761–771, 2013.

Peter Hall and Ker-Chau Li. On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, 21(2):867–889, 06 1993. doi: 10.1214/aos/1176349155. URL `http://dx.doi.org/10.1214/aos/1176349155`.

Yoni Halpern, Steven Horng, Youngduck Choi, and David Sontag. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association*, 23(4):731–740, 2016.

Chuan Hong, Katherine P Liao, and Tianxi Cai. Semi-supervised validation of multiple surrogate outcomes with application to electronic medical records phenotyping. *Biometrics*, 75(1):78–89, 2019.

Yuan Jiang, Yunxiao He, and Heping Zhang. Variable selection with prior information for generalized linear models via the prior LASSO method. *Journal of the American Statistical Association*, 111(513):355–376, 2016. doi: 10.1080/01621459.2015.1008363.

Isaac S Kohane. Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics*, 12(6):417–428, 2011.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *arXiv preprint arXiv:2008.01064*, 2020.

Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *The Annals of Statistics*, 17(3):1009–1052, 09 1989. doi: 10.1214/aos/1176347254. URL `http://dx.doi.org/10.1214/aos/1176347254`.

Katherine P Liao, Dorothée Diogo, Jing Cui, Tianxi Cai, Yukinori Okada, Vivian S Gainer, Shawn N Murphy, Namrata Gupta, Daniel Mirel, Ashwin N Ananthakrishnan, et al. Association between low density lipoprotein and rheumatoid arthritis genetic factors with low density lipoprotein levels in rheumatoid arthritis and non-rheumatoid arthritis controls. *Annals of the rheumatic diseases*, 73(6):1170–1175, 2014.

Katherine P Liao, Tianxi Cai, Guergana K Savova, Shawn N Murphy, Elizabeth W Karlson, Ashwin N Ananthakrishnan, Vivian S Gainer, Stanley Y Shaw, Zongqi Xia, Peter Szolovits, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *bmj*, 350:h1885, 2015.

Matthew McDermott, Tom Yan, Tristan Naumann, Nathan Hunt, Harini Suresh, Peter Szolovits, and Marzyeh Ghassemi. Semi-supervised biomedical translation with cycle wasserstein regression gans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in neural information processing systems*, pages 1348–1356, 2009.

Katherine M Newton, Peggy L Peissig, Abel Ngo Kho, Suzette J Bielinski, Richard L Berg, Vidhu Choudhary, Melissa Basford, Christopher G Chute, Iftikhar J Kullo, Rongling Li, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the emerge network. *Journal of the American Medical Informatics Association*, 20(e1):e147–e154, 2013.

Hoh Suk Noh and Byeong U Park. Sparse varying coefficient models for longitudinal data. *Statistica Sinica*, pages 1183–1202, 2010.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of*

the VLDB Endowment. International Conference on Very Large Data Bases, volume 11, page 269. NIH Public Access, 2017.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

Sara A. van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 04 2008. doi: 10.1214/009053607000000929. URL `http://dx.doi.org/10.1214/009053607000000929`.

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009. doi: 10.1214/09-EJS506. URL `https://doi.org/10.1214/09-EJS506`.

Hai Wang and Hoifung Poon. Deep probabilistic logic: A unifying framework for indirect supervision. *arXiv preprint arXiv:1808.08485*, 2018.

RA Wilke, H Xu, JC Denny, DM Roden, RM Krauss, CA McCarty, RL Davis, Todd Skaar, J Lamba, and G Savova. The emerging role of electronic medical records in pharmacogenomics. *Clinical Pharmacology & Therapeutics*, 89(3):379–386, 2011.

Sheng Yu, Katherine P Liao, Stanley Y Shaw, Vivian S Gainer, Susanne E Churchill, Peter Szolovits, Shawn N Murphy, Isaac S Kohane, and Tianxi Cai. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *Journal of the American Medical Informatics Association*, 22(5): 993–1000, 2015.

Lingjiao Zhang, Xiruo Ding, Yanyuan Ma, Naveen Muthu, Imran Ajmal, Jason H Moore, Daniel S Herman, and Jinbo Chen. A maximum likelihood approach to electronic health record phenotyping using positive and unlabeled patients. *Journal of the American Medical Informatics Association*, 27(1):119–126, 2020.

Yichi Zhang, Tianrun Cai, Sheng Yu, Kelly Cho, Chuan Hong, Jiehuan Sun, Jie Huang, Yuk-Lam Ho, Ashwin N Ananthakrishnan, Zongqi Xia, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (phecap). *Nature protocols*, 14(12):3426–3444, 2019.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. doi: 10.1198/016214506000000735.

Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733–1751, 08 2009. doi: 10.1214/08-AOS625. URL `http://dx.doi.org/10.1214/08-AOS625`.