Semi-supervised logistic learning based on exponential tilt mixture models

Xinwei Zhang | Zhiqiang Tan

Consider semi-supervised learning for classification, where both labelled and unlabelled data are available for training. The goal is to exploit both datasets to achieve higher prediction accuracy than just using labelled data alone. We develop a semi-supervised logistic learning method based on exponential tilt mixture models by extending a statistical equivalence between logistic regression and exponential tilt modelling. We study maximum nonparametric likelihood estimation and derive novel objective functions that are shown to be Fisher probability consistent. We also propose regularized estimation and construct simple and highly interpretable expectation–maximization (EM) algorithms. Finally, we present numerical results that demonstrate the advantage of the proposed methods compared with existing methods.

KEYWORDS
empirical likelihood, expectation–maximization algorithm, exponential tilt model, Fisher consistency, logistic regression, semi-supervised learning

1 INTRODUCTION

Semi-supervised learning for classification involves exploiting a large amount of unlabelled data and a relatively small amount of labelled data to build better classifiers. This approach can potentially be used to achieve higher accuracy, with a limited budget for obtaining labelled data. Various methods have been proposed, including expectation–maximization (EM) algorithms, transductive support vector machines (TSVMs), and regularized methods (e.g. Chapelle, Schölkopf, & Zien, 2006; Zhu, 2005). Typically, semi-supervised methods are developed by invoking a cluster assumption that features (or covariates) within different classes are distributed in clusters, separable by low-density regions. Variations of this assumption are also known as smoothness or manifold assumptions (Chapelle et al., 2006, Section 1.2).

For supervised classification, there are various objective (or loss) functions which are Fisher consistent in the following sense: optimization of the population and nonparametric version of a loss function leads to the true conditional probability function of labels given features as for the logistic loss or to the Bayes classifier as for the hinge loss (Bartlett, Jordan, & McAuliffe, 2006; Lin, 2002). For distinction if needed, the two properties are referred to as Fisher probability consistency or sign-consistency. To handle finite sample data, a common statistical paradigm is to use a penalized objective function, defined by adding a regularization term to a Fisher consistent loss.

A subtle issue that we notice for semi-supervised classification is that Fisher consistency seems to be lacking for various existing objective functions. Examples include the objective functions in TSVMs (Vapnik, 1998; Joachims, 1999) and various regularized methods (Belkin, Niyogi, & Sindhwani, 2006; Grandvalet & Bengio, 2005; Krishnapuram et al., 2005; Mann & McCallum, 2007). See Data S1 Section III for objective functions of the methods studied in our numerical experiments. In fact, a typical objective function from these semi-supervised methods can be decomposed into two parts. The first part is a supervised part, based on labelled data only and associated with a functional complexity penalty (such as the norm in a reproducing kernel Hilbert space). The second part is an unsupervised part, based on unlabelled data and possibly labelled data and associated with a semi-supervised tuning parameter which controls the balance between the first and second parts. Then Fisher consistency (or, more specifically, infinite-sample consistency) can be stated as follows: for any fixed semi-supervised tuning parameter, minimization of the two parts of the objective function (excluding any functional complexity penalty) based on population data would yield the true functions of interest. This property seems to be in general not satisfied by the aforementioned methods, unless the semi-supervised tuning parameter is set to 0 and the objective function reduces to the supervised part only. Accordingly, large-sample consistency of these methods can be shown as the semi-supervised tuning parameter tends to 0 under suitable conditions, which is analogous to consistency of supervised nonparametric regression with a smoothing parameter tending to 0. Here, infinite-sample consistency and large-sample consistency are related in a similar manner as in the
classical setting for maximum likelihood estimation where infinite-sample consistency means that minimization of the Kullback–Leibler divergence leads to the true data-generating distribution and is a prerequisite for large-sample consistency. While existing semi-supervised methods may potentially be tuned to achieve good performance, the lack of Fisher consistency may contribute to unstable performances (e.g. Li & Zhou, 2014).

Another restriction in existing methods is that the class proportions in labelled and unlabelled data are typically assumed to be the same. For example, the TSVMs (Joachims, 1999) involves an explicit constraint that the predicted class proportions in unlabelled data are identical to the observed class proportions in labelled data. When this assumption fails, the performance can deteriorate considerably (Li & Zhou, 2014). Even when the assumption holds, such a constraint can be undesirable because labelled data are typically of limited sizes.

We develop a semi-supervised extension of logistic regression based on exponential tilt mixture models (Qin, 1999; Tan, 2009; Zou, Fine, & Yandell, 2002), which requires a functional specification of the density ratio of class-specific feature distributions but without invoking a cluster assumption. Moreover, the class proportions in the unlabelled data are allowed to be different from those in the labelled data. Parenthetically, our approach can also be combined with smoothness assumptions, and such investigation is of interest for future work.

A basic motivation for our approach is the statistical equivalence between logistic regression for the conditional probability of a label given features and exponential tilt modelling for the density ratio between the feature distributions within different classes (Anderson, 1972; Prentice & Pyke, 1979). Our work involves two main contributions: (i) we derive novel objective functions which are shown not only to be Fisher probability consistent but also to lead to asymptotically more efficient estimation than based on labelled data only, and (ii) we propose regularized estimation and construct computationally and conceptually desirable EM algorithms. From numerical experiments, our methods achieve a substantial advantage over existing methods when the class proportions in unlabelled data differ from those in labelled data. A possible explanation is that while the class proportions in unlabelled data are estimated as unknown parameters in our methods, they are implicitly assumed to be the same as in labelled data for existing methods including TSVMs (Joachims, 1999) and entropy regularization (Grandvalet & Bengio, 2005).

2 | ILLUSTRATION

In this section, we provide some illustration in a simple bivariate Gaussian setting to highlight the comparison between the proposed semi-supervised learning method and the supervised counterpart, logistic regression. Here, we focus on comparisons when class proportions in labelled and unlabelled data are different and the classification target is unlabelled data or out-of-sample data from the same mixture distribution as unlabelled data. See Data S1 (Section I) for results when class proportions in labelled and unlabelled data are the same.

First, we compare decision boundaries from supervised logistic regression and our method based on a random dataset, as shown in Figure 1. The dataset consists of 90 labelled data, 30 from $G_0$ and 60 from $G_1$, and 750 unlabelled data, 500 from $G_0$ and 250 from $G_1$, where $G_0$ and $G_1$ are bivariate Gaussian with mean $(-6, -6)$ and diagonal variance matrix $\begin{pmatrix} 5^2 & 15^2 \\ 15^2 & 5^2 \end{pmatrix}$ and mean $(6, 6)$ and diagonal variance matrix $\begin{pmatrix} 5^2 & 0^2 \\ 0^2 & 5^2 \end{pmatrix}$. In this case, the class proportions in labelled data, 1:2, differ from those in unlabelled data, 2:1. The dot-dash line and dash line are the decision boundaries from logistic regression and our method, respectively. As a reference, the oracle decision boundary is estimated by logistic regression based on an independent labelled sample of size 15,000 with the same class proportions, 2:1, as in unlabelled data.

As seen from Figure 1, the dash line from our method is closer to the oracle line in both intercept and slope, which confirms Fisher probability consistency for our method. On the other hand, in spite of a relatively large labelled size, the dot-dash line from logistic regression is noticeably different from the oracle line. This shows that our method can exploit unlabelled data together with labelled data to achieve a better classification on the unlabelled data.

In Figure 2, we provide a heat map to compare out-of-sample prediction accuracy between our method (penalized version) and ridge logistic regression, over a grid of labelled data sizes and separation degrees, measured by the prediction accuracy from oracle logistic regression based on independent labelled samples as in Figure 1. See Data S1 for further details. The two-class distributions are bivariate Gaussian as in Figure 1.
but with means \((- \mu, \mu)\) and \((\mu, \mu)\) varied to achieve different separation degrees. The class proportions in labelled data are fixed at the ratio 1:2, and the unlabelled data are drawn with a fixed size 750 and class proportions 2:1, similarly as in the Flip Prop scheme for numerical experiments in Section 6.

Figure 2 shows that our method can achieve improvement (positive difference) over ridge logistic regression in prediction accuracy up to 15%. The improvement becomes larger as the labelled size increases. This is because with more labelled data, our method can estimate the class proportions in unlabelled data as well as the regression coefficients more accurately to yield better classification results. Moreover, the improvement by our method is greater (or smaller) when the separation degree is lower (or higher). This indicates that there is more benefit in using our method to exploit unlabelled data when the two-class data are less separable.

3 | BACKGROUND: LOGISTIC REGRESSION AND EXPONENTIAL TILT MODEL

For supervised classification, the training data consist of a sample \(\{(y_i, x_i): i = 1, \ldots, n\}\) of \((y, x)\), where \(x \in \mathbb{R}^p\) and \(y \in \{0, 1\}\) represent a feature vector and an associated label, respectively. Consider a logistic regression model

\[
P(y = 1|x) = \frac{\exp(\beta_0^c + \beta_1^c x)}{1 + \exp(\beta_0^c + \beta_1^c x)},
\]

where \(\beta_i\) is a coefficient vector associated with \(x\) and \(\beta_0^c\) is an intercept, with superscript \(c\) indicating classification or conditional probability of \(y = 1\) given \(x\). The maximum likelihood estimator \((\hat{\beta}_0^c, \hat{\beta}_1^c)\) is defined as a maximizer of the log (conditional) likelihood:

\[
\sum_{i=1}^n \left[ y_i (\beta_0^c + \beta_1^c x_i) - \log(1 + \exp(\beta_0^c + \beta_1^c x_i)) \right].
\]

In general, nonlinear functions of \(x\) can be used in place of \(\beta_1^c x\), and a penalty term can be incorporated into the log-likelihood such as the ridge penalty \(\|\beta_1\|^2\) or the squared norm of a reproducing kernel Hilbert space of functions of \(x\). See further discussion in Sections 4.3 and 7.

Interestingly, logistic regression on \(P(y|x)\) can be made equivalent to an exponential tilt model on \(P(x|y)\) (Anderson, 1972; Prentice & Pyke, 1979; Qin, 1998). Denote by \(G_0\) or \(G_1\) the conditional distribution \(P(x|y = 0)\) or \(P(x|y = 1)\), respectively, and \(\alpha = P(y = 1)\). By the Bayes rule, Model (1) is equivalent to the exponential tilt model

\[
\frac{dP(y = 1|x)}{dP(y = 0|x)} = \frac{dG_1}{dG_0} = e^{\hat{\beta}_0^c + \hat{\beta}_1^c x},
\]

where \(dG_1/dG_0\) denotes the density ratio between \(G_1\) and \(G_0\) with respect to a dominating measure, and \(\beta_0 = \beta_0^c + \log((1 - \alpha)/\alpha)\). Model (3) is explicitly a semi-parametric model, where \(G_0\) is an infinitely dimensional parameter and \((\beta_0, \beta_1)\) are finitely dimensional parameters. In fact, logistic model (1) is also semi-parametric, where the marginal distribution of \(x\) is an infinitely dimensional parameter, and \((\beta_0, \beta_1)\) are finitely dimensional parameters. Furthermore, the maximum likelihood estimator \((\beta_0, \beta_1)\) in Model (1) can be related to the following estimator \((\hat{\beta}_0, \hat{\beta}_1)\) in Model (3) by the method of nonparametric likelihood (Kiefer & Wolfowitz, 1956) or empirical likelihood (Owen, 2001). Formally, \((\beta_0, \beta_1, \hat{G}_0)\) are defined as a maximizer of the log-likelihood,

\[
\sum_{i=1}^n \left[(1 - y_i) + y_i (\beta_0 + \beta_1^c x_i) + \log G_0((x_i)) \right],
\]

over all possible \((\beta_0, \beta_1, G_0)\) such that \(G_0\) is a probability measure supported on the pooled data \(\{x_i: i = 1, \ldots, n\}\) with \(\int \exp(\beta_0 + \beta_1^c x) dG_0 = 1\).

Analytically, it can be shown that \(\beta_1 = \hat{\beta}_1, \beta_0^c = \beta_0 + \log(\hat{\delta}/(1 - \hat{\delta}))\), where \(\hat{\delta} = \sum_{i=1}^n y_i/n\). See Qin (1998) and references therein.
By the foregoing discussion, we see that there are two statistically distinct but equivalent approaches for supervised classification: logistic regression or exponential tilt models. It is such a relationship that we aim to exploit in developing a new method for semi-supervised classification.

4 | THEORY AND METHODS

For semi-supervised classification, the training data consist of labelled data \( S^\prime = \{(x, y) : i = 1, \ldots, n\} \) and unlabelled data \( S^u = \{x : i = n+1, \ldots, N\} \), for which the associated labels \( \{y : i = n+1, \ldots, N\} \) are unobserved. For clarity, it is helpful to denote \( S^{u*} = \{(x, y) : i = n+1, \ldots, N\} \), as the full data underlying the unlabelled data. Typically for existing methods including TSVMs, the two samples \( S^\prime \) and \( S^{u*} \) are assumed to be from a common population of \((y, x)\). However, we only require that \( S^\prime \) and \( S^{u*} \) be drawn from two joint distributions, with the same conditional distributions of \( x \) given \( y \) but possibly different marginal probabilities of \( y \).

**Assumption 1.** Assume that \( S^\prime \) and \( S^{u*} \) are drawn from joint distributions \( P^\prime(x, y) \) and \( P^{u*}(x, y) \), respectively, such that the conditional distributions \( P^\prime(x|y) \) and \( P^{u*}(x|y) \) are identical, denoted as \( P(x|y) \), whereas the marginal probabilities \( P^\prime(y = 1) \) and \( P^{u*}(y = 1) \) may be different.

4.1 | Exponential tilt mixture model

Although it seems difficult at first look to extend logistic model (1) for semi-supervised learning, we realize that both the labelled data \( S^\prime \) and the unlabelled data \( S^u \) can be taken of by a natural extension of the exponential tilt model (3), called an exponential tilt mixture model (Qin, 1999; Tan, 2009; Zou et al., 2002). Denote

\[
S_1 = \{x_i : y_i = 0, i = 1, \ldots, n\} \text{ drawn from } P_1(x) = P(x|y = 0),
\]

\[
S_2 = \{x_i : y_i = 1, i = 1, \ldots, n\} \text{ drawn from } P_2(x) = P(x|y = 1),
\]

\[
S_3 = \{x_i : i = n+1, \ldots, N\} \text{ drawn from } P_3(x) = P^u(x).
\]

An exponential tilt mixture model for the three samples \((S_1, S_2, S_3)\) postulates that

\[
dP_1(x) = dG_0(x), \quad \text{(5)}
\]

\[
dP_2(x) = dG_1(x), \quad \text{(6)}
\]

\[
dP_3(x) = (1 - \rho)dG_0(x) + \rho dG_1(x), \quad \text{(7)}
\]

where \( G_0 \) or \( G_1 \) represents the conditional distribution of \( x \) given \( y = 0 \) or \( y = 1 \), respectively, in both the labelled and unlabelled data such that

\[
\frac{dG_1}{dG_0} = e^{\theta_1 x^T}, \quad \text{(8)}
\]

and \( \rho = P^\prime(y = 1) \) is the proportion of \( y = 1 \) underlying the unlabelled data. These equations can be understood as follows. Equations (5)–(6) merely give definitions of \( G_0 \) and \( G_1 \). Equation (7) says that the feature distribution in the unlabelled sample is a mixture of \( G_0 \) and \( G_1 \), which is a direct consequence of the structural restriction stated as Assumption 1, that is, the conditional distribution \( P(x|y) \) is invariant between the labelled and unlabelled samples. Equation (8) imposes a functional restriction on the density ratio between \( G_0 \) and \( G_1 \), similarly as in (3).

The exponential tilt mixture model, defined by (5–8), is a semi-parametric model, with an infinitely dimensional parameter \( \rho \) and \( \beta = (\beta_0, \beta_1)^T \). We briefly summarize maximum nonparametric likelihood estimation previously studied (Qin, 1999; Zou et al., 2002; Tan, 2009). For notational convenience, rewrite the sample \( S_1 \) as \( \{x_i : i = 1, \ldots, n_1\} \), where \( n_1 = n - n_2 \), \( n_2 = \sum_{i=1}^{n} y_i \), and \( n_3 = N - n \). Equations (5)–(7) can be expressed as

\[
dP_j = (1 - \rho_j)dG_0 + \rho_j dG_1, \quad j = 1, 2, 3,
\]

where \( \rho_1 = 0, \rho_2 = 1, \) and \( \rho_3 = \rho \). For any fixed \( (\rho, \beta) \), the average profile log-likelihood of \( (\rho, \beta) \) is defined as \( pl(\rho, \beta) = \max_{G_0} l(\rho, \beta, G_0) \) with

\[
l(\rho, \beta, G_0) = \frac{1}{N} \sum_{j=1}^{n} \sum_{i=1}^{n_j} \log \left[ 1 - \rho + \rho \exp(\beta_0 + \beta_1^T x_i) \right] + \log P_j(x_i), \quad \text{(9)}
\]

over all possible \( G_0 \), which is a probability measure supported on the pooled data \( \{x_j : i = 1, \ldots, n_j, j = 1, 2, 3\} \) with \( \int \exp(\beta_0 + \beta_1^T x) dG_0 = 1 \). Denote

\[
k(\rho, \beta, a) = \frac{1}{N} \sum_{j=1}^{n} \sum_{i=1}^{n_j} \log \left[ \frac{1 - \rho + \rho \exp(\beta_0 + \beta_1^T x_i)}{1 - a + a \exp(\beta_0 + \beta_1^T x_i)} \right] - \log N,
\]

which can be easily shown to be concave in \( \rho \in (0, 1) \) and convex in \( a \in (0, 1) \). Then Proposition 1 in (Tan, 2009) leads to the following result.
**Lemma 1.** The average profile log-likelihood of \((\rho, \beta)\) can be determined as \(p(\rho, \beta) = \min_{\alpha \in (0,1]} \kappa(\rho, \beta, \alpha) = \kappa(\rho, \beta, \hat{\alpha}(\beta))\), where \(\hat{\alpha}(\beta)\) is a minimizer of \(\kappa(\rho, \beta, \alpha)\) over \(\alpha\), satisfying the stationary condition (free of \(\rho\))

\[
1 = \frac{1}{N} \sum_{j=1}^{3} \sum_{i=1}^{n_j} \frac{1}{1 - \alpha + \alpha \exp(\beta_0 + \beta_1^T x_j)}.
\]  

(10)

The maximum likelihood estimator of \((\rho, \beta)\) is then defined by maximizing the profile log-likelihood, that is, \((\hat{\rho}, \hat{\beta}) = \arg\max_{\rho, \beta} p(\rho, \beta)\). From Lemma 1, we notice that the estimators \((\hat{\rho}, \hat{\beta}, \hat{\alpha}(\beta))\) jointly solve the saddle-point problem:

\[
\max_{(\rho, \beta)} \min_{\alpha} \kappa(\rho, \beta, \alpha).
\]  

(11)

Large-sample theory of \((\hat{\rho}, \hat{\beta})\) has been studied in Qin (1999) under standard regularity conditions as \(N \to \infty\) and \(n_j/N \to \eta_j\) with some constant \(\eta_j > 0\) for \(j = 1, 2, 3\). The theory shows the existence of a local maximizer of \(p(\rho, \beta)\), which is consistent and asymptotically normal provided the exponential tilt mixture model (5–8) is correctly specified. However, there remain subtle questions. It seems unclear whether the population version of the average profile log-likelihood \(\kappa(\rho, \beta, \hat{\alpha}(\beta))\) attains a global maximum at the true values of \((\rho, \beta)\) under a correctly specified exponential tilt mixture model. Moreover, what property can be deduced for \((\hat{\rho}, \hat{\beta})\) under a misspecified exponential tilt mixture model?

### 4.2 Semi-supervised logistic regression

We derive a new classification model with parameters \((\rho, \beta)\) for the three samples \((S_1, S_2, S_3)\) such that a maximum likelihood estimator of \((\rho, \beta)\) in the new model coincides with a maximum likelihood estimator \((\hat{\rho}, \hat{\beta})\) in the exponential tilt mixture model, and vice versa. Let \(z_i = 1 + y_i\) if \(i = 1, \ldots, n\) and \(z_i = 3\) if \(i = n + 1, \ldots, N\). Consider a conditional probability model for predicting the label \(z_i\) from \(x_i\):

\[
P(z = 1|x) = \frac{n_1}{N} \frac{1-\rho + \rho \exp(\beta_0 + \beta_1^T x_i)}{1 - \rho + \rho \exp(\beta_0 + \beta_1^T x_i)},
\]  

(12)

\[
P(z = 2|x) = \frac{n_2}{N} \frac{\exp(\beta_0 + \beta_1^T x_i)}{1 - \rho + \rho \exp(\beta_0 + \beta_1^T x_i)},
\]  

(13)

\[
P(z = 3|x) = \frac{n_3}{N} \frac{1 - \rho + \rho \exp(\beta_0 + \beta_1^T x_i)}{1 - \rho + \rho \exp(\beta_0 + \beta_1^T x_i)},
\]  

(14)

where \(\hat{\alpha}(\rho) = \sum_{j=1}^{3} (n_j/N)\rho_j = (n_2 + n_3\rho)/N\), which ensures that \(\sum_{i=1}^{3} P(z = j|x) = 1\). The model, defined by (12–14), will be called a semi-supervised logistic regression model. The average log-likelihood function of \((\rho, \beta)\) with the data \([z_i, x_i]: i = 1, \ldots, N\) in Models (12–14) can be written, up to an additive constant free of \((\rho, \beta)\), as

\[
\kappa(\rho, \beta, \hat{\alpha}(\rho)) = \frac{1}{N} \sum_{j=1}^{3} \sum_{i=1}^{n_j} \log \left\{ \frac{1 - \rho + \rho \exp(\beta_0 + \beta_1^T x_j)}{1 - \rho + \rho \exp(\beta_0 + \beta_1^T x_j)} \right\} - \log N.
\]

**Proposition 1.** If and only if \((\hat{\rho}, \hat{\beta})\) is a local (or global) maximizer of the average log-likelihood \(\kappa(\rho, \beta, \hat{\alpha}(\rho))\) in semi-supervised logistic regression model (12–14), then it is a local (or respectively global) maximizer of the average profile log-likelihood \(\kappa(\rho, \beta, \hat{\alpha}(\beta))\) in exponential tilt mixture model (5–8).

Proposition 1 shows an equivalence between maximum nonparametric likelihood estimation in exponential tilt mixture model (5–8) and usual maximum likelihood estimation in semi-supervised logistic regression model (12–14), even though the objective functions \(\kappa(\rho, \beta, \hat{\alpha}(\beta))\) and \(\kappa(\rho, \beta, \hat{\alpha}(\rho))\) are not equivalent. This differs from the equivalence between logistic regression (1) and exponential tilt model (3) with labelled data only, where the log-likelihood (2) and the profile log-likelihood from (4) are equivalent (Prentice & Pyke, 1979). From another angle, this result says that saddle-point problem (11) can be equivalently solved by directly maximizing \(\kappa(\rho, \beta, \hat{\alpha}(\rho))\). This transformation is nontrivial, because a saddle-point problem in general cannot be converted into optimization with a closed-form objective.

By the identification of \(\kappa(\rho, \beta, \hat{\alpha}(\rho))\) as a usual log-likelihood function, we show that the objective functions \(\kappa(\rho, \beta, \hat{\alpha}(\beta))\) and \(\kappa(\rho, \beta, \hat{\alpha}(\rho))\), with the linear predictor \(\beta_0 + \beta_1^T x\) replaced by an arbitrary function \(h(x)\), are Fisher probability consistent; i.e. maximization of the population versions of the objective functions leads to the true log density ratio function \(h'(x) = \log(dG_1/dG_0)\). This seems to be the first time Fisher consistency of a loss function is established for semi-supervised classification. By some abuse of notation, denote

\[
\kappa(\rho, h, a) = \frac{1}{N} \sum_{j=1}^{3} \sum_{i=1}^{n_j} \log \left\{ \frac{1 - \rho + \rho \exp(h(x_j))}{1 - a + a \exp(h(x_j))} \right\} - \log N.
\]
Proposition 2. Suppose that $S_j = (x_i : i = 1, \ldots, n_j)$ is drawn from $P_j$ in (5)-(7) for $j = 1, 2, 3$, with $\rho = \rho^*$ and $dG_j/dG_0 = \exp(h^*(x))$ for some fixed value $\rho^* \in (0, 1)$ and function $h^*(x)$. Denote $\kappa^*(\rho, h, a) = E[\kappa(\rho, h, a)]$. For any $\rho \in (0, 1)$ and function $h(x)$, we have

$$
\min_{\hat{\rho}, \hat{\beta}, \hat{\lambda}} \kappa^*(\rho, h, \hat{a}(\rho)) \leq \kappa^*\{\rho^*, h^*, \hat{a}(\rho^*)\},
$$

where both equalities hold if $\rho = \rho^*$ and $h = h^*$. Hence, the population objective functions $\kappa^*\{\rho, h, \hat{a}(\rho)\}$ and $\min_{\rho \in (0, 1)} \kappa^*\{\rho, h, a\}$ are maximized at the true value $\rho^*$ and function $h^*(x)$.

Proposition 2 fills existing gaps in understanding maximum likelihood estimation in exponential tilt mixture model (5–8), through its equivalence with that in semi-supervised model (12–14). If the exponential tilt mixture model is correctly specified, then the population version of $\kappa\{\rho, \beta, \hat{a}(\beta)\}$ has a global maximum at the true values of $(\rho, \beta)$, and hence a global maximizer $(\hat{\beta}, \hat{\rho})$ is consistent under suitable regularity conditions. If the exponential tilt mixture model is misspecified, then by theory of estimation with misspecified models (Manski, 1988; White, 1982), the maximum likelihood estimator $(\hat{\beta}, \hat{\rho})$ converges in probability to a limit value which minimizes the difference between $\kappa^*\{\rho, h, \hat{a}(\rho)\}$ with $h = h_0 + \beta_0^2x$ and $\kappa^*\{\rho^*, h^*, \hat{a}(\rho^*)\}$. This difference, as shown in Data S1 (Section II.2), is the expected Kullback–Leibler divergence.

Proposition 3. Denote by $\hat{\beta}'$ the estimator of $\beta$ obtained by maximizing $\kappa\{\rho', \beta, \hat{a}(\rho')\}$ or equivalently by logistic regression based on the labelled data only. Then the asymptotic variance matrix of the maximum likelihood estimator $\hat{\beta}$ from exponential tilt mixture model (5–8) is no greater (in the usual order on positive-definite matrices) than that of $\hat{\beta}'$ under standard regularity conditions.

### 4.3 Regularized estimation and EM algorithm

The results in Section 4.2 provide theoretical support for the use of the objective functions $\kappa\{\rho, \beta, \hat{a}(\beta)\}$ and $\kappa\{\rho, \beta, \hat{a}(\rho)\}$. In real applications, the maximum likelihood estimator $(\hat{\beta}, \hat{\rho})$ may not behave satisfactorily as predicted by standard asymptotic theory for various reasons. The labelled sample size may not be sufficiently large. The dimension of the feature vector or the complexity of functions of features may be too high, compared with the labelled and unlabelled data sizes. Therefore, we propose regularized estimation by adding suitable penalties to the objective functions. It is important to emphasize that, unlike existing regularized methods (Section 5), our approach follows the usual statistical paradigm of adding a regularization term to a Fisher-consistent loss function, to improve finite-sample performance.

For the coefficient vector $\beta$, we employ a ridge penalty $\lambda∥\beta∥_2^2$, although alternative penalties can also be allowed including a lasso penalty. For the mixture proportion $\rho$, we use a penalty in the form of the log density of a Beta distribution, $\tau_1 \log(1 - \rho) + \tau_2 \log \rho$, where $\tau_1 = \gamma (1 - \rho^0)n_2/N$ and $\tau_2 = \gamma \rho^0n_2/N$ for a 'centre' $\rho^0 \in (0, 1)$ and a 'scale' $\gamma \geq 0$. This choice is motivated by conceptual and computational simplicity in the EM algorithm to be discussed. Combining these penalties with $\kappa\{\rho, \beta, \hat{a}(\beta)\}$ gives the following penalized objective function:

$$
k\{\rho, \beta, \hat{a}(\beta)\} - \lambda∥\beta∥_2^2 + \gamma(1 - \rho^0)(n_2/N) \log(1 - \rho) + \gamma \rho^0(n_2/N) \log \rho.
$$

Similarly, the penalized objective function based on $\kappa\{\rho, \beta, \hat{a}(\rho)\}$ is

$$
k\{\rho, \beta, \hat{a}(\rho)\} - \lambda∥\beta∥_2^2 + \gamma(1 - \rho^0)(n_2/N) \log(1 - \rho) + \gamma \rho^0(n_2/N) \log \rho.
$$

Maximization of (15) or (16) will be called profile or direct semi-supervised logistic regression, respectively. The two methods in general lead to different estimates of $(\rho, \beta)$ when $\gamma > 0$, although they can be shown to be equivalent similarly as in Proposition 1 when $\gamma = 0$. In fact, as $\gamma \to \infty$ (i.e. $\rho$ is fixed as $\rho^0$), the estimator of $\beta$ from profile semi-supervised logistic regression is known to be asymptotically more efficient than from direct semi-supervised logistic regression (Tan, 2009).
We construct EM algorithms (Dempster, Laird, & Rubin, 1977) to numerically maximize (15) and (16). Of particular interest is that these algorithms shed light on the effect of the regularization introduced. Various other optimization techniques can also be exploited, because \( \kappa (\rho, \beta, \alpha (\rho)) \) is directly of a closed form, and \( \kappa (\rho, \beta, \alpha (\rho)) \) is defined only after univariate minimization in \( \alpha \).

We describe some details about the EM algorithm for profile semi-supervised logistic regression. See the Appendix for the corresponding algorithm for direct semi-supervised logistic regression. We return to the nonparametric log-likelihood (9) and introduce the following data augmentation. For \( j = 1, 2, 3 \), let \( u_j \sim \text{Bernoulli} (\rho_j) \) such that \( (x_j | u_j = 0) \sim G_0 \) and \( (x_j | u_j = 1) \sim G_1 \). Recall that \( \rho_1 = 0 \) and \( \rho_2 = 1 \), and hence, \( u_{j1} = 0 \) and \( u_{j2} = 1 \) fixed. Denote the penalty term in (15) or (16) as \( \text{pen}(\rho, \beta) \).

**E-step.** The expectation of the augmented objective given the current estimates \( (\rho^0, \beta^0) \) is

\[
Q^{(1)}(\rho, \beta, G_0) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_j} \left[ (1 - E^{u_j} u_j (1 - \rho)_{G_0} (x_j)) + E^{u_j} (1 - \rho) \exp(\beta_0 + \beta_1 x_j) G_0 (x_j)) \right] + \text{pen}(\rho, \beta),
\]

where \( E^{u_j} u_j = \rho_j^{(1)} \exp(\beta_0 + \beta_1 x_j) / (1 - \rho_j^{(1)} + \rho_j^{(1)} \exp(\beta_0 + \beta_1 x_j)) \).

**M-step.** The next estimates \( (\rho^{t+1}, \beta^{t+1}) \) are obtained as a maximizer of the expected objective (17) with \( G_0 \) profiled out, that is, \( pQ^{(1)}(\rho, \beta) = \max_{G_0} Q^{(1)}(\rho, \beta, G_0) \) over all possible \( G_0 \), which is a probability measure supported on the pooled data \( \{x_j : i = 1, \ldots, n_j, j = 1, 2, 3\} \) with \( \int \exp(\beta_0 + \beta_1 x) dG_0 = 1 \). In correspondence to \( \kappa (\rho, \beta, \alpha) \), denote

\[
\kappa^{(1)}(\rho, \beta, \alpha) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_j} \left[ (1 - E^{u_j} u_j (1 - \rho) \log \left( \frac{1 - \rho_j}{1 - \rho_j + \alpha \exp(\beta_0 + \beta_1 x_j)} \right) + \text{pen}(\rho, \beta) \right].
\]

Instead of maximizing \( pQ^{(1)}(\rho, \beta) \) directly, we find a simple scheme for computing \( (\rho^{t+1}, \beta^{t+1}) \).

**Proposition 4.** Let

\[
\rho^{(t+1)} = \frac{n_3^{(t)} \sum_{i=1}^{n_3} E^{u_j} u_j + \gamma \rho^{(t)}}{1 + \gamma} \quad \text{and} \quad \beta^{(t+1)} = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_j} E^{u_j} u_j.
\]

If and only if \( \beta^{(t+1)} \) is a local (or global) maximizer of \( \kappa^{(1)}(\beta^{(t+1)}, \alpha^{(t+1)}) \), then \( (\rho^{(t+1)}, \beta^{(t+1)}) \) is a local (or respectively global) maximizer of \( pQ^{(1)}(\rho, \beta) \).

Proposition 4 is useful both computationally and conceptually. First, \( \beta^{(t+1)} \) is of a closed form, as a weighted average, with the weight depending on the scale \( \gamma \), between the prior centre \( \rho^0 \) and the empirical estimate \( \sum_{i=1}^{n_3} E^{u_j} u_j \), which would be obtained with \( \gamma \to \infty \) or \( \gamma = 0 \), respectively. Moreover, \( \beta^{(t+1)} \) can be equivalently computed by maximizing the objective function

\[
\frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_j} \left[ E^{u_j} u_j (\beta_0 + \beta_1 x_j) - \log (1 - \alpha^{(t+1)} + \alpha^{(t+1)} \exp(\beta_0 + \beta_1 x_j)) \right] - \lambda \| \beta_1 \|_2^2.
\]

which is concave in \( \beta \) and of a similar form to the log-likelihood (2) with a ridge penalty for logistic regression. Each imputed probability \( E^{u_j} u_j \) serves as a pseudo response.

In our implementation, the prior centre \( \rho^0 \) is fixed as \( \rho' = n_3 / n \), and the proportion of \( \gamma = 1 \) in the labelled sample, and the scales \( (\alpha, \gamma) \) are treated as tuning parameters, to be selected by cross validation. Numerically, this procedure allows an adaptive interpolation between the two extremes: a fixed choice \( \rho' \) or an empirical estimate by maximum likelihood. For direct semi-supervised logistic regression (but not profile semi-supervised logistic regression), our adaptive procedure reduces to and hence, accommodates logistic regression with labelled data only at one extreme with \( \gamma \to \infty \). See the Appendix for further discussion.

5 | RELATED WORK

There is a vast literature on semi-supervised learning. See, for example, Chapelle et al. (2006) and Zhu (2005). For space limitation, we only discuss directly related work to ours.

**Generative models and EM.** A generative model can be postulated for \( (y, x) \) jointly such that \( p(y, x; \rho, \theta) = p(y; \rho) p(x; \theta) \), where \( \rho \) denotes the label proportion and \( \theta \) denotes the parameters associated with the feature distributions given labels (e.g. Nigam, McCallum, Thrun, & Mitchell, 2000). In our notation, a generative model corresponds to Equations (5)–(7), but with both \( G_0 \) and \( G_1 \) parametrically specified. For training by EM
algorithms, the expected objective in the E-step is similar to $Q^\theta(\eta, \beta, G)$ in (17), except that $G_k([x_j])$ is replaced by $\rho_j(x_j|y_j=k; \theta)$ for $k = 0$ or 1. The performance of generative modelling can be sensitive to whether the model assumptions are correct or not (Cozman, Cohen, & Cirelo, 2003). In this regard, our approach based on exponential tilt mixture models is attractive in only specifying a parametric form (8) for the density ratio between $G_0$ and $G_1$ while leaving the distribution $G_0$ nonparametric.

**Logistic regression and EM.** There are various efforts to extend logistic regression in an EM-style for semi-supervised learning. Notably, Amini and Gallinari (2002) proposed a classification EM algorithm using logistic regression (1), which can be described as follows:

- **E-step:** Compute $E^t u_{ij} = \{1 + \exp(-\beta_0^{(t)} x_i - \beta_1^{(t)} x_{ij})\}^{-1}$ for each $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. Fix $E^0 u_{ij} = 0$ and $E^0 u_{ij} = 1$.
- **C-step:** Let $u_{ij}^{(0)} = 1$ if $E^t u_{ij} > 0.5$ and 0 otherwise. Fix $u_{ij}^{(0)} = 0$ and $u_{ij}^{(0)} = 1$.
- **M-step:** Compute $(\beta_0^{(t+1)}, \beta_1^{(t+1)})$ by maximizing the objective $\sum_{i=1}^n \sum_{j=1}^{n_i} [u_{ij}^{(t)} (\beta_0^{(t)} x_i + \beta_1^{(t)} x_{ij}) - \log(1 + \exp(\beta_0^{(t)} + \beta_1^{(t)} x_{ij}))].$

Although convergence of classification EM was studied for clustering (Celeux & Govaert, 1992), it seems unclear what objective function is optimized by the preceding algorithm. A worrisome phenomenon we notice is that if soft classification is used instead of hard classification, and then the algorithm merely optimizes the log-likelihood of logistic regression with the labelled data only. By comparing (19) and (20), this modified algorithm can be shown to reduce to our EM algorithm with $\rho_j^{(t)}$ and $a_j^{(t)}$ clamped at $\rho' = n_2/n_1$, the proportion of $y=1$ in the labelled sample.

**Proposition 5.** If the objective in the M-step is modified with $u_{ij}^{(t)}$ replaced by $E^t u_{ij}$ as

$$\sum_{i=1}^n \sum_{j=1}^{n_i} [E^t u_{ij} (\beta_0^{(t)} x_i + \beta_1^{(t)} x_{ij}) - \log(1 + \exp(\beta_0^{(t)} + \beta_1^{(t)} x_{ij}))].$$

then $(\beta_0^{(t)}, \beta_1^{(t)})$ converges as $t \to \infty$ to the maximum likelihood estimator of logistic regression based on the labelled data only.

We notice that the conclusion also holds if (20) is replaced by the cost function proposed in Wang, Shen, and Pan (2009), Equation (2), when the logistic loss is used as the cost function on labelled data.

**Regularized methods.** Various methods have been proposed by introducing a regularizer depending on unlabelled data to the logistic likelihood loss or the hinge loss with labelled data. Examples include entropy regularization (Grandvalet & Bengio, 2005), manifold regularization (Belkin et al., 2006), expectation regularization (Mann & McCallum, 2007), and graph-based priors (Krishnapuram et al., 2005).

As discussed in Section 1, an important difference from our methods is that these regularized objective functions are in general not Fisher consistent based on infinite population data with a fixed semi-supervised tuning parameter, unless the semi-supervised tuning parameter is set to 0 and the objective functions reduce to the supervised parts based on labelled data only. For another difference, the class proportions in unlabelled training data and the test set are the same. To allow different class proportions between labelled and unlabelled data, we consider two schemes: the class proportions in the labelled data are close to those of the original dataset (‘Homo Prop’), or larger (or smaller) than the latter by an odds ratio of 4 (‘Flip Prop’) if the odds of positive versus negative labels is $\leq 1$ (or respectively $> 1$) in the original dataset. Hence, the class balance constraint as used in TSVM is satisfied in the first scheme but misspecified in the second scheme.

To make fair comparisons, all the methods are applied without knowledge about which of the two proportion schemes is used. Care is needed to define classifiers on test data. For all the four existing methods, classifiers are defined as usual, i.e. as the sign of $\hat{\beta}_0 + \hat{\beta}_1 x$ with $(\hat{\beta}_0, \hat{\beta}_1)$ in Model 1. This amounts to an implicit assumption of same class proportions in labelled training data and the test set. For our methods, classifiers are defined by using the estimated class proportions $\hat{\rho}_0$ (i.e. the sign of $\log(\hat{\rho}/(1 - \hat{\rho}))$) and $\hat{\beta}_0 + \hat{\beta}_1 x$. The classification accuracies are reported in Table 1. In addition, we also compare all methods with the classifiers adjusted by assuming 1:1 class proportions in test data. This assumption is often violated in the datasets but seems neutral when the actual class proportions in test data are unknown. The classification accuracies using the assumption of 1:1 class proportions can be found in Data S1 (Section IV).

---

**6 | NUMERICAL EXPERIMENTS**

We report experiments on 15 benchmark datasets including 11 UCI datasets and 4 SSL benchmark datasets. We compare our methods, profile and direct semi-supervised logistic regressions, with two supervised methods, ridge logistic regression and SVM, and two semi-supervised methods, entropy regularization (Grandvalet & Bengio, 2005) and TSVM (Joachims, 1999). For each method, only linear predictors are studied. All tuning parameters are selected by five-fold cross validation. See Data S1 (Section III) for details about the datasets and implementations.

For each dataset except SPAM, a training set is obtained as follows: 100 labelled data are sampled with prespecified class proportions, and then unlabelled data are sampled such that the labelled and unlabelled data combined are 2/3 of the original dataset. The remaining 1/3 of the dataset is used as a test set. For SPAM, the preceding procedure is applied to a subsample of size 750 from the original dataset. By this design, the class proportions in unlabelled training data and the test set are the same. To allow different class proportions between labelled and unlabelled data, we consider two schemes: the class proportions in the labelled data are close to those of the original dataset (‘Homo Prop’), or larger (or smaller) than the latter by an odds ratio of 4 (‘Flip Prop’) if the odds of positive versus negative labels is $\leq 1$ (or respectively $> 1$) in the original dataset. Hence, the class balance constraint as used in TSVM is satisfied in the first scheme but misspecified in the second scheme.
As seen from Table 1, ridge logistic regression performs the best in the Homo Prop scheme, whereas entropy regularization, the two semi-supervised logistic regression methods, and SVM achieve close to the best performances. Although the class balance constraint is satisfied in the Homo Prop scheme, TSVM is noticeably worse than other methods. While unstable performances of TSVM were previously noticed (e.g. Li & Zhou, 2014), such good performances of ridge logistic regression on these benchmark datasets appear not to have been reported before.

It should be stressed that to not invoke knowledge of the Homo Prop scheme, our methods are applied without using the ratio of class proportions in labelled data, \( n_2/n_1 \), which is a better estimator than \( \hat{\rho} \) for the class proportions in unlabelled data in the Homo Prop scheme. If \( n_2/n_1 \) were used instead of \( \hat{\rho} \) in our methods, the average accuracy would increase to 83.51 and 83.45, and the number of accuracies within 1\% of the highest would increase to 12, same as ridge logistic regression. The Homo Prop scheme is more favourable to ridge logistic regression as well as other existing methods than to our methods using estimated class proportions, because the assumption of same class proportions in labelled training and the test set is valid.

In the Flip Prop scheme, our methods, direct and profile semi-supervised logistic regression, achieve the best two performances, with considerable improvements over other methods. Comparatively, profile semi-supervised regression is better than the direct method. A possible explanation is that the profile method estimates the class proportions better. If the two methods are applied by assuming 1:1 class proportions in the test set, then the direct method would achieve a better performance than the profile method. See the additional results in Data S1.
7 | CONCLUSION

We develop an extension of logistic regression for semi-supervised learning, with strong support from statistical theory, algorithms, and numerical results, including theory on Fisher efficiency and asymptotic efficiency and EM algorithms for regularized estimation. Our numerical results indicate strong performance of the proposed methods, particular in the case where the class proportions in the unlabelled data differ from those in the labelled data. There are various questions of interest for future work. Our approach can be readily extended by employing nonlinear predictors such as kernel representations or neural networks. Moreover, it can be useful to incorporate cluster or smoothness assumptions into our approach. Further experiments with such extensions are desired, as well as applications to more complex text and image classification.

ORCID

Zhiqiang Tan https://orcid.org/0000-0003-1780-6839

REFERENCES


SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.
APPENDIX: EM ALGORITHM FOR DIRECT SEMI-SUPERVISED LOGISTIC REGRESSION

We present an EM algorithm to numerically maximize (16) for direct semi-supervised logistic regression, based on the semi-supervised logistic regression model defined by (12–14). We introduce the following data augmentation. Given the pooled data \((z_i, x_i) : i = 1, \ldots, N\), let

\[
u_i | (z_i = j, x_i) \sim \text{Bernoulli} \left( \frac{\rho_j \exp(\beta_0 + \beta_j^T x_i)}{1 - \rho_j + \rho_j \exp(\beta_0 + \beta_j^T x_i)} \right).
\]

(A.1)

Equivalently, \(u_i; i = 1, \ldots, N\) can be denoted as \(u_i; i = 1, \ldots, n_j, j = 1, 2, 3\), such that \(u_i | x_i \sim \text{Bernoulli}(\rho_j \exp(\beta_0 + \beta_j^T x_i)/(1 - \rho_j + \rho_j \exp(\beta_0 + \beta_j^T x_i)))\) for \(j = 1, 2, 3\). Similarly as in Section 4.3, \(u_0; i = 0\) and \(u_3; i = 1\) fixed, because \(\rho_1 = 0\) and \(\rho_2 = 1\).

\textbf{E-step.} The expectation of the average penalized log-likelihood from the augmented data, given the current estimates \((\rho^{(t)}, \beta^{(t)})\) is, up to an additive constant free of \((\rho, \beta)\),

\[
\begin{align*}
Q^{(t)}(\rho, \beta) = \frac{1}{N} \sum_{j=1}^{n} \sum_{i=1}^{n_j} \left[(1 - \rho)^j \log(1 - \rho) + \rho^j u_j \log(\rho) + (\beta_0 + \beta_j^T x_i)\right] \\
- \log(1 - \bar{a}(\rho) + \bar{a}(\rho) \exp(\beta_0 + \beta_j^T x_i))] + \text{pen}(\rho, \beta),
\end{align*}
\]

where \(\text{pen} = \rho^{(t)} \exp(\beta_0^{(t)} + \beta_j^{(t)T} x_i)/(1 - \rho^{(t)} + \beta_0^{(t)} + \beta_j^{(t)T} x_i)\).

\textbf{M-step.} The next estimates \((\rho^{(t+1)}, \beta^{(t+1)})\) are obtained as a maximizer of the expected objective \(Q^{(t)}(\rho, \beta)\). Recall that \(\kappa_Q^{(t)}(\rho, \beta, a)\) defined in Section 4.3 is

\[
\kappa_Q^{(t)}(\rho, \beta, a) = \frac{1}{N} \sum_{j=1}^{n} \sum_{i=1}^{n_j} \left[(1 - E^{(t)} u_j) \log \left( \frac{1 - \rho}{1 - a + a \exp(\beta_0 + \beta_j^T x_i)} \right) \right]
\]

\[
+ \text{E}^{(t)} u_j \log \left( \frac{\rho \exp(\beta_0 + \beta_j^T x_i)}{1 - a + a \exp(\beta_0 + \beta_j^T x_i)} \right) \right) - \log N + \text{pen}(\rho, \beta).
\]

It directly follows that \(Q^{(t)}(\rho, \beta) = \kappa_Q^{(t)}(\rho, \beta, \bar{a}(\beta))\) up to an additive constant. Therefore, the expected objective \(Q^{(t)}(\rho, \beta)\) is related to \(pQ^{(t)}(\rho, \beta) = \kappa_Q^{(t)}(\rho, \beta, \bar{a}(\beta))\) in the profile method, in a similar manner as the average log-likelihood \(k(\rho, \beta, a^{(t)})(\beta)\) in the semi-supervised logistic regression model is related to the average profile log-likelihood \(p(\rho, \beta)\) in the exponential tilt mixture model before data augmentation.

Unfortunately, when \(\rho\) is penalized with \(\gamma > 0\), there is no simple, closed-form expression for computing \(\rho^{(t+1)}\) as in Proposition 4. Nevertheless, we show that \(\rho^{(t+1)}\) can be obtained as a solution to a simple equation, independently of \(\beta^{(t+1)}\).

\textbf{Proposition 6.} The estimate \(\hat{\rho} = \rho^{(t+1)}\) satisfies

\[
\hat{\rho} = \frac{\sum_{j=1}^{n} \text{E}^{(t+1)} u_{3j}}{n_3} \frac{\psi(\hat{\beta})}{\psi(\hat{\beta}) + \gamma} + \rho^0 \frac{\gamma}{\psi(\hat{\beta}) + \gamma},
\]

where \(\psi(\hat{\beta}) = 1 - n_2(1 - \hat{\beta})/((N - n)(1 - \bar{a}(\beta))) \in (0, 1)\) because \(\bar{a}(\hat{\beta})(1 - \bar{a}(\hat{\beta})) > (n_3/N)\bar{a}(1 - \hat{\beta})\) for any \(\hat{\rho} \in (0, 1)\) as shown in the proof of Proposition 1.

Formula (A.2) shows that \(\hat{\rho} = \rho^{(t+1)}\) implicitly remains a weighted average of the prior centre \(\rho^0\) and the empirical estimate \(n_3^{-1} \sum_{j=1}^{n} \text{E}^{(t+1)} u_{3j}\), with the weight depending on \(\gamma\). If \(\gamma = 0\), then \(\rho^{(t+1)}\) reduces to \(n_3^{-1} \sum_{j=1}^{n} \text{E}^{(t+1)} u_{3j}\), and hence, the EM iterations \((\rho^{(t)}, \beta^{(t)})\) coincide with those for profile semi-supervised logistic regression in Section 4.3. If \(\gamma \rightarrow \infty\), then \(\rho^{(t+1)}\) becomes fixed at \(\rho^0\), and then \(\beta^{(t)}\) converges to a maximizer of \(k(\rho^0, \beta, \bar{a}(\beta)) - \lambda ||\beta||_2^2\), the ridge estimator of \(\beta\) in the semi-supervised logistic regression model (12–14) with \(\rho = \rho^0\) fixed. When \(\rho^0\) is set to \(\rho' = n_3/n\), this estimator of \(\beta\) is identical to that from ridge logistic regression with labelled data only, except for an intercept shift.

In contrast, if \(\rho = \rho^0\) is fixed in the EM algorithm for profile semi-supervised logistic regression, then \(\beta^{(t)}\) converges to a maximizer of \(k(\rho^0, \beta, \bar{a}(\beta)) - \lambda ||\beta||_2^2\), the ridge estimator of \(\beta\) in the exponential tilt mixture model (5–8).