The Effect of Model Misspecification on Semi-Supervised Classification

Ting Yang and Carey E. Priebe, *Senior Member*, *IEEE*

Abstract—Semi-supervised classification—training both on labeled and unlabeled observations—can yield improved performance compared to the classifier based on only the labeled observations. Unlabeled observations are always beneficial to classification if the model we assume is correct. However, they may degrade the classifier performance when the model is misspecified. In the classical classification problem setting, many factors affect the semi-supervised performance, including training data, model specification, estimation method, and the classifier itself. For concreteness, we consider maximum likelihood estimation in finite mixture models and the Bayes plug-in classifier, due to their ubiquitousness and tractability. In this specific setting, we examine the effect of model misspecification on semi-supervised classification performance and shed some light on when and why performance degradation occurs.

Index Terms—Semi-supervised classification, finite mixture model, Bayes plug-in classifier.

1 Introduction

1.1 Probabilistic Model

Let $(X, Y) \sim F_{XY}$. The feature observation X is an $\mathbb{R}^d$-valued random variable. The nature of the observation is called a class label, denoted by Y and taking values in a finite set $\{1, 2, \ldots, K\}$. For $j = 1, \ldots, K$, denote the class conditional distributions by $F_j = F_{X|Y=j}$, and assume we are in the continuous case, so the class conditional densities f_j exist. Let $\pi_j = P\{Y = j\}$ be class priors, which can also be referred to as component coefficients.

We assume that the class label Y is not observed, and our goal is to classify the feature observation X with small classification error $L(g) = P\{g(X) \neq Y\}$. Suppose we are interested in learning probabilistic classifiers from semi-supervised data given by

$$(X_1,Y_1),(X_2,Y_2),\ldots,(X_\ell,Y_\ell) \sim F_{XY} \ \ (i.i.d.),$$
$$X_{\ell+1}, X_{\ell+2}, \ldots, X_{\ell+u} \sim F_X \ \ (i.i.d.),$$

where u and ℓ are the numbers of unlabeled and labeled observations, respectively, and the two types of data are independent. We will assume a parametric model $\mathcal{F}$ indexed by some parameter θ for the density f of X. The set of all parameter points Θ is called the parameter space for the model. The true density f_0 belongs to the model $\mathcal{F}_0$ with parameter space Θ_0. $\theta_0 \in \Theta_0$ denotes the true parameter. If f_0 is not an element of the assumed model $\mathcal{F}$, then we say the model is misspecified.

• *The authors are with the Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218. E-mail: {ting.yang, cep}@jhu.edu.*

Manuscript received 3 May 2010; revised 14 Sept. 2010; accepted 20 Dec. 2010; published online 1 Mar. 2011.
Recommended for acceptance by D. Schuurmans.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-05-0342.
Digital Object Identifier no. 10.1109/TPAMI.2011.45.

Let θ^*_{sup} be the limit of the supervised MLE of θ, and θ^*_{unsup} the limit of the unsupervised MLE of θ. Both θ^*_{sup} and θ^*_{unsup} exist under mild regularity conditions.

For simplicity, we consider two-class classification problems throughout this paper. The joint density of (X, Y), $f_{(X,Y)}(x, y)$, can be written as

$$\pi_1 f(x \mid 1)\mathbb{1}\{y = 1\} + (1 - \pi_1) f(x \mid 2)\mathbb{1}\{y = 2\}.$$

1.2 Previous Work

The questions of value and risk of semi-supervised learning has been investigated by many authors. In this section, we review four methods that have been used in the past to analyze the performance of semi-supervised classifiers.

1. J. Ratsaby and S. Venkatesh [1]:
   A two-class Gaussian mixture classification problem is studied in this paper to determine how the error rate depends on the labeled sample size ℓ, unlabeled sample size u, and the dimensionality d. Using Chernoff bounds and rate of convergence in a uniform strong law for bounded functions, the authors consider the effect of introducing n unlabeled examples. Their method is to look at the reduction in the size ℓ of the labeled samples needed to obtain a given error rate and use that as a rough measure of the worth of each labeled example in terms of the number of unlabeled examples that are needed to compensate for it. They conclude that the introduction of unlabeled examples has reduced the demands on the number of labeled examples and a rough measure is given as a ratio of the unlabeled examples over the reduction in labeled examples. This ratio also shows that the efficacy of the unlabeled examples diminishes as the dimensionality d increases. Moreover, the authors show that the semi-supervised classification error rate deviates from the Bayes optimal error rate by $\mathcal{O}(d^{3/5}u^{-1/5}) + \mathcal{O}(e^{-c\ell})$, indicating that the

error rate contribution from the labeled examples decreases exponentially fast in ℓ, while the error rate contribution from the unlabeled examples decreases only as an inverse polynomial in u.

The authors point out that because the answer to the question is not immediate, even for the archetypal problem of two d-dimensional Gaussian distributed pattern classes, for definiteness they focus only on the case of Gaussian mixtures in this paper, also making the following assumptions: correct model assumption, equiprobable pattern classes assumption, and equal unit covariance matrices assumption.

2. Castelli and Cover [2], [3]:

These two papers study the relative value of labeled and unlabeled observations in a two-class classification problem. To set up the problem, it is pointed out that for an identifiable family of mixtures $\mathcal{F}$, the mixture

$$f_0(x) = \pi_1 f_1(x) + \pi_2 f_2(x)$$

can be estimated from unlabeled observations. Labeled observations are needed to label the classes. The notation $L(\ell, u)$ is used to denote the error rate of Bayes plug-in classifier resulted from ℓ labeled observations and u unlabeled observations.

In the first paper, the first theorem is about the value of labeled observations. Specifically, the error rate of the Bayes plug-in classifier with one labeled observation and an infinite number of unlabeled observation is

$$L(1, \infty) = 2L^*(1 - L^*) < 2L^*,$$

where L^* is the Bayes error.

The second theorem states that $L(\ell, \infty)$ converges to L^* as $\ell \to \infty$ in the following way:

$$-\lim \frac{1}{\ell} \log(L(\ell, \infty) - L^*) = -\log\left(2\sqrt{\pi_1 \pi_2} \int \sqrt{f_1(x) f_2(x)} dx\right).$$

The quantity $-\log(\int \sqrt{f_1(x) f_2(x)} dx)$ is the Bhattacharyya distance between the densities $f_1(x)$ and $f_2(x)$. Hence,

$$L(\ell, \infty) - L^* = \exp\left\{\ell \log\left(2\sqrt{\pi_1 \pi_2} \int \sqrt{f_1(x) f_2(x)} dx\right)\right\} + o(\ell).$$

This proves that labeled samples have an exponential value in reducing the probability of error.

In the second paper, the authors assume that the class conditional densities are known, but not labeled. The mixing coefficient is unknown. Considering finite ℓ and u, they show that under some specific limiting conditions, labeled observations are exponentially more valuable than unlabeled observations. In both papers, unlabeled observations benefit the semi-supervised classification, although not as much as labeled observations. This is due to the correct model assumption.

3. Zhang and Oles [4]:

To understand the value of using unlabeled data under correct parametric model, the authors consider binary classification problems and use Fisher information matrices to judge the asymptotic value of unlabeled data. In other words, they judge the value of unlabeled data by evaluating its impact on the efficiency of parameter estimation. Since, in a correct model,

$$I_{labeled+unlabeled}(\theta) = I_{labeled}(\theta) + I_{unlabeled}(\theta),$$

the conclusion is that including unlabeled data always helps because Fisher information is increased.

4. Cozman and Cohen [5], [6]:

The key theorem in this paper is that the semi-supervised maximum likelihood estimator (MLE) converges to a parameter value that maximizes a convex combination of the supervised and unsupervised expected log-likelihood functions. Hence under some regularity conditions on the density, the limit of the semi-supervised maximum likelihood estimator can be proven to travel on a continuous path connecting the two limits: the limit of the supervised maximum likelihood estimator and the limit of the unsupervised maximum likelihood estimator. The position of the semi-supervised limit on the path depends on the ratio of the numbers of labeled and unlabeled observations. If the fully supervised limit misclassification rate is no greater than the fully unsupervised limit misclassification rate, then adding more unlabeled observation would improve performance when these two limits are the same (i.e., when the model is correct) and *may* degrade the performance when the two limits are different (i.e., when the model is incorrect).

The authors write "regardless of the approach that is used, semi-supervised learning is affected by modeling assumptions in rather complex ways." "Results in this paper can be extended in several directions. It should be interesting to find necessary and sufficient conditions for a model to suffer performance degradation with unlabeled data" [5], [6].

1.3 Our Work

Unlabeled observations *may* degrade the classification performance when the model is misspecified. In this paper, we establish the relationship between model misspecification and performance degradation in semi-supervised classification for a restricted case. We show that under some conditions, the probability that semi-supervised classification results in performance degradation is determined by the two MLE limits, θ^*_{sup} and θ^*_{unsup}.

2 Preliminaries

We focus our study and present examples in the area of classification using finite mixture models. The classification task will be handled by Bayes plug-in classifier with maximum likelihood estimation.

2.1 Bayes Plug-in Classifier

Let $f_0 = f_{\theta_0}$ be the true mixture density. Suppose $\hat{\pi}_1, \ldots, \hat{\pi}_K$ are estimates of the true component coefficients $\pi_{01}, \ldots, \pi_{0K}$, and $\hat{f}_1, \ldots, \hat{f}_K$ are estimates of the true class conditional densities $f_{01}, \ldots, f_{0K}$. Then the *Bayes plug-in classifier* is given by

$$g_{\hat{f}}(x) = \underset{j}{\operatorname{argmax}}\ \hat{\pi}_j \hat{f}_j(x).$$

There is a best possible classifier, g^*, which is defined by

$$g^* = \underset{g}{\operatorname{argmin}}\ L(g).$$

The minimal probability of error is called the Bayes error and is denoted by $L^* = L(g^*) = L(g_{f_0})$, where

$$g_{f_0}(x) = \underset{j}{\operatorname{argmax}}\ \pi_{0j} f_{0j}.$$

But since f_0 is unknown in practice, L^* is unavailable.

2.2 MLE for Misspecified Models

When the model is misspecified under mild regularity conditions, the MLE $\hat{\theta}$ is a strongly consistent estimator for θ^*, the parameter vector which minimizes the Kullback-Leibler divergence (KL) between the actual density of X and the postulated parametric family $\mathcal{F}$, where

$$KL(f_0 \parallel f) = E_{\boldsymbol{\theta}_0}[\log[f_0(X)/f(X)]].$$

See [7] for the general conditions for the existence of such MLE, its consistency, and asymptotic normality (Assumptions A1-A6).

2.3 KL Divergence between Two Joint Densities

Let $\mathcal{F}_{joint} = \{f(X, Y \mid \theta); \theta \in \Theta\}$ be a model, with the discrepancy between the true joint density $f_0(X, Y)$ and each element in the model defined by the Kullback-Leibler divergence, that is, the discrepancy between any two densities with parameter values θ and θ' is

$$KL(\ f_\theta \parallel f_{\theta'}\ ) = \int log \frac{f_\theta(x, y)}{f_{\theta'}(x, y)} dF_\theta(x, y).$$

2.4 Gaussian Mixture Models

Gaussian mixture modeling is at the heart of many classification problems, such as computer vision, brain image segmentation, speech recognition, etc. Although our theoretical results are more general, the semi-supervised classification examples in this paper are based on parametric family of finite Gaussian mixture models, i.e., families of probability density functions of the form

$$X_i \sim f_0 \in \mathcal{F}_K = \left\{ f_{\boldsymbol{\theta}} = \sum_{j=1}^{K} \pi_j \phi(x_i|\theta_j) : \pi_j > 0, \right.$$
$$\left. \sum_{j=1}^{K} \pi_j = 1, \boldsymbol{\theta} = \left(\pi_1, \ldots, \pi_{K-1}, \mu_1, \sigma_1^2, \ldots \mu_K, \sigma_K^2\right) \in \boldsymbol{\Theta} \right\},$$

where

$$\boldsymbol{\Theta} = (0, 1)^{K-1} \times (-\infty, \infty)^K \times (0, \infty)^K \subset \mathbb{R}^{3K-1}$$

and $\phi(\cdot \mid \theta_j)$ denotes a Gaussian density with parameter $\theta_j = (\mu_j, \sigma_j^2)$ [8].

It is important to consider identifiability here in order for estimation procedures to be well defined. Borrowing notation from [9], given an arbitrary g, we define the index of the economical representation of g as

$$m(g) = \min\{m : g \in \mathcal{F}_m\}.$$

This means we always use the most parsimonious Gaussian mixture model representation for a finite mixture of Gaussians g. It also has the benefit of estimating fewer parameters. For example, even though a mixture of two Gaussians can be represented as an element in $\mathcal{F}_3$, we would prefer to use $\mathcal{F}_2$ as the model for estimation.

The likelihood under this model is unbounded. The MLE of $\boldsymbol{\theta}$ as a global maximizer of the likelihood function does not exist. Singularities occur at certain points on the boundary of the parameter space. In our examples, we carefully design our experiments and pick initial values that would in most cases keep us away from these singularities and obtain a sensible local likelihood maximum with desirable asymptotic properties. Full theoretical results and regularity conditions are available in [10].

3 Bias and Variance Tradeoff

The performance of the Bayes plug-in classifier relies on how good the estimates for the true parameters are. The overall error rate comes from model bias (approximation error) and estimation error. In this section, we conceptually study semi-supervised misclassification rate from the point of view of bias and variance tradeoff.

Let f_{opt} denote the best estimator of f_0 in $\mathcal{F}$, in the sense that

$$g_{f_{opt}} = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ L(g_f).$$

Let the parameter associated with $g_{f_{opt}}$ be θ_{opt}. Let $\hat{f} \in \mathcal{F}$ be the learned estimator with parameter $\hat{\theta}$. The dashed circle (we are unclear about its real shape) is the collection of all Bayes plug-in rules based on the model $\mathcal{F}$ and is denoted by $g_{\mathcal{F}}$. Each element in $g_{\mathcal{F}}$ represents a Bayes plug-in classification rule. We demonstrate the various errors associated with classifiers in Fig. 1, a modified version of [11, Fig. 12.1]. The model bias is measured by $L(g_{f_{opt}}) - L^*$ and the estimation error $L(g_{\hat{f}}) - L(g_{f_{opt}})$.

3.1 Correct Parametric Model

If $f_0 \in \mathcal{F}$, then the model bias is 0, i.e., $L(g_{f_{opt}}) - L^* = 0$. The estimation error is the only thing that contributes to the classification error rate. In the parametric setting, suppose we use mean squared error on the parameter space to measure the estimation error, then we have

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta_{opt})^2] = (E[\hat{\theta}] - \theta_{opt})^2 + Var(\hat{\theta}). \quad (1)$$

The term $(E[\hat{\theta}] - \theta_{opt})^2$ is a form of bias. In a correct parametric model case, both supervised MLE and semi-supervised MLE converge to the same parameter value θ_0. We have $(E[\hat{\theta}] - \theta_{opt})^2 = (E[\hat{\theta}] - \theta_0)^2$ since $\theta_{opt} = \theta_0$. MLE is consistent (asymptotically unbiased and variance is reduced

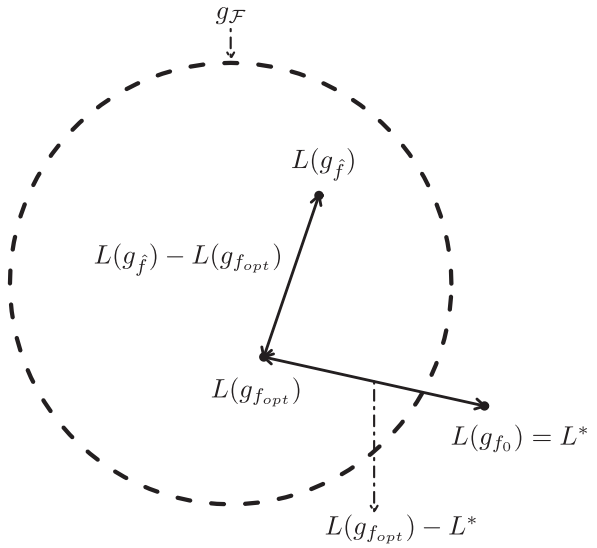


Fig. 1. Various errors in empirical classifier selection.

by adding more observations) under mild regularity conditions. The classification rate will be reduced as the sample size increases. In this case, semi-supervised Bayes plug-in classifier is consistent. That is, $L_{\ell+u} \to L^*$ in probability as $u \to \infty$.

3.2 Incorrect Parametric Model

If $f_0 \notin \mathcal{F}$, then $L(g_{f_{opt}}) - L^* > 0$. The change in the training set, however, only results in the change in estimation error. We again have (1). Adding unlabeled observations still reduces the estimation variance. However, since the model is misspecified, the supervised MLE and the semi-supervised MLE may converge to different parameter values. Given a fixed labeled training set, more unlabeled observations may cause a larger estimation bias, that is,

$$(E[\hat{\theta}_{\ell+u}] - \theta_{opt})^2 > (E[\hat{\theta}_{\ell}] - \theta_{opt})^2,$$

where $\hat{\theta}_\ell$ and $\hat{\theta}_{\ell+u}$ are the supervised and semi-supervised estimates. In the case where the increase in the estimation bias is more significant than the decrease in the estimation variance in terms of classification error, semi-supervised learning degrades the classification performance. Whether degradation or improvement occurs is determined by the bias and variance trade-off.

3.3 Semi-Parametric Model

In parametric models, degradation of performance is directly related to incorrect modeling assumptions. We conjecture that if we increase the model dimensionality (slowly) as a function of the unlabeled sample size u (with labeled sample size ℓ held constant) by using semi-parametric models, we will in some cases avoid this problem. Moreover, in semi-parametric models, by allowing the nuisance parameter space to be infinite-dimensional, we are placing fewer restrictions on the probability model from which the data were generated as compared to a parametric model. Therefore, estimators of parameters of interest in semi-parametric models may have greater applicability and greater robustness. However, along with a larger model which gives smaller model bias, we face the possibility of larger estimation error. The influence of the model change on classifier performance needs to be studied.

It is our hope that conceptually understanding where the change in semi-supervised classification error rate comes from will show the way to improvement of the design of semi-supervised algorithms in practical settings. In any case, it would be ideal to establish a quantitative relationship between model misspecification and performance degradation in semi-supervised classification. That is what we will present next.

4 Semi-Supervised Degradation Theorem

4.1 On Idealization

Any theory necessarily involves idealization. Let $L(\hat{f})$ be the error rate of Bayes plug-in classifier based on maximum likelihood estimate $\hat{f}$, and $KL(f_{\theta^*_{sup}} \parallel \hat{f})$ be the Kullback-Leibler divergence between the supervised limit density and the estimated density $\hat{f}$. Our first idealization concerns the relationship between $L(\hat{f})$ and $KL(f_{\theta^*_{sup}} \parallel \hat{f})$. To formulate the theory, consider error rate as a function of the joint distribution of (X, Y), i.e., $L : \mathcal{F}_{joint} \to [0, 1]$, with Kullback-Leibler divergence featuring as a measure of divergence and Fisher information taking the role of curvature on $\mathcal{F}_{joint}$. Assume identifiability (one-to-one parameterization), then the discussion can be carried out on the model $\mathcal{F}_{joint}$ as well as on the parameter space. Among all of the parameter values we can possibly obtain by using maximum likelihood estimation, we have a good reason to believe that in terms of accuracy, in some cases, the supervised MLE limit θ^*_{sup} is the best parameter value in a small neighborhood of parameter space of the assumed model (in some cases we even have $\theta^*_{sup} = \theta_{opt}$). Note we do not deny the possibility that it is not. As a matter of fact, it is very possible that there are other parameter values that give a lower or the same error rate as θ^*_{sup}. However, in the scenario of performing maximum likelihood estimation under some smoothness conditions, these values are unlikely to be reached; and even if they are reached by chance, they are relatively less significant in the sense that in our discussion of the asymptotic behavior of semi-supervised maximum likelihood estimator, as long as there are only finite number of such values, they are not what the estimator will ultimately converge to and hence will not affect the asymptotic theory significantly. Therefore, treating θ^*_{sup} as the best parameter value either in a small neighborhood or in the whole the parameter space (if that's the case) still stands a chance of giving us a good approximation of the truth. Furthermore, we argue that since the performance of Bayes plug-in classifier depends on how good the parameter estimates are, given that θ^*_{sup} is the best parameter value in a small neighborhood of the parameter space in the assumed model, with the Bayes plug-in error rate being a continuous function in θ, parameter values closer to θ^*_{sup} in that small neighborhood should lead to better classification performance. With this idealization, we have a hope of dealing with the problem at hand and simplifying the discussion from a functional space to the parameter space. For the completeness of the theory, we will discuss under what assumptions it is true and how robust it is in practice when the assumptions are violated.

4.2 Intuition

Define a KL ball to be

$$B_{KL}(\theta^*, \epsilon) = \{\theta \in \Theta : KL(f_{\theta^*} \parallel f_\theta) < \epsilon\}.$$

Let

$$\sup L(B_{KL}(\theta^*, \epsilon)) = \sup\{L(g_\theta) \mid \theta \in B_{KL}(\theta^*, \epsilon)\},$$
$$\inf L(B_{KL}(\theta^*, \epsilon)) = \inf\{L(g_\theta) \mid \theta \in B_{KL}(\theta^*, \epsilon)\}.$$

If $L(g_{f_{\theta^*_{sup}}}) < L(g_{f_{\theta^*_{unsup}}})$, then as $\ell \to \infty$, and $\ell/u \to 0$, there should exist $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that

$$\sup L(B_{KL}(\theta^*_{sup}, \epsilon_1)) < \inf L(B_{KL}(\theta^*_{unsup}, \epsilon_2)).$$

That is, if $L(g_{f_{\theta^*_{sup}}}) < L(g_{f_{\theta^*_{unsup}}})$, then intuitively there should exist small neighborhoods around the two MLE limit densities such that in terms of error rate, the worst density in the neighborhood around $f_{\theta^*_{sup}}$ is still better than the best density in the neighborhood around $f_{\theta^*_{unsup}}$. This suggests that semi-supervised learning eventually degrades the performance in this case. And the occurrence of degradation can be closely predicted by the discrepancy between the estimated densities and supervised limit density. This intuition gives rise to our theorem in the next section.

4.3 Lemma and Theorem

For the lemma, theorem, and corollaries, we assume that θ^*_{sup} and θ^*_{unsup} exist under the current model $\mathcal{F}$ as limits of MLE. All the results are stated only under the theoretical setting of the paper.

Lemma. *For any fixed finite ℓ or $\ell \to \infty$, as $\ell/u \to 0$, the limit of the maxima of semi-supervised likelihood function is the fully unsupervised MLE limit θ^*_{unsup}. That is, let $\hat{\theta}_{(l+u)}$ be the semi-supervised MLE when the sample size is ℓ for labeled data and u for unlabeled data, then as $\ell/u \to 0$,*

$$\{\hat{\theta}_{(l+u)}\}_u \xrightarrow{p} \theta^*_{unsup} \quad \forall \ell.$$

Proof. Suppose in semi-supervised learning the samples are realizations of labeled samples (X, Y) with probability λ and of unlabeled samples X with probability $1 - \lambda$. Then the theorem in [6] states that the limiting value θ^* of MLE is

$$\underset{\theta}{\operatorname{argmax}} \; (\lambda \mathbf{E}_{f(X,Y)}[\log f(X, Y \mid \theta)] + (1 - \lambda)\mathbf{E}_{f(X,Y)}[\log f(X \mid \theta)]),$$

a convex combination of the supervised and unsupervised expected log-likelihood functions. For an arbitrary finite value of ℓ, as $\ell/u \to 0$, $\lambda = \ell/u \to 0$, indicating in the limit, θ^* maximizes $\mathbf{E}_{f(X,Y)}[\log p(X \mid \theta)]$, thereby by definition is θ^*_{unsup}. By a similar argument, as $\ell \to \infty$, and $\ell/u \to 0$, the limit of the maxima of semi-supervised learning likelihood function is θ^*_{unsup}. □

Semi-supervised Degradation Theorem. *If $L(f_{\theta^*_{sup}}) < L(f_{\theta^*_{unsup}})$, then for $\ell, u > 0$, as $\ell \to \infty$ and $\ell/u \to 0$,*

$$\mathbb{1}\{L(f_{\hat{\theta}_\ell}) < L(f_{\hat{\theta}_{(\ell+u)}})\} - \mathbb{1}\{KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_\ell}) < KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_{\ell+u}})\} \xrightarrow{p} 0.$$

And we have

$$\lim_{\ell\to\infty, \ell/u\to 0} P\{L(f_{\hat{\theta}_\ell}) < L(f_{\hat{\theta}_{(\ell+u)}})\} = \lim_{\ell\to\infty} P\{KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_\ell}) < KL(f_{\theta^*_{sup}} \parallel f_{\theta^*_{unsup}})\}.$$

Proof. As $\ell \to \infty$, the supervised estimator converges to θ^*_{sup}. We also have, by lemma, as $\ell \to \infty$, and $\ell/u \to 0$, the semi-supervised estimator converges to θ^*_{unsup}. As $\ell \to \infty$, and $\ell/u \to 0$, by the continuous mapping theorem, we have

$$\mathbb{1}\{L(f_{\hat{\theta}_\ell}) < L(f_{\hat{\theta}_{(\ell+u)}})\} \xrightarrow{p} \mathbb{1}\{L(f_{\theta^*_{sup}}) < L(f_{\theta^*_{unsup}})\},$$

and

$$\mathbb{1}\{KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_\ell}) < KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_{\ell+u}})\} \xrightarrow{p} \mathbb{1}\{KL(f_{\theta^*_{sup}} \parallel f_{\theta^*_{sup}}) < KL(f_{\theta^*_{sup}} \parallel f_{\theta^*_{unsup}})\}.$$

Since

$$\mathbb{1}\{L(f_{\theta^*_{sup}}) < L(f_{\theta^*_{unsup}})\} = 1,$$

and

$$\mathbb{1}\{KL(f_{\theta^*_{sup}} \parallel f_{\theta^*_{sup}}) < KL(f_{\theta^*_{sup}} \parallel f_{\theta^*_{unsup}})\} = 1,$$

by the Slutsky's theorem, we have

$$\mathbb{1}\{L(f_{\hat{\theta}_\ell}) < L(f_{\hat{\theta}_{(\ell+u)}})\} - \mathbb{1}\{KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_\ell}) < KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_{\ell+u}})\} \xrightarrow{p} 0.$$

Taking expectations on both sides, the rest of the theorem follows. □

Corollary 1. *If $L(f_{\theta^*_{sup}}) < L(f_{\theta^*_{unsup}})$, then for a given misspecified model, $\exists \ell$, s.t.*

$$\lim_{u\to\infty} P\{L(f_{\hat{\theta}_\ell}) < L(f_{\hat{\theta}_{(\ell+u)}})\} > 0.$$

That is, semi-supervised classification yields degradation with positive probability as $u \to \infty$.

Proof. Given ℓ, $\exists \epsilon_1$ s.t. $\hat{\theta}_\ell \in B_{KL}(\theta^*_{sup}, \epsilon_1)$. As $u \to \infty$, $\exists \epsilon_2$ s.t.

$$\sup L(B_{KL}(\theta^*_{sup}, \epsilon_1)) < \inf L(B_{KL}(\theta^*_{unsup}, \epsilon_2)).$$

Therefore, we have $\lim_{u\to\infty} P\{L(f_{\hat{\theta}_\ell}) < L(f_{\hat{\theta}_{(\ell+u)}})\} > 0$ for such ℓ. □

In the finite case, we do not claim that either of the following is true under general conditions.

$$\mathbb{1}\{L(f_{\hat{\theta}_\ell}) < L(f_{\hat{\theta}_{(\ell+u)}})\} = \mathbb{1}\{KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_\ell}) < KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_{\ell+u}})\},$$
$$P\{L(f_{\hat{\theta}_\ell}) < L(f_{\hat{\theta}_{(\ell+u)}})\} = P\{KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_\ell}) < KL(f_{\theta^*_{sup}} \parallel f_{\theta^*_{unsup}})\}.$$

They are true in the limit, that is, as $\ell \to \infty$, and $\ell/u \to 0$, or when error rate L as a function of the joint density f_θ is convex, uniquely minimized at θ^*_{sup}. The convexity implies

$$KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_1}) < KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_2}) \Longrightarrow L(f_{\hat{\theta}_1}) < L(f_{\hat{\theta}_2})$$

and it suffices to show that the above two equations hold by a simple argument under this condition.

Corollary 2. *If error rate L as a function of the joint density f_θ is convex, uniquely minimized at θ^*_{sup}, then for a given model, if $\ell_1 < \ell_2$, then we have*

$$\lim_{u\to\infty} P\{L(f_{\hat{\theta}_{\ell_1}}) < L(f_{\hat{\theta}_{(\ell 1+u)}})\} < \lim_{u\to\infty} P\{L(f_{\hat{\theta}_{\ell 2}}) < L(f_{\hat{\theta}_{(\ell_2+u)}})\}.$$

Proof. For $\ell_1 < \ell_2$, we have

$$P\{\|\hat{\theta}_{\ell_1} - \theta^*_{sup}\| < \|\theta^*_{unsup} - \theta^*_{sup}\|\} < P\{\|\hat{\theta}_{\ell_2} - \theta^*_{sup}\| < \|\theta^*_{unsup} - \theta^*_{sup}\|\},$$

where $\|\hat{\theta}_\ell - \theta^*_{sup}\| = (\hat{\theta}_\ell - \theta^*_{sup})^T I(\theta^*_{sup})(\hat{\theta}_\ell - \theta^*_{sup})$. Geometrically, $\{\hat{\theta}_\ell : (\hat{\theta}_\ell - \theta^*_{sup})^T I(\theta^*_{sup})(\hat{\theta}_\ell - \theta^*_{sup}) < 1\}$ is the ellipsoid around θ^*_{sup} whose axis points in the direction of the eigenvectors and whose axis lengths are given by the square roots of the inverse eigenvalues of $I(\theta^*_{sup})$. Therefore, by the approximation of the KL ball by a rescaled L2-norm neighborhood [12], we have

$$P\{KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_{\ell 1}}) < KL(f_{\theta^*_{sup}} \parallel f_{\theta^*_{unsup}})\} < P\{KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_{\ell 2}}) < KL(f_{\theta^*_{sup}} \parallel f_{\theta^*_{unsup}})\}.$$

(We will discuss the approximation of the KL ball by a rescaled L2-norm neighborhood in more detail in Section 5.3.) This implies

$$\lim_{u\to\infty} P\{L(f_{\hat{\theta}_{\ell_1}}) < L(f_{\hat{\theta}_{(\ell_1+u)}})\} < \lim_{u\to\infty} P\{L(f_{\hat{\theta}_{\ell_2}}) < L(f_{\hat{\theta}_{(\ell_2+u)}})\}.$$

□

Corollary 2 implies that as u increases, compared to a smaller labeled data samples size ℓ_1, the larger labeled data sample size ℓ_2 situation has a greater probability of semi-supervised degradation.

5 Simulation Study

In this and the next section, we consider a semi-supervised classification example where the true density f_0 is essentially a mixture of three Gaussians considered as a two-component mixture density. The number of simulations is 1,000 for all examples. Note that the error rates denoted by L are in fact estimates $\hat{L}_{\ell+u}$ under 1,000 simulations. The simulations were carried out using the R statistical programming environment [13].

Suppose the model $\mathcal{F}$ for estimation is the collection of mixtures of two Gaussians, giving us a semi-supervised classification problem with a misspecified model when $\mu_2 \neq \mu_3$.

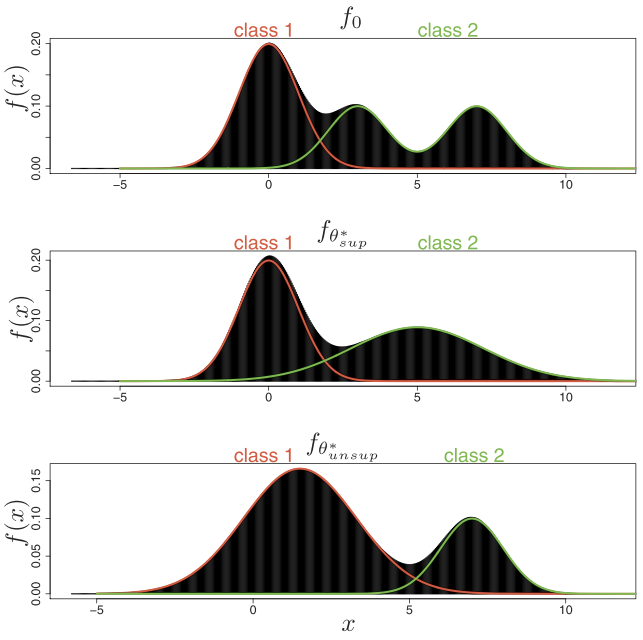


Fig. 2. The truth, supervised classifying, and unsupervised classifying processes.

$$f_0(x) = \frac{1}{2}\phi(x \mid 0,1) + \frac{1}{2}\left[\frac{1}{2}\phi(x \mid 3,1) + \frac{1}{2}\phi(x \mid 7,1)\right],$$

$$\mathcal{F} = \Big\{ f_\theta = \pi_1 \phi\big(x \mid \mu_1, \sigma_1^2\big) + (1-\pi_1)\phi\big(x \mid \mu_2, \sigma_2^2\big), \theta = \big(\pi_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2\big) \in \Theta \Big\},$$

where

$$\Theta = [0,1] \times (-\infty, \infty) \times (0, \infty) \times (-\infty, \infty) \times (0, \infty) \subset \mathbb{R}^5.$$

5.1 Assumption Verification

There are three Gaussian densities. With supervised information, we can correctly categorize the second and third Gaussian density as class 2. Since the misspecified model allows only mixtures of two Gaussians, the learning algorithm will combine the last two Gaussians to be one Gaussian. In unsupervised learning, we conjecture that without supervised information, the learning algorithm is likely to combine the first two Gaussians as one Gaussian and categorize it as class 1, while leaving the third Gaussian to be class 2. The mistake unsupervised learning algorithm makes is caused by the fact that the first two Gaussians are closer compared to the last two Gaussians. This would not have happened if the mean of the third Gaussian μ_{03} is smaller, say 4. Fig. 2 shows our conjecture as to what will happen in these two learning processes.

From observing Fig. 2, the supervised result gives an error rate very close to L^* for the intersection point of the two Gaussian curves is very close to the true intersection point. It is obvious from Fig. 2 that the more we go from supervised classification to unsupervised classification, the more the intersection point will move to the right, causing larger error rates. That is, $L(f_{\theta^*_{sup}}) < L(f_{\theta^*_{unsup}})$.

To explicitly verify the condition $L(f_{\theta^*_{sup}}) < L(f_{\theta^*_{unsup}})$, the key is to determine θ^*_{sup} and θ^*_{unsup}. The calculation of θ^*_{sup} is straightforward.

$$\theta^*_{sup} = (\pi^*_{1sup}, \mu^*_{1sup}, \sigma^{*2}_{1sup}, \mu^*_{2sup}, \sigma^{*2}_{2sup}) = (0.5, 0, 1, 5, 5).$$

The calculation of θ^*_{unsup} needs to be done numerically and is rather difficult, but there are several ways to approximate it.

- The heuristic argument above suggests that θ^*_{unsup} could be close to the following value

$$\theta^*_{unsup} \approx (\pi^*_{1unsup}, \mu^*_{1unsup}, \sigma^{*2}_{1unsup}, \mu^*_{2unsup}, \sigma^{*2}_{2unsup}) = (0.75, 1.5, 3.25, 7, 1).$$

  This calculation is done based on the third subfigure in Fig. 2. Note that holding all other parameters constant, whether this is exactly how the unsupervised classifying result turns out depends on μ_{03}. The larger the value of μ_{03}, the closer the unsupervised classifying result is to what happens in the third subfigure in Fig. 2.
- If we estimate it using one simulated training set, then the EM algorithm gives approximation

$$\hat{\theta}_{0+50,000} \approx (0.75, 1, 3, 7, 1.1).$$

Our calculation shows that

$$KL(f_0 \parallel f_{(0.75,1.5,3.25,7,1)}) > KL(f_0 \parallel f_{(0.75,1,3,7,1.1)}).$$

Based on the fact that θ^*_{unsup} minimizes the KL divergence between the truth f_0 and the model $\mathcal{F}$, we can conclude that $(0.75, 1, 3, 7, 1.1)$ is a closer approximation for θ^*_{unsup}. This indicates that in the case of $\mu_3 = 7$, the semi-supervised estimation process is not as extreme as that suggested by Fig. 2 (it is imaginable that the process will be as extreme if, say, $\mu_{03} = 10$ or even larger).

Better estimates can certainly be reached with better optimization algorithms. However, for the purpose of verifying the condition of the theorem, what we really need to be sure of is not the exact value of θ^*_{unsup}, but whether $L(f_{\theta^*_{sup}}) < L(f_{\theta^*_{unsup}})$. Knowing that $L^* \approx 0.046, L(f_{\theta^*_{sup}}) \approx 0.048, L(f_{(0.75,1,3,7,1.1)}) \approx 0.248$, and that $(0.75, 1, 3, 7, 1.1)$ is a close approximation of θ^*_{unsup}, we feel safe to conclude that $L(f_{\theta^*_{sup}}) < L(f_{\theta^*_{unsup}})$ and that the main condition of the theorem is satisfied.

5.2 Asymptotic Result

Let $\Delta_L = \mathbb{1}\{L(f_{\hat{\theta}_\ell}) < L(f_{\hat{\theta}_{(\ell+u)}})\}$ and $\Delta_{KL} = \mathbb{1}\{KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_\ell}) < KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_{\ell+u}})\}$. According to the theorem, as $\ell \to \infty$ and $\ell/u \to 0$, $\Delta_L - \Delta_{KL}$ converges to 0 in probability. This is demonstrated in the top two plots in Fig. 3. The two blue curves represent kernel density estimates of $\Delta_L - \Delta_{KL}$. As ℓ increases from 15 to 25 (from the left column to the right column) and at the same time the ratio $\frac{\ell}{u}$ decreases from 0.3 to 0.0125, corresponding to the limit conditions $\ell \to \infty$ and $\ell/u \to 0$, the estimated density function of $\Delta_L - \Delta_{KL}$ is concentrated at 0 for the first case, suggesting $\Delta_L - \Delta_{KL} = 0$ with high probability, and becomes even more so in the second case.

5.3 Normal Approximation

The reason we have been using Kullback-Leibler divergence as a measure of discrepancy is its connection with maximum likelihood estimation. But Kullback-Leibler divergence is not the only option. In fact, from the theorem, it is quite clear that what is important is the "closeness" to $f_{\theta^*_{sup}}$, whether the "closeness" is measured by Kullback-Leibler divergence on a functional space or something else on another space is not strictly fixed. Since the calculation of probabilities involving KL is particularly difficult, here we simplify the discussion from a functional space to the parameter space by approximating the Kullback-Leibler divergence by a rescaled L2-norm.

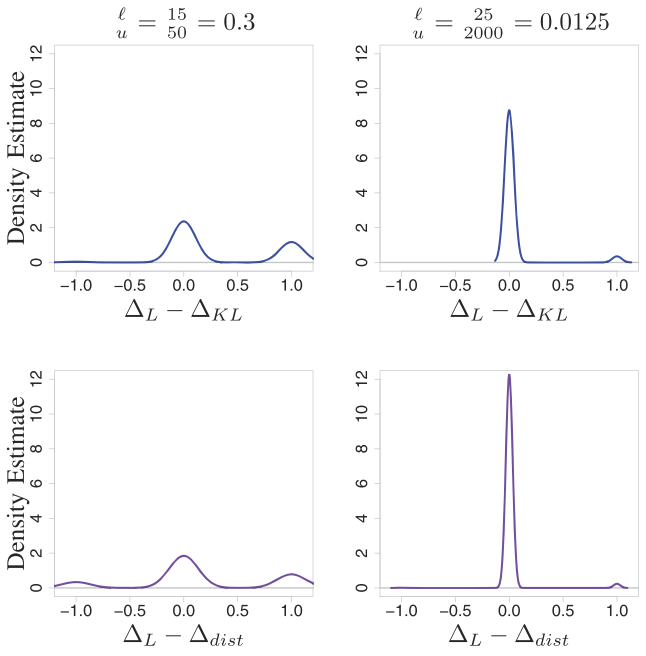


Fig. 3. Empirical asymptotic distributions of $\Delta_L - \Delta_{KL}$, and $\Delta_L - \Delta_{dist}$.

We have defined a Kullback-Leibler divergence ball as

$$B_{KL}(\theta^*, \epsilon) = \{\theta \in \Theta : KL(f_{\theta^*} \parallel f_\theta) < \epsilon\}.$$

"For sufficiently regular models, in a small enough region around the 'true' distribution f_{θ^*}, the Kullback-Leibler divergence as a function of θ behaves like 'rescaled' Euclidean distance with the particular rescaling depending on θ^*" [12]. That is, a KL ball can be well approximated by a ball defined by a rescaled L2-norm neighborhood

$$B(\theta^*, \epsilon) = \{\theta \in \Theta : (\theta^* - \theta)^T I(\theta^*)(\theta^* - \theta) < \epsilon\},$$

where $I(\theta^*)$ is the Fisher information matrix evaluated at θ^*. With this in mind, now we derive a formula to approximate the probability of degradation for large samples. Even though this formula is for our specific example, it can be easily generalized to be used for problems of any mixtures of exponential family densities. Assume $L(f_{\theta^*_{sup}}) < L(f_{\theta^*_{unsup}})$, then with $\ell \gg 1$, we have,

$$\begin{aligned}
&\lim_{\ell\to\infty, \ell/u\to 0} P[L(\hat{\theta}_\ell) < L(\hat{\theta}_{\ell+u})] \\
&= P\{KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_\ell}) < KL(f_{\theta^*_{sup}} \parallel f_{\theta^*_{unsup}})\} \\
&\approx \int \mathbb{1}\{KL(f_{\theta^*_{sup}} \parallel f_t) < KL(f_{\theta^*_{sup}} \parallel f_{\theta^*_{unsup}})\} \cdot f_{\hat{\theta}_\ell}(t)\, dt \\
&\approx \int \mathbb{1}\{(t - \theta^*_{sup})^T I(\theta^*_{sup})(t - \theta^*_{sup}) \\
&< (\theta^*_{unsup} - \theta^*_{sup})^T I(\theta^*_{sup})(\theta^*_{unsup} - \theta^*_{sup})\} \cdot f_{\hat{\theta}_\ell}(t)\, dt.
\end{aligned}$$

Let N and MVN denote the normal distribution and multivariate normal distribution respectively. For our example, the large sample approximation distribution of $\hat{\theta}_\ell = (\hat{\pi},\ \hat{\mu}_1,\ \hat{\sigma}_1^2,\ \hat{\mu}_2,\ \hat{\sigma}_2^2)$ is given by

$$\hat{\pi} \sim N\left(\pi_1, \frac{\pi_1(1-\pi_1)}{\ell_1}\right),$$

$$\begin{bmatrix}\hat{\mu}_1 \\ \hat{\sigma}_1^2\end{bmatrix} \sim MVN\left(\begin{bmatrix}\mu_1 \\ \sigma_1^2\end{bmatrix}, \begin{bmatrix}\frac{\sigma_1^2}{\ell_1} & 0 \\ 0 & \frac{2\sigma_1^4}{\ell_1}\end{bmatrix}\right),$$

$$\begin{bmatrix}\hat{\mu}_2 \\ \hat{\sigma}_2^2\end{bmatrix} \sim MVN\left(\begin{bmatrix}\mu_2^* \\ \sigma_2^{2*}\end{bmatrix}, \frac{C(\mu_2^*, \sigma_2^{2*})}{\ell_2}\right),$$

where

$$C(\mu_2^*, \sigma_2^{2*}) := \begin{bmatrix}\frac{1}{\sigma_2^{2*}} & \frac{\beta_1}{2\sigma_2^{3*}} \\ \frac{\beta_1}{2\sigma_2^{3*}} & \frac{(\beta_2-1)}{4\sigma_2^{4*}}\end{bmatrix},$$

and β_1 and β_2 are the skewness and kurtosis, respectively. For X following a finite mixture of Gaussian distribution $f_\psi = \sum_{j=1}^K \pi_j f(x|\theta_j)$, the skewness and kurtosis are given by

$$\beta_1 = \frac{1}{\sigma^3}\sum_{j=1}^K \pi_j(\mu_j - \mu)\cdot\left[3\sigma_j^3 + (\mu_j-\mu)^2\right],$$

$$\beta_2 = \frac{1}{\sigma^3}\sum_{j=1}^K \pi_j\left[3\sigma_j^4 + 6(\mu_j-\mu)^2\sigma_j^2 + (\mu_j-\mu)^4\right],$$

where μ and σ are the overall mean and overall standard deviation. Putting everything together, the large sample approximation distribution of $\hat{\theta}_\ell = (\hat{\pi},\ \hat{\mu}_1,\ \hat{\sigma}_1^2,\ \hat{\mu}_2,\ \hat{\sigma}_2^2)$ is given by

$$\begin{bmatrix}\hat{\pi} \\ \hat{\mu}_1 \\ \hat{\sigma}_1^2 \\ \hat{\mu}_2, \\ \hat{\sigma}_2^2\end{bmatrix} \sim MVN\left(\begin{bmatrix}\pi_1 \\ \mu_1 \\ \sigma_1^2 \\ \mu_2^* \\ \sigma_2^{2*}\end{bmatrix}, \hat{\Sigma}\right),$$

where

$$\mu_2^* = .5*(\mu_2+\mu_3),$$

$$\sigma_2^* = \sqrt{.5*\left(\sigma_2^2+\sigma_3^2\right) + .25*\mu_2^2 + .25*\mu_3^2 - .5*\mu_2*\mu_3},$$

and

$$\hat{\Sigma} = \begin{bmatrix}\frac{\pi_1(1-\pi_1)}{\ell_1} & 0 & 0 & 0 & 0 \\ 0 & \frac{\sigma_1^2}{\ell_1} & 0 & 0 & 0 \\ 0 & 0 & \frac{2\sigma_1^4}{\ell_1} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sigma_2^{2*}} & \frac{\beta_1}{2\sigma_2^{3*}} \\ 0 & 0 & 0 & \frac{\beta_1}{2\sigma_2^{3*}} & \frac{(\beta_2-1)}{4\sigma_2^{4*}}\end{bmatrix}.$$

TABLE 1
Approximation of the Probability of Semi-Supervised Degradation

μ_3	3	4	5	6	7	8	9
$\Delta_L/10^3$	0.19	0.18	0.22	0.30	0.99	1.00	1.00
$\Delta_{KL}/10^3$	0.00	0.00	0.00	0.01	0.97	1.00	1.00
$\Delta_{dist}/10^3$	0.01	0.01	0.01	0.04	0.98	1.00	1.00

Here, $(\hat{\theta}_\ell - \theta_{sup}^*)^T I(\theta_{sup}^*)(\hat{\theta}_\ell - \theta_{sup}^*)$ is considered as a good local estimate of $KL(f_{\theta_{sup}^*} \parallel f_{\hat{\theta}_\ell})$ as well as a scaled variance-covariance matrix.

Let $\Delta_{dist} = \mathbb{1}\{(\hat{\theta}_\ell - \theta_{sup}^*)^T I(\theta_{sup}^*)(\hat{\theta}_\ell - \theta_{sup}^*) < (\hat{\theta}_{\ell+u} - \theta_{sup}^*)^T I(\theta_{sup}^*)(\hat{\theta}_{\ell+u} - \theta_{sup}^*)\}$. The bottom two plots in Fig. 3 show the result of normal approximation. The two purple curves represent kernel density estimates of $\Delta_L - \Delta_{dist}$. Again, as ℓ increases from 15 to 25 (from the left column to the right column), and at the same time the ratio $\frac{\ell}{u}$ decreases from 0.3 to 0.0125, corresponding to the limit conditions $\ell \to \infty$ and $\ell/u \to 0$, the estimated density function of $\Delta_L - \Delta_{dist}$ is concentrated at 0 for the first case, suggesting $\Delta_L - \Delta_{dist} = 0$ with high probability, and becomes even more so in the second case. Δ_{dist} gives a pretty good estimate for Δ_{KL}, and it too captures the degradation behavior of Bayes plug-in classifier in the semi-supervised learning setting.

5.4 Probability of Semi-Supervised Degradation

The Semi-Supervised Degradation Theorem, together with the normal approximation, provides two different ways to approximate the probability of semi-supervised degradation, the Kullback-Leibler divergence method and the rescaled L2-norm method.

With the same example, we consider $\mu_3 \in [3, 9]$. According to the theorem, a large ℓ and a small ℓ/u ratio is required to achieve a good approximation. Table 1 shows the approximation result from 1,000 simulations with $\ell = 25$ and $u = 2{,}000$.

The approximation result shows that the Kullback-Leibler divergence method and the rescaled L2-norm method approximate the probability of semi-supervised degradation well for large μ_3. The poor approximation for other cases is due to the fact that when θ_{sup}^* and θ_{unsup}^* are close to each other, a good approximation can only be reached by using a significantly larger ℓ and a significantly smaller ℓ/u ratio (in order to shrink the $B_{KL}(\theta_{sup}^*, \epsilon_1)$ and $B_{KL}(\theta_{unsup}^*, \epsilon_2)$ to be extremely small relative to the distance between θ_{sup}^* and θ_{unsup}^*).

Note that when θ_{sup}^* and θ_{unsup}^* are close to each other, degradation occurs with only a small probability. Therefore in this experiment, even though it seems that the approximation can be bad for some cases, it is accurate for the cases that really matter, i.e., when semi-supervised classification degrades the performance, which are exactly the cases we need to detect. The values of ℓ and ℓ/u needed for a good approximation depend on the problem at hand.

6 Discussion on Finite Sample Cases

The conditions of the theorem involve taking limits. In this section, we discuss what happens with finite samples, as we see in practice.

6.1 Finite Sample Semi-Supervised Classification

Consider the example from Section 5 with a slight modification: Instead of fixing μ_3 as 7, let it take the values 4, 7, and 10. We investigate the semi-supervised error rates for these three cases. Different ℓ values are 15, 50, 100, 300, and u is increased from 0 to 2,000 in increments of 50;

$$f_0(x) = \frac{1}{2}\phi(x \mid 0,1) + \frac{1}{2}\left[\frac{1}{2}\phi(x \mid 3,1) + \frac{1}{2}\phi(x \mid \mu_3, 1)\right].$$

By the theorem, if $L(f_{\theta^*_{sup}}) < L(f_{\theta^*_{unsup}})$, semi-supervised classification will degrade the performance for sure as $\ell \to \infty$, and $\ell/u \to 0$. However, in the case where the limit conditions are not satisfied, we expect to see that for any fixed ℓ, increasing the number of unlabeled samples will either improve or degrade the classifier performance, depending on the bias and variance trade-off. For the case where $\mu_3 = 4$, the phenomenon that semi-supervised classification mistakenly combines the first and second Gaussian densities and categorizes them as class 1 will not happen. The reason is obvious; $\mu_2 = 3$ is much closer to $\mu_3 = 4$ than to $\mu_1 = 0$, whereas for the other two cases where $\mu_3 = 7$ and $\mu_3 = 10$, semi-supervised classification makes that mistake and it is more true for $\mu_3 = 10$ than for $\mu_3 = 7$. Therefore, when $\mu_3 = 4$, unlabeled observations are more likely to help the classification since, given a fixed model bias, the classifier benefits from the reduced estimation variance brought by a larger sample size more than it gets hurt by a slight estimation bias. In the other two cases, $\mu_3 = 7$ and $\mu_3 = 10$, unlabeled observations have a larger probability of degrading the classification performance since, compared to the estimation bias caused by unlabeled data, the reduced estimation variance is less significant. All this is demonstrated by plots in Fig. 4.

6.2 Semi-Supervised Degradation Point

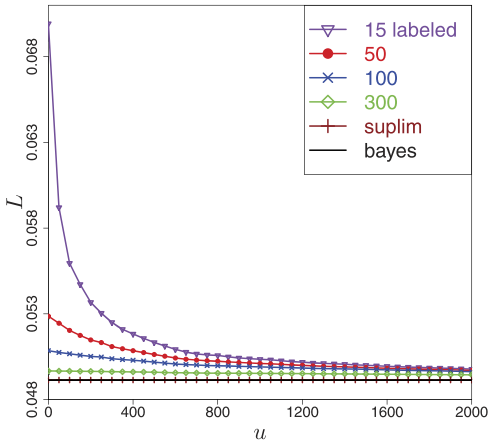
From the previous experiment, we see that as we keep the same model $\mathcal{F}$ and vary the truth from being "not so wrongly approximated by $\mathcal{F}$" to "very wrongly approximated by $\mathcal{F}$," semi-supervised classification breaks down. Clearly there is a μ_3 value at which semi-supervised classification reaches the degradation point.

Consider $\mu_3 \in [3, 10]$. Since the estimation is based on the mixtures of the two Gaussian model, the model is correct for $\mu_3 = 3$, but incorrect for the rest of the μ_3 values. Suppose the labeled training set is fixed to be of size 15, a series of Bayes plug-in classifiers are learned with samples (labeled and unlabeled) distributed according to f_0 with μ_3 varying from 3 to 10.

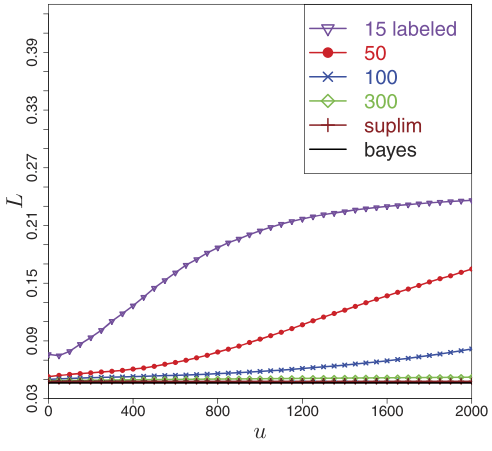
In Fig. 5 we plot the following three curves: The curve at the bottom (brown) is the supervised limit error rate, i.e., $L(f_{\theta^*_{sup}})$, the curve in the middle and the one at the top (purple) are the supervised error rate with $\ell = 15$, i.e., $L(f_{\hat{\theta}_{15}})$, and the semi-supervised error rate with $\ell = 15$ and $u = 2{,}000$, i.e., $L(f_{\hat{\theta}_{15+2,000}})$. For the sake of a more meaningful comparison, we plot the difference between these error rates and the Bayes error L^* (the purpose is to capture the pattern in error rates with the change in the true density filtered out).

This investigation empirically reveals a critical value $\tilde{\mu}_3$ for which, when $\mu_3 < \tilde{\mu}_3$, semi-supervised learning improves the classifier performance regardless of the model assumption incorrectness, while semi-supervised learning degrades the classifier performance when $\mu_3 \geq \tilde{\mu}_3$.

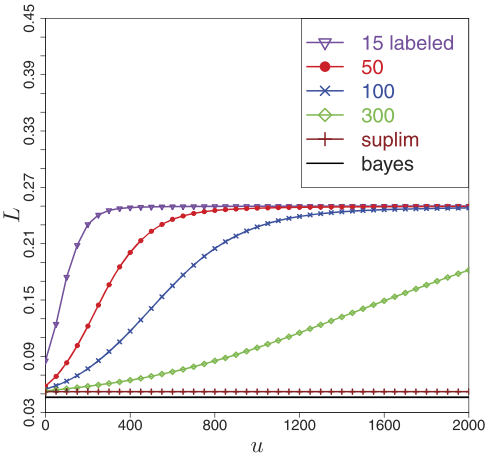
Intuitively, $\mu_3 = 6$ makes $\mu_2 = 3$ equally close to $\mu_1 = 0$ and $\mu_3 = 6$, which suggests that degradation will start to happen for some value slightly bigger than $\mu_3 = 6$, and when exactly the semi-supervised classification breaks down should also depend on ℓ. We will call $\mu_3 = 6$ the



(a) $\mu_3 = 4$



(b) $\mu_3 = 7$



(c) $\mu_3 = 10$

Fig. 4. Semi-supervised misclassification rate.

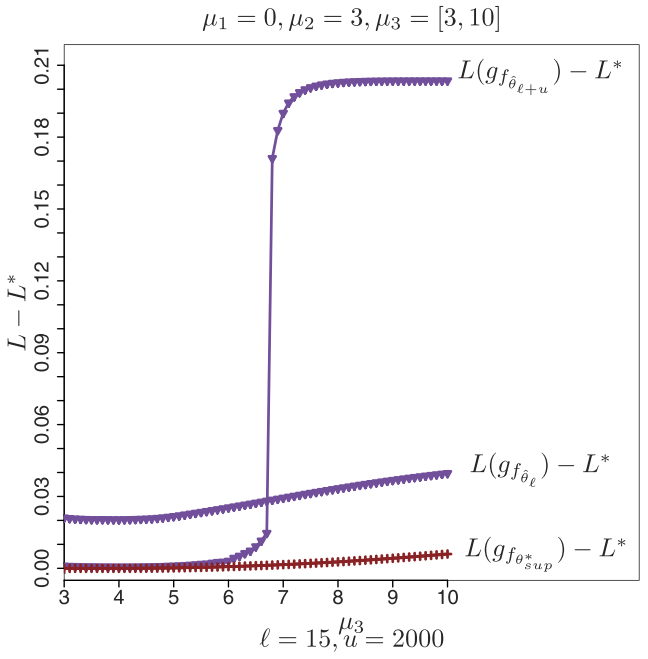


Fig. 5. Semi-supervised degradation point.

semi-supervised degradation point for this example. Fig. 5 confirms this intuition: Degradation starts to happen at $\mu_3 = 6.8$.

In this experiment, $L(f_{\theta^*_{sup}}) < L(f_{\theta^*_{unsup}})$, and in fact, $L(f_{\theta^*_{sup}})$ is close to L^*. Under the same model $\mathcal{F}$, the more wrong the model is to approximate the truth, the further away the two MLE limits are, and from one point on, semi-supervised degradation happens with high probability. If the Kullback-Leibler divergence between the truth f_0 and supervised limit density $f_{\theta^*_{sup}}$ (approximately optimal in the model) is used to measure the degree of model mismatch, then we have

$$KL(f_{\theta_{01}} \parallel f_{\theta^*_{sup1}}) < KL(f_{\theta_{02}} \parallel f_{\theta^*_{sup2}}) \Longrightarrow \parallel \theta^*_{sup1} - \theta^*_{unsup1} \parallel < \parallel \theta^*_{sup2} - \theta^*_{unsup2} \parallel .$$

The further away the two limits are, the higher the semi-supervised probability of degradation is (given fixed ℓ and u).

In general, we believe that for every situation, there exists a semi-supervised degradation point and a corresponding degree of model mismatch. For example, at $\mu_3 = 6$, the degree of model mismatch at which semi-supervised classification reaches the degradation point is given by

$$KL(f_{\theta_0} \parallel f_{\theta^*_{sup}}) = 0.026.$$

We can use the degree of model mismatch to determine "when" degradation will happen, but for every different situation, the exact value of $KL(f_{\theta_0} \parallel f_{\theta^*_{sup}})$ at the semi-supervised degradation point should be different since it depends on the complexity of the distributions in discussion.

7 Conclusion

In this paper, the performance of semi-supervised classification is investigated for the classical method: generative models, independent observations, Bayes plug-in classifier, and maximum likelihood classification. Under the framework introduced in this paper, in the limit ($\ell \to \infty$ and $\ell/u \to 0$), a necessary and sufficient condition for a model to suffer performance degradation with unlabeled data is $L(f_{\theta^*_{sup}}) < L(f_{\theta^*_{unsup}})$. With ℓ large enough, and ℓ/u small enough at the same time, the probability of degradation $P\{L(f_{\hat{\theta}_\ell}) < L(f_{\hat{\theta}_{(\ell+u)}})\}$ can be well approximated by $P\{KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_\ell}) < KL(f_{\theta^*_{sup}} \parallel f_{\hat{\theta}_{\ell+u}})\}$ provided the two MLE limits, θ^*_{sup} and θ^*_{unsup}, differ significantly.

It is obvious that when the true density f_0 is unknown, none of the quantities in the theorem is calculable; therefore, the theorem cannot be directly used in practice. The philosophy we are trying to convey here is that by taking one step back from practice, it is possible to learn the behavior of a particular classifier in a particular model in the semi-supervised learning setting through the theorem and simulated examples, in which we know the truth f_0. Having this clear understanding of the model and classifier used before applying them to real data semi-supervised analysis is only a solid first step towards using them well.

Because of the restricted framework used in the paper, the necessary and sufficient condition does not necessarily generalize to other approaches. However, due to the complexity of the problem, our realistic goal has been to provide some analytical tool or even simply some theoretical perspective that sheds some light on when and why semi-supervised learning degrades the performance in misspecified models. A similar approach/perspective may be applied to other frameworks, with the modifications conforming to the specific framework at hand.

References

[1] J. Ratsaby and S. Venkatesh, "Learning from a Mixture of Labeled and Unlabeled Examples with Parametric Side Information," *Proc. Eighth Ann. Conf. Computational Learning Theory*, p. 417, 1995.

[2] V. Castelli and T.M. Cover, "On the Exponential Value of Labeled Samples," *Pattern Recognition Letters*, vol. 16, pp. 105-111, 1995.

[3] V. Castelli and T. Cover, "The Relative Value of Labeled and Unlabeled Samples in Pattern Recognition with an Unknown Mixing Parameter," *IEEE Trans. Information Theory*, vol. 42, no. 6, pp. 2102-2117, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=556600, Nov. 1996.

[4] T. Zhang and F.J. Oles, "A Probability Analysis on the Value of Unlabeled Data for Classification Problems," *Proc. 17th Int'l Conf. Machine Learning*, pp. 1191-1198, 2000.

[5] F. Cozman and I. Cohen, "Semi-Supervised Learning of Classifiers: Theory, Algorithms, and Their Application to Human-Computer Interaction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 12, pp. 1553-1567, Dec. 2004.

[6] F. Cozman and I. Cohen, "Risks of Semi-Supervised Learning: How Unlabeled Data Can Degrade Performance of Generative Classifiers," *Semi-Supervised Learning*, vol. 4, pp. 57-72, 2006.

[7] H. White, "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, vol. 50, no. 1, pp. 1-25, 1982.

[8] D.M. Titterington, A.F.M. Smith, and U.E. Makov, *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, 1987.

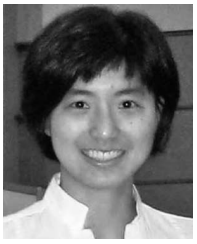
[9] L.F. James, C.E. Priebe, and D.J. Marchette, "Consistent Estimation of Mixture Complexity," *The Annals of Statistics*, vol. 29, no. 5, pp. 1281-1296, 2001.

[10] R. Redner and H. Walker, "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Rev.*, vol. 26, no. 2, pp. 195-239, 1984.

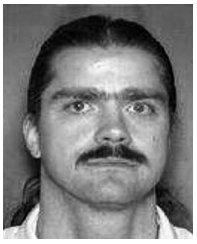
[11] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, first ed. Springer-Verlag, 1996.

[12] P.D. Grünwald, *The Minimum Description Length Principle*, first ed. The MIT Press, 2007.

[13] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, http://www.R-project.org, 2009.



Ting Yang received the BS degree in management information systems from Beijing Materials University in 2003, the MS degree in mathematics from East Carolina University in 2005, and enrolled in the applied mathematics and statistics PhD program at Johns Hopkins University in 2005.



Carey E. Priebe received the BS degree in mathematics from Purdue University in 1984, the MS degree in computer science from San Diego State University in 1988, and the PhD degree in information technology (computational statistics) from George Mason University in 1993. From 1985 to 1994, he worked as a mathematician and scientist in the US Navy Research and Development Laboratory system. Since 1994 he has been a professor in the Department of Applied Mathematics and Statistics, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland. At Johns Hopkins, he holds joint appointments in the Department of Computer Science, the Department of Electrical and Computer Engineering, the Center for Imaging Science, the Human Language Technology Center of Excellence, and the Whitaker Biomedical Engineering Institute. He is a a past president of the Interface Foundation of North America-Computing Science and Statistics, a past chair of the American Statistical Association Section on Statistical Computing, a past vice president of the International Association for Statistical Computing, and on the editorial boards of the *Journal of Computational and Graphical Statistics*, *Computational Statistics and Data Analysis*, and *Computational Statistics*. His research interests include computational statistics, kernel and mixture estimates, statistical pattern recognition, statistical image analysis, dimensionality reduction, model selection, and statistical inference for high-dimensional and graph data. He is a research professor in the National Security Institute at the Naval Postgraduate School, and was named one of six inaugural National Security Science and Engineering faculty fellows. He is a senior member of the IEEE, a lifetime member of the Institute of Mathematical Statistics, an elected member of the International Statistical Institute, and a fellow of the American Statistical Association.