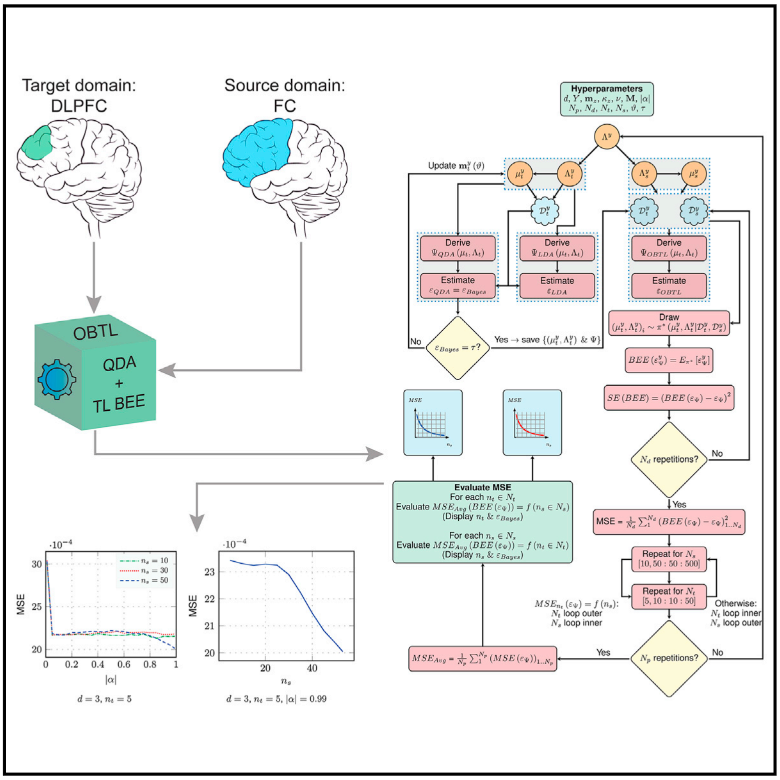
Article

Patterns

Robust importance sampling for error estimation in the context of optimal Bayesian transfer learning

Graphical abstract


Target domain:
DLPFC
Source domain:
FC
OBTL
QDA
+
TL BEE
Hyperparameters
$d, Y, \mathbf{m}_z, \kappa_z, \nu, \mathbf{M}, |\alpha|$
$N_p, N_d, N_t, N_s, \vartheta, \tau$
Λ^y
Update $\mathbf{m}_t^y(\vartheta)$
μ_t^y
Λ_t^y
Λ_s^y
μ_s^y
$\mathcal{D}_t^y$
$\mathcal{D}_s^y$
Derive $\Psi_{QDA}(\mu_t, \Lambda_t)$
Estimate $\varepsilon_{QDA} = \varepsilon_{Bayes}$
Derive $\Psi_{LDA}(\mu_t, \Lambda_t)$
Estimate ε_{LDA}
Derive $\Psi_{OBTL}(\mu_t, \Lambda_t)$
Estimate ε_{OBTL}
No
$\varepsilon_{Bayes} = \tau$?
Yes → save $\{(\mu_t^y, \Lambda_t^y)\ \&\ \Psi\}$
Draw $(\mu_t^y, \Lambda_t^y)_i \sim \pi^*(\mu_t^y, \Lambda_t^y | \mathcal{D}_t^y, \mathcal{D}_s^y)$
$BEE(\varepsilon_\Psi^y) = E_{\pi^*}[\varepsilon_\Psi^y]$
$SE(BEE) = (BEE(\varepsilon_\Psi) - \varepsilon_\Psi)^2$
N_d repetitions?
No
Yes
MSE = $\frac{1}{N_d}\sum_1^{N_d}(BEE(\varepsilon_\Psi) - \varepsilon_\Psi)^2_{1..N_d}$
Repeat for N_s $[10, 50 : 50 : 500]$
Repeat for N_t $[5, 10 : 10 : 50]$
$MSE_{n_t}(\varepsilon_\Psi) = f(n_s)$:
N_t loop outer
N_s loop inner
Otherwise:
N_t loop inner
N_s loop outer
N_p repetitions?
Yes
No
$MSE_{Avg} = \frac{1}{N_p}\sum_1^{N_p}(MSE(\varepsilon_\Psi))_{1..N_p}$
MSE
n_s
Evaluate MSE
For each $n_t \in N_t$
Evaluate $MSE_{Avg}(BEE(\varepsilon_\Psi)) = f(n_s \in N_s)$
(Display n_t & ε_{Bayes})
For each $n_s \in N_s$
Evaluate $MSE_{Avg}(BEE(\varepsilon_\Psi)) = f(n_t \in N_t)$
(Display n_s & ε_{Bayes})
$\cdot 10^{-4}$
MSE
30
25
20
0
0.2
0.4
0.6
0.8
1
$|\alpha|$
$n_s = 10$
$n_s = 30$
$n_s = 50$
$d = 3, n_t = 5$
$\cdot 10^{-4}$
MSE
23
22
21
20
20
40
n_s
$d = 3, n_t = 5, |\alpha| = 0.99$

Highlights

- A transfer learning (TL) framework for Bayesian error estimation (BEE) is proposed
- Relatedness between domains is modeled by a joint prior in a Bayesian paradigm
- TL-based BEE can leverage data from other relevant domains to improve accuracy
- Data from domains with moderate to high relatedness can improve BEE outcomes

Authors

Omar Maddouri, Xiaoning Qian, Francis J. Alexander, Edward R. Dougherty, Byung-Jun Yoon

Correspondence

bjyoon@ece.tamu.edu

In brief

Accurate estimation of classification error is challenging in scientific domains, where available data are limited. Although transfer of data and knowledge from relevant domains can alleviate this issue, previous studies on transfer learning have mostly focused on improving the learned models rather than enhancing the performance analysis. In this paper, we propose a transfer learning scheme for Bayesian error estimation that can leverage data from relevant domains to enhance the estimation of classification error in the domain of interest.

Maddouri et al., 2022, Patterns *3*, 100428
March 11, 2022 © 2022 The Authors.
https://doi.org/10.1016/j.patter.2021.100428

Patterns

CellPress
OPEN ACCESS

Article

Robust importance sampling for error estimation in the context of optimal Bayesian transfer learning

Omar Maddouri,[1] Xiaoning Qian,[1,2] Francis J. Alexander,[2] Edward R. Dougherty,[1] and Byung-Jun Yoon[1,2,3,*]
[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA
[2]Computational Science Initiative, Brookhaven National Laboratory, Upton, NY 11973, USA
[3]Lead contact
*Correspondence: bjyoon@ece.tamu.edu

https://doi.org/10.1016/j.patter.2021.100428

THE BIGGER PICTURE In scientific domains with limited data availability, accurate classification error estimation is practically challenging. Although transfer learning (TL) may provide a promising solution under such circumstances by learning from data available in other relevant domains, it has not been explored for enhancing error estimation. Here, we place the problem of estimating the classification error in a Bayesian paradigm and introduce a TL-based error estimator that can significantly enhance the accuracy and robustness of error estimates under data scarcity. We demonstrate that our proposed TL-based Bayesian error estimation framework effectively models and exploits the relatedness between different domains to improve error estimation. Experimental results based on both synthetic data as well as real-world data show that our proposed error estimator clearly outperforms existing error estimators, especially in a small sample setting, by tapping into the data from other relevant domains.

1 2 3 4 5 **Development/Pre-production:** Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Classification has been a major task for building intelligent systems because it enables decision-making under uncertainty. Classifier design aims at building models from training data for representing feature-label distributions—either explicitly or implicitly. In many scientific or clinical settings, training data are typically limited, which impedes the design and evaluation of accurate classifiers. Atlhough transfer learning can improve the learning in target domains by incorporating data from relevant source domains, it has received little attention for performance assessment, notably in error estimation. Here, we investigate knowledge transferability in the context of classification error estimation within a Bayesian paradigm. We introduce a class of Bayesian minimum mean-square error estimators for optimal Bayesian transfer learning, which enables rigorous evaluation of classification error under uncertainty in small-sample settings. Using Monte Carlo importance sampling, we illustrate the outstanding performance of the proposed estimator for a broad family of classifiers that span diverse learning capabilities.

INTRODUCTION

Transfer learning (TL) provides promising means to repurpose the data and/or scientific knowledge available in other relevant domains for new applications in a given domain. The ability to transfer relevant data/knowledge across different domains practically enables learning effective models in target domains with limited data. Classifier design can take advantage of TL to address small-sample challenges we often face in various scientific applications. However, rigorous error estimators that can leverage such transferred data/knowledge for better estimation of classification error have been missing to date, which makes the design framework epistemologically incomplete.[1] Generally, the scientific validity of any predictive model is assessed by the ability to generalize outside the observed training sample. However, the available sample is often too small in many scientific applications (e.g., bio-marker discovery) to hold out sufficient data just for testing purpose, which makes the reuse of training data for both classifier design and error estimation inevitable. While various error estimation schemes exist to date, their accuracy and reliability in a small-sample setting are often questioned.[2] For instance, in Dalton and Dougherty[3] many classification studies of cancer gene expression data have been listed where the performance was assessed by cross-validation (CV) based on small-size training datasets. Analyses in Braga-Neto and Dougherty[4] have shown that CV error estimators derived based on small-size samples show large variance, which explains the controversy across many biological studies that relied on data-driven CV.[5] Model-based error estimation also faces practical challenges as non-informative modeling assumptions may mislead the error estimators in case of model mismatch.

The ability for accurate error estimation based on small samples is also critical in other contexts, an example being continual learning,[6] where a series of labeled datasets are sequentially fed to the learner as in realistic learning scenarios. In recent years, continual learning regained attention as a promising strategy for avoiding "catastrophic forgetting" that may arise when the training data are split for a series of small learning operations called tasks.[7] Such a continual learning setting is becoming prevalent these days, where retaining the observed training data is either undesirable (confidentiality) or intractable (high-throughput systems), and developing reliable task-specific error estimators is indispensable. For instance, an intuitive approach to continual learning from a Bayesian perspective is to leverage the posterior of the current task to update the prior of the next task.[8] However, analysis in Farquhar and Gal[9] has shown that evaluation approaches for this prior-focused setup suffer from severe bias in realistic scenarios, particularly for finely partitioned data. Recent work in Goodfellow et al.[10] provided a solution for test data scarcity by reusing the same test set in the context of a continuously evolving classification problem. To avoid overfitting the test data, the authors employed a reusable holdout mechanism based on the area under the receiver operating characteristic curve metric. Nevertheless, this approach remains contingent on the availability of an independent test set. For these reasons, there is a pressing need to develop novel error estimators that can effectively overcome data scarcity limitations. For assessing different classification models in the context of small-size training datasets, having an accurate error estimator with TL capabilities that can take advantage of relevant datasets in other domains would be highly beneficial. Such an estimator would be readily applicable to continual learning as cross-task datasets can be seen as related source-target samples.

In the next sections, we provide a brief review of the standard error estimation techniques along with prevalent TL scenarios. A more comprehensive review can be found in the supplemental information, sections 3 and 5.

For unknown feature-label distributions, the classification error of a given classifier is typically estimated by leveraging a large sample collected from the true distribution. However, limiting factors, such as the excessive cost of large-scale data acquisition, make it often infeasible to collect and hold out large test sets. Consequently, the available small-size sample may have to be used for both training and evaluating the classifier, and researchers have strived to devise practical methods for accurate error estimation. Existing error estimation schemes can be broadly categorized into parametric and non-parametric methods. Non-parametric estimators compute the error rate by counting the misclassified points, where widely used estimators include the resubstitution, CV, and bootstrap estimators. Parametric methods include the popular plug-in estimator that naively estimates the true error from an empirical model. The Bayesian minimum mean-square error estimator (BEE) proposed in Dalton and co-workers[3,11] is another benchmark parametric estimator that significantly enhances the robustness by computing the expected true error with respect to the posterior of the model parameters. The BEE has shown notable improvements over standard estimators as it effectively handles the uncertainty about the underlying feature-label distribution.[3,11]

Recently, TL has emerged as an alternative to provide remedies for pitfalls caused by training data scarcity in a target domain by utilizing available data from different yet relevant source domains.[12] Based on the properties of source and target domains, two scenarios of TL may arise. The first one, commonly known as "homogeneous TL," occurs when the source and target domains share the same feature space. The second scenario is called "heterogeneous TL" and is considered when differences exist between domains in terms of their feature space or data dimensionality. In practice, the most common setting for TL, known also as domain adaptation, assumes similar families of feature-label distributions across domains.

In this study, we propose a TL framework for robust estimation of classification error based on a rigorous Bayesian paradigm. To the best of our knowledge, this study is the first work on TL-based BEE, which can significantly enhance our understanding of transferability across domains in the context of error estimation. Building on the Bayesian transfer learning framework proposed in Karbalayghareh et al.,[13] we introduce a TL-based BEE estimator that can enhance the error estimation accuracy in the target domain by utilizing the data available in a relevant source domain based on the joint prior of their feature-label distributions. We present a rigorous study of error estimation in the context of Bayesian TL and show that our proposed TL-based BEE effectively represents and exploits the relatedness (or dependency) between different domains to improve error estimation in a challenging small-sample setting, where the number of observed data points from the target domain of interest is in the range of 5–50. For applicability of the proposed TL-based BEE estimator in real-world problems for arbitrary classifiers, we introduce an efficient and robust importance sampling setup with control variates where the importance density and the control variates function are carefully defined to reduce the variance of the estimator while keeping the overall sampling process computationally feasible and scalable. For this purpose, we utilize Laplace approximations for fast evaluation of matrix-variate confluent and Gauss hypergeometric functions. The performance of the TL-based BEE estimator is extensively evaluated using both synthetic datasets as well as real-world biological datasets. As our main focus in this study is the estimation of classification error, we consider a variety of existing classifiers with different levels of learning capabilities to demonstrate the general applicability of our TL-based BEE estimation scheme. We also show the outstanding performance of the proposed estimator with respect to standard error estimation techniques that are commonly used.

RESULTS AND DISCUSSION

Overview of the proposed Bayesian error estimation via TL

We propose a class of Bayesian minimum mean-square error (MMSE) estimators for TL where the observed sample is a mixture of source and target data. The basic classification setting and a brief review of the standard BEE estimator are presented in the supplemental information, sections 2 and 4. For symbols and notations, see Table S1.

Rooted in signal estimation, the BEE has been motivated by optimal filtering for functions of random variables.[3] For a function of two random variables $g(X, Y)$, the optimal estimator $\widehat{g}(Y)$ of a filter $g(Y)$ after observing only Y in the mean-square sense is given by

$$\widehat{g}(Y) = E_X[g(X, Y)|Y]. \qquad \text{(Equation 1)}$$

Replacing X with the parameter vector θ of the feature-label distribution and Y by the sample S_n (of size n), leads to the standard BEE that has been introduced in Dalton and Dougherty[3] as

$$\widehat{\varepsilon}(S_n) = E_{\boldsymbol{\theta}}[\varepsilon_n(\boldsymbol{\theta}, S_n|S_n)]. \qquad \text{(Equation 2)}$$

In TL, the sample S_n is a mixture of source and target data such that $S_n = (\mathcal{D}_s \cup \mathcal{D}_t)_n$, with $n = N_s + N_t$, and the classifier ψ_n is designed either on $\mathcal{D}_t$, $\mathcal{D}_s$, or $\mathcal{D}_s \cup \mathcal{D}_t$. We note that $\mathcal{D}_s$ and $\mathcal{D}_t$ are two labeled datasets from the source and target domains with sizes N_s and N_t, respectively (see Bayesian TL framework for binary classification, for generation details). This requires close attention as the TL-based BEE is valid only for fixed classifiers given the sample. This assumption carries limitations. For instance, classifiers that are only fixed given $\mathcal{D}_t$ but not $\mathcal{D}_s$, are not deterministic for every set of parameters estimated based on $\mathcal{D}_s \cup \mathcal{D}_t$. In this paper, we introduce the TL-based BEE defined as

$$\widehat{\varepsilon}\left((\mathcal{D}_s \cup \mathcal{D}_t)_n\right) = E_{\boldsymbol{\theta}}\left[\varepsilon_n\left(\boldsymbol{\theta}, (\mathcal{D}_s \cup \mathcal{D}_t)_n\right) \middle| (\mathcal{D}_s \cup \mathcal{D}_t)_n\right], \qquad \text{(Equation 3)}$$

where $\theta = [\theta_t, \theta_s]$ denotes the parameter vector of the joint model formed by the target parameters θ_t and the source parameters θ_s. For a fixed classifier given $(\mathcal{D}_s \cup \mathcal{D}_t)_n$, this estimator is optimal on average in the mean-square sense and unbiased when averaged over all parameters and samples. For classification in the target domain, the posterior density $\pi^*(\theta)$ reduces to the posterior of the target parameters after observing the target and source data and takes the form

$$\pi^*(\theta_t) = \pi^*(\theta_t|\mathcal{D}_s, \mathcal{D}_t), \qquad \text{(Equation 4)}$$

where $\pi^*(\theta_t|\mathcal{D}_s, \mathcal{D}_t)$ is obtained by marginalizing out the source domain parameters. Ultimately, the BEE for TL takes the form

$$\begin{aligned} \widehat{\varepsilon}\left((\mathcal{D}_s \cup \mathcal{D}_t)_n\right) &= E_{\boldsymbol{\theta}_t}\left[\varepsilon_n\left(\boldsymbol{\theta}_t, (\mathcal{D}_s \cup \mathcal{D}_t)_n\right) \middle| (\mathcal{D}_s \cup \mathcal{D}_t)_n\right] \\ &= E_{\pi^*(\boldsymbol{\theta}_t)}\left[\varepsilon_n\left(\boldsymbol{\theta}_t, (\mathcal{D}_s \cup \mathcal{D}_t)_n\right)\right]. \end{aligned} \qquad \text{(Equation 5)}$$

For the sake of simplicity we write

$$\widehat{\varepsilon} = E_{\pi^*}[\varepsilon_n], \qquad \text{(Equation 6)}$$

where $\pi^* = \pi^*(\boldsymbol{\theta}_t|\mathcal{D}_t, \mathcal{D}_s)$ denotes the posterior of the target parameters after observing the hybrid sample $\mathcal{D}_t \cup \mathcal{D}_s$.

Experiments and datasets

To evaluate the performance of the proposed error estimator, we consider the mean-square error (MSE) as a performance measure to understand the joint behavior of the classification error ε_n and its estimate $\widehat{\varepsilon}$. For the random vector $(\varepsilon_n, \widehat{\varepsilon})$, the MSE is defined as

$$\text{MSE}(\widehat{\varepsilon}) = \text{E}\left[\left|\widehat{\varepsilon} - \varepsilon_n\right|^2\right]. \qquad \text{(Equation 7)}$$

In what follows, we present an overview of the experimental setup for demonstrating the performance of the proposed TL-based BEE based on three different types of classifiers (see experimental procedures, sections 4.5 and 4.6 for more details) applied to both synthetic data as well as real-world biological datasets.

Bayesian TL framework for binary classification

We consider a binary classification problem in the context of supervised TL where there are two common classes in each domain. Let $\mathcal{D}_s$ and $\mathcal{D}_t$ be two labeled datasets from the source and target domains with sizes N_s and N_t, respectively. We are interested in the scenario where $N_t \ll N_s$. Let $\mathcal{D}_s^y = \{\mathbf{x}_{s,1}^y, \mathbf{x}_{s,2}^y, \cdots, \mathbf{x}_{s,n_s}^y\}$, $y \in \{0,1\}$, where n_s^y denotes the size of source data in class y. Likewise, let $\mathcal{D}_t^y = \{\mathbf{x}_{t,1}^y, \mathbf{x}_{t,2}^y, \cdots, \mathbf{x}_{t,n_t}^y\}$, $y \in \{0,1\}$, where n_t^y denotes the size of target data in class y. We consider a d-dimensional homogeneous transfer learning scenario where $\mathcal{D}_s$ and $\mathcal{D}_t$ are normally distributed and separately sampled from the source and target domains, respectively.

$$\mathbf{x}_z^y \sim \mathcal{N}\left(\boldsymbol{\mu}_z^y, \left(\boldsymbol{\Lambda}_z^y\right)^{-1}\right), y \in \{0,1\}, \qquad \text{(Equation 8)}$$

where $z \in \{s,t\}$, $\boldsymbol{\mu}_z^y$ is a $(d \times 1)$ mean vector in domain z for class y, and $\boldsymbol{\Lambda}_z^y$ is a $(d \times d)$ precision matrix (inverse of covariance) in domain z for label y. An augmented feature vector $\mathbf{x}^y = \begin{bmatrix} \mathbf{x}_t^y \\ \mathbf{x}_s^y \end{bmatrix}$ is a joint sample point from two related source and target domains given by

$$\mathbf{x}^y \sim \mathcal{N}\left(\boldsymbol{\mu}^y, \left(\boldsymbol{\Lambda}^y\right)^{-1}\right), y \in \{0,1\}, \qquad \text{(Equation 9)}$$

with

$$\boldsymbol{\mu}^y = \begin{bmatrix} \boldsymbol{\mu}_t^y \\ \boldsymbol{\mu}_s^y \end{bmatrix}, \boldsymbol{\Lambda}^y = \begin{bmatrix} \boldsymbol{\Lambda}_t^y & \boldsymbol{\Lambda}_{ts}^y \\ \boldsymbol{\Lambda}_{ts}^{yT} & \boldsymbol{\Lambda}_s^y \end{bmatrix}, \qquad \text{(Equation 10)}$$

where X^T denotes the transpose of matrix X. This sampling is enabled through a joint prior distribution for $\boldsymbol{\Lambda}_s^y$ and $\boldsymbol{\Lambda}_t^y$ that marginalizes out the off-diagonal block matrix $\boldsymbol{\Lambda}_{ts}^y$. Using a Gaussian-Wishart distribution as the joint prior for mean and precision matrices, the joint model factorizes as

$$p\left(\boldsymbol{\mu}_s^y, \boldsymbol{\mu}_t^y, \boldsymbol{\Lambda}_s^y, \boldsymbol{\Lambda}_t^y\right) = p\left(\boldsymbol{\mu}_s^y, \boldsymbol{\mu}_t^y \middle| \boldsymbol{\Lambda}_s^y, \boldsymbol{\Lambda}_t^y\right) p\left(\boldsymbol{\Lambda}_s^y, \boldsymbol{\Lambda}_t^y\right). \qquad \text{(Equation 11)}$$

For conditionally independent mean vectors given the covariances, the joint prior in (Equation 11) further factorizes into

Table 1. Independent schizophrenia RNA-seq datasets sampled from two different brain tissues

Disease	No. of samples		Brain region	Dataset
	Case	Control		
Schizophrenia	53	53	frontal cortex	syn4590909[14]
	262	293	DLPFC	syn2759792[16]
Total	315	346		

$$p\left(\boldsymbol{\mu}_s^y, \boldsymbol{\mu}_t^y, \boldsymbol{\Lambda}_s^y, \boldsymbol{\Lambda}_t^y\right) = p\left(\boldsymbol{\mu}_s^y \middle| \boldsymbol{\Lambda}_s^y\right) p\left(\boldsymbol{\mu}_t^y \middle| \boldsymbol{\Lambda}_t^y\right) p\left(\boldsymbol{\Lambda}_s^y, \boldsymbol{\Lambda}_t^y\right). \quad \text{(Equation 12)}$$

The block diagonal precision matrices $\boldsymbol{\Lambda}_z^y$ for $z \in \{t, s\}$ are obtained after sampling $\boldsymbol{\Lambda}^y$ from a predefined joint Wishart distribution as defined in Karbalaygphareh et al.[13] such that $\boldsymbol{\Lambda}^y \sim W_{2d}(\mathbf{M}^y, \nu^y)$, where ν^y is a hyperparameter for the degrees of freedom that satisfies $\nu^y \geq 2d$ and $\mathbf{M}^y$ is a $(2d \times 2d)$ positive definite scale matrix of the form

$$\mathbf{M}^y = \begin{pmatrix} \mathbf{M}_t^y & \mathbf{M}_{ts}^y \\ \mathbf{M}_{ts}^{yT} & \mathbf{M}_s^y \end{pmatrix}. \quad \text{(Equation 13)}$$

$\mathbf{M}_t^y$ and $\mathbf{M}_s^y$ are also positive definite scale matrices and $\mathbf{M}_{ts}$ denotes the off-diagonal component that models the interaction between source and target domains. Given $\boldsymbol{\Lambda}_z^y$, and assuming normally distributed mean vectors, we get

$$\boldsymbol{\mu}_z^y \sim \mathcal{N}\left(\mathbf{m}_z^y, \left(\kappa_z^y \boldsymbol{\Lambda}_z^y\right)^{-1}\right), z \in \{s, t\} \text{ and } y \in \{0, 1\}, \quad \text{(Equation 14)}$$

where $\mathbf{m}_z^y$ is the $(d \times 1)$ mean vector of the mean parameter $\boldsymbol{\mu}_z^y$ and κ_z^y is a positive scalar hyperparameter. The joint prior distribution $p(\boldsymbol{\Lambda}_s^y, \boldsymbol{\Lambda}_t^y)$ as derived in Karbalaygphareh et al.[13] acts like a channel through which the useful knowledge transfers from the source to the target domain, causing the posterior of the target parameters of the underlying feature-label distribution to be distributed more narrowly around the true values.

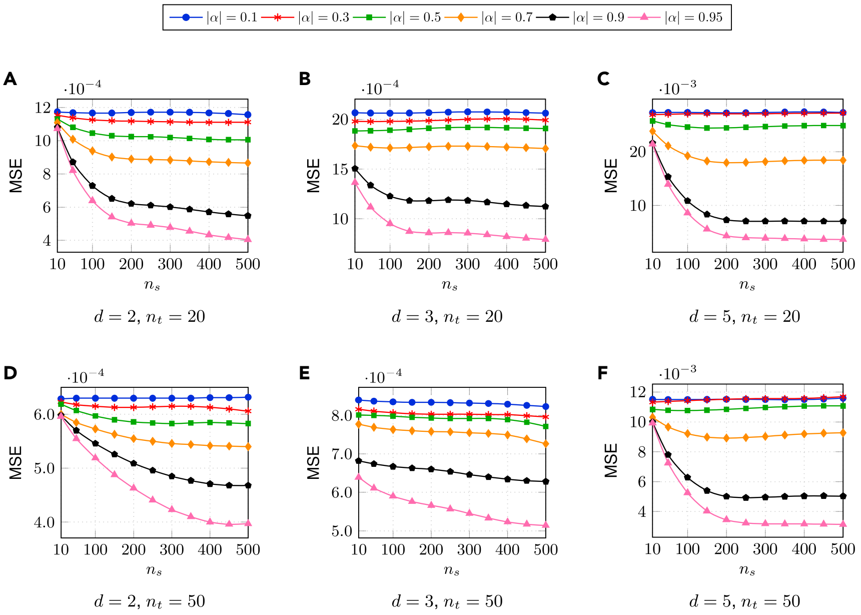
Synthetic datasets

To simulate and verify the extent of knowledge transferability across domains, we consider a wide range of joint prior densities that model the different levels of relatedness between the source and target domains. The proposed setup is as follows. We consider a binary classification problem in the context of homogeneous TL with dimensions 2, 3, and 5. In the simulated datasets, the number of source data points per class varies between 10 and 500 and between 5 and 50 for target datasets. This mimics realistic settings of small-size sample conditions (especially in the target domain) as reported in the literature.[3] We set up the data distributions as follows. $\nu = \nu^y = d + 20$, $\kappa_t = \kappa_t^y = 100$, $\kappa_s = \kappa_s^y = 100$, $\mathbf{m}_t^0 = 0_d$, $\mathbf{m}_t^1 = \vartheta \times 1_d$, $\mathbf{m}_s^0 = \mathbf{m}_t^0 + 10 \times 1_d$, $\mathbf{m}_s^1 = \mathbf{m}_t^1 + 10 \times 1_d$, where ϑ is an adjustable scalar used to control the Bayes error in the target domain, and 0_d and 1_d are $d \times 1$ all-zero and all-one vectors, respectively. For the scale matrices of Wishart distributions we set $\mathbf{M}_t^y = k_t \mathbf{I}_d$, $\mathbf{M}_s^y = k_s \mathbf{I}_d$, and $\mathbf{M}_{ts}^y = k_{ts} \mathbf{I}_d$, where $\mathbf{I}_d$ is the identity matrix of rank d. To ensure that the joint scale matrix $\mathbf{M}^y = \begin{pmatrix} \mathbf{M}_t^y & \mathbf{M}_{ts}^y \\ \mathbf{M}_{ts}^{yT} & \mathbf{M}_s^y \end{pmatrix}$ is positive definite $\forall y \in \{0, 1\}$, we set $k_{ts} = \alpha \sqrt{k_t k_s}$ with $k_t > 0$, $k_s > 0$, and $|\alpha| < 1$. As in Karbalaygphareh et al.,[13] the value of $|\alpha|$ controls the amount of relatedness between the source and target domains (see experimental procedures, section 4.6, for more details). To control the level of relatedness by adjusting only $|\alpha|$ without involving other confounding factors, we set $k_t = k_s = 1$ such that $\mathbf{M}_{ts}^y = \alpha \mathbf{I}_d$. In this setting, the correlation between the features across source and target domains are governed by $|\alpha|$, where small values of $|\alpha|$ correspond to poor relatedness between source and target domains while larger values imply stronger relatedness. To sample from the joint prior, we first sample from a non-singular Wishart distribution $W_{2d}(\mathbf{M}^y, \nu)$ to get a block partitioned sample of the form $\boldsymbol{\Lambda}^y = \begin{pmatrix} \boldsymbol{\Lambda}_t^y & \boldsymbol{\Lambda}_{ts}^y \\ \boldsymbol{\Lambda}_{ts}^{yT} & \boldsymbol{\Lambda}_s^y \end{pmatrix}$ from which we extract $(\boldsymbol{\Lambda}_t^y, \boldsymbol{\Lambda}_s^y)$. Afterward, we sample $\mu_z^y \sim \mathcal{N}(\mathbf{m}_z^y, (\kappa_z^y \boldsymbol{\Lambda}_z^y)^{-1})$ for $z \in \{s, t\}$ and $y \in \{0, 1\}$. In our simulations we use two types of datasets: training datasets that contain samples from both domains and testing datasets that contain only samples from the target domain. In all the simulations we consider testing datasets of 1,000 data points per class and we assume equal prior probabilities for the classes.

RNA sequencing datasets

To evaluate the performance of the TL-based BEE on real-world data, we consider classifying patients diagnosed with schizophrenia using transcriptomic profiles collected from psychiatric disorder studies.[14] Based on two RNA sequencing (RNA-seq) datasets listed in Table 1, we selected the transcriptomic profiles of three genes, based on a stringent feature selection procedure comprising the analysis of differential gene expression, clustering of gene-gene interactions, and statistical testing for multivariate normality. More specifically, we focus on analyzing the astrocyte-related cluster of differentiation 4, found to be significantly upregulated in subjects with schizophrenia.[14] We select the top three hub genes that collectively satisfy the Royston's multivariate normality test applied to the full datasets for both classes at a significance level of 99%. The identified genes satisfying all the aforementioned criteria include SOX9, AHCYL1, and CLDN10, with an average module centrality of 0.86 measured by genes' module membership (kME).[14] In addition to normalization and quality control performed in Gandal et al.,[14] the selected features in both datasets have been further standardized to zero means and unit variances across both classes as in Karbalaygphareh and co-workers.[13,15]

We consider the dataset syn2759792, sampled from the brain dorsolateral prefrontal cortex (DLPFC) area, as a target dataset and syn4590909, sampled from the frontal cortex (FC) region, as a source dataset. Among 555 postmortem brain samples in syn2759792, we randomly draw 5 samples per class as training data and we use the remaining samples to evaluate the classification error. This process is repeated 10,000 times to estimate the average MSE deviation of the TL-based BEE from the true error. To determine the model hyperparameters, we assume shared values for case and control samples in source and target domains and we set $\nu = 10 \times d = 30$, $n_t = 5$. As $|\alpha|$ represents a cross-domain property, we employ the TL-based BEE to conduct an exhaustive greedy search for $|\alpha| \in \{0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.99\}$ in the task of estimating the true



$d = 2$, $n_t = 20$

$d = 3$, $n_t = 20$

$d = 5$, $n_t = 20$

$d = 2$, $n_t = 50$

$d = 3$, $n_t = 50$

$d = 5$, $n_t = 50$

Figure 1. Effect of source data on the performance of the TL-based BEE for quadratic classifiers

MSE deviation from true error for Gaussian distributions with respect to source sample size. The Bayes error is fixed at 0.2 in all subfigures. For direct evaluation and higher dimensions, see Figures S2 and S3.

classification error by leveraging data points from a source domain dataset. In our hyperparameter tuning experiments, we consider source datasets of different sizes ($n_s \in \{10, 30, 50\}$) and we retain the value of $|\alpha|$ that leads to the smallest MSE deviation from the true error across all the experiments. At each iteration, we randomly permute the source samples for statistical significance. The remaining parameters are set as follows: $\kappa_t = n_t$, $\kappa_s = n_s$, and $k_t = k_s = \frac{1}{\nu}$, such that the mean of the Wishart precision matrices will be equal to the identity matrix, which matches the normal standardization. For mean vectors $\mathbf{m}_t$ and $\mathbf{m}_s$, we pool all case and control samples in each domain and consider their means, respectively.

Performance on synthetic datasets

We start by evaluating the performance of the proposed TL-based BEE in estimating the Bayes error, which corresponds to the true error of the quadratic discriminant analysis (QDA) (see experimental procedures, section 4.5) in the target domain, for different levels of $|\alpha|$ and different size combinations of the utilized source and target datasets.

In Figure 1, we investigate the behavior of the TL-based BEE when the target data are fixed while we vary the size of the source data. We show the results for $d = 2$ in the first column, the results for $d = 3$ in the second column, and the results for $d = 5$ in the last column. The rows correspond to the results for target datasets with different sizes: $n_t = 20$ on the top and $n_t = 50$ on the bottom. The MSE curves show similar trends for all three values of d, where we can see that the deviation of the error estimate from the true error significantly decreases when highly related source data are employed. This behavior diminishes as the relatedness between the two domains decreases. Notably, using large source datasets ($n_s \geq 200$) of moderate to small relatedness values ($|\alpha| \leq 0.7$) does not negatively impact the performance of the estimator for low dimensions ($d \in \{2, 3\}$) as shown in the first and second columns of Figure 1. As the dimensionality further increases ($d = 5$), relying on large source datasets with moderate or poor relatedness to the target domain slightly increases the deviation of the estimated error from the true error (i.e., $|\alpha| = 0.7$ in the third column). This tiny asymptotic deviation is explained by potential undesirable effects of relying on large source datasets of modest relatedness. However, it is important to note that the proposed TL-based BEE in the context of the given Bayesian TL framework suppresses this behavior, as it does not directly depend on the source data but the information transfer occurs through the joint prior. The joint prior acts like a bridge through which the useful knowledge passes from the source to the target domain. Effects of using source data in different TL settings (especially, a non-Bayesian setting) may require further investigation. Moreover, the simulation results in different columns show that the MSE deviation decreases as we rely on larger target datasets. However, the gain in performance as we use additional source data is reduced when target data are more abundant. This is illustrated by the slope of the MSE graphs that flattens as n_t increases. Finally, Figure 1 shows that, for higher dimensions, the MSE deviation tends to increase. This is expected as increasing the dimensionality generally leads to a more difficult error estimation problem.

Next, Figure 2 shows the MSE deviation with respect to the size of the target dataset for dimensions 2, 3, and 5. The first row corresponds to the case of using source datasets of size $n_s = 50$ and the second row shows the results for $n_s = 200$. The performance of the TL-based BEE estimator improves with the increasing availability of target data. We can also clearly see that the MSE deviation from the true error asymptotically converges to comparable values for all relatedness levels. When highly related source data are available, the TL-based estimator yields accurate estimation results even when the target dataset is small. These results consolidate the findings in Figure 1 about the redundancy of source data in the presence of abundant target data. Across all graphs in Figure 2, we can see that a relatedness coefficient $|\alpha| = 0.95$ results in a nearly constant deviation from the true error as a function of target data size, which suggests that highly related source data $|\alpha| > 0.95$ act almost identically like the target data, regardless of the shift across the domains in terms of their means. Similar to the trends shown in Figure 1, results across different columns of Figure 2 demonstrate that the error estimation difficulty increases with the increase of dimensionality. This is clearly reflected in the MSE deviation from the true error in Figure 2, which shows that, as the dimension increases from $d = 2$ (first column) to

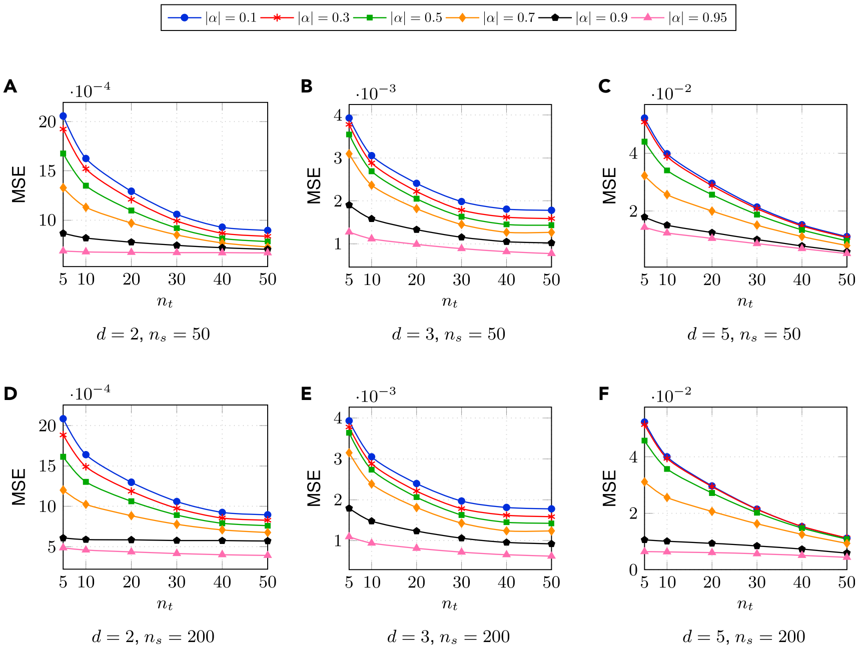


Figure 2. Effect of target data on the performance of the TL-based BEE for quadratic classifiers

MSE deviation from true error for Gaussian distributions with respect to target sample size. The Bayes error is fixed at 0.2 in all subfigures.

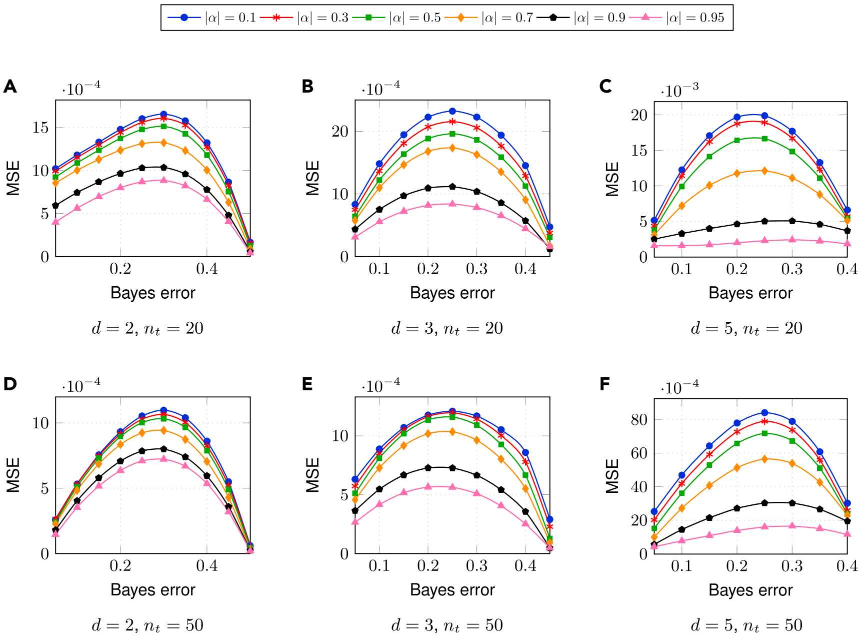
$d = 5$ (last column), the MSE increases by one order of magnitude.

Now, we aim at investigating the effect of classification complexity on the performance of the proposed TL-based BEE. To this end, we conduct simulations, in which we vary the Bayes error through a wide range of possible values and evaluate the TL-based BEE at each given Bayes error for different sizes of target data while using source datasets of a fixed size $n_s = 200$. In binary classification, the Bayes error has an upper bound specified by the true error of random classification, which is 0.5, as every data point can be randomly assigned one of the class labels. Ideally, we would vary the Bayes error across the interval $[0, 0.5]$ as in Dalton and Dougherty.[11] However, in our setup, we do not impose any structure on the covariance matrices, nor do we assume that they are scaled identities. This makes the control of the Bayes error much more difficult. In addition, the joint sampling setup within our Bayesian TL framework inhibits any modification of the randomized parameters. Consequently, the only practical way to adjust the Bayes error is to tune the mean vector parameters $\mathbf{m}_t^y$ that specify the means for the class mean vectors μ_t^y with $y \in \{0, 1\}$. In our experiments, we were able to fully control the Bayes error for $d = 2$ and we considered the following values $[0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5]$. Achieving the same range of values for $d = 3$ and $d = 5$ was more challenging, and our implemented heuristic did not converge for high values of Bayes error as setting $\mathbf{m}_t^0 = \mathbf{m}_t^1$ did not help in increasing the Bayes error. However, we were able to vary the Bayes error for $d = 3$ within the range $[0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45]$, and for $d = 5$, within $[0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4]$, sufficient for observing the trends.

Figure 3 shows the MSE deviation with respect to the Bayes error for dimensions 2, 3, and 5. Results in the first row are obtained using target datasets of size 20 and those in the second row are obtained using target datasets of size 50. We can see that the Bayesian MMSE estimator performs best when using source data of high relatedness to the target domain as expected. For Bayes error in the range $[0.25, 0.35]$, the MSE deviation from the true error is very high, which makes this range of Bayes error as the most challenging setting for error estimation. For a Bayes error of 0.2, the MSE deviation is average across all the experiments, which confirms the validity of our previous assumption in selecting this value to investigate classification problems of moderate difficulty. We note that the TL-based BEE shifts the performance in favor of low and high Bayes error levels. Indeed, the TL-based BEE performs well in this case because the estimated target parameters are sufficiently accurate, even with a small target sample.

In addition to investigating the effect of different relatedness levels between source and target domains, in Figure 4 we have examined the performance of the TL-based BEE for the case when the source class means are swapped between the two classes, such that they show opposite trends compared with the class means in the target domain. For this purpose, we reproduced the experiments in Figure 1 after flipping the class means of source datasets with respect to the target classes (i.e., $\mathbf{m}_s^y = \mathbf{m}_t^{1-y}$, for $y \in \{0, 1\}$). In the first row of Figure 4, we use the generated source datasets as observed samples from the source domain. Interestingly, the obtained results match those observed in Figure 1. This postulates that the knowledge transfer across source and target domains in the context of the studied Bayesian TL framework does not depend on the arrangement of the class means in the source and target domains but only rests on the level of relatedness between the two domains. For verification, we have intentionally considered the same source datasets in the previous experiment as target datasets for estimating the TL-based BEE and we plotted the obtained results in the second row of Figure 4. Clearly, the TL-based BEE veers away from the true error as we consider additional source data points. This deviation is worse with poorly related source data ($|\alpha| = 0.1$). These results confirm previous findings in Karbalayghareh et al.[13] that the joint prior model in the utilized Bayesian TL framework acts like a bridge that distills the useful knowledge from the source domain and effectively transfers it to the target domain.

Results from the second set of experiments that use a linear discriminant analysis (LDA) classifier (see experimental procedures, section 4.5) were similar to the ones obtained using the QDA classifier except for some differences in the performance of the TL-based BEE with respect to the Bayes error that we report in Figure 5 (see supplemental information, section 8, for additional results). The TL-based BEE performance has similar



$d = 2, n_t = 20$

$d = 3, n_t = 20$

$d = 5, n_t = 20$

$d = 2, n_t = 50$

$d = 3, n_t = 50$

$d = 5, n_t = 50$

Figure 3. Effect of the classification complexity on the performance of the TL-based BEE for quadratic classifiers

MSE deviation from QDA true error with respect to Bayes error. Source sample size was set to $n_s = 200$ in all subfigures.

trends with respect to small and moderate Bayes errors when compared with the presented results obtained using the QDA classifier. A notable difference here is observed for large values of Bayes error where the TL-based BEE shows decreased performance in terms of MSE deviation from the true error, which is due to the fact that the employed LDA classifier is sub-optimal compared with the Bayes classifier. This is expected as linear decision boundaries tend to be more sensitive to deviations from true model parameters for highly overlapping class-conditional distributions. In our final set of experiments using synthetic datasets, we compare the performance of the proposed TL-based BEE to standard error estimators for different dimensions and various source datasets of relatedness level $|\alpha| = 0.9$ to the target domain for an optimal Bayesian transfer learning (OBTL) classifier (see experimental procedures, section 4.5). In Figure 6, we show the MSE deviation with respect to different target dataset size. As clearly shown, our proposed TL-based BEE significantly outperforms all other standard error estimators by a substantial margin. In agreement with previous findings in the literature, the standard error estimators perform comparably for low dimensions (i.e., $d = 2$), where the bootstrap may show a slight advantage. As the dimensionality increases (i.e., $d = 5$), the performance shift of the studied estimators becomes more apparent. For example, the resubstitution estimator performs poorly in the small-sample regime while the bootstrap estimator outperforms leave-one-out cross validation (LOO) and CV. Furthermore, we noticed that increasing the size of the source dataset does not lead to any apparent performance improvement for the standard estimators. This is because these estimators do not directly depend on the source data for error estimation (as they are incapable of taking advantage of data from different yet relevant domains). However, providing additional source data to the TL-based BEE considerably reduces the MSE deviation from the true error for all dimensions as shown in Figure 6.

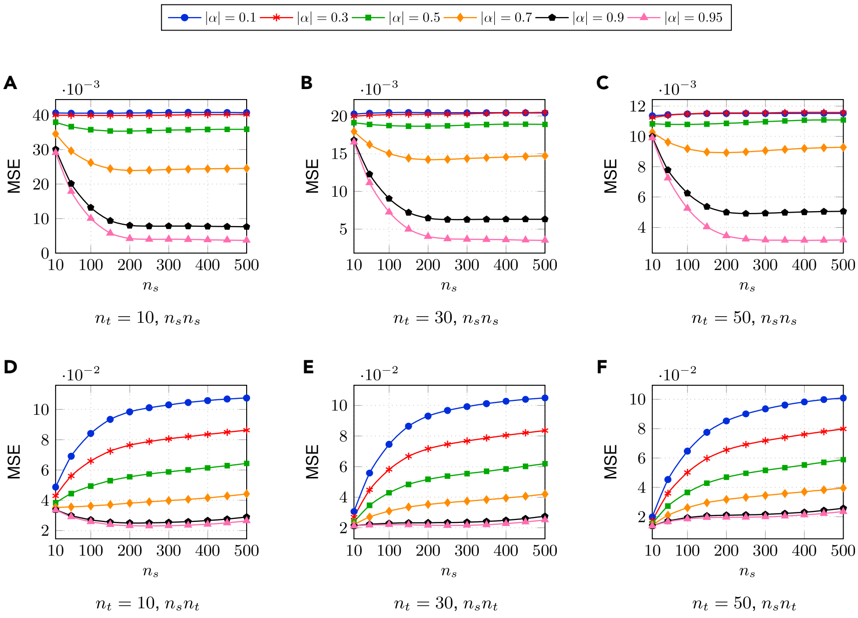
Performance on real-world RNA-seq datasets

To analyze the performance of the TL-based BEE on real-world data, we have trained a QDA classifier on a small target dataset that consists of five sample points per class extracted from syn2759792 in Table 1. Using different source datasets collected from syn4590909, we show in Figure 7A the MSE deviation of the TL-based BEE from the true error with respect to $|\alpha|$.

For all combinations and different sizes of source datasets, the FC brain region showed high relatedness to the DLPFC brain area where the optimal MSE deviation from the true error was obtained for $|\alpha| = 0.99$. Interestingly, findings in Gandal et al.[14] also confirm that syn4590909 and syn2759792 are highly related, as independent gene expression assays for both brain regions have consistently replicated the gradient of transcriptomic severity observed for three different types of psychiatric disorders, including bipolar disorder and schizophrenia.[14] We note that the significant decrease in the MSE deviation from the true error in Figure 7A corresponds to the boost in performance caused by increasing $|\alpha|$ from 0.01 to larger values. This can be explained by the high relatedness between the two studied domains. Indeed, assuming very poor relatedness (i.e., $|\alpha| = 0.01$) between the domains, deviating from the ground truth of high relatedness results in a very large MSE. We show in Figure 7B the increasing gain in accuracy of the TL-based BEE in estimating the classification error after using additional labeled observations from the source domain. These results again confirm the efficacy and advantages of our TL-based error estimation scheme, compared with other standard error estimation methods, when additional data are available from different source domains that are nevertheless relevant to the target domain. From a practical perspective, our proposed TL-based BEE has the potential to facilitate the analysis of real-world datasets in the context of small-sample classification. Challenges of designing and evaluating classifiers (e.g., for clinical diagnosis or prognosis) in a small-sample setting are prevalent in scientific studies in life sciences and physical sciences due to the formidable cost, time, and effort required for data acquisition. This is certainly the case for the example that we consider in this section, where invasive brain biopsies would be needed to get the data.

Insights gained

In this section, we summarize the insights gained from our analyses, which demonstrate the potential advantages of applying



$n_t = 10, n_s n_s$

$n_t = 30, n_s n_s$

$n_t = 50, n_s n_s$

$n_t = 10, n_s n_t$

$n_t = 30, n_s n_t$

$n_t = 50, n_s n_t$

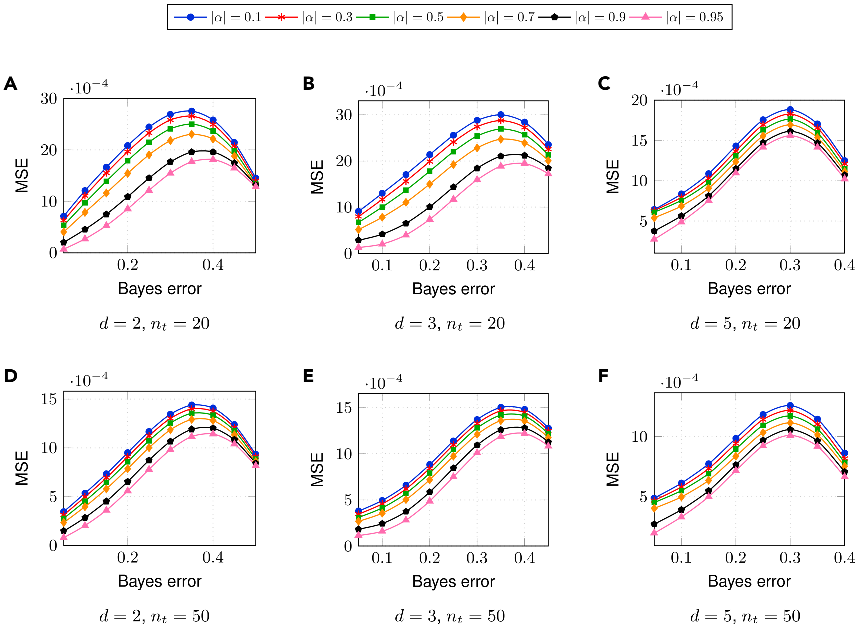
Figure 4. Effect of the arrangement of the class means in the source and target domains on the performance of the TL-based BEE

MSE deviation from true error with respect to source sample size. The source class means are flipped with respect to target classes ($\mathbf{m}_s^y = \mathbf{m}_t^{1-y}$, for $y \in \{0, 1\}$). In the first row, the source datasets are correctly considered as source samples. In the second row, the source datasets are intentionally considered as target samples. The Bayes error is fixed at 0.2 and $d = 5$.

TL to the estimation of classification errors. Our results have shown that incorporating data and knowledge from relevant source domains is helpful to significantly enhance the classification error estimation accuracy. When an appropriate source domain is identified, the efficiency of the knowledge transfer process depends on the correlation of the features across domains, rather than the class-conditional mean values of the features, with our problem setups. From an error estimation perspective, our investigation has revealed that, unlike classifier design, the most challenging setting for error estimation arises in classification problems of moderate complexity in terms of Bayes error. When source datasets that are at least modestly relevant to the target domain of interest are available, knowledge transfer to the target domain by appropriate modeling of the joint prior could enhance both the accuracy and the reliability of the error estimation. This was validated in our current study, where the joint prior acts like a "channel" as well as a "filter," through which useful relevant knowledge is passed from the source domain to the target domain. Our results have shown that using at least 200 data points from a relevant source domain, whose relatedness level is above 0.7, enables an accurate error estimation even with small target data (less than 50 sample points). Using real-world biological data (RNA-seq data), we have shown that the relatedness level can be empirically determined by exploring the range of possible values.

Limitations of the study

This section discusses the limitations of our current work in modeling assumptions, computational cost, and scalability to



$d = 2, n_t = 20$

$d = 3, n_t = 20$

$d = 5, n_t = 20$

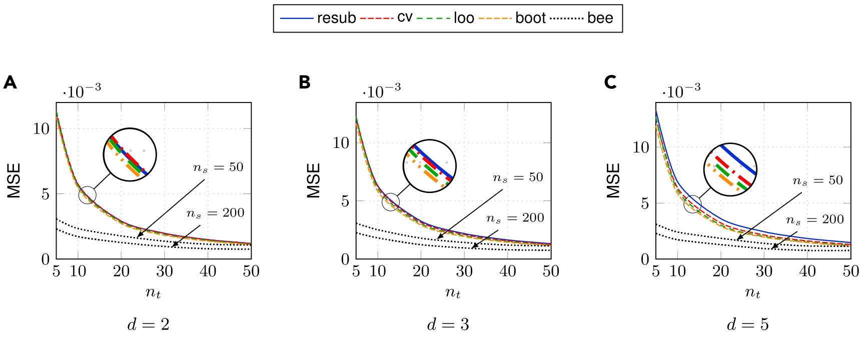
$d = 2, n_t = 50$

$d = 3, n_t = 50$

$d = 5, n_t = 50$

Figure 5. Effect of the classification complexity on the performance of the TL-based BEE for linear classifiers

MSE deviation from LDA true error for Gaussian distributions with respect to Bayes error. Source sample size is set to $n_s = 200$ in all subfigures. See also Figures S4 and S5.



$d = 2$ $d = 3$ $d = 5$

Figure 6. Comparative analysis of the performance of the TL-based BEE with respect to standard error estimators

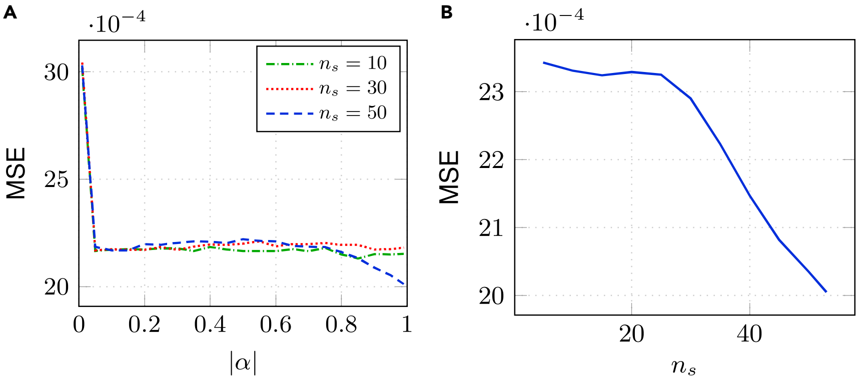
MSE deviation from true error with respect to target data size. The proposed TL-based BEE is compared with other widely used estimators. In all subfigures, the Bayes error is fixed at 0.2, and $|\alpha| = 0.9$.

higher dimensions. Despite the precise mathematical definition of our error estimator, accurate estimation of the classification error is contingent on whether predictive posterior densities are available in closed forms or can be approximated in an effective manner. While such densities are available for Gaussian models (e.g., assuming joint Wishart priors), one may need to derive them for different priors for non-Gaussian distributions. The computational complexity to accurately estimate the proposed TL-based BEE through direct sampling methods can be excessive and may scale poorly for higher dimensions. However, we efficiently overcame this limitation by developing a robust importance sampling setup that has shifted all the computational overhead related to the TL process from Monte Carlo sampling to the numerical evaluation of the importance likelihood. Developing similar statistical methods for TL-based BEE would be needed for different modeling assumptions. While the definition of the TL-based BEE and the proposed robust importance sampling scheme are general and applicable to higher dimensions, controlling the Bayes error for synthetic datasets for dimensions higher than 5 can be challenging, which was the main reason for choosing the dimensions $d = 2, \ldots, 5$ in this study. However, this is not an issue in practice, as the classification complexity in real-world applications (reflected by the Bayes error) is an inherent property of a given classification problem governed by the underlying feature-label distribution, and not a design choice. Technically, the proposed TL-based BEE can be applied to classification problems based on high-dimensional features as long as the required computational resources are available. Furthermore, we can also consider classifier design and error estimation based on a lower-dimensional representation of the original feature space—e.g., using principal-component analysis or auto-encoders—to make the computational cost manageable.

Conclusions

In this study, we have introduced a Bayesian MMSE estimator that draws from concepts and theories in TL to enable accurate estimation of classification error in the (target) domain of interest by utilizing samples from other closely related (source) domains. We have developed an efficient and robust importance sampling setup that can be used for accurate error estimation in small-sample scenarios that often arise in many real-world scientific problems. Extensive performance analysis based on both synthetic and real-world biological data demonstrates the outstanding performance of the proposed TL-based BEE clearly outperforming conventional estimators.

In our proposed framework, Laplace approximations were used to alleviate the complexity associated with the exact evaluation of generalized hypergeometric functions that appear in the posterior distribution of the target parameters. Beyond the Gaussian model assumed in the validation experiments, we also provide a general mathematical definition for the TL-based BEE that can directly be extended to applications with non-Gaussian distributions where the model parameters can be inferred through Markov chain Monte Carlo (MCMC) methods. In



$d = 3$, $n_t = 5$ $d = 3$, $n_t = 5$, $|\alpha| = 0.99$

Figure 7. Performance of the TL-based BEE on real-world RNA-seq datasets

MSE deviation from QDA true error for normally distributed brain gene expression data with respect to $|\alpha|$ and n_s. (A) Gene features from the FC brain region demonstrate high relatedness with those from DPLFC area ($|\alpha| = 0.99$). (B) Utilizing the data from source domain significantly reduces the MSE of the TL-based BEE in the target domain.

this study, target and source domains were related through the joint prior of the model parameters that transfers useful knowledge across domains. A key property of the proposed TL-based BEE is its elegant ability to handle the uncertainty about the model parameters by integrating this prior with data, deducing robust estimates by accounting for all possible parameter values.

Paramount practical challenges for the TL-based BEE include the identification of suitable source domains that share similar families of distributions as the target domain of interest. This is crucial as the relatedness across domains is mathematically modeled assuming the similarity of the feature-label distributions across domains. Furthermore, learning the joint prior for the distributions and modeling the relatedness between different domains may also present an engineering challenge. While techniques for knowledge-driven prior construction have been developed,[17,18] such techniques have yet to be developed for joint prior construction for relevant domains, which is an important future research direction.

An important aspect enabled by the proposed TL-based BEE is optimal data acquisition from multiple domains that aims at maximally enhancing the error estimation capability based on a finite budget for data acquisition. For example, if one has a fixed budget to acquire additional data from either the source or target domain, what would be the most cost-effective strategy for data acquisition? In typical TL scenarios, data acquisition cost may be relatively cheaper in the source domain than in the target domain, although the data acquired in the target domain might be more impactful. A natural question is how one can maximize the "return-on-investment" for data acquisition given the available budget. Such strategies for optimal experimental design[19–24] and active learning[25–27] have been actively studied in a Bayesian paradigm that enables objective-based uncertainty quantification via mean objective cost of uncertainty.[28,29] While this is beyond the scope of this current study, it opens up interesting directions for future research.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Dr. Byung-Jun Yoon is the lead contact for this study and can be reached at bjyoon@ece.tamu.edu.

Materials availability

This study did not generate any physical materials.

Data and code availability

All RNA-seq datasets that have been utilized in this study are publically available. All original code has been deposited at https://github.com/omarmaddouri/TL_BEE, archived in Zenodo under the https://doi.org/10.5281/zenodo.5594476, and are publicly available as of the date of publication. In addition to the proposed importance sampling estimate, we also provide implementation of the direct evaluation using the predictive posterior density of target parameters.

Bayesian TL for error estimation

The advantage of the mathematical formulation that underlies the proposed TL-based BEE (and also the original TL Bayesian framework in Karbalayghareh et al.[13]) is that it articulates a unified Bayesian inference model that assumes a specified prior distribution governing the parameter vector θ_t and acting like a bridge to help update $\pi^*(\theta_t)$ after observing $\mathcal{D}_t$ and $\mathcal{D}_s$. From this standpoint, the derivation of the TL-based BEE for TL depends on determining $\pi^*(\theta_t)$. To determine the TL-based BEE in the context of the presented Bayesian transfer learning framework we evoke the following theorem.

Theorem 1:[13] given the target $\mathcal{D}_t$ and source $\mathcal{D}_s$ data, the posterior distribution of target mean μ_t^y and the target precision matrix $\mathbf{\Lambda}_t^y$ for the classes $y\in\{0,1\}$ has Gaussian-hypergeometric function distribution given by

$$p\left(\mu_t^y,\mathbf{\Lambda}_t^y\middle|\mathcal{D}_t^y,\mathcal{D}_s^y\right)=A^y\left|\mathbf{\Lambda}_t^y\right|^{\frac{1}{2}}\exp\left(-\frac{\kappa_{t,n}^y}{2}\left(\mu_t^y-\mathbf{m}_{t,n}^y\right)^T\mathbf{\Lambda}_t^y\left(\mu_t^y-\mathbf{m}_{t,n}^y\right)\right)\times\left|\mathbf{\Lambda}_t^y\right|^{\frac{\nu^y+n_t^y-d-1}{2}}\operatorname{etr}\left(-\frac{1}{2}\left(\mathbf{T}_t^y\right)^{-1}\mathbf{\Lambda}_t^y\right){}_1F_1\left[\begin{matrix}\frac{\nu^y+n_s^y}{2}\\\frac{\nu^y}{2}\end{matrix};\frac{1}{2}\mathbf{F}^{\mathbf{y}}\,\mathbf{\Lambda}_{\mathbf{t}}^{\mathbf{y}}\,\mathbf{F}^{\mathbf{y}^{\mathbf{T}}}\,\mathbf{T}_{\mathbf{s}}^{\mathbf{y}}\right], \quad \text{(Equation 15)}$$

where A^y is a constant of proportionality given by

$$\left(A^y\right)^{-1}=\left(\frac{2\pi}{\kappa_{t,n}^y}\right)^{\frac{d}{2}}2^{\frac{d\left(\nu^y+n_t^y\right)}{2}}\Gamma_d\left(\frac{\left(\nu^y+n_t^y\right)}{2}\right)\left|\mathbf{T}_t^y\right|^{\frac{\left(\nu^y+n_t^y\right)}{2}}{}_2F_1\left[\begin{matrix}\frac{\nu^y+n_s^y}{2},\frac{\nu^y+n_t^y}{2}\\\frac{\nu^y}{2}\end{matrix};\mathbf{T}_s^y\mathbf{F}^y\mathbf{T}_t^y\mathbf{F}^{y^T}\right], \quad \text{(Equation 16)}$$

and

$$\begin{aligned}&\kappa_{t,n}^y=\kappa_t^y+n_t^y,\\&\mathbf{m}_{t,n}^y=\frac{\kappa_t^y\mathbf{m}_t^y+n_t^y\overline{\mathbf{x}}_t^y}{\kappa_t^y+n_t^y},\\&\left(\mathbf{T}_t^y\right)^{-1}=\left(\mathbf{M}_t^y\right)^{-1}+\mathbf{F}^{y^T}\mathbf{C}^y\mathbf{F}^y+\mathbf{S}_t^y+\frac{\kappa_t^y n_t^y}{\kappa_t^y+n_t^y}\left(\mathbf{m}_t^y-\overline{\mathbf{x}}_t^y\right)\left(\mathbf{m}_t^y-\overline{\mathbf{x}}_t^y\right)^T,\\&\left(\mathbf{T}_s^y\right)^{-1}=\left(\mathbf{C}^y\right)^{-1}+\mathbf{S}_s^y+\frac{\kappa_s^y n_s^y}{\kappa_s^y+n_s^y}\left(\mathbf{m}_s^y-\overline{\mathbf{x}}_s^y\right)\left(\mathbf{m}_s^y-\overline{\mathbf{x}}_s^y\right)^T,\end{aligned} \quad \text{(Equation 17)}$$

with $\mathbf{C}^y$ and $\mathbf{F}^y$ given by

$$\mathbf{C}^y=\mathbf{M}_s^y-\mathbf{M}_{ts}^{y^T}\left(\mathbf{M}_t^y\right)^{-1}\mathbf{M}_{ts}^y$$

$$\mathbf{F}^y=\left(\mathbf{C}^y\right)^{-1}\mathbf{M}_{ts}^{y^T}\left(\mathbf{M}_t^y\right)^{-1}$$

and sample means and covariances for $z\in\{s,t\}$ given by

$$\overline{\mathbf{x}}_z^y=\frac{1}{n_z^y}\sum_{i=1}^{n_z^y}\mathbf{x}_{z,i}^y$$

$$\mathbf{S}_z^y=\sum_{i=1}^{n_z^y}\left(\mathbf{x}_{z,i}^y-\overline{\mathbf{x}}_z^y\right)\left(\mathbf{x}_{z,i}^y-\overline{\mathbf{x}}_z^y\right)^T.$$

${}_1F_1\left[\begin{matrix}a\\b\end{matrix};\mathbf{X}\right]$ and ${}_2F_1\left[\begin{matrix}a,b\\c\end{matrix};\mathbf{X}\right]$ are, respectively, the confluent and Gauss matrix-variate hypergeometric functions reviewed in the supplemental information, section 6. Now, using Theorem 1 and assuming that the class-0 prior probability c, θ_t^0, and θ_t^1 are independent prior to observing $\mathcal{D}_t$ and $\mathcal{D}_s$, the BEE for TL is given by

$$\widehat{\varepsilon}=E_{\pi^*}[c]E_{\pi^*}\left[\varepsilon_n^0\right]+\left(1-E_{\pi^*}[c]\right)E_{\pi^*}\left[\varepsilon_n^1\right], \quad \text{(Equation 18)}$$

where

$$E_{\pi^*}\left[\varepsilon_n^y\right]=\int_{\Theta_t^y}\varepsilon_n^y\left(\theta_t^y\right)\pi^*\left(\theta_t^y\right)d\theta_t^y, \quad \text{(Equation 19)}$$

with Θ_t^y being the parameter space that contains all possible values for θ_t^y.

Computing TL-based BEE for arbitrary classifiers

Computing the TL-based BEE for an arbitrary classifier ψ_n involves the evaluation of the integral in (Equation 19). Even when we have an analytic expression for the true error of the studied classifier, the closed-form expression for the

TL-based BEE cannot be easily derived due to the complex expression of the target posterior in the presence of the matrix-variate hypergeometric functions. With non-linear classifiers, this becomes practically impossible as no closed-form expression exists for the true error itself. The standard way to approximate the true error in this case is to consider the test error. For a specified parameter θ_t, a large test set is generated from $f_{\theta_t}(\mathbf{x}, y)$, and the performance of ψ_n is evaluated on that test set. This requires sampling from $\pi^*(\theta_t^y)$ so that the integral in (Equation 19) can be approximated by a finite sum. Suppose we have N posterior sample points $\theta_{t,i}^y \sim \pi^*(\theta_t^y), i = 1 \cdots N$. Then the approximation is given by

$$E_{\pi^*}\left[\varepsilon_n^y\right] \approx \frac{1}{N}\sum_{i=1}^{N}\varepsilon_n^y\left(\theta_{t,i}^y\right). \qquad \text{(Equation 20)}$$

Because of the generalized confluent and Gauss hypergeometric functions in the expression of π^*, sampling directly from the posterior is very laborious and the computational cost of applying MCMC methods is exorbitant as the execution may take several weeks even on high-performance computing clusters. To address this issue, in the next section we propose an efficient self-normalized importance sampling setup with control variates that provides accurate estimates for the TL-based BEE and significantly reduces the computation time to make the proposed TL-based BEE feasible.

Self-normalized importance sampling with control variates

Importance sampling

Importance sampling (IS) is a variance reduction technique that provides a remedy to sampling from complex distributions.[30] To estimate $E_{\pi^*}[\varepsilon_n^y]$, IS makes a multiplicative adjustment to ε_n^y to compensate for sampling from an alternative importance distribution Φ^* instead of π^*. If Φ^* is a positive probability density function on Θ_t^y, we can write

$$\begin{aligned} E_{\pi^*}\left[\varepsilon_n^y\right] &= \int_{\Theta_t^y}\varepsilon_n^y\left(\theta_t^y\right)\pi^*\left(\theta_t^y\right)d\theta_t^y \\ &= \int_{\Theta_t^y}\frac{\varepsilon_n^y\left(\theta_t^y\right)\pi^*\left(\theta_t^y\right)}{\Phi^*\left(\theta_t^y\right)}\Phi^*\left(\theta_t^y\right)d\theta_t^y \\ &= E_{\Phi^*}\left[\frac{\varepsilon_n^y\left(\theta_t^y\right)\pi^*\left(\theta_t^y\right)}{\Phi^*\left(\theta_t^y\right)}\right]. \end{aligned} \qquad \text{(Equation 21)}$$

Achieving an accurate IS estimation is contingent on selecting an appropriate importance density that is nearly proportional to $\varepsilon_n^y(\theta_t^y)\pi^*(\theta_t^y)$. By analogy to Gordon and co-workers,[31,32] a plausible and cogent candidate for Φ^* emanates as the posterior of target parameters upon observation of target-only data. Obviously, both distributions are tracking the same model parameters in the target domain upon observation of data. To determine $\Phi^*(\theta_t^y) = p(\mu_t^y, \Lambda_t^y | \mathcal{D}_t^y)$ we require the following lemma:

Lemma 1:[33] if $\mathcal{D} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ where $\mathbf{x}_i$ is a $d \times 1$ vector and $\mathbf{x}_i \sim \mathcal{N}(\mu, (\Lambda)^{-1})$, for $i = 1, \cdots, n$, and (μ, Λ) has a Gaussian-Wishart prior, such that $\mu\big|\Lambda \sim \mathcal{N}(\mathbf{m}, (\kappa\Lambda)^{-1})$ and $\Lambda \sim \mathbf{W_d}(\mathbf{M}, \mathbf{v})$, then the posterior of (μ, Λ) upon observing $\mathcal{D}$ is also a Gaussian-Wishart distribution such that

$$\mu\Big|\Lambda, \mathcal{D} \sim \mathcal{N}\left(\mathbf{m}_n, (\kappa_n\Lambda)^{-1}\right); \ \Lambda\Big|\mathcal{D} \sim W_d(\mathbf{M}_n, \nu_n), \qquad \text{(Equation 22)}$$

where

$$\begin{aligned} &\kappa_n = \kappa + n, \quad \nu_n = \nu + n, \\ &\mathbf{m}_n = \frac{\kappa\mathbf{m} + n\overline{\mathbf{x}}}{\kappa + n}, \text{ and} \\ &\mathbf{M}_n^{-1} = \mathbf{M}^{-1} + \mathbf{S} + \frac{\kappa n}{\kappa + n}(\mathbf{m} - \overline{\mathbf{x}})(\mathbf{m} - \overline{\mathbf{x}})^T, \end{aligned} \qquad \text{(Equation 23)}$$

depending on the sample mean and covariance matrix

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \quad \mathbf{S} = \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T. \qquad \text{(Equation 24)}$$

Using Lemma 1 we now get the expression of the importance density Φ^* given by

$$\begin{aligned} p\left(\mu_t^y, \Lambda_t^y \middle| \mathcal{D}_t^y\right) = &\left(\frac{2\pi}{\kappa_{t,n}^y}\right)^{-\frac{d}{2}} 2^{-\frac{d\left(\nu^y + n_t^y\right)}{2}}\Gamma_d^{-1}\left(\frac{\left(\nu^y + n_t^y\right)}{2}\right)\left|\mathbf{M}_{t,n}^y\right|^{-\frac{\left(\nu^y + n_t^y\right)}{2}}\left|\Lambda_t^y\right|^{\frac{1}{2}} \\ \times \exp&\left(-\frac{\kappa_{t,n}^y}{2}\left(\mu_t^y - \mathbf{m}_{t,n}^y\right)^T\Lambda_t^y\left(\mu_t^y - \mathbf{m}_{t,n}^y\right)\right)\left|\Lambda_t^y\right|^{\frac{\nu^y + n_t^y - d - 1}{2}} \operatorname{etr}\left(-\frac{1}{2}\left(\mathbf{M}_{t,n}^y\right)^{-1}\Lambda_t^y\right), \end{aligned} \qquad \text{(Equation 25)}$$

where

$$\begin{aligned} &\kappa_{t,n}^y = \kappa_t^y + n_t^y, \\ &\mathbf{m}_{t,n}^y = \frac{\kappa_t^y\mathbf{m}_t^y + n_t^y\overline{\mathbf{x}}_t^y}{\kappa_t^y + n_t^y}, \\ &\left(\mathbf{M}_{t,n}^y\right)^{-1} = \left(\mathbf{M}_t^y\right)^{-1} + \mathbf{S}_t^y + \frac{\kappa_t^y n_t^y}{\kappa_t^y + n_t^y}\left(\mathbf{m}_t^y - \overline{\mathbf{x}}_t^y\right)\left(\mathbf{m}_t^y - \overline{\mathbf{x}}_t^y\right)^T, \end{aligned} \qquad \text{(Equation 26)}$$

with sample mean and covariance given by

$$\overline{\mathbf{x}}_t^y = \frac{1}{n_t^y}\sum_{i=1}^{n_t^y}\mathbf{x}_{t,i}^y$$

$$\mathbf{S}_t^y = \sum_{i=1}^{n_t^y}\left(\mathbf{x}_{t,i}^y - \overline{\mathbf{x}}_t^y\right)\left(\mathbf{x}_{t,i}^y - \overline{\mathbf{x}}_t^y\right)^T.$$

After simplifications, the expression of the TL-based BEE in (Equation 21) takes the form

$$E_{\pi^*}\left[\varepsilon_n^y\right] = E_{\Phi^*}\left[\varepsilon_n^y\left(\theta_t^y\right)\mathcal{L}\left(\theta_t^y\right)\right], \qquad \text{(Equation 27)}$$

where $\theta_t^y = (\mu_t^y, \Lambda_\mathbf{t}^\mathbf{y})$ and $\mathcal{L}(\theta_t^y)$ is the likelihood ratio given by

$$\begin{aligned} \mathcal{L}(\mu_t^y, \Lambda_\mathbf{t}^\mathbf{y}) = \operatorname{etr}&\left(-\frac{1}{2}\left[\left(\mathbf{T}_t^y\right)^{-1}\right.\right. \\ &\left.\left. - \left(\mathbf{M}_{t,n}^y\right)^{-1}\right]\Lambda_t^y\right)\left|\frac{\mathbf{M}_{t,n}^y}{\mathbf{T}_t^y}\right|^{\frac{\left(\nu^y + n_t^y\right)}{2}} \frac{{}_1F_1\left[\begin{array}{l}\frac{\nu^y + n_s^y}{2} \\ \frac{\nu^y}{2}\end{array}; \frac{1}{2}\,\mathbf{F}^y\,\Lambda_t^y\,\mathbf{F}^{y^T}\,\mathbf{T}_s^y\right]}{{}_2F_1\left[\begin{array}{l}\frac{\nu^y + n_s^y}{2}, \frac{\nu^y + n_t^y}{2} \\ \frac{\nu^y}{2}\end{array}; \mathbf{T_s^y}\,\mathbf{F^y}\,\mathbf{T_t^y}\,\mathbf{F^{y^T}}\right]}. \end{aligned} \qquad \text{(Equation 28)}$$

Although the likelihood ratio has a simplified expression, computing the hypergeometric functions involves the computation of series of zonal polynomials, which is computationally expensive and not scalable to high dimensions. To mitigate this limitation, we use the Laplace approximations of these functions (see Figure S1 and supplemental information, section 6). To rectify possible disproportionalities in likelihood ratios due to approximations, we consider the self-normalized IS estimate given by

$$\widehat{E}_{\Phi^*}\left[\varepsilon_n^y\right] \approx \frac{\sum_{i=1}^{N}\varepsilon_n^y\left(\theta_{t,i}^y\right)\mathcal{L}\left(\theta_{t,i}^y\right)}{\sum_{i=1}^{N}\mathcal{L}\left(\theta_{t,i}^y\right)} \qquad \text{(Equation 29)}$$

with $\theta_{t,i}^y \sim \Phi^*(\theta_t^y), i = 1 \cdots N$.

Control variates

For more stable and efficient estimates, we further combine IS with control variates. Using control variates in conjunction with IS is a variance reduction technique, in particular when a significant portion of a model for estimating the expectation can be solved explicitly. In our case, a useful control variates function (CVF) $\mathcal{V}(\theta_t^y)$ satisfies

$$E_{\Phi^*}\left[\mathcal{V}(\theta_t^y)\right] = \int_{\Theta_t^y} \mathcal{V}(\theta_t^y)\Phi^*(\theta_t^y)\, d\theta_t^y = \delta, \tag{Equation 30}$$

where δ is a constant. Under such circumstances, a more stable estimate for the TL-based BEE can be derived as

$$\tilde{E}_{\Phi^*}\left[\varepsilon_n^y\right] \approx \frac{\sum_{i=1}^N \varepsilon_n^y\left(\theta_{t,i}^y\right)\mathcal{L}\left(\theta_{t,i}^y\right)}{\sum_{i=1}^N \mathcal{L}\left(\theta_{t,i}^y\right)} - \frac{1}{N}\sum_{i=1}^N \frac{\beta\mathcal{V}\left(\theta_{t,i}^y\right)}{\Phi^*\left(\theta_{t,i}^y\right)} + \beta\delta, \tag{Equation 31}$$

where $\theta_{t,i}^y \sim \Phi^*(\theta_t^y), i = 1\cdots N$ and β is a weighting coefficient tuned to reduce the variance of the estimate. The optimal value of β is given by

$$\beta_{opt} = \frac{\mathrm{cov}\left[\zeta_n^y(\theta_t^y), \mathcal{V}(\theta_t^y)\right]}{\mathrm{var}\left[\mathcal{V}(\theta_t^y)\right]}, \tag{Equation 32}$$

with

$$\zeta_n^y\left(\theta_t^y\right) = \frac{\varepsilon_n^y\left(\theta_t^y\right)\mathcal{L}\left(\theta_t^y\right)}{\frac{1}{N}\sum_{i=1}^N \mathcal{L}\left(\theta_{t,i}^y\right)} \tag{Equation 33}$$

and cov[•, •] and var[•] denote covariance and variance, respectively (see supplemental information, section 7.3, for more details). In practice, it is not likely that we know β_{opt} beforehand, but it is estimated from the Monte Carlo sample. It turns out that $\tilde{E}_{\Phi^*}$ has lower variance than $\widehat{E}_{\Phi^*}$ by a factor of $(1 - \mathrm{corr}[\zeta_n^y(\theta_t^y), \mathcal{V}(\theta_t^y)])$, where corr[**a**, **b**] denotes the correlation coefficient between **a** and **b** and given by

$$\mathrm{corr}[\mathbf{a},\mathbf{b}] = \frac{\mathrm{cov}[\mathbf{a},\mathbf{b}]}{\sqrt{\mathrm{var}[\mathbf{a}]}\sqrt{\mathrm{var}[\mathbf{b}]}}. \tag{Equation 34}$$

To select an appropriate CVF we need to consider two criteria. First, its expectation with respect to Φ^* should have an exact evaluation. Second, it has to be correlated with the estimated error. A favorable candidate is the analytic true error of linear classifiers. In this study, we consider a CVF given by the true error of an LDA classifier defined by $g_{N_t}(\mathbf{x}) = \mathbf{a}_{N_t}^T\mathbf{x} + b_{N_t}$ where $\mathbf{a}_{N_t} = \mathbf{S}_t^{-1}(\overline{\mathbf{x}}_t^1 - \overline{\mathbf{x}}_t^0)$, $b_{N_t} = -\frac{1}{2}\mathbf{a}^T(\overline{\mathbf{x}}_t^1 + \overline{\mathbf{x}}_t^0) + \ln\frac{n_t^1}{n_t^0}$, and the pooled covariance $\mathbf{S}_t$ is given by

$$\mathbf{S}_t = \frac{(n_t^0 - 1)\mathbf{S}_t^0 + (n_t^1 - 1)\mathbf{S}_t^1}{N_t - 2}. \tag{Equation 35}$$

$\overline{\mathbf{x}}_t^y$ and $\mathbf{S}_t^y$ are the empirical estimates utilized in (Equation 26). Thus, the CVF is given by

$$\mathcal{V}(\mu_t^y, \boldsymbol{\Lambda}_t^y) = \boldsymbol{\Phi}\left(\frac{(-1)^y g_{N_t}(\mu_t^y)}{\sqrt{\mathbf{a}_{N_t}^T\left(\boldsymbol{\Lambda}_t^y\right)^{-1}\mathbf{a}_{N_t}}}\right), \tag{Equation 36}$$

with $\boldsymbol{\Phi}$ denoting the standard normal Gaussian cumulative distribution function. Now it remains only to determine $E_{\Phi^*}[\mathcal{V}(\mu_t^y, \boldsymbol{\Lambda}_t^y)]$ in closed-form to fully define the estimation setup. We can show after simplifications and using results from[11] that

$$E_{\Phi^*}[\mathcal{V}(\mu_t^y, \boldsymbol{\Lambda}_t^y)] = \frac{1}{2} + \frac{\mathrm{sgn}(A)}{2}\mathcal{I}\left(\frac{A^2}{A^2 + \mathbf{a}_{N_t}{}^T\left[\mathbf{M}_{t,n}^y\right]^{-1}\mathbf{a}_{N_t}}; \frac{1}{2}, \frac{\nu^y + n_t^y - d + 1}{2}\right), \tag{Equation 37}$$

where sgn(•) is the sign function,

$$A = (-1)^y g_{N_t}(\mathbf{m}_{t,n}^y)\sqrt{\frac{\kappa_{t,n}^y}{1 + \kappa_{t,n}^y}}, \tag{Equation 38}$$

and $\mathcal{I}(\bullet;\bullet,\bullet)$ denotes the regularized incomplete beta function given by

$$\mathcal{I}(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^x t^{a-1}(1-t)^{b-1}dt, \tag{Equation 39}$$

with Γ(•) being the regular univariate gamma function. Details for simplifying $E_{\Phi^*}[\mathcal{V}(\mu_t^y, \boldsymbol{\Lambda}_t^y)]$ are covered in supplemental information, section 7.4.

The complete specification of the CVF concludes our IS setup. We enumerate some advantages of the proposed setup over direct sampling methods. First, the importance density Φ^* is much simpler than the nominal density π^*, which involves matrix-variate hypergeometric functions. Second, our setup successfully combines two variance reduction techniques that enable accurate estimation. Last, and most importantly, the independence of the generated Monte Carlo samples w.r.t source data permits the reuse of the sampled parameters with various source datasets for fixed models. This reusability significantly reduces the computational cost of sampling from Φ^* and makes the utilization of advanced MCMC methods amenable as the whole process could be accelerated by a factor of 10–20, which also grows with the dimensionality and the number of used source datasets (see supplemental information, sections 7.5 and 7.6, for more details). For efficient sampling from Φ^*, we use Hamiltonian Monte Carlo (HMC), proven to have a superior performance to standard MCMC samplers.[34] For this purpose, we utilize the STAN software, which offers a full Bayesian statistical inference framework with HMC.[34]

Classifier design

For a comprehensive evaluation of our TL-based error estimator, we design and perform a set of experiments. The proposed TL-based estimator is applied to a collection of classifiers with different levels of learning capacities and tested under various scenarios. To separate error estimation from classifier design, we start by analyzing the performance of the TL-based BEE estimator for fixed classifiers that do not depend on training data. This setup distinctly reveals the major characteristics of the TL-based BEE, excluding any confounding factors that may stem from classifier design and the performance of the resulting classifier.

Next, we also conduct a comparative study of the TL-based BEE performance with respect to other widely used error estimators, which include resubstitution, CV, LOO, and the 0.632-bootstrap estimators. As these popular data-driven estimators involve classifier design on the training data, we will also consider a TL-based classifier designed on target and source data that operates in the target domain for comparison. For this, we employ the OBTL classifier introduced in Karbalayghareh et al.,[13] which shares the same Bayesian framework on which our TL-based BEE is developed. In what follows, we recall the definition of each classifier considered in our evaluations and also present the details of the evaluation experiments performed in this study.

In the first set of experiments, we employ a fixed quadratic classifier assuming we know beforehand the true target parameters. For normally distributed data, this quadratic classifier corresponds also to the Bayes classifier that is optimal for the given feature-label distributions. Using QDA, we define $\Psi_{QDA}(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$, where

$$\begin{gathered}\mathbf{A} = -\frac{1}{2}\left(\boldsymbol{\Lambda}_\mathbf{t}^\mathbf{1} - \boldsymbol{\Lambda}_\mathbf{t}^\mathbf{0}\right),\ \mathbf{b} = \boldsymbol{\Lambda}_\mathbf{t}^\mathbf{1}\mu_t^1 - \boldsymbol{\Lambda}_\mathbf{t}^\mathbf{0}\mu_t^0, \\ c = -\frac{1}{2}\left(\mu_t^{1^T}\boldsymbol{\Lambda}_\mathbf{t}^\mathbf{1}\mu_t^1 - \mu_t^{0^T}\boldsymbol{\Lambda}_\mathbf{t}^\mathbf{0}\mu_t^0\right) - \frac{1}{2}\ln\left(\frac{\left|\boldsymbol{\Lambda}_\mathbf{t}^\mathbf{0}\right|}{\left|\boldsymbol{\Lambda}_\mathbf{t}^\mathbf{1}\right|}\right).\end{gathered} \tag{Equation 40}$$

The error estimation problem turns out to be an estimation of the Bayes error that coincides here with the true error of the designed QDA. Obviously, this classifier is independent from any observed sample as it is fixed assuming known true model parameters. Without loss of generality, we apply the TL-based BEE using labeled observations from a compound dataset compiled from target and source domains.

In the second set of experiments we investigate the behavior of the TL-based BEE within the class of sub-optimal classifiers. To this end, we consider a linear classifier derived through LDA and we define $\Psi_{LDA}(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$ where $\mathbf{a} = \mathbf{S}_t^{-1}(\mu_t^1 - \mu_t^0)$, $b = -\frac{1}{2}\mathbf{a}^T(\mu_t^1 + \mu_t^0)$, and the average covariance $\mathbf{S}_t$ is given by

$$\mathbf{S}_t = \frac{\left(\boldsymbol{\Lambda}_t^0\right)^{-1} + \left(\boldsymbol{\Lambda}_t^1\right)^{-1}}{2}. \tag{Equation 41}$$

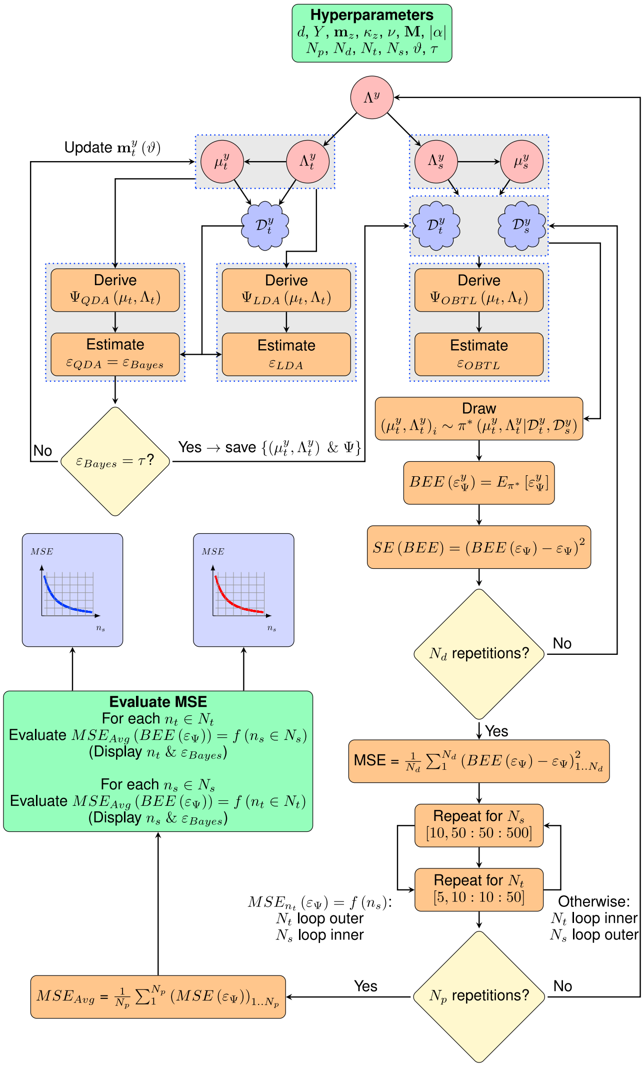

Hyperparameters
$d, Y, \mathbf{m}_z, \kappa_z, \nu, \mathbf{M}, |\alpha|$
$N_p, N_d, N_t, N_s, \vartheta, \tau$
Λ^y
Update $\mathbf{m}_t^y(\vartheta)$
μ_t^y
Λ_t^y
Λ_s^y
μ_s^y
$\mathcal{D}_t^y$
$\mathcal{D}_t^y$
$\mathcal{D}_s^y$
Derive $\Psi_{QDA}(\mu_t, \Lambda_t)$
Derive $\Psi_{LDA}(\mu_t, \Lambda_t)$
Derive $\Psi_{OBTL}(\mu_t, \Lambda_t)$
Estimate $\varepsilon_{QDA} = \varepsilon_{Bayes}$
Estimate ε_{LDA}
Estimate ε_{OBTL}
Draw $(\mu_t^y, \Lambda_t^y)_i \sim \pi^*(\mu_t^y, \Lambda_t^y | \mathcal{D}_t^y, \mathcal{D}_s^y)$
No
$\varepsilon_{Bayes} = \tau$?
Yes → save $\{(\mu_t^y, \Lambda_t^y)\ \&\ \Psi\}$
$BEE(\varepsilon_\Psi^y) = E_{\pi^*}[\varepsilon_\Psi^y]$
$SE(BEE) = (BEE(\varepsilon_\Psi) - \varepsilon_\Psi)^2$
MSE
n_s
MSE
n_s
N_d repetitions?
No
Yes
Evaluate MSE
For each $n_t \in N_t$
Evaluate $MSE_{Avg}(BEE(\varepsilon_\Psi)) = f(n_s \in N_s)$
(Display n_t & ε_{Bayes})
For each $n_s \in N_s$
Evaluate $MSE_{Avg}(BEE(\varepsilon_\Psi)) = f(n_t \in N_t)$
(Display n_s & ε_{Bayes})
MSE = $\frac{1}{N_d}\sum_1^{N_d}(BEE(\varepsilon_\Psi) - \varepsilon_\Psi)^2_{1..N_d}$
Repeat for N_s $[10, 50 : 50 : 500]$
Repeat for N_t $[5, 10 : 10 : 50]$
$MSE_{n_t}(\varepsilon_\Psi) = f(n_s)$:
N_t loop outer
N_s loop inner
Otherwise:
N_t loop inner
N_s loop outer
$MSE_{Avg} = \frac{1}{N_p}\sum_1^{N_p}(MSE(\varepsilon_\Psi))_{1..N_p}$
Yes
N_p repetitions?
No

Figure 8. Simulation diagram using synthetic data

Flow chart illustrating the simulation setup based on synthetic datasets.

Our goal is then to approximate the true error of this sub-optimal classifier using TL.

Next, we evaluate the performance of the TL-based BEE for the OBTL classifier that can take advantage of both source and target domain data. The OBTL classifier is defined by

$$\Psi_{OBTL}(\mathbf{x}) = \arg\max_{y \in \{0,1\}} \mathcal{O}_{OBTL}(\mathbf{x}|y), \qquad \text{(Equation 42)}$$

where the objective function $\mathcal{O}_{OBTL}(\mathbf{x}|y)$ denotes the effective class-conditional density $p(\mathbf{x}|y)$ given by the following theorem:

Theorem 2:[13] the effective class-conditional density, denoted by $p(\mathbf{x}|y) = \mathcal{O}_{OBTL}(\mathbf{x}|y)$, in the target domain is given by

$$\mathcal{O}_{OBTL}(\mathbf{x}|y) = \pi^{-\frac{d}{2}}\left(\frac{\kappa_{t,n}^y}{\kappa_x^y}\right)^{\frac{d}{2}} \Gamma_d\left(\frac{\nu^y + n_t^y + 1}{2}\right)\Gamma_d^{-1}\left(\frac{\nu^y + n_t^y}{2}\right)\left|\mathbf{T}_x^y\right|^{\frac{\nu^y + n_t^y + 1}{2}}\left|\mathbf{T}_t^y\right|^{-\frac{\nu^y + n_t^y}{2}}$$
$$\times\ {}_2F_1\left[\begin{matrix}\frac{\nu^y + n_s^y}{2}, \frac{\nu^y + n_t^y + 1}{2} \\ \frac{\nu^y}{2}\end{matrix}; \mathbf{T_s^y F^y T_x^y F^{y^T}}\right]\ {}_2F_1\left[\begin{matrix}\frac{\nu^y + n_s^y}{2}, \frac{\nu^y + n_t^y}{2} \\ \frac{\nu^y}{2}\end{matrix}; \mathbf{T}_s^y\mathbf{F}^y\mathbf{T}_t^y\mathbf{F}^{y^T}\right]^{-1}, \qquad \text{(Equation 43)}$$

where

$$\kappa_x^y = \kappa_{t,n}^y + 1 = \kappa_t^y + n_t^y + 1,$$
$$\left(\mathbf{T}_x^y\right)^{-1} = \left(\mathbf{T}_t^y\right)^{-1} + \frac{\kappa_{t,n}^y}{\kappa_{t,n}^y + 1}\left(\mathbf{m}_{t,n}^y - \mathbf{x}\right)\left(\mathbf{m}_{t,n}^y - \mathbf{x}\right)^T. \qquad \text{(Equation 44)}$$

Simulation setup

Figure 8 provides a combined illustration of the simulation setup for all three classifiers. For rigorous evaluation of the performance of the proposed TL-based BEE, we primarily focus our experiments on assessing the impact of using different types and amounts of source data. This is enabled by the joint prior imposed over the model parameters and controlled by the relatedness coefficient $|\alpha|$ that dictates the extent of interaction between the features in the two domains. For this purpose, we repeatedly conducted experiments following the flow chart in Figure 8 with different relatedness values ($|\alpha| = [0.1, 0.3, 0.5, 0.7, 0.9, 0.95]$), where $|\alpha| = 0.1$ corresponds to the lowest relatedness between the two domains and $|\alpha| = 0.95$ reflects the highest relatedness within the range of studied values.

In the first set of experiments, we start by drawing a joint sample $(\mathbf{\Lambda}_t^y, \mathbf{\Lambda}_s^y)$ for each class $y \in \{0, 1\}$, as described previously. Next, we iterate over the values of the hyperparameter ϑ to control $\mathbf{m}_t(\vartheta)$ through a dichotomic search to get a desired value τ of the Bayes error. This is achieved by drawing a sample $\mu_t^y \sim \mathcal{N}(\mathbf{m}_t(\vartheta), (\kappa_t^y \mathbf{\Lambda}_t^y)^{-1})$ and then generating a test set based on the joint sample $(\mu_t^y, \mathbf{\Lambda}_t^y)$. Using this test set, we determine the true error of the optimal QDA derived from $(\mu_t^y, \mathbf{\Lambda}_t^y)$. If the desired Bayes error (true error of the designed QDA) is attained then the iteration stops, otherwise we update ϑ and reiterate. In our experiments, unless otherwise specified, we set $\tau = 0.2$ to mimic a moderate level of classification complexity. This step is indeed crucial as it maintains the same level of complexity across the experiments and guarantees a fair comparison across different levels of relatedness. We note that this procedure is valid for general covariances as it acts only on updating the value of the mean parameter without altering the structure of the covariances nor the random mean vectors. Obviously, this approach to specify the Bayes error maintains the Bayesian TL framework intact. However, it is not guaranteed to find values of target parameters that correspond to the desired Bayes error, especially for high dimensions and complex classification (large Bayes error) as we discuss in Performance on synthetic datasets. Once the problem complexity is set and the classifier is fixed, we generate $N_d = 10,000$ training datasets that we use to evaluate the MSE of the TL-based BEE as depicted in Figure 8. To estimate the TL-based BEE, we employ the IS setup described previously and we draw 1,000 MC samples from the importance density using HMC sampler.

In the second set of experiments, we follow a similar setup using an LDA classifier designed based on the true model parameters. As before, we employ QDA to determine the Bayes error to maintain the same complexity level across different experiments. As in the first set of experiments, we use the TL-based BEE to estimate the true error of the designed LDA classifier.

In the last set of experiments on synthetic datasets, we conduct a comparative analysis study using an OBTL classifier designed using training datasets generated from the model parameters specified by the Bayes error. The error estimation task, in this scenario, aims at approximating the true error of the designed OBTL classifier determined using a large test set generated from the true feature-label distributions. As illustrated in Figure 8, QDA and LDA classifiers are fixed and derived from the true model parameters while the OBTL classifier is designed based on training datasets collected from the underlying feature-label distributions that correspond to the specified Bayes error. In all simulations, the designed classifiers are fixed given the observed samples and the TL-based BEE estimator is safely applied. Finally, regarding synthetic datasets, we note that the flow chart in Figure 8 is valid for all classifiers (QDA, LDA, and OBTL) and the notation Ψ designates the classifier of interest in the corresponding set of experiments. For instance, in the second set of experiments, Ψ refers to Ψ_{LDA}.

In addition to this in-depth analysis of the performance, behavior, and characteristics of our proposed TL-based BEE based on synthetic datasets, we also performed additional validation based on real-world biological datasets. By using RNA-seq datasets syn2759792 and syn4590909 taken from different brain regions for studying brain disorders, we train a QDA classifier using the target data from the RNA-seq dataset syn2759792, and we leverage the source data from syn4590909 to evaluate the performance of the proposed TL-based BEE.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100428.

ACKNOWLEDGMENTS

This work was supported in part by the Department of Energy (DOE) award DE-SC0019303 and the National Science Foundation (NSF) award 1835690. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing (HPRC).

AUTHOR CONTRIBUTIONS

Conceptualization, X.Q., E.R.D., and B.-J.Y.; methodology, O.M., X.Q., E.R.D., and B.-J.Y.; software, O.M.; validation, O.M.; formal analysis, O.M., X.Q., F.J.A., E.R.D., and B.-J.Y.; investigation, O.M.; resources, X.Q., F.J.A., E.R.D., and B.-J.Y.; writing – original draft, O.M.; writing – review & editing, O.M., X.Q., F.J.A., E.R.D., and B.-J.Y.; visualization, O.M.; supervision, X.Q., E.R.D., and B.-J.Y.; project administration, E.R.D. and B.-J.Y.; funding acquisition, X.Q., F.J.A., E.R.D., and B.-J.Y.

DECLARATION OF INTERESTS

Dr. Byung-Jun Yoon is a member of the advisory board of *Patterns*.

Received: September 17, 2021
Revised: October 13, 2021
Accepted: December 16, 2021
Published: January 25, 2022

REFERENCES

1. Dougherty, E.R., and Braga-Neto, U.M. (2006). Epistemology of computational biology: mathematical models and experimental prediction as the basis of their validity. Biol. Syst. *14*, 65–90. https://doi.org/10.1142/S0218339006001726.
2. Diamandis, E.P. (2010). Cancer biomarkers: can we turn recent failures into success? J. Natl. Cancer Inst. *102*, 1462–1467. https://doi.org/10.1093/jnci/djq306.
3. Dalton, L.A., and Dougherty, E.R. (2011). Minimum mean-square error estimation for classification error—Part I: definition and the Bayesian MMSE error estimator for discrete classification. IEEE Trans. Signal Process. *59*, 115–129. https://doi.org/10.1109/TSP.2010.2084572.
4. Braga-Neto, U.M., and Dougherty, E.R. (2004). Is cross-validation valid for small-sample microarray classification? Bioinformatics *20*, 374–380. https://doi.org/10.1093/bioinformatics/btg419.
5. Song, P.P., Xia, J.F., Inagaki, Y., Hasegawa, K., Sakamoto, Y., Kokudo, N., and Tang, W. (2016). Controversies regarding and perspectives on clinical utility of biomarkers in hepatocellular carcinoma. World J. Gastroenterol. *22*, 262–274. https://doi.org/10.3748/wjg.v22.i1.262.
6. Schlimmer, J.C., and Fisher, D. (1986). A case study of incremental concept induction. In Proceedings of the Fifth AAAI National Conference on Artificial Intelligence, pp. 496–501.
7. Goodfellow, I.J., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. (2014). An empirical investigation of catastrophic forgetting in gradient-based neural networks. arXiv, 1312.6211.
8. Farquhar, S., and Gal, Y. (2018). A unifying Bayesian view of continual learning. arXiv, 1902.06494.
9. Farquhar, S., and Gal, Y. (2018). Towards robust evaluations of continual learning. arXiv, 1805.09733.
10. Gossmann, A., Pezeshk, A., Wang, Y.-P., and Sahiner, B. (2021). Test data reuse for the evaluation of continuously evolving classification algorithms using the area under the receiver operating characteristic curve. SIAM J. Mathematics Data Sci. *3*, 692–714. https://doi.org/10.1137/20M1333110.
11. Dalton, L.A., and Dougherty, E.R. (2011). Bayesian minimum mean-square error estimation for classification error—Part II: linear classification of Gaussian models. IEEE Trans. Signal Process. *59*, 130–144. https://doi.org/10.1109/TSP.2010.2084573.
12. Pan, S.J., and Yang, Q. (2010). A survey on transfer learning. IEEE Trans. Knowl. Data Eng. *22*, 1345–1359. https://doi.org/10.1109/TKDE.2009.191.
13. Karbalayghareh, A., Qian, X., and Dougherty, E.R. (2018). Optimal Bayesian transfer learning. IEEE Trans. Signal Process. *66*, 3724–3739. https://doi.org/10.1109/TSP.2018.2839583.
14. Gandal, M.J., Haney, J.R., Parikshak, N.N., Leppa, V., Ramaswami, G., Hartl, C., Schork, A.J., Appadurai, V., Buil, A., Werge, T.M., et al. (2018). Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. Science *359*, 693–697. https://doi.org/10.1126/science.aad6469.
15. Hoffman, J., Rodner, E., Darrell, T., Donahue, J., and Saenko, K. (2013). Efficient Learning of Domain-Invariant Image Representations. arXiv, 1301.3224.
16. Fromer, M., Roussos, P., Sieberts, S.K., Johnson, J.S., Kavanagh, D.H., Perumal, T.M., Ruderfer, D.M., Oh, E.C., Topol, A., Shah, H.R., Klei, L.L., et al. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat. Neurosci. *19*, 1442–1453. https://doi.org/10.1038/nn.4399.
17. Boluki, S., Esfahani, M.S., Qian, X., and Dougherty, E.R. (2017). Constructing pathway-based priors within a Gaussian mixture model for Bayesian regression and classification. IEEE/ACM Trans. Comput. Biol. Bioinform. *16*, 524–537. https://doi.org/10.1109/TCBB.2017.2778715.
18. Boluki, S., Esfahani, M.S., Qian, X., and Dougherty, E.R. (2017). Incorporating biological prior knowledge for Bayesian learning via maximal knowledge-driven information priors. BMC Bioinformatics *18*, 552. https://doi.org/10.1186/s12859-017-1893-4.
19. Dehghannasiri, R., Yoon, B.-J., and Dougherty, E.R. (2014). Optimal experimental design for gene regulatory networks in the presence of uncertainty. IEEE ACM Trans. Comput. Biol. Bioinformatics *12*, 938–950. https://doi.org/10.1109/TCBB.2014.2377733.
20. Broumand, A., Esfahani, M.S., Yoon, B.-J., and Dougherty, E.R. (2015). Discrete optimal Bayesian classification with error-conditioned sequential sampling. Pattern Recogn. *48*, 3766–3782. https://doi.org/10.1016/j.patcog.2015.03.023.
21. Dehghannasiri, R., Yoon, B.-J., and Dougherty, E.R. (2015). Efficient experimental design for uncertainty reduction in gene regulatory networks. BMC Bioinformatics *16*, 1–18. https://doi.org/10.1186/1471-2105-16-S13-S2.
22. Zhao, G., Qian, X., Yoon, B.-J., Alexander, F.J., and Dougherty, E.R. (2020). Model-based robust filtering and experimental design for stochastic differential equation systems. IEEE Trans. Signal Process. *68*, 3849–3859. https://doi.org/10.1109/TSP.2020.3001384.
23. Hong, Y., Kwon, B., and Yoon, B.-J. (2021). Optimal experimental design for uncertain systems based on coupled differential equations. IEEE Access *9*, 53804–53810. https://doi.org/10.1109/ACCESS.2021.3071038.
24. Woo, H.-M., Hong, Y., Kwon, B., and Yoon, B.-J. (2021). Accelerating optimal experimental design for robust synchronization of uncertain Kuramoto oscillator model using machine learning. IEEE Trans. Signal Process. https://doi.org/10.1109/TSP.2021.3130967.
25. Zhao, G., Dougherty, E.R., Yoon, B.-J., Alexander, F.J., and Qian, X. (2020). Uncertainty-aware active learning for optimal Bayesian classifier. In International Conference on Learning Representations (ICLR).
26. Zhao, G., Dougherty, E.R., Yoon, B.-J., Alexander, F.J., and Qian, X. (2021). Bayesian active learning by soft mean objective cost of uncertainty. International Conference on Artificial Intelligence and Statistics (AISTATS) *130*, 3970–3978.

27. Zhao, G., Dougherty, E.R., Yoon, B.-J., Alexander, F.J., and Qian, X. (2021). Efficient Active Learning for Gaussian Process Classification by Error Reduction. 35th Conference on Neural Information Processing Systems (NeurIPS).

28. Yoon, B.-J., Qian, X., and Dougherty, E.R. (2013). Quantifying the objective cost of uncertainty in complex dynamical systems. IEEE Trans. Signal Process. *61*, 2256–2266. https://doi.org/10.1109/TSP.2013.2251336.

29. Yoon, B.-J., Qian, X., and Dougherty, E.R. (2021). Quantifying the multi-objective cost of uncertainty. IEEE Access *9*, 80351–80359. https://doi.org/10.1109/ACCESS.2021.3085486.

30. Robert, C., and Casella, G. (2004). Monte Carlo Statistical Methods (Springer).

31. Gordon, N., Salmond, J., and Smith, A. (1993). A novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEEE Proc. Radar Signal Process. 107–113. https://doi.org/10.1049/ip-f-2.1993.0015.

32. Ackerberg, D.A. (2001). A new use of importance sampling to reduce computational burden in simulation estimation. Quant Mark Econ. *7*, 343–376. https://doi.org/10.1007/s11129-009-9074-z.

33. Muirhead, R.J. (2009). Aspects of Multivariate Statistical Theory (Wiley).

34. Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: a probabilistic programming language. J. Stat. Softw. *76*, 1–32. https://doi.org/10.18637/jss.v076.i01.