The Cost of Dichotomizing Continuous Labels for Binary Classification Problems: Deriving a Bayesian-Optimal Classifier

Soroosh Mariooryad, *Student Member, IEEE,* and Carlos Busso, *Senior Member, IEEE,*

Abstract—Many pattern recognition problems involve characterizing samples with continuous labels instead of discrete categories. While regression models are suitable for these learning tasks, these labels are often discretized into binary classes to formulate the problem as a conventional classification task (e.g., classes with low versus high values). This methodology brings intrinsic limitations on the classification performance. The continuous labels are typically normally-distributed, with many samples close to the boundary threshold, resulting in poor classification rates. Previous studies only use the discretized labels to train binary classifiers, neglecting the original, continuous labels. This study demonstrates that, even in binary classification problems, exploiting the original labels before splitting the classes can lead to better classification performance. This work proposes an optimal classifier based on the *Bayesian maximum a posterior* (MAP) criterion for these problems, which effectively utilizes the real-valued labels. We derive the theoretical average performance of this classifier, which can be considered as the expected upper bound performance for the task. Experimental evaluations on synthetic and real data sets show the improvement achieved by the proposed classifier, in contrast to conventional classifiers trained with binary labels. These evaluations clearly demonstrate the optimality of the proposed classifier, and the precision of the expected upper bound obtained by our derivation.

Index Terms—Bayesian decision classifier, maximum a posteriori.

1 Introduction

Many machine learning problems rely on continuous labels describing the intensity of an attribute. These real-valued descriptors, referred to as *continuous labels*, aim to characterize an inherently fuzzy attribute. Examples include labels derived from psychometric Likert scale questionnaires or similar rating schemes (e.g., sliding bars between two extremes). These continuous labels are popular descriptors for quantifying the degree of emotion of a given stimuli [1], intensity of attribute-based emotional traces such as arousal and valence [2]–[7], distraction level of drivers [8] and conflict level in conversations [9], [10]. Other recognition tasks using continuous labels include likability (pleasantness) of someone's voice [11], [12], speaking rate and speaking liveliness [13], and perceived personality traits in the voice [14], [15]. These continuous labels are used to train machine learning models to automatically estimate these attributes. While regression frameworks are suitable for modeling these continuous labels [16], [17], current studies commonly formulate the problem as a multi-class classification task with discrete classes derived after discretizing the continuous labels [11], [12], [18]. This methodology introduces intrinsic limitations on the classification performance.

• *This study was funded by* National Science Foundation *(NSF) grants (IIS-1217104, IIS-1329659) and a NSF CAREER award (IIS-1453781).*
The authors are with the Multimodal Signal Processing (MSP) laboratory, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: soroosh.ooryad@utdallas.edu, busso@utdallas.edu).

This study explores the practice of dichotomizing continuous attributes in machine learning problems. The study reviews the costs and benefits associated with this practice, highlighting the inevitable loss of information caused by transforming continuous attributes into discrete classes [19]–[22]. We discuss alternative frameworks for binary classification that allow us to incorporate the original continuous attributes, including classifiers with weighted sum loss function and regression-based binary classifiers. Furthermore, we present a novel derivation of a Bayesian-optimal classifier based on the *maximum a posterior* (MAP) criterion for cases where binary labels are discretized from continuous attributes. The derivation assumes that the continuous attribute, and the features of the classifiers follow Gaussian distributions. Although non-Gaussian distributions can also be incorporated in the proposed model, they might not result in mathematically tractable classifiers. In contrast to previous studies, the proposed classifier uses the original, continuous values of the attribute, resulting in an optimum linear classifier. For similar classification tasks, the proposed Bayesian classifier achieves statistically significant improvements over baseline models trained with conventional classifiers. Furthermore, we derive a closed form expression for the expected upper bound performance for such classification tasks. We demonstrate the performance of the proposed classifier, and validate the expected upper bound performance with evaluations using synthetic and real data sets.

The rest of the paper is organized as follows. Section 2 reviews studies that evaluated the cost and benefits

associated with using dichotomized labels. Most of these studies concluded that this practice produced inevitable loss of information. This section presents alternative machine learning approaches for binary classification that consider the original labels. Section 3 describes the proposed Bayesian-optimal classifier and also derives its expected accuracy, which can be considered as an upper bound performance for other conventional classifiers. Section 4 presents experimental results using synthetic data, where we evaluate the loss in performance caused by dichotomized labels. The section also evaluates the robustness of the proposed classifier when we vary the underlying assumptions. Section 5 evaluates the Bayesian-optimal classifier with real data, in emotion recognition experiments. Section 6 concludes the paper, discussing contributions of the proposed model, and research directions for future improvements.

2 Dichotomization of Continuous Attributes

The use of dichotomized labels derived from continuous attributes is a common approach in social and behavioral sciences, developmental and clinical psychology, and psychiatric and criminological research. In the context of statistical analysis, several studies have reported benefits in using this approach [23]. Using dichotomized labels simplifies data analysis. Common statistical tests such as ANOVA require discrete groups or factors. Since many researchers are trained with these tools, studies tend to transform continuous variables into discrete groups (e.g., grouping people's age into discrete classes) [24]. In many cases, the relevant information in the study is whether a behavior or trait is "normal" or "abnormal." Does a continuous measurement exceed a given threshold? (e.g., legal limit of percentage of alcohol in blood while driving). In these cases, having categorical levels is more appropriate, simplifying the understanding of the results between people with different training (e.g., law enforcers, health practitioners). Sometimes, the cut points between categories are defined by theoretically meaningful values or by thresholds used in previous studies [20]. Another common justification for using dichotomized labels is when the data is highly skewed [23], or when relevant classes have different, separated modes. Studies have argued that when the number of samples is small, the use of dichotomized labels increases the robustness of these models [24].

After surveying researchers who used dichotomized labels in their work, DeCoster et al. [20] categorized the benefits of dichotomizing continuous variables into three broad groups: underlying distribution of the variables, simplification of the analysis, and existence of meaningful cuts. They conducted Monte Carlo simulations to demonstrate that in most of the cases, using continuous attributes instead of dichotomized labels provides better results, emphasizing the underlying loss of information in dichotomizing continuous labels [19], [21], [22].

2.1 Cost of Dichotomization

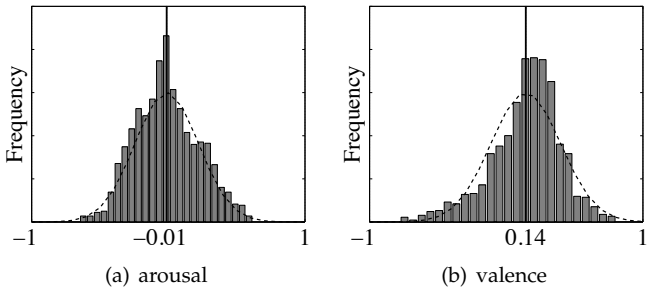
Using simple examples, Cohen [21] demonstrated that dichotomization significantly reduces the dependency between dependent and independent variables. Using Gaussian distributions, that study shows that the variance accounted for the original continuous attributes is reduced by more than 20% after dichotomization. This loss of information is equivalent to discard one third of the samples. The loss of information is even higher when both dependent and independent variables are discretized. MacCallum et al. [19] presented a similar study, where they evaluated the main advantages of using dichotomized labels, concluding that in most of the cases using the continuous labels is a better option. Dichotomization reduces the correlation and the power of statistical tests. It creates significant results that are not observed in the original values, especially for cases where the correlation between dependent and independent variables is small [19]. The selection of the boundaries also creates fluctuations that prevent the results across studies to be compared [22]. This approach also overlooks nonlinear relationship between dependent and independent variables that are only apparent in the original attributes [19]. The loss of information is caused by treating every sample within a group as having the same underlying properties, regardless of the actual values (i.e., samples in the low/high extremes versus samples close to the cutting points). For these reasons, the conclusion of these studies is that there is no good justification for using dichotomized labels.

From modeling perspective, post processing the original annotations may result in noisy labels. If the data had been originally annotated with the dichotomized classes, the resulting classes would have probably been different from the post-processed labels. This can be avoided by constructing the annotation process using the same descriptors that the intended classifier is supposed to use.

Most of the studies analyzing the cost of dichotomization have focussed on statistical tests such as regression and correlation [19]–[24]. To the best knowledge of the authors, studying the use of dichotomization labels in binary or multi class machine learning problems has not been addressed. This study addresses this problem for binary classification tasks.

2.2 Classifiers for Dichotomized Labels

Training classifiers with classes obtained after dichotomizing continuous attributes is a common practice in several affective, paralinguistic, and social behavioral problems. The continuous labels are commonly derived from subjective evaluations conducted by multiple annotators. The scores are averaged across evaluators providing a mean value associated with a stimulus. The central limit theorem says that the *cumulative distribution function* (CDF) of the normalized sum of large number of mutually independent samples tends to a Normal



(a) arousal (b) valence

Fig. 1. Distribution of annotations of arousal and valence as emotion descriptors in the SEMAINE database [6].

CDF. As a result, most of the samples are close to the median/mean of the distribution, and few samples lie on the tails of the distribution. When the continuous labels are discretized into binary classes, the separating threshold often lies close to the median value (e.g., median splits). Therefore, the confusion of the samples in this region reduces the performance of the classifiers trained with these labels.

Figure 1 gives the distributions of the average values for the emotional attributes corresponding to arousal (calm versus active) and valence (negative versus positive), for the utterances in the SEMAINE database [6]. These labels are derived across multiple evaluators. The figures show that the labels have a single mode, where most of the samples are perceived in the middle of the distributions. If binary classes are defined by estimating the means of the values as the separating thresholds (solid vertical lines in Figs. 1(a) and 1(b)), the binary classifiers will have low performance on the samples close to the thresholds. Certain samples lying in different classes close to the boundaries are closer than samples within their own class [19]. As a result, the classification accuracy for samples near the threshold will be close to random. Experimental evaluations with this framework achieved very low classification accuracies with this corpus [25].

In spite of the problems associated with discretizing continuous labels into discrete categories, this approach has been widely used in various problems including recognizing likable versus non-likable voices [11], sleepy versus alert voices [12], positive versus negative expressive behaviors [25]. Different discretization approaches have been proposed to create discrete classes including defining equal-width intervals [26], [27], applying mean [25] or median [11], [28] as the separating threshold, and using the K-means algorithm (K=2) [18], [29]. After defining the discrete classes, the original continuous labels are not considered. Ignoring these continuous labels limits the classifiers' performance. A potential approach to mitigate this problem is to only select samples in the extremes by defining a margin between the classes, discarding samples close to the mode (we have used this approach in Abdelwahab and Busso [30]). This approach is not always possible, given the limited number of samples available in many related problems.

While using dichotomized labels in classification problems introduces inevitable loss of information, there are alternatives that can be implemented to mitigate this problem. These approaches consider the original, continuous labels, in addition to the dichotomized labels. Two of these approaches are described below.

Regression-based binary classifiers: An intuitive approach to incorporate continuous labels in classification problems with dichotomized labels consists in training regression models where the dependent variable is the original continuous metric. During testing, the binary classes are formed by setting the required thresholds over the predicted values of the dependent variable. This approach was used in Wollmer et al. [18]. This study presents an implementation of this approach as baseline, where we train a *support vector regression* (SVR) model to predict the continuous attributes.

Weighted sum loss function: A key problem highlighted by MacCallum et al. [19] is when we assume that each sample from the discrete classes has the same relevance. Given the underlying confusion between samples from different classes close to the boundary, an intuitive solution consists in introducing reliability weights to the samples during training. The continuous labels are transformed into weights for the training samples. For example, we can assign higher weights to samples closer to the extremes, and lower weights to samples closer to the cutting boundaries. A potential implementation of this approach is changing the cost function in *support vector machine* (SVM) by applying these weights to account for the relevance of misclassified samples. We also implement this approach as baseline.

Although these approaches seem intuitive, to the best knowledge of the authors, these approaches are seldom used in previous studies. This study aims to raise awareness of this problem, highlighting the benefits of considering continuous labels during training. Furthermore, we present the derivation of a novel, elegant, Bayesian-optimal classifier that uses the continuous labels to mitigate the loss of information caused by dichotomized labels (Section 3).

3 Bayesian-Optimal Classifier for Dichotomized Labels

We present the novel derivation of the proposed Bayesian-optimal classifier for dichotomized labels. We denote matrices, random variables and random vectors with bold, uppercase letters (e.g., $\mathbf{X}$). We denote realizations of the random variables with lower case letters (e.g., y). We do not distinguish between vectors and scalar, but we specify the dimension of the variables.

3.1 Problem Formulation and Assumptions

Given a feature vector $\mathbf{X}$ of dimension N, and a target attribute $\mathbf{Y}$ (continuous labels), the classification task

consists in classifying the samples according to the discretized labels $Q_{\mathbf{Y}}$ of $\mathbf{Y}$. The study assumes that the continuous attribute is discretized into binary classes. Extension of the approach to cases with more than two discrete classes is left as future work. The discretization process separates the upper half (positive class) from the lower half (negative class) of the distribution by using the mean or median of $\mathbf{Y}$ as the boundary threshold.

To derive a closed-form solution for the proposed classifier and its expected performance, we make the following simplifications. We assume that the feature vector $\mathbf{X}$ follows a multivariate Gaussian distribution $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{XX}})$, where $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{XX}}$ are the mean vector and covariance matrix of $\mathbf{X}$, respectively. Using Gaussian distribution to model the observations is a common approach in many practical problems. Likewise, we assume that the attribute $\mathbf{Y}$ also follows a Gaussian distributions $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\Sigma}_{\mathbf{YY}})$, with mean $\boldsymbol{\mu}_{\mathbf{Y}}$ and variance $\boldsymbol{\Sigma}_{\mathbf{YY}}$. As mentioned in Section 2.2, this assumption is justified by the common approach used to estimate the attribute values consisting in averaging the results across multiple annotations for a given sample. Finally, we assume that the joint *probability density function* (PDF) $P(\mathbf{X}, \mathbf{Y})$ is also Gaussian with covariance matrix $\boldsymbol{\Sigma}$ and mean vector $\boldsymbol{\mu}$ given by equations 1 and 2. $\boldsymbol{\Sigma}_{\mathbf{XY}}$ is the cross-covariance matrix of variables $\mathbf{X}$ and $\mathbf{Y}$.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{XX}} & \boldsymbol{\Sigma}_{\mathbf{XY}} \\ \boldsymbol{\Sigma}_{\mathbf{YX}} & \boldsymbol{\Sigma}_{\mathbf{YY}} \end{bmatrix} \quad (1)$$

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{X}} \\ \boldsymbol{\mu}_{\mathbf{Y}} \end{bmatrix} \quad (2)$$

Without loss of generality, we assume all the variables are zero-mean (i.e., $\boldsymbol{\mu}_{\mathbf{X}} = 0$ and $\boldsymbol{\mu}_{\mathbf{Y}} = 0$). A simple mean subtraction step can achieve this requirement.

3.2 Proposed Classifier

Under the assumptions described in Section 3.1, the conditional PDF of $\mathbf{Y}$ given $\mathbf{X}$ is also Gaussian (Eq. 3), with parameters $\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}}$ (Eq. 4) and $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ (Eq. 5).

$$P(\mathbf{Y}|\mathbf{X}) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}) \quad (3)$$

$$\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Sigma}_{\mathbf{YX}}{\boldsymbol{\Sigma}_{\mathbf{XX}}}^{-1}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) \quad (4)$$

$$\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YX}}{\boldsymbol{\Sigma}_{\mathbf{XX}}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}} \quad (5)$$

For zero-mean, multidimensional Gaussian distributions, the mean and the median of $\mathbf{Y}$ are both zero. Therefore, samples are assigned to the lower class ($\mathcal{L}$) if $\mathbf{Y} < 0$, and to the upper class ($\mathcal{U}$) if $\mathbf{Y} \geq 0$. Under this formulation, we use the Bayesian decision rule with zero-one cost [31], which is known as the *maximum a posteriori* (MAP) classifier (Eq. 6). This classifier selects the label that maximizes the posterior probability $P(Q_{\mathbf{Y}}|\mathbf{X})$ resulting in the following decision boundary:

$$\frac{P(\mathbf{Y} \geq 0|\mathbf{X})}{P(\mathbf{Y} < 0|\mathbf{X})} \underset{\mathcal{L}}{\overset{\mathcal{U}}{\gtrless}} 1. \quad (6)$$

Given the conditional distribution in Equation 3, the decision boundary of the classifier can be obtained starting from Equation 7. We simplify the expression using the error function $\mathrm{erf}(x) = \int_0^x \exp(-\frac{t^2}{2})\mathrm{d}t$. By changing variables and simplifying terms, we obtain a compact expression for the decision boundary given in Equation 8.

$$\frac{\int_0^{+\infty} \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}|}} \exp\left(-\frac{\left(y-\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}}\right)^2}{2\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}\right) \mathrm{d}y}{\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}|}} \exp\left(-\frac{\left(y-\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}}\right)^2}{2\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}\right) \mathrm{d}y} = 1 \quad (7)$$

$$2\,\mathrm{erf}\left(\frac{\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}}}{\sqrt{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}}\right) = 0 \quad (8)$$

Since the error function is only zero when it is evaluated at the origin ($\mathrm{erf}(0) = 0$), Equation 8 implies that $\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}} = 0$ (Eq. 9). Given the expression for the conditional mean in Equation 4, and the assumption that $\boldsymbol{\mu}_{\mathbf{X}} = 0$ and $\boldsymbol{\mu}_{\mathbf{Y}} = 0$, we derive the closed form solution for the decision boundary in Equation 10.

$$\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{X}} = 0 \quad (9)$$

$$\boldsymbol{\Sigma}_{\mathbf{YX}}{\boldsymbol{\Sigma}_{\mathbf{XX}}}^{-1}\mathbf{X} = 0 \quad (10)$$

Notice that deriving the MAP classifier on the discretized binary classes resulted in a linear classifier $\mathbf{W}^T\mathbf{X} = \mathbf{b}$, where $\mathbf{W}^T = \boldsymbol{\Sigma}_{\mathbf{YX}}{\boldsymbol{\Sigma}_{\mathbf{XX}}}^{-1}$ and $\mathbf{b} = 0$. In case of non-zero mean vector for $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\mu}_{\mathbf{Y}}$, we only need to assign $\mathbf{b} = \boldsymbol{\Sigma}_{\mathbf{YX}}{\boldsymbol{\Sigma}_{\mathbf{XX}}}^{-1}\boldsymbol{\mu}_{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{Y}}$. This linear classifier is a hyper-plane in the feature space, where its coefficients depend on the joint covariance matrix $\boldsymbol{\Sigma}$ (Eq. 1).

3.2.1 Training the Classifier

The training step consists in estimating $\boldsymbol{\Sigma}_{\mathbf{XX}}$, $\boldsymbol{\Sigma}_{\mathbf{YY}}$, $\boldsymbol{\Sigma}_{\mathbf{YX}}$ from the features ($\mathbf{X}$) and the continuous label ($\mathbf{Y}$). In contrast to conventional classifiers which are trained with the discretized classes, the proposed classifier utilizes the original continuous labels to estimate $\boldsymbol{\Sigma}_{\mathbf{YX}}$. This is a key feature of the proposed classifier, which contrasts with conventional machine learning solutions trained with binary classes derived from continuous attributes.

3.2.2 Testing the Classifier

The testing step uses the linear decision boundary in Equation 10 to classify the given samples according to:

$$Label(X) = \begin{cases} \mathcal{L}, & \text{if } \boldsymbol{\Sigma}_{\mathbf{YX}}{\boldsymbol{\Sigma}_{\mathbf{XX}}}^{-1}\mathbf{X} < 0. \\ \mathcal{U}, & \text{if } \boldsymbol{\Sigma}_{\mathbf{YX}}{\boldsymbol{\Sigma}_{\mathbf{XX}}}^{-1}\mathbf{X} \geq 0. \end{cases} \quad (11)$$

3.3 Performance Analysis

The study presents a closed form expression for the expected accuracy of the proposed classifier. Let's consider the 1×1 variable

$$\mathbf{V} = \mathbf{\Sigma_{YX}}\mathbf{\Sigma_{XX}}^{-1}\mathbf{X}. \tag{12}$$

Assuming that the upper class $\mathcal{U}$ (i.e., $\mathbf{Y} \geq 0$) is the target class in the binary classification problem, then $P(\mathbf{Y} > 0, \mathbf{V} > 0)$ is the probability of *true positive* (TP), and $P(\mathbf{Y} < 0, \mathbf{V} < 0)$ is the probability of *true negative* (TN). We can combine both probabilities to estimate the expected accuracy of the classifier:

$$E[Acc] = 100 \cdot [P(\mathbf{Y} > 0, \mathbf{V} > 0) + P(\mathbf{Y} < 0, \mathbf{V} < 0)] \tag{13}$$

Since $\mathbf{V}$ is a linear transformation of the feature vector $\mathbf{X}$, and $\mathbf{\Sigma_{XX}}$ is a nonsingular matrix, $\mathbf{V}$ also follows a Gaussian distribution, $\mathbf{V} \sim \mathcal{N}(0, \mathbf{\Sigma_{YX}}\mathbf{\Sigma_{XX}}^{-1}\mathbf{\Sigma_{XY}})$. To estimate the true positive probability, we use the Sheppard's theorem on median dichotomy [32], [33], which gives a closed form expression for the integral over the first quadrant of the joint distribution of two Gaussian variables (in this case $\mathbf{Y}$ and $\mathbf{V}$). The expression for the true positive probability is given by Equation 14, where $\rho_{\mathbf{VY}}$ is the normalized correlation coefficient between $\mathbf{Y}$ and $\mathbf{V}$.

$$P(\mathbf{Y} > 0, \mathbf{V} > 0) = \frac{1}{4} + \frac{\arcsin \rho_{\mathbf{VY}}}{2\pi} \tag{14}$$

$$\rho_{\mathbf{VY}} = \frac{E[\mathbf{VY}]}{\sqrt{E[\mathbf{VV}]E[\mathbf{YY}]}}$$

$$\rho_{\mathbf{VY}} = \sqrt{\frac{\mathbf{\Sigma_{YX}}\mathbf{\Sigma_{XX}}^{-1}\mathbf{\Sigma_{XY}}}{\mathbf{\Sigma_{YY}}}} \tag{15}$$

Given the symmetry of the problem, the true negative probability equals the true positive probability. Combining Equations 13, 14, and 15, we derive the expected accuracy of the proposed classifier:

$$E[Acc] = 50 + 100\frac{\arcsin\left(\sqrt{\frac{\mathbf{\Sigma_{YX}}\mathbf{\Sigma_{XX}}^{-1}\mathbf{\Sigma_{XY}}}{\mathbf{\Sigma_{YY}}}}\right)}{\pi} \tag{16}$$

Under the assumptions of the models (see Sec. 3.1), this accuracy can be considered as an upper bound on the expected performance for other classifiers, given the optimality of the Bayesian classifier. It is important to stress that Equation 16 provides the expected performance. It is possible that a classifier under certain conditions achieves better classification results, as we will discuss in Sections 4 and 5

3.4 Case Study for Single Dimensional Feature

An interesting case arises when the feature vector $\mathbf{X}$ consists of a single variable. In this case, the expected accuracy in Equation 16 can be simplified by Equation 17, where $\rho_{\mathbf{XY}}$ is the normalized correlation coefficient between $\mathbf{X}$ and $\mathbf{Y}$.

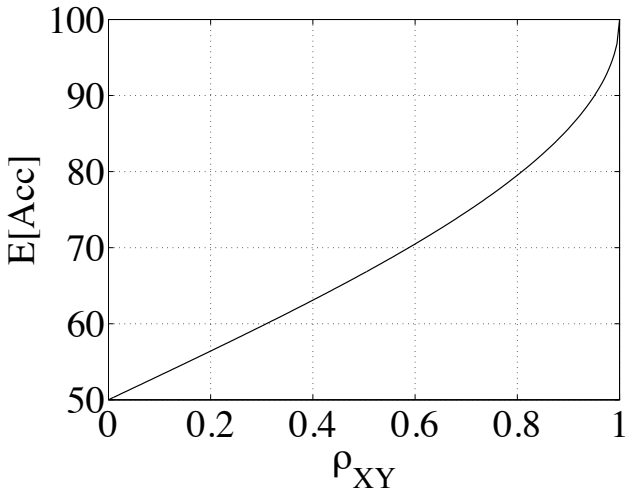
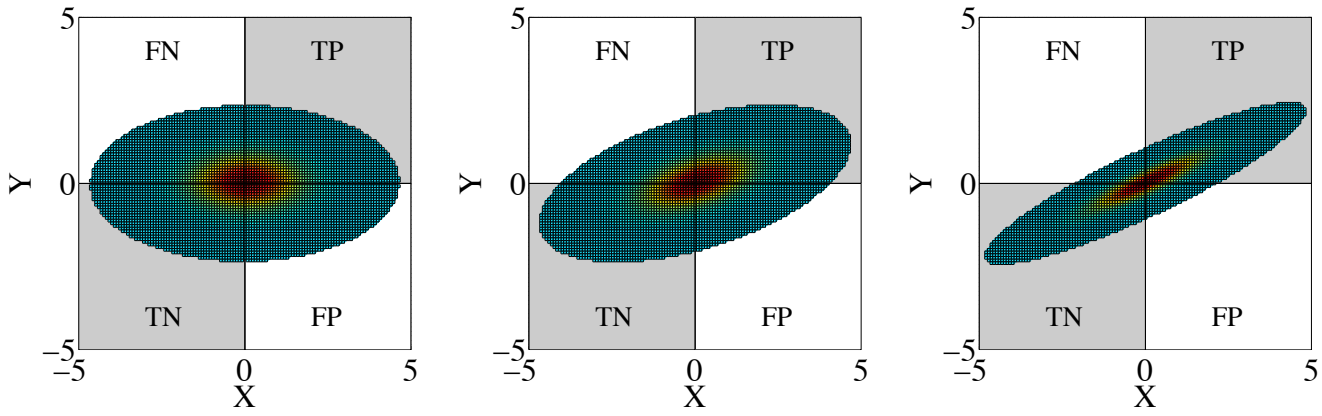


Fig. 2. Expected accuracy of predicting target label $Q_{\mathbf{Y}}$ with a single feature $\mathbf{X}$ as a function of $\rho_{\mathbf{XY}}$ (see Eq. 17).

$$E[Acc] = 50 + 100\frac{\arcsin(\rho_{\mathbf{XY}})}{\pi} \tag{17}$$

We will consider the case with a single feature to illustrate the intrinsic limitation in accuracy imposed by using binary labels derived from continuous attributes. We use equation 17 to plot the expected performance of the proposed classifier with respect to $\rho_{\mathbf{XY}}$. Figure 2 gives the results which show that the correlation between the feature $\mathbf{X}$ and the label $\mathbf{Y}$ should be very high to achieve high performance. For instance, we expect to achieve only 80% accuracy when $\rho_{\mathbf{XY}}$=0.8. When the correlation between the feature and target attribute is zero, the expected accuracy is 50% which is the performance of a random classifier.

Figure 3 gives the joint distribution of a single feature $\mathbf{X}$ and a label $\mathbf{Y}$ for three levels of correlation. From Equation 12, we observe that when $\mathbf{X}$ is a one dimensional feature, $\mathbf{V} = (\mathbf{\Sigma_{YX}}/\mathbf{\Sigma_{XX}})\mathbf{X}$ ($\mathbf{\Sigma_{YX}}$ and $\mathbf{\Sigma_{XX}}$ are scalars). As long as $\mathbf{\Sigma_{YX}} > 0$, the decision rule in Equation 11 is equivalent to evaluating the sign of $\mathbf{X}$ (i.e., $\mathbf{X} > 0$ implies that $\mathbf{V} > 0$, and, therefore, the sample belong to $\mathcal{U}$). The classification errors for *false negative* (FN) and *false positive* (FP) cases correspond to the second and four quadrants, respectively. This figure shows that higher correlation values reduce the areas for FN and FP. When $\rho_{\mathbf{YX}} \neq 1$, the figure illustrates that using the median or mean of $\mathbf{Y}$ to define the binary classes intrinsically introduces an upper bound performance on the accuracy, given the errors on the corners (i.e., FP and FN). This upper bound performance is analytically derived in Equation 16 based on the performance of the Bayesian-optimal classifier.



(a) $\rho_{\mathbf{YX}} = 0$ (b) $\rho_{\mathbf{YX}} = 0.5$ (c) $\rho_{\mathbf{YX}} = 0.9$

Fig. 3. Illustration of joint distribution of a single predictor feature ($\mathbf{X}$) and continuous target label ($\mathbf{Y}$) for different correlation between $\mathbf{X}$ and $\mathbf{Y}$ (TP: true positive, TN: true negative, FP: false positive, FN: false negative).

4 Experimental Evaluation with Synthetic Datasets

This section evaluates the proposed classifier with synthetic datasets, where we can carefully control the conditions of the experiments (e.g., whether the underlying assumptions hold). In particular, we consider cases where the features follow Gaussian (Sec. 4.1) and non-Gaussian distributions (Sec. 4.2).

We compare the proposed Bayesian-optimal classifier with conventional classifiers: *linear discriminant classifier* (LDC), *logistic regression classifier* (LRC), and *support vector machine* (SVM), with linear and *radial basis function* (RBF) kernels. The LDC is derived by assuming normal densities with equal covariance for each class [31]. The LRC is trained with maximum-likelihood estimation. For the SVM, we set its complexity parameter c to 0.1, following the results from our previous study [34]. All these baseline classifiers are trained using the dichotomized labels after median split.

We also implement the weighted-SVM classifier and regression-based binary classifier approach using SVR discussed in Section 2.2. During testing, the regression model predicts the continuous labels, assigning the samples to the positive class when the predicted scores are greater or equal to the cutting threshold. In contrast to other baseline methods, these approach uses the original continuous values.

4.1 Synthetic Datasets with Gaussian Distribution

In this section, we create machine learning problems satisfying the assumptions made on our derivation. We use the generative model in Equation 18 to create the synthetic data. $\mathbf{U}$ is an $(N+1)\times 1$ vector, where its first N elements are considered as the features $\mathbf{X}$, and the last element is considered as the continuous label $\mathbf{Y}$ (see Eq. 22). The $(N+1)\times 1$ vector $\mathbf{Z}$ is generated by sampling a zero-mean multidimensional white Gaussian distribution (i.e., $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I})$). $\mathbf{T}$ is the spectral shaping matrix of the transformation, where each of its elements is fixed, but randomly set by sampling the standard normal distribution. The $(N+1)\times 1$ vector $\mathbf{S}$ is created by sampling a multidimensional Gaussian distribution with covariance matrix $\sigma^2\mathbf{I}$ (i.e., $\mathbf{S} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$). For simplicity, σ is assumed constant, generated by sampling a normal distribution and taking its absolute value.

$$\mathbf{U} = \mathbf{TZ} + \mathbf{S} \quad (18)$$

$$\mathbf{Z}_{(N+1)\times 1} = \begin{bmatrix} \mathbf{Z}_{\mathbf{X}_{N\times 1}} \\ \mathbf{Z}_{\mathbf{Y}_{1\times 1}} \end{bmatrix} \quad (19)$$

$$\mathbf{S}_{(N+1)\times 1} = \begin{bmatrix} \mathbf{S}_{\mathbf{X}_{N\times 1}} \\ \mathbf{S}_{\mathbf{Y}_{1\times 1}} \end{bmatrix} \quad (20)$$

$$\mathbf{T}_{(N+1)\times(N+1)} = \left[\begin{array}{c|c} \mathbf{A}_{N\times N} & \mathbf{B}_{N\times 1} \\ \hline \mathbf{C}_{1\times N} & \mathbf{D}_{1\times 1} \end{array}\right] \quad (21)$$

$$\mathbf{U} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{Z}_{\mathbf{X}} \\ \mathbf{Z}_{\mathbf{Y}} \end{bmatrix} + \begin{bmatrix} \mathbf{S}_{\mathbf{X}} \\ \mathbf{S}_{\mathbf{Y}} \end{bmatrix} \quad (22)$$

With these definitions, the theoretical covariance matrices $\mathbf{\Sigma_{XX}}$, $\mathbf{\Sigma_{YY}}$ and cross-covariance matrix $\mathbf{\Sigma_{XY}}$ are given by Equations 23-25, respectively.

$$\mathbf{\Sigma_{XX}} = \mathbf{AA}^T + \mathbf{BB}^T + \sigma^2\mathbf{I} \quad (23)$$

$$\mathbf{\Sigma_{YY}} = \mathbf{CC}^T + \mathbf{DD}^T + \sigma^2 \quad (24)$$

$$\mathbf{\Sigma_{XY}} = \mathbf{AC}^T + \mathbf{BD}^T \quad (25)$$

There are some important remarks about this generative model. First, the cross-covariance matrix $\mathbf{\Sigma_{XY}}$ depends only on the values of the transformation matrix $\mathbf{T}$ (see Eq. 25). Its entries determine the correlation between features and the continuous label. Second, the vector $\mathbf{S}$ in Equation 18 introduces uncorrelated variability in the features and the label. Therefore, the higher its variance σ^2, the lower the normalized correlation between the features and the label. As a result, increasing the values of σ reduces the expected upper bound accuracy of the proposed classifier.

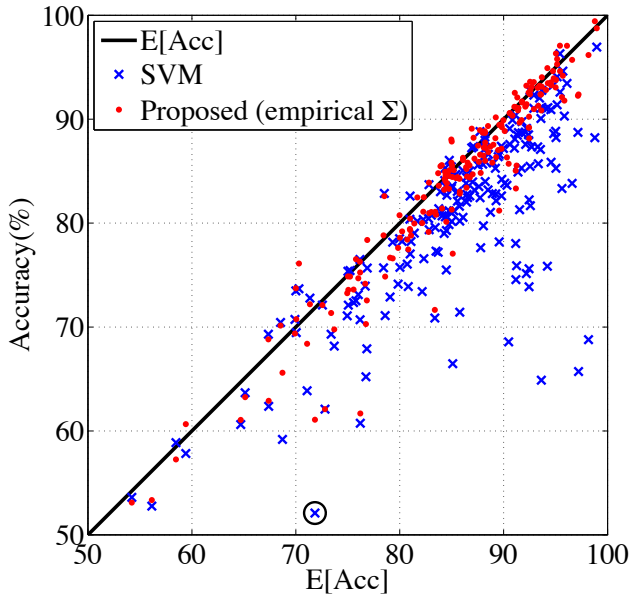


Fig. 4. Accuracy of the proposed method and SVM classifier on the synthetic datasets. The solid line corresponds to cases where the actual accuracy matches the expected upper bound performance.

For the classification experiments, we aim to cover different scenarios. We randomly set the number of features $N \in [2-50]$ and number of samples $M \in [100-1000]$. This experiment is repeated 200 times. In each run, we randomly generate $\mathbf{T}$ and σ to create the feature vectors $\mathbf{X}$ and the continuous labels $\mathbf{Y}$. Over the 200 runs, the median and 95-percentile absolute correlation between the generated features and labels is 0.12 and 0.37 respectively. We randomly partition the data into two balanced sets for training the classifiers and testing their performance. We evaluate two conditions. First, we use the actual covariance matrices using Equations 23-25. We refer to this condition as the *theoretical* case. We also use the theoretical covariance matrices to estimate the expected upper bound performance using Equation 16. Second, we use the synthesized data to estimate the sample covariance matrices. We refer to this condition as the *empirical* case.

Table 1 reports the average results of this experiment across the 200 runs. The expected upper bound performance estimated with Equation 16 using the theoretical covariances is 85%. The performances of the conventional linear classifiers is around 81% (with the exception of SVM with RBF kernel). The proposed classifier using the empirical covariance matrices yields an average accuracy of 84.1%. This is a statistically significant improvement over the conventional linear classifiers, based on the population mean hypothesis test (p-value$< 1e-10$). On average, the proposed method achieves accuracies that are 5.7% (relative) higher than the ones achieved by other classifiers. Although the machine learning problems are the same in terms of the features and labels, the extra information hidden in the continuous values provides useful information to obtain this optimal classifier. This information is not utilized by conventional classifiers which only use the discretized version of the labels. The SVR-based and weighted-SVM classifiers utilize the continuous values, but their performance is lower than the proposed method. Table 1 also reports the accuracy of the classifier built with the theoretical covariance, which shows a performance equal to the expected upper bound. The theoretical covariance matrices are not available in practice. Nevertheless, this result indicates that using more reliable estimations of the covariance matrices improves the recognition rates.

Figure 4 depicts the accuracy of the proposed classifier and the SVM classifier with linear kernel against the upper bound performance for each of the 200 runs. The $x-$axis shows the expected upper bound performance for each condition (Eq. 16), derived from the theoretical covariance matrices (Eqs. 23-25). The $y-$axis represents the actual performance given by the classifiers. For example, the point highlighted with a circle in the figure has an accuracy of 52.1%, while its expected upper bound performance is 71.8%. The solid line represents cases where the actual performance equals the expected upper bound performance. The accuracies achieved by the proposed classifier (red dots) lie closer to the solid line, revealing that they are closer to the average upper bound performance. Since $E[Acc]$ is the expected upper bound performance, not the upper bound itself, certain cases can achieve accuracies above $E[Acc]$ (points above the solid line). The accuracies for the SVM classifier (blue crosses) are further from the line.

Table 1 reports the average classification performance over 200 runs, where we consider different number of features and samples. To analyze the effect of number of features on the performance of the proposed classifier, we conducted a similar experiment by fixing the number of samples to 500. Then, we increase the number of features from 2 to 50, and we evaluate the classification performance. Figure 5 shows the mean and standard deviation of the accuracy obtained with the proposed

TABLE 1
Evaluation of the proposed Bayesian-optimal classifier on synthetic datasets with Gaussian distributed features. The results are compared with LDC, LRC and SVM with linear and RBF kernels. We also evaluate the SVR-based and weighted-SVM classifiers described in Section 2.2.

Classifier	Accuracy [%]
Upper bound ($E[Acc]$)	85.4
SVM (linear kernel)	80.5
SVM (RBF kernel)	50.2
LDC	80.9
LRC	81.0
Proposed (empirical $\mathbf{\Sigma}$)	84.1
Proposed (theoretical $\mathbf{\Sigma}$)	85.4
SVR-based classifier	78.7
Weighted-SVM classifier	81.1

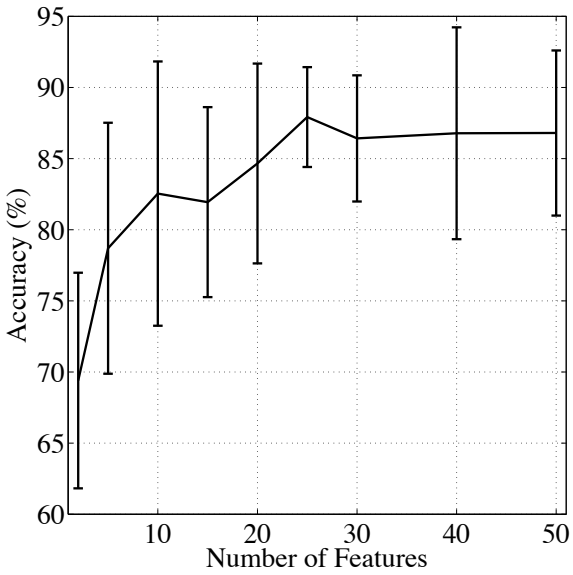


Fig. 5. Effect of increasing the number of features on the average accuracy of the proposed classifier. The number of samples is set to 500.

classifier, which reveals that adding more features improves the classification performance.

4.2 Synthetic Datasets with Non-Gaussian Distributions

The previous section shows the effectiveness of the proposed derivation when the assumptions described in Section 3.1 hold. This section evaluates the proposed approach with non-Gaussian features to assess the robustness against violations of this key assumption. Instead of defining the joint distribution of the feature set $\mathbf{X}$, we generate individual features with known marginal distributions. Equation 26 describes the approach to generate the dataset, where X_i is an individual feature and $\mathbf{Y}$ is the continuous label. This equation imposes a dependency between the features and the label as a function of a weighting factor a_i and a random signal S_i. We consider two cases. First, we assume that $\mathbf{Y}$ and S_i follow exponential distributions. In this case, X_i is a weighted sum of exponential densities, resulting in a hyper-exponential distribution. Second, we assume that $\mathbf{Y}$ and S_i follow Gamma distributions. Therefore, X_i also follows a Gamma distribution. The linear combination coefficient (a_i) and the parameters of the distributions of $\mathbf{Y}$ and S_i are randomly selected.

$$X_i = a_i \mathbf{Y} + S_i \tag{26}$$

Table 2 summarizes the results of this evaluation. With exponential distributions for $\mathbf{Y}$ and X_i, the proposed method still outperforms other classifiers. However, the improvement is not as high as the improvements observed for features and labels following Gaussian distributions (Table 1). When $\mathbf{Y}$ and X_i follow a Gamma distribution, the performance of the proposed method drops. These results show that when the labels and the features do not follow a Gaussian distribution, the proposed approach may or may not provide optimum performance. Table 2 only shows robustness against exponential distributions. While the exponential distribution is a special form of Gamma distribution, varying the parameters of Gamma distribution generates more diverse distributions affecting the performance of the proposed system. This evaluation demonstrates that the proposed approach is sensitive to the assumptions, only when the distributions of the labels and features radically deviate from Gaussian distributions.

5 Experimental Evaluation with Real Datasets: Emotion Recognition

We report classification evaluations with real data sets to validate the proposed classifier and the expected upper bound performance for binary classification problems trained with discretized labels. In particular, the study focuses on detecting low and high levels of the emotional dimensions arousal (i.e., calm versus active) and valence (i.e., negative versus positive).

5.1 Database, Labels and Features

This study uses the *sustained emotionally coloured machine-human interaction using nonverbal expressions* (SEMAINE) database [6]. This is an audiovisual corpus of dyadic interactions between users and operators. The operators act the role of four characters with different personalities (happy, gloomy, angry and pragmatic) to induce various emotional reactions in the users. We consider 44 sessions recorded from nine subjects.

The expressive behaviors of the users are annotated by multiple annotators (2-8) in terms of degree of arousal, and valence. The perceptual evaluation included other emotional dimensions that are not considered in this study. These continuous attributes are collected along the duration of the videos, with the FEELTRACE toolkit [35]. This interface presents the video stimuli to the subjects, while continuously recording the mouse cursor location on a two-dimensional coordinate system, in which

TABLE 2
Evaluation of the proposed Bayesian-optimal classifier on synthetic datasets with non-Gaussian distributed features (Exponential and Gamma).

Distribution	Exponential	Gamma
Classifier	Accuracy [%]	
SVM (linear)	78.8	79.8
SVM (RBF)	63.1	73.7
LDC	77.8	79.1
LRC	79.4	79.5
Proposed (empirical $\mathbf{\Sigma}$)	79.5	68.8
SVR-based classifier	63.9	78.8
Weighted-SVM classifier	76.2	73.2

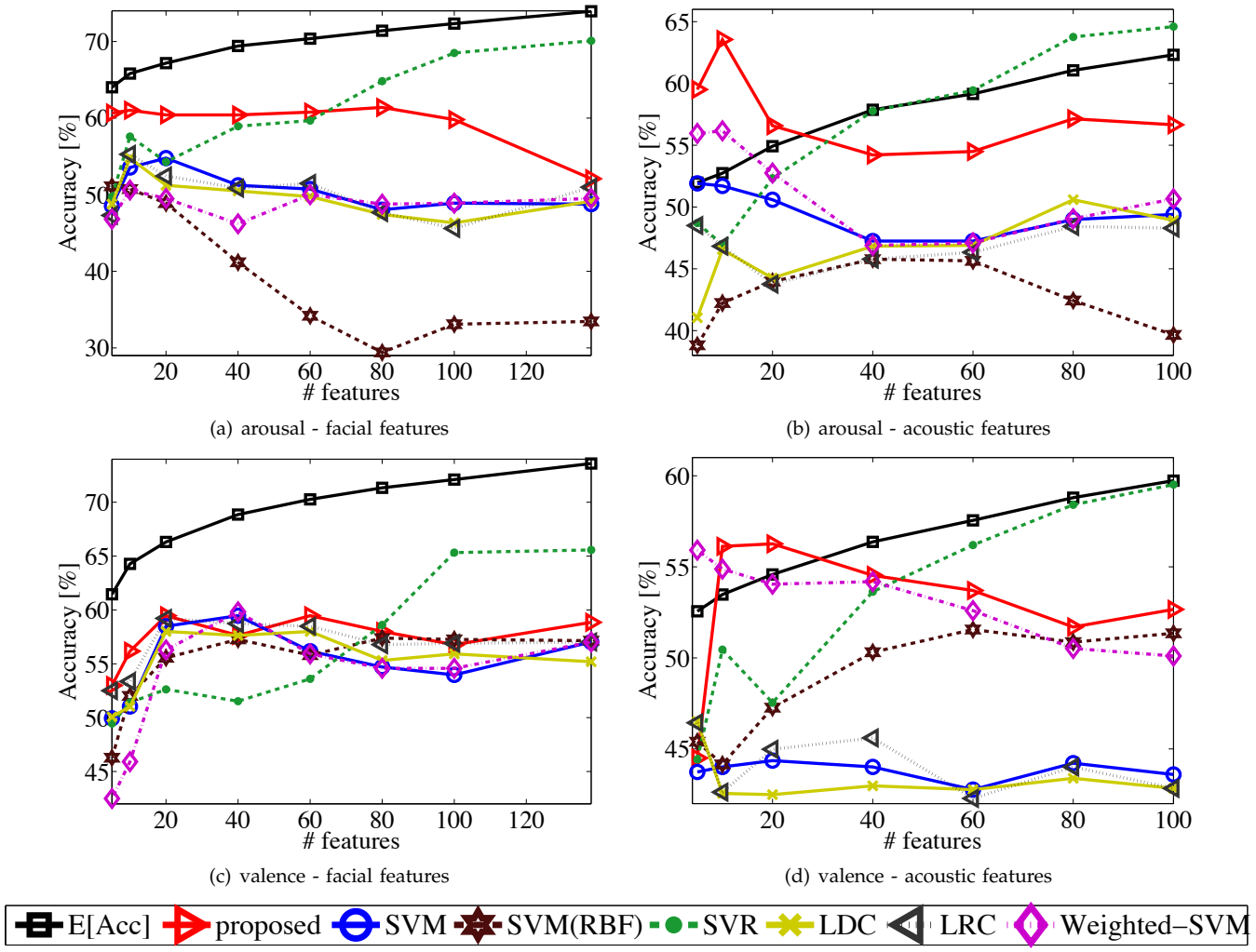


Fig. 6. Accuracy of the proposed method, other baseline methods and the upper bound performance for arousal (a-b) and valence (c-d) recognition (SEMAINE database). Results are given in terms of the number of features for speech and face emotion recognition.

its axes represent these two attributes. The annotators reflect their perceived emotional content over time by moving the mouse cursor to the location that better captures the emotional attribute. The collected traces are mapped into the interval [-1, +1] for each dimension. In our previous study, we have identified a reaction lag in the evaluators' responses, as they watched the video, perceived the emotional content, and responded by moving the mouse cursor [34]. We have compensated for this delay in the annotations using a mutual-information based technique introduced in Mariooryad and Busso [36]. The data is segmented into turns and the average value across the duration of the turn is assigned to it. Figure 1 depicts the histogram of these two emotional descriptors. The figure also describes the two binary classes, where the mean of the attributes are used as the separating thresholds.

This study explores speech emotion recognition during the users' speaking turns (1435 segments), and facial emotion recognition during the users' listening turns (816 segments). We do not analyze the facial expressions of the users during speaking turns, since the lexical information introduces variability, affecting the performance of the classifiers [37], [38]. The speech features are the feature set proposed for the INTERSPEECH 2011 speaker state challenge [12]. The set contains 4368 features describing prosodic, spectral and voice quality features, and it is extracted using the OpenSMILE toolkit [39]. The facial features are extracted using the *computer expression recognition toolkit* (CERT) [40], which includes a set of 20 *action units* (AUs) and three head rotation parameters. For each turn, we estimate the following six statistics: mean, median, lower quartile, upper quartile, 1%-percentile (∼min) and 99%-percentile (∼max). We represent facial expression with this 138D feature vector (20 AUs + 3 head pose $\times$ 6).

5.2 Classification Experiments

The proposed approach requires the estimation of the covariance and cross-covariance matrices ($\mathbf{\Sigma_{XX}}$, $\mathbf{\Sigma_{YY}}$,

TABLE 3
Performance of the Bayesian-Optimal classifier for different number of features evaluated on the SEMAINE database. The table lists the values for $\rho_{\mathbf{VY}}$ (see Eq. 15). Acc: Accuracy, TP: true positive, TN: true negative, FP: false positive, FN: false negative.

Problem	Stat	Number of features					
		10	20	40	60	80	100
Arousal facial features	$\rho_{\mathbf{VY}}$	0.48	0.51	0.57	0.60	0.62	0.65
	Acc	62%	60%	61%	60%	61%	60%
	TP	38%	36%	38%	37%	40%	41%
	TN	24%	24%	23%	23%	21%	19%
	FP	29%	28%	30%	29%	31%	33%
	FN	10%	11%	10%	10%	7%	7%
Arousal acoustic features	$\rho_{\mathbf{VY}}$	0.09	0.15	0.25	0.28	0.34	0.38
	Acc	63%	56%	54%	55%	57%	57%
	TP	35%	29%	22%	24%	27%	28%
	TN	28%	27%	32%	31%	30%	29%
	FP	25%	26%	21%	23%	24%	25%
	FN	11%	17%	24%	22%	19%	18%
Valence facial features	$\rho_{\mathbf{VY}}$	0.43	0.49	0.56	0.60	0.62	0.64
	Acc	56%	60%	57%	59%	58%	57%
	TP	53%	54%	54%	51%	47%	46%
	TN	3%	6%	3%	8%	11%	11%
	FP	40%	37%	40%	34%	32%	32%
	FN	4%	3%	3%	6%	10%	11%
Valence acoustic features	$\rho_{\mathbf{VY}}$	0.11	0.14	0.20	0.24	0.27	0.30
	Acc	56%	57%	54%	54%	51%	53%
	TP	45%	38%	36%	35%	29%	30%
	TN	11%	19%	18%	19%	22%	23%
	FP	33%	26%	27%	26%	22%	22%
	FN	10%	18%	19%	21%	26%	25%

and $\mathbf{\Sigma_{XY}}$). However, the large dimension of the feature sets will produce unreliable estimation. Hence, we reduce the feature set to 100. For speech features, we rank the features according to their information gain with respect to the discretized labels. Since the set of facial features are significantly smaller, we use *forward feature selection* (FFS) maximizing the performance of the SVM classifier to rank all the 138 features. We also report the results using smaller feature sets. We implement the classification experiments using *leave-one-subject-out* (LOSO) cross-validation, where the data of each subject is either in the training or testing partitions. For each cross-validation fold, we normalize the features by subtracting their mean estimated from the training set (i.e., zero mean feature vector).

Table 3 and Figure 6 depict the performances of the proposed method for facial and acoustic features in recognizing low and high levels of arousal and valence. The figures and the table report the performance for different number of features. In most of the cases the proposed classifier clearly outperforms conventional linear classifiers. Notice that in some cases the baseline classifiers perform even lower than 50%, which shows the difficulty of the task. Similar results have been reported on this corpus for word-level recognition experiments [25].

The evaluations in Figure 6 clearly show that the approaches that utilize continuous labels (i.e., proposed method, SVR-based classifier, and weighted-SVM) outperform conventional classifiers in most of the cases.

TABLE 4
Statistical significance test to verify if the proposed method outperforms (✓) SVM or not (✗) using proportion hypothesis test with $p - value = 0.05$.

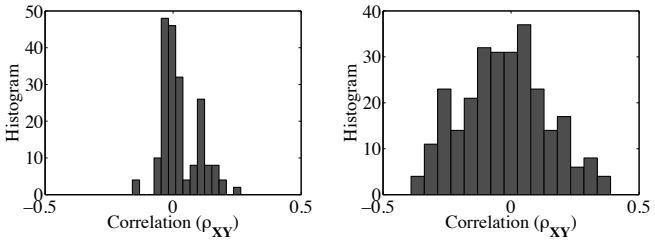
Number of features	10	20	40	60	80	100
arousal - facial features	✓	✓	✓	✓	✓	✓
arousal - acoustic features	✓	✓	✓	✓	✓	✓
valence - facial features	✓	✗	✗	✗	✗	✗
valence - acoustic features	✓	✓	✓	✓	✓	✓

Also, the figures suggest that with lower number of features the proposed Bayesian-optimal classifier generally provides the best performance. As we add more features, the SVR-based classifier becomes the system with the best accuracy. Notice that adding more features decreases the validity of joint Gaussian distribution assumption on the features. In fact, we use the *multivariate normality* (MVN) package [41] to test the normality assumption as we increase the number of selected features. The three tests of this package [42]–[44] consistently indicate that the distribution of the feature set deviates from the multivariate Gaussian distribution. The p-values for all these tests are always less than 0.05, and they become near zero when the number of features exceeds 10. In these cases, the SVR-based classifier outperforms the proposed method. SVR-based classifier also uses the continuous labels without making assumptions on the distribution of features.

It is important to highlight that the proposed method outperforms conventional classifiers trained with the dichotomized labels. We compare the proposed Bayesian-optimal classifier against the SVM classifier that does not consider the continuous labels. We use the proportion hypothesis test asserting significance at p-value=0.05. We consider cases where both classifiers are trained with a given number of features. Tables 4 shows the results, where a check mark (✓) indicates that our proposed method is significantly better than the SVM classifier. With the exception of detecting valence with facial features, the analysis shows that the proposed method outperforms the SVM classifier.

In theory, adding more features should increase the performance (see Fig. 5). Figure 6 shows that adding more features does not always increase the performance of the Bayesian-optimal classifier, indicating that a more sophisticated feature selection approach may provide better performance. We use information gain and FFS, which can provides local optimum.

The upper bound ($E[Acc]$) in Figure 6 is estimated using Equation 16 based on the cross-covariance of features and continuous label. The figure clearly demonstrates the optimality of the expected upper bound ($E[Acc]$). In all cases, the expected upper bound performance is higher than the performances of the conventional classifiers. Only the classifiers that make use of continuous labels (i.e., proposed method, SVR-based classifier, and weighted-SVM) exceed the expected upper bound



(a) facial features (b) acoustic features

Fig. 7. Histogram of correlation between continuous emotion descriptors (arousal and valence) and individual features ($\rho_{\mathbf{XY}}$).

level in some cases with acoustic features. Notice that this is expected, since the $E[Acc]$ is the expected upper bound, not an actual upper bound value. Similar to the results shown in Figure 5 for synthetic data, the accuracy increases as we add extra features. Notice that estimating $E[Acc]$ is straightforward with Equation 16 after estimating the joint covariance matrix of the features and the continuous label. This value provides a practical reference performance that can be achieved for a given problem, which can guide the feature selection process.

5.3 Analysis of Features and Labels

We study the correlation between the individual features and the continuous labels ($\rho_{\mathbf{XY}}$) for either arousal or valence (Eq. 17). Figure 7 gives the histogram of the correlation observed for all the selected facial and acoustic features (we combine results for arousal and valence). The figure suggests that the features are not highly correlated with the continuous label, highlighting the challenges in this classification problem.

We also consider the correlation between all the selected facial/speech features and the continuous labels by estimating $\rho_{\mathbf{VY}}$ with variable $\mathbf{V}$ (see Eqs.12 and 15). Table 3 lists the correlation for different number of features. It is interesting that $\rho_{\mathbf{VY}}$ increases as we add more features, replicating the results observed with synthetic data (see Fig. 5). We hypothesize that using a more robust feature selection algorithm that prevents local optimum may give even better results than the information gain and FFS algorithms used in this study.

6 Conclusions and Discussion

This work raised awareness of the limitations of using dichotomized labels in machine learning problems. We presented systematic analyses showing that using the original continuous labels, in addition to the binary classes, can mitigate the loss of information caused by dichotomizing the labels. In particular, we presented a Bayesian-optimal classifier for binary problems where the discrete labels are obtained by discretizing continuous, normally-distributed labels. The derivation of the classifier included the expected performance of this classifier. Given the optimality of Bayesian classifiers, the derived expression serves as the expected upper bound performance for other classifiers. The experiments on synthetic and real data sets showed the optimality of the classifier. The proposed method, which considers the original continuous label to estimate the covariance and cross-covariance matrices of the features and labels, yields better classification performance. This information is commonly neglected in conventional approaches, where only the discretized labels are used. As a result, they achieve lower classification performance.

The results from the synthetic data set showed that using the actual covariance matrices improves the classification performance of the proposed classifier (see Table 1). These results suggest that estimating accurate sample covariance matrices is a key problem. This is particularly important in speech emotion recognition, where hundreds of features are commonly used. We are exploring robust covariance estimation approaches especially when the number of samples is limited [45]. Other options include covariance estimation with smallest determinant to remove outliers and covariance adaptation from universal background model. By improving the estimation of the sampled covariances, we expect to reduce the gap in the classification performance observed when we use either the theoretical or the empirical covariance matrices.

Another interesting research direction is to analyze the generalization of the framework with mismatched conditions (e.g., cross-corpus or cross-language experiments), or with noisy audio. For these cases, the actual and estimated covariance matrices may differ, reducing the classification performance. We expect that adaptation schemes to modify $\mathbf{\Sigma_{YX}}$ and $\mathbf{\Sigma_{XX}}$ can increase the robustness and generalization of the proposed approach.

One limitation of the approach is the assumptions made on the features and label. This work assumed that the features follow a multidimensional Gaussian distribution, which is reasonable for many machine learning problems. However, the proposed classifier can produce lower performance when this assumption is not realistic. In those cases, a closed form solution for the decision boundary similar to Equation 10 may not be available.

The derived expected accuracy of the optimal classifier for a single dimensional feature increases when the correlation between the feature and the continuous label increases, as shown in Equation 17 and Figures 2 and 3. In the general case with more than one feature, maximizing the upper bound performance requires maximizing the quadratic form $\mathbf{\Sigma_{YX}\Sigma_{XX}}^{-1}\mathbf{\Sigma_{XY}}$ (see Eq. 16 – $\arcsin$ and square root are monotonically increasing functions). Understanding this relationship between the cross-covariance coefficients in $\mathbf{\Sigma_{YX}}$ and inter-feature cross-covariance coefficients in $\mathbf{\Sigma_{XX}}$ can yield a systematic framework for feature selection.

This study clearly shows the limitations in the performance of discrete machine learning problems formulated

after discretizing continuous attributes. In contrast to binary classification, other problem formulations such as regression [9], [10], [16], [17] or three-way classification [13], [29] (i.e., low range, mid range and high range) can be explored for specific applications. Notice that in most of these problems the majority of the population tend to have average behaviors. Hence, detecting instances in either tails of the distribution might be more appealing and useful for practical applications (i.e., discarding samples in the middle of the distribution).

ACKNOWLEDGMENTS

We thank the MPLab at UCSD for providing CERT.

REFERENCES

[1] T. Rahman and C. Busso, "A personalized emotion recognition system using an unsupervised feature adaptation scheme," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 5117–5120.

[2] H. Schlosberg, "Three dimensions of emotion," *Psychological review*, vol. 61, no. 2, p. 81, March 1954.

[3] J. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, September 1977.

[4] J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, December 2007.

[5] D. Grandjean, D. Sander, and K. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and cognition*, vol. 17, no. 2, pp. 484–495, June 2008.

[6] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, January-March 2012.

[7] B. Vlasenko and A. Wendemuth, "Determining the smallest emotional unit for level of arousal classification," in *Affective Computing and Intelligent Interaction (ACII 2013)*, Geneva, Switzerland, September 2013, pp. 734–739.

[8] N. Li and C. Busso, "Predicting perceived visual and cognitive distractions of drivers with multimodal features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 51–65, February 2015.

[9] S. Kim, M. Filippone, F. Valente, and A. Vinciarelli, "Predicting the conflict level in television political debates: An approach based on crowdsourcing, nonverbal communication and Gaussian processes," in *ACM International Conference on Multimedia (MM 2012)*, Nara, Japan, October-November 2012, pp. 793–796.

[10] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[11] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, ""Would you buy a car from me?" - on the likability of telephone voices," in *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, Florence, Italy, August 2011, pp. 1557–1560.

[12] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, Florence, Italy, August 2011, pp. 3201–3204.

[13] S. Mariooryad, A. Kannan, D. Hakkani-Tür, and E. Shriberg, "Automatic characterization of speaking styles in educational videos," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, Florence, Italy, May 2014, pp. 4881–4885.

[14] A. Rosenberg and J. Hirschberg, "Acoustic/prosodic and lexical correlates of charismatic speech," in *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 513–516.

[15] E. Strangert and J. Gustafson, "What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations," in *Interspeech 2008*, Brisbane, Australia, September 2008, pp. 1688–1691.

[16] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012: the continuous audio/visual emotion challenge," in *International conference on Multimodal interaction (ICMI 2012)*, Santa Monica, CA, USA, October 2012, pp. 449–456.

[17] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge (avec 2013)," in *ACM International Workshop on Audio/Visual Emotion Challenge*, Barcelona, Spain, October 2013, pp. 3–10.

[18] M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, and R. Cowie, "Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 1595–1598.

[19] R. C. MacCallum, S. Zhang, K. J. Preacher, and D. Rucker, "On the practice of dichotomization of quantitative variables," *Psychological Methods*, vol. 7, no. 1, pp. 19–40, March 2002.

[20] J. DeCoster, A. Iselin, and M. Gallucci, "A conceptual and empirical examination of justifications for dichotomization," *Psychological Methods*, vol. 14, no. 4, pp. 349–366, December 2009.

[21] J. Cohen, "The cost of dichotomization," *Applied Psychological Measurement*, vol. 7, no. 3, pp. 249–253, June 1983.

[22] D. Altman and P. Royston, "The cost of dichotomising continuous variables," *The BMJ: Practice*, vol. 332, p. 1080, May 2006.

[23] D. P. Farrington and R. Loeber, "Some benefits of dichotomization in psychiatric and criminological research," *Criminal Behaviour and Mental Health*, vol. 10, no. 2, pp. 100–122, June 2000.

[24] J. DeCoster, M. Gallucci, and A. Iselin, "Best practices for using median splits, artificial categorization, and their continuous alternatives," *Journal of Experimental Psychopathology*, vol. 2, no. 2, pp. 197–209, 2011.

[25] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011- the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction (ACII 2011)*, ser. Lecture Notes in Computer Science, S. DMello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Memphis, TN, USA: Springer Berlin / Heidelberg, October 2011, vol. 6975/2011, pp. 415–424.

[26] A. Wong and D. Chiu, "Synthesizing statistical knowledge from incomplete mixed-mode data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 6, pp. 796–805, November 1987.

[27] D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," *Journal of Artificial Intelligence Research*, vol. 6, no. 1, pp. 1–34, January 1997.

[28] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 speaker trait challenge," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 254–257.

[29] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, April-June 2012.

[30] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*, Brisbane, Australia, April 2015.

[31] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York, NY, USA: Wiley-Interscience, 2000.

[32] W. F. Sheppard, "On the application of the theory of error to cases of normal distribution and normal correlation," *Philosophical Transactions of the Royal Society of London*, vol. 192, pp. 101–167,531, January 1899.

[33] S. M. Kendall and A. Stuart, *The Advanced Theory of Statistics*. New York, NY, USA: Macmillan Publishing Co, Inc., October 1977, vol. 1.

[34] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 183–196, April-June 2013.

[35] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Newcastle, Northern Ireland, UK: ISCA, September 2000, pp. 19–24.
[36] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *Affective Computing and Intelligent Interaction (ACII 2013)*, Geneva, Switzerland, September 2013, pp. 85–90.
[37] ——, "Feature and model level compensation of lexical content for facial emotion recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2013)*, Shanghai, China, April 2013.
[38] L. Chen and T. Huang, "Emotional expressions in audiovisual human computer interaction," in *IEEE International Conference on Multimedia and Expo (ICME 2000)*, vol. 1, New York City, NY, USA, July-August 2000, pp. 423–426.
[39] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.
[40] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, September 2006.
[41] S. Korkmaz, D. Goksuluk, and G. Zararsiz, "MVN: An R package for assessing multivariate normality," *The R Journal*, vol. 6, no. 2, pp. 151–161, December 2014.
[42] K. V. Mardia, "Measures of multivariate skewness and kurtosis with applications," *Biometrika*, vol. 57, no. 3, pp. 519–530, December 1970.
[43] N. Henze and B. Zirkler, "A class of invariant consistent tests for multivariate normality," *Communications in Statistics - Theory and Methods*, vol. 19, no. 10, pp. 3595–3617, September 1990.
[44] P. Royston, "Approximating the Shapiro-Wilk W-Test for non-normality," *Statistics and Computing*, vol. 2, no. 3, pp. 117–119, September 1992.
[45] J. Hoffbeck and D. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 763–767, July 1996.

Soroosh Mariooryad (S'12) received his B.S degree (2007) with high honors in computer engineering from Ferdowsi University of Mashhad, and his M.S degree (2010) in computer engineering (artificial intelligence) from Sharif University of Technology (SUT), Tehran, Iran. He is currently pursuing his Ph.D. degree in Electrical Engineering at the University of Texas at Dallas (UTD), Richardson, Texas, USA. From 2008 to 2010, he was a member of the Speech Processing Lab (SPL) at SUT. In 2010, he joined as a research assistant the Multimodal Signal Processing (MSP) laboratory at UTD. In summer 2013, he interned at Microsoft Research working on analyzing speaking style characterstics. His research interests are in speech and video signal processing, probabilistic graphical models and multimodal interfaces. His current research includes modeling and analyzing human non-verbal behaviors, with applications to speech-driven facial animations and emotion recognition. He has also worked on statistical speech enhancement and fingerprint recognition.

Carlos Busso (S'02-M'09-SM'13) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is an associate professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He was selected by the School of Engineering of Chile as the best electrical engineer graduated in 2003 across Chilean universities. At USC, he received a provost doctoral fellowship from 2003 to 2005 and a fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interests include digital signal processing, speech and video processing, and multimodal interfaces. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, modeling and synthesis of verbal and nonverbal behaviors, sensing human interaction, in-vehicle active safety system, and machine learning methods for multimodal processing. He is a member of ISCA, AAAC, and ACM, and a senior member of the IEEE.