Zollanvari and Dougherty *EURASIP Journal on Bioinformatics and Systems Biology* (2016) 2016:2
DOI 10.1186/s13637-016-0036-y

EURASIP Journal on Bioinformatics and Systems Biology
a SpringerOpen Journal

RESEARCH Open Access

Incorporating prior knowledge induced from stochastic differential equations in the classification of stochastic observations

Amin Zollanvari[1*] and Edward R. Dougherty[2]

Abstract

In classification, prior knowledge is incorporated in a Bayesian framework by assuming that the feature-label distribution belongs to an uncertainty class of feature-label distributions governed by a prior distribution. A posterior distribution is then derived from the prior and the sample data. An optimal Bayesian classifier (OBC) minimizes the expected misclassification error relative to the posterior distribution. From an application perspective, prior construction is critical. The prior distribution is formed by mapping a set of mathematical relations among the features and labels, the prior knowledge, into a distribution governing the probability mass across the uncertainty class. In this paper, we consider prior knowledge in the form of stochastic differential equations (SDEs). We consider a vector SDE in integral form involving a drift vector and dispersion matrix. Having constructed the prior, we develop the optimal Bayesian classifier between two models and examine, via synthetic experiments, the effects of uncertainty in the drift vector and dispersion matrix. We apply the theory to a set of SDEs for the purpose of differentiating the evolutionary history between two species.

Keywords: Classification, Gaussian processes, Stochastic differential equations, Optimal Bayesian classifier

1 Introduction

A purely data-driven classifier design with small samples encounters a fundamental conundrum: since the error rate of a classifier quantifies its predictive accuracy, the salient epistemic attribute of any classifier and resampling strategies such as cross-validation and bootstrap is generally very inaccurate on small samples due to excessive variance and lack of regression with the true error [1]. The inability to satisfactorily estimate the error with model-free methods with small samples implies that classifier error estimation is virtually impossible without the use of prior information. Prior knowledge can be incorporated in a Bayesian framework by assuming that the feature-label distribution belongs to an uncertainty class of feature-label distributions governed by a prior distribution [2, 3]. Given the latter, in conjunction with sample data, one can optimally estimate the error of any classifier, relative to the mean square error (MSE) between the true and estimated errors, where expectations are taken with respect to a posterior distribution derived from the prior distribution and the data [4, 5]. Hence, optimality is with respect to our prior knowledge and the data. Furthermore, one can derive an optimal classifier relative to the expected error of the classifier over the posterior distribution, this being called the *optimal Bayesian classifier* (OBC) [6, 7]. Closed-form solutions have been developed for multinomial and Gaussian models. In other situations, Markov Chain Monte Carlo (MCMC) methods can be used [8].

Having developed the statistical theory, one is confronted with an engineering problem: transform scientific knowledge given in some mathematical form into a prior distribution. Intuitively, given a set of mathematical relations among the features and labels, these relations constrain the uncertainty class of feature-label distributions that could potentially govern the classification and the relative strengths of the relations can be transformed so as to determine the probability mass of the prior distribution. For instance, in phenotype classification based on gene expression, genetic regulatory pathways constitute

*Correspondence: amin.zollanvari@nu.edu.kz
[1] Department of Electrical and Electronic Engineering, Nazarbayev University, Astana, 010000, Kazakhstan
Full list of author information is available at the end of the article

graphical prior knowledge and this prior knowledge can be employed to formulate a prior distribution governing the uncertainty class of feature-label distributions [9, 10]. Another genomic application involves using prior knowledge concerning RNA-seq data to form sequence-based classifiers [8].

From a general perspective, when using Bayesian methods, prior construction constitutes the highest hurdle. A half century ago, E. T. Jaynes remarked,

> Bayesian methods, for all their advantages, will not be entirely satisfactory until we face the problem of finding the prior probability squarely [11].

The aim of this paper is to utilize prior knowledge in the form of stochastic differential equations (SDEs) to classify time-series data. Although we will confine ourselves to a Gaussian problem so that we can take advantage of existing closed-form OBC representations, one can envision further applications using MCMC methods. Hence, the approach taken in the present paper may lead to utilizing SDEs across a number of time-series classification problems, keeping in mind that SDEs play a major role in many disciplines including physics, biology, finance, and chemistry. Vector SDEs, our concern here, have various applications. Not only do they arise naturally in many systems with vector value states, but they also arise in many systems where the process is restricted to lie on certain manifolds [12].

In the stochastic setting, training data are collected over time processes. Given certain Gaussian assumptions, classification in the SDE setting takes the same form as ordinary classification in the Gaussian model and we can apply the optimal Bayesian classification theory once we have a prior distribution constructed in accordance with known stochastic equations. In this paper, we provide the mathematical framework to synthesize an OBC in the presence of prior knowledge induced in the form of SDEs governing the dynamics of the system. We consider a vector SDE in integral form involving a drift vector and dispersion matrix, develop the OBC between two models, and examine via synthetic experiments the effects of uncertainty in the drift vector and dispersion matrix.

We compare the performance of the OBC with quadratic discriminant analysis (QDA), a classical approach to building classifiers in the Gaussian model (see Additional file 1: Section I for definition of QDA). Such comparisons are useful because, even though the OBC is optimal given the uncertainty, its optimality is *on average* across the uncertainty class, so that its performance advantage varies for different feature-label distributions in the uncertainty class (and can be disadvantageous for some distributions, although these will have small probability mass in the posterior distribution). Comparison to QDA is instructive because, as we will explain in the next section, QDA is a sample-based approximation to the optimal classifier for the true feature-label distribution. In addition to synthetic experiments, we apply optimal Bayesian classification using a form of the Ornstein-Uhlenbeck process that has been employed for modeling the evolutionary change of species; specifically, we use a set of SDEs to construct a classifier to differentiate the evolutionary history between two species.

2 Background

2.1 Classification

In a two-class classification, the population is characterized by a feature-label distribution F for a random pair $(\mathbf{X}, Y)$, where $\mathbf{X}$ is a vector of p features and Y is the binary label, 0 or 1, of the class containing $\mathbf{X}$. The *prior class probabilities* are defined by $c_j = P(Y = j)$ and the *class-conditional densities* by $p_j(\mathbf{x}) = p(\mathbf{x} \mid Y = j)$, for $j = 0, 1$. To avoid trivialities, we assume $\min\{c_0, c_1\} \neq 0$. A *classifier* is a function $\psi(\mathbf{X})$ assigning a binary label to each feature vector $\mathbf{X}$. The error, $\varepsilon[\psi]$, of ψ is the probability $P(\psi(\mathbf{X}) \neq Y)$, which can be decomposed into $\varepsilon = c_0\varepsilon^0 + c_1\varepsilon^1$, where $\varepsilon^j = P(\psi(\mathbf{X}) = 1 - j | Y = j)$, for $j = 0, 1$. A classifier with minimum error among all classifiers is known as a *Bayes classifier* for F. The minimum error is called the *Bayes error*. Epistemologically, the error is the key issue since it quantifies the predictive capacity.

In practice, F is unknown and a *classification rule* Ψ is used to design a classifier ψ_n from a random sample $S_n = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)\}$ of pairs drawn from F. If feature selection is involved, then it is part of the classification rule. Since the true classifier error $\varepsilon[\psi_n]$ depends on F, which is unknown, $\varepsilon[\psi_n]$ is unknown. The true error must be estimated by an *estimation rule*, Ξ. Thus, the random sample S_n yields a classifier $\psi_n = \Psi(S_n)$ and an error estimate $\hat{\varepsilon}[\psi_n] = \Xi(S_n)$ (see Additional file 1: Section II for more information).

When a large amount of data is available, the sample can be split into independent training and test sets, the classifier being designed on the training data and its error being estimated by the proportion of errors on the test data; however, when data are limited, the sample cannot be split without leaving too little data to design a good classifier. Hence, training and error estimation must take place on the same data set. As noted in Section 1, accurate error estimation is virtually impossible with small samples in the absence of distributional assumptions.

2.2 Optimal Bayesian classification

Distributional assumptions can be imposed by defining a prior distribution over an uncertainty class of feature-label distributions. This results in a Bayesian approach with the uncertainty class being given a prior distribution and the data being used to construct a posterior distribution.

Let Π_0 and Π_1 denote the class-0 and class-1 conditional distributions, respectively; let c be the probability of a point coming from Π_0 (the "mixing" probability); and let Π_0 and Π_1 be parameterized by θ_0 and θ_1, respectively. The overall model is parameterized by $\theta = (c, \theta_0, \theta_1)$ with prior distribution $\pi(\theta)$. Given a random sample, S_n, a classifier ψ_n is designed and we wish to minimize the MSE between its true error, ε, and an error estimate, $\widehat{\varepsilon}$. The minimum mean square error (MMSE) error estimator is the expected true error, $\widehat{\varepsilon}(\psi_n, S_n) = \mathrm{E}_\theta[\,\varepsilon(\psi_n, \theta)|S_n]$. The expectation given the sample is over the posterior density of θ, denoted by $\pi^*(\theta)$. Thus, we write the Bayesian MMSE error estimator as $\widehat{\varepsilon} = \mathrm{E}_{\pi^*}[\,\varepsilon]$.

The Bayesian error estimate is not guaranteed to be the optimal error estimate for any particular feature-label distribution but optimal for a given sample, and assuming the parameterized model and prior probabilities, it is both optimal on average with respect to MSE and unbiased when averaged over all parameters and samples. To facilitate analytic representations, we assume c, θ_0, and θ_1 are all mutually independent prior to observing the data. Denote the marginal priors of c, θ_0, and θ_1 by $\pi(c)$, $\pi(\theta_0)$, and $\pi(\theta_1)$, respectively, and suppose data are used to find each posterior, $\pi^*(c)$, $\pi^*(\theta_0)$, and $\pi^*(\theta_1)$, respectively. Independence is preserved, i.e., $\pi^*(c, \theta_0, \theta_1) = \pi^*(c)\pi^*(\theta_0)\pi^*(\theta_1)$ [4].

If ψ_n is a trained classifier given by $\psi_n(\mathbf{x}) = 0$ if $\mathbf{x} \in R_0$ and $\psi_n(\mathbf{x}) = 1$ if $\mathbf{x} \in R_1$, where R_0 and R_1 are measurable sets partitioning the sample space, then the Bayesian MMSE error estimator can be found from *effective class-conditional densities*, which are derived by taking the expectations of the individual class-conditional densities with respect to the posterior distribution,

$$f(\mathbf{x}|y) = \int_{\Theta_y} f_{\theta_y}(\mathbf{x}|y)\,\pi^*\left(\theta_y\right) d\theta_y. \tag{1}$$

Using these [6] (see Additional file 1: Section III for more information),

$$\widehat{\varepsilon}(\psi_n, S_n) = \mathrm{E}_{\pi^*}[c]\int_{R_1} f(\mathbf{x}|0)\,d\mathbf{x} + (1-\mathrm{E}_{\pi^*}[c])\int_{R_0} f(\mathbf{x}|1)\,d\mathbf{x}. \tag{2}$$

In the context of a prior distribution, an optimal Bayesian classifier, ψ_{OBC}, is any classifier satisfying

$$\mathrm{E}_{\pi^*}[\varepsilon(\psi_{\mathrm{OBC}}, \theta)] \le \mathrm{E}_{\pi^*}[\varepsilon(\psi, \theta)] \tag{3}$$

for all $\psi \in \mathcal{C}$, where $\mathcal{C}$ is an arbitrary family of classifiers. Under the Bayesian framework, this is equivalent to minimizing the probability of error,

$$\begin{aligned}\mathrm{P}(\psi_n(\mathbf{X}) \neq Y|S_n) &= \mathrm{E}_{\pi^*}[P(\psi_n(\mathbf{X}) \neq Y|\theta, S_n)] \\ &= \widehat{\varepsilon}(\psi_n, S_n).\end{aligned} \tag{4}$$

If $\mathcal{C}$ is the set of all classifiers with measurable decision regions (which we will assume), then an optimal Bayesian classifier, ψ_{OBC}, satisfying (3) for all $\psi \in \mathcal{C}$ exists and is given pointwise by [6]

$$\psi_{\mathrm{OBC}}(\mathbf{x}) = \begin{cases} 0 \text{ if } \mathrm{E}_{\pi^*}[c] f(\mathbf{x}|0) \ge (1 - \mathrm{E}_{\pi^*}[c])f(\mathbf{x}|1), \\ 1 \text{ otherwise.} \end{cases} \tag{5}$$

In many applications, especially in biomedicine, the sample S_n is obtained by first deciding how many sample points will be taken from each class and then randomly sampling from each class separately, the resulting sample said to be "separately sampled." With separate sampling, the data cannot be used to generate a posterior distribution for c, so that c must be known. Stratified sampling is a special case of separate sampling in which the sample is drawn so that the proportion of sample points from class 0 is equal to c. In such a case, there is no posterior $\mathrm{E}_{\pi^*}[c]$ and $\mathrm{E}_{\pi^*}[c]$ is replaced by c in (5). We will utilize stratified sampling in our examples.

3 Binary classification of Gaussian processes

In this section, we frame the setting in which we are working and then define the problem of binary classification in the context of Gaussian processes. To begin with, a collection $\{\mathbf{X}_t : t \in \mathbf{T}\}$ of $\mathbb{R}^p$-valued random variables defined on a common probability space $(\Omega, \mathcal{F}, P)$ indexed by a parameter $t \in \mathbf{T} \subset \mathbb{R}$ (here assumed to be time) and $\mathcal{F}$ being a σ-algebra of subsets of the sample space Ω (events) constitutes a stochastic process $\mathbf{X}$ with state space $\mathbb{R}^p$. Throughout this work, we consider $\mathcal{F}$ as the σ-algebra of Borel subsets of $\mathbb{R}^p$. A stochastic process $\mathbf{X}$ is *adapted* to an increasing family of σ-algebra $\{\mathcal{F}_t : t \ge 0\}$ (a filtration) if for each $t \ge 0$, $\mathbf{X}_t$ is $\mathcal{F}_t$-measurable.

We study classification in the context of multivariate Gaussian processes (see Additional file 1: Section IV for a review of literature pertaining to classification of stochastic processes). Consider the p-dimensional column random vectors $\mathbf{X}_{t_1}$, $\mathbf{X}_{t_2}$, ..., $\mathbf{X}_{t_N}$. A random process $\mathbf{X}$ is a multivariate Gaussian process if any finite-dimensional vector $\left[\mathbf{X}_{t_1}^T, \mathbf{X}_{t_2}^T, ..., \mathbf{X}_{t_N}^T\right]^T$ possesses a multivariate normal distribution $\mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{t}_N}, \boldsymbol{\Sigma}_{\mathbf{t}_N}\right)$, where

$$\boldsymbol{\mu}_{\mathbf{t}_N} = \left[\boldsymbol{\mu}_{t_1}^T, \boldsymbol{\mu}_{t_2}^T, ..., \boldsymbol{\mu}_{t_N}^T\right]_{Np\times 1}^T, \tag{6}$$

with $\boldsymbol{\mu}_{t_i} = E[\mathbf{X}_{t_i}]$, and $\boldsymbol{\Sigma}_{\mathbf{t}_N}$ is the $Np \times Np$ covariance matrix dependent on $\mathbf{t}_N = [t_1, t_2, ..., t_N]^T$ and structured as

$$\boldsymbol{\Sigma}_{\mathbf{t}_N} = \begin{bmatrix} \boldsymbol{\Sigma}_{t_1,t_1} & \boldsymbol{\Sigma}_{t_1,t_2} & ... & \boldsymbol{\Sigma}_{t_1,t_N} \\ \boldsymbol{\Sigma}_{t_2,t_1} & \boldsymbol{\Sigma}_{t_2,t_2} & ... & \boldsymbol{\Sigma}_{t_2,t_N} \\ ... & ... & ... & ... \\ \boldsymbol{\Sigma}_{t_N,t_1} & \boldsymbol{\Sigma}_{t_N,t_2} & ... & \boldsymbol{\Sigma}_{t_N,t_N} \end{bmatrix}_{Np\times Np}, \tag{7}$$

where

$$\mathbf{\Sigma}_{t_i,t_j} = E\left[\left(\mathbf{X}_{t_i} - E(\mathbf{X}_{t_i})\right)(\mathbf{X}_{t_j} - E(\mathbf{X}_{t_j})^T\right)\right]. \tag{8}$$

We refer to $\mathbf{t}_N$ as the *observation time vector*. For any fixed $\omega \in \Omega$, a *sample path* is a collection $\{\mathbf{X}_t(\omega) : t \in \mathbf{t}\}$. We denote a realization of $\mathbf{X}$ at sample path ω and time vector $\mathbf{t}_N$ by $\mathbf{x}_{\mathbf{t}_N}(\omega)$.

We consider a general framework, referred to as *binary classification of Gaussian processes (BCGP)*. Consider two independent multivariate Gaussian processes $\mathbf{X}^0$ and $\mathbf{X}^1$, where for any $\mathbf{t}_N$, $\mathbf{X}^0$ and $\mathbf{X}^1$ possess mean and covariance $\boldsymbol{\mu}^0_{\mathbf{t}_N}$ and $\mathbf{\Sigma}^0_{\mathbf{t}_N}$, and $\boldsymbol{\mu}^1_{\mathbf{t}_N}$ and $\mathbf{\Sigma}^1_{\mathbf{t}_N}$, respectively. For $y = 0, 1$, $\boldsymbol{\mu}^y_{\mathbf{t}_N}$ is defined similarly to (6) with $\boldsymbol{\mu}^y_{t_i} = E\left[\mathbf{X}^y_{t_i}\right]$ and $\mathbf{\Sigma}^y_{\mathbf{t}_N}$ is defined similarly to (7) with

$$\mathbf{\Sigma}^y_{t_i,t_j} = E\left[\left(\mathbf{X}^y_{t_i} - E(\mathbf{X}^y_{t_i})\right)\left(\mathbf{X}^y_{t_j} - E(\mathbf{X}^y_{t_j})\right)^T\right]. \tag{9}$$

Let $\mathbf{S}^y_{\mathbf{t}_N}$ denote a set of n^y *sample paths* from process $\mathbf{X}^y$ at $\mathbf{t}_N$,

$$\mathbf{S}^y_{\mathbf{t}_N} = \left\{\mathbf{x}^y_{\mathbf{t}_N}(\omega_1), \mathbf{x}^y_{\mathbf{t}_N}(\omega_2), \ldots, \mathbf{x}^y_{\mathbf{t}_N}(\omega_{n^y})\right\}. \tag{10}$$

We assume that $\mathbf{t}_N$ is the same for both classes. Let $\mathbf{x}^y_{\mathbf{t}_N}(\omega_s)$ denote a future test sample path observed on the same observation time vector as the training sample paths, where $y \in \{0, 1\}$ indicates the label of the class-conditional process the sample path is coming from, either $\mathbf{X}^0$ or $\mathbf{X}^1$. Note that, as compared with the classical probabilistic definition of classification where the sample points are observations of p-dimension, here we define stochastic-process classification in connection with a set of sample paths, which can be considered as observations of Np dimension. A classification problem arises from the fact that the experimenter is blind to the class label of $\mathbf{x}^y_{\mathbf{t}_N}(\omega_s)$, i.e., to y, and desires a discriminant $\psi_{\mathbf{t}_N}(.)$ such that

$$y = \begin{cases} 0, & \text{if } \psi_{\mathbf{t}_N}\left(\mathbf{x}^y_{\mathbf{t}_N}(\omega_s)\right) > 0 \\ 1, & \text{otherwise.} \end{cases}. \tag{11}$$

Other types of classification could be defined. For example, one might be interested in classifying a test sample path $\mathbf{x}^y_{\mathbf{t}_{N+M}}(\omega_s)$ where the observation time vector of the test sample path is obtained by augmenting $\mathbf{t}_N$ by another vector $[t_{N+1}, t_{N+2}, ..., t_{N+M}]$, where M is a positive integer. In this case, the time of observation for the future sample path is extended. Similarly, one may define problems where the future time of observation is shrunken to a subset of time points in $\mathbf{t}_N$ or problems where the future observation time vector is a set of time points totally or partially different from time points in $\mathbf{t}_N$. Throughout this work, we are mainly concerned with solving the classification problem as defined in (11), which we refer to as the *standard* type, and we discuss the feasibility of solving other cases when possible.

3.1 General presentation of stochastic differential equations (SDEs)

To define SDEs, we consider a diffusion process, the most fundamental being the Wiener process. For a general definition of a q-dimensional Wiener process, see the Appendix. Let $\mathbf{W} = \{\mathbf{W}_t : t \geq 0\}$ be a q-dimensional Wiener process. For each sample path and for $0 \leq t_0 \leq t \leq T$, we consider a vector SDE in the integral form as follows:

$$\begin{aligned}\mathbf{X}_t(\omega) = \mathbf{X}_{t_0}(\omega) &+ \int_{t_0}^{t} \mathbf{f}\left(s, \mathbf{X}_s(\omega)\right) ds \\ &+ \int_{t_0}^{t} \mathbf{G}\left(s, \mathbf{X}_t(\omega)\right) d\mathbf{W}_s(\omega),\end{aligned} \tag{12}$$

where $\mathbf{f} : [0, T] \times \Omega \to \mathbb{R}^p$ (the p-dimensional drift vector) and $\mathbf{G} : [0, T] \times \Omega \to \mathbb{R}^{p \times q}$ (the $p \times q$ dispersion matrix). The first integral in (12) is an ordinary Lebesgue integral, and throughout this work, we assume an Itô integration for the second integral. With slightly more work, the results can be extended to Stratonovich integration. Let $\mathcal{L}$ be the σ-algebra of Lebesgue subsets of $\mathbb{R}$. A function $h(t, \omega)$ defined on a probability space $(\Omega, \mathcal{F}, P)$ belongs to $\mathcal{L}^{\omega}_T$ if it is jointly $\mathcal{L} \times \mathcal{F}$ measurable, $h(t, .)$ is $\mathcal{F}_t$-measurable for each $t \in [0, T]$, and with probability 1, $\int_0^T h(s, \omega)^2 ds < \infty$. Let f^i and $g^{i,j}$ denote the components of $\mathbf{f}$ and $\mathbf{G}$, respectively. If we assume $\mathbf{X}_0(\omega)$ is $\mathcal{F}_0$-measurable and if $\sqrt{|f^i|} \in \mathcal{L}^{\omega}_T$ and $g^{i,j} \in \mathcal{L}^{\omega}_T$, then each component of the p-dimensional process $\mathbf{X}_t(\omega)$ is $\mathcal{F}_t$-measurable [12]. The $\mathcal{F}_t$-measurability of $\mathbf{X}_t(\omega)$ along with the martingale property of $\mathbf{W}$ indicates "nonanticipativeness" of $\mathbf{X}_t(\omega)$ in general.

The integral Eq. (12) is commonly written in a symbolic form as

$$\mathrm{d}\mathbf{X}_t = \mathbf{f}(t, \mathbf{X}_t)\mathrm{d}t + \mathbf{G}(t, \mathbf{X}_t)\mathrm{d}\mathbf{W}_t, \tag{13}$$

which is the representation of a vector SDE.

4 SDE prior knowledge in the BCGP model

Prior knowledge in the form of a set of stochastic differential equations constrains the possible behavior of the dynamical system to an uncertainty class. If such prior knowledge is available, then it can be used in the BCGP model to improve classification performance. The core underlying assumption of the BCGP model is that the data are generated from two Gaussian processes for which binary classification is desired. In this regard, we define *valid prior knowledge* (in the form of SDEs) as a set of SDEs with a unique solution that does not contradict the Gaussianity assumption of the dynamics of the model. For nonlinear $\mathbf{f}(t, \mathbf{X}_t)$ and $\mathbf{G}(t, \mathbf{X}_t)$ (w.r.t. to state $\mathbf{X}_t$), the solution of SDE (13) is generally a non-Gaussian process. Fortunately, under a wide class of linear functions, the SDE solutions are Gaussian. To wit, the SDEs become

valid prior knowledge for each class-conditional process defined in the BCGP model. Henceforth, we focus on this type of SDE.

For class label $y = 0, 1$, the linear classes of SDEs that we consider are defined by replacing

$$\begin{aligned}\mathbf{f}^y(t, \mathbf{X}_t) &= \mathbf{A}^y(t)\mathbf{X}_t^y + \mathbf{a}^y(t),\\ \mathbf{G}^y(t, \mathbf{X}_t) &= \mathbf{B}^y(t),\end{aligned} \tag{14}$$

in (13) with $\mathbf{A}^y(t)$ (a $p \times p$ matrix), $\mathbf{a}^y(t)$ (a $p \times 1$ vector), and $\mathbf{B}^y(t)$ (a $p \times q$ matrix), these being measurable and bounded on $[t_0, T]$. This results in

$$\mathrm{d}\mathbf{X}_t^y = (\mathbf{A}^y(t)\mathbf{X}_t^y + \mathbf{a}^y(t))\mathrm{d}t + \mathbf{B}^y(t)\mathrm{d}\mathbf{W}_t^y, \;\; \mathbf{X}_{t_0}^y(\omega) = c^y. \tag{15}$$

This initial value problem has a unique solution that is a Gaussian stochastic process if and only if the initial conditions c^y are constant or normally distributed (Theorem 8.2.10 [13]). Note that in this model, $\mathbf{G}^y(t, \mathbf{X}_t)$ (i.e. $\mathbf{B}^y(t)$) is independent of ω. Under this model, it can be shown that the mean (at a time index t_i) and the covariance matrix (at t_i and t_j) of the Gaussian process $\mathbf{X}_t^y$ are given by [13]

$$\mathbf{m}_{t_i}^y = E\left[\mathbf{X}_{t_i}^y\right] = \mathbf{\Phi}^y(t_i)\left(E[c^y] + \int_{t_0}^{t_i} \mathbf{\Phi}^y(s)^{-1}\mathbf{a}^y(s)\mathrm{d}s\right) \tag{16}$$

and

$$\begin{aligned}\mathbf{\Psi}_{t_i,t_j}^y &= E\left[\left(\mathbf{X}_{t_i}^y - E\left[\mathbf{X}_{t_i}^y\right]\right)\left(\mathbf{X}_{t_j}^y - E\left[\mathbf{X}_{t_j}^y\right]\right)^T\right]\\ &= \mathbf{\Phi}^y(t_i)\left(E\left[\left(c^y - E[c^y]\right)\left(c^y - E[c^y]\right)^T\right]\right.\\ &\quad \left. + \int_{t_0}^{t_i} \mathbf{\Phi}^y(u)^{-1}\mathbf{B}^y(u)\mathbf{B}^y(u)^T\left(\mathbf{\Phi}^y(u)^{-1}\right)^T \mathrm{d}u\right)\mathbf{\Phi}^y(t_j)^T,\end{aligned} \tag{17}$$

where $t_0 \le t_i \le t_j \le T$ and $\mathbf{\Phi}^y(t_i)$ is the fundamental matrix of the deterministic equation

$$\dot{\mathbf{X}}_t^y = \mathbf{A}^y(t)\mathbf{X}_t^y. \tag{18}$$

4.1 SDEs as perfect representatives for the dynamics of class-conditional processes

If the SDE model presented in (15) could perfectly represent the dynamics of the underlying stochastic processes of the BCGP model, then there would be no need for training sample paths. To see this, note that in this case $\boldsymbol{\mu}_t^y$ and $\mathbf{\Sigma}_{t_i,t_j}^y$ defined in (6) and (7) are obtained by

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{t}_N}^y &= \mathbf{m}_{\mathbf{t}_N}^y\\ \mathbf{\Sigma}_{\mathbf{t}_N}^y &= \mathbf{\Psi}_{\mathbf{t}_N}^y\end{aligned}, \tag{19}$$

where

$$\mathbf{m}_{\mathbf{t}_N}^y = \left[\mathbf{m}_{t_1}^{y\,T}, \mathbf{m}_{t_2}^{y\,T}, ..., \mathbf{m}_{t_N}^{y\,T}\right]_{Np\times 1}^T \tag{20}$$

and

$$\mathbf{\Psi}_{\mathbf{t}_N}^y = \begin{bmatrix} \mathbf{\Psi}_{t_1,t_1}^y & \mathbf{\Psi}_{t_1,t_2}^y & \cdots & \mathbf{\Psi}_{t_1,t_N}^y \\ \mathbf{\Psi}_{t_2,t_1}^y & \mathbf{\Psi}_{t_2,t_2}^y & \cdots & \mathbf{\Psi}_{t_2,t_N}^y \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{\Psi}_{t_N,t_1}^y & \mathbf{\Psi}_{t_N,t_2}^y & \cdots & \mathbf{\Psi}_{t_N,t_N}^y \end{bmatrix}_{Np\times Np}, \tag{21}$$

where $\mathbf{m}_{t_i}^{y\,T}$ and $\mathbf{\Psi}_{t_i,t_j}^y$ are obtained from (16) and (17), respectively. Therefore, one can obtain the exact (or at least approximately exact) values of the means and auto-covariances used to characterize the Gaussian processes involved in the BCGP model. To obtain $\mathbf{m}_{t_i}^y$ and $\mathbf{\Psi}_{t_i,t_j}^y$, two approaches can be taken. First, one may analytically solve (18) where possible and then use numerical methods to evaluate the integrations presented in (16) and (17). For example, if $\mathbf{A}^y(t) = \mathbf{A}^y$, i.e., being independent of t, the solution of (18) is given by a matrix exponential as

$$\mathbf{\Phi}^y(t) = e^{\mathbf{A}^y(t-t_0)}, \tag{22}$$

which can be used in (16) and (17). In general, where one may not be able to analytically solve (18), numerical methods such as the Euler-Maruyama scheme [14] can be used to directly solve for $\mathbf{X}_t^y(\omega)$ and obtain

$$\begin{aligned}\hat{\mathbf{m}}_{\mathbf{t}_N}^y &= \frac{1}{l^y}\sum_{i=1}^{l^y} \mathbf{x}_{\mathbf{t}_N}^{y,\mathrm{SDE}}(\omega_i),\\ \hat{\mathbf{\Psi}}_{\mathbf{t}_N}^y &= \frac{1}{l^y - 1}\sum_{i=1}^{l^y}\left(\mathbf{x}_{\mathbf{t}_N}^{y,\mathrm{SDE}}(\omega_i) - \bar{\mathbf{x}}_{\mathbf{t}_N}^{y,\mathrm{SDE}}\right)\left(\mathbf{x}_{\mathbf{t}_N}^{y,\mathrm{SDE}}(\omega_i) - \bar{\mathbf{x}}_{\mathbf{t}_N}^{y,\mathrm{SDE}}\right)^T,\end{aligned} \tag{23}$$

where $\mathbf{x}_{\mathbf{t}_N}^{y,\mathrm{SDE}}(\omega_i), i = 1, 2, ..., l^y$, are the generated sample paths obtained from solving SDEs. Since there is no restriction on generating an arbitrary number of sample paths from $\mathbf{X}_t^y(\omega)$, one can take $l^y >> Np$ to have a positive definite $\hat{\mathbf{\Psi}}_{\mathbf{t}_N}^y$ and, at the same time, obtain an accurate estimate of the actual values of $\mathbf{m}_{\mathbf{t}_N}^y$ and $\mathbf{\Psi}_{\mathbf{t}_N}^y$. In this approach, the knowledge of (16) and (17) is used in the existence of the limits $\lim_{l^y\to\infty} \hat{\mathbf{m}}_{\mathbf{t}_N}^y$ and $\lim_{l^y\to\infty} \hat{\mathbf{\Psi}}_{\mathbf{t}_N}^y$, i.e., justifies generating more sample paths as $\lim_{l^y\to\infty} \hat{\mathbf{m}}_{\mathbf{t}_N}^y = \mathbf{m}_{\mathbf{t}_N}^y$ and $\lim_{l^y\to\infty} \hat{\mathbf{\Psi}}_{\mathbf{t}_N}^y = \mathbf{\Psi}_{\mathbf{t}_N}^y$.

In any case, we can assume exact (approximately exact) values of $\mathbf{m}_{t_i}^0$, $\mathbf{m}_{t_i}^1$, $\mathbf{\Psi}_{t_i,t_j}^0$, and $\mathbf{\Psi}_{t_i,t_j}^1$ are available. The optimal discriminant in this case is obtained by using the conventional quadratic discriminant analysis (QDA), which is now defined by using the following statistic in (11):

$$\begin{aligned}\psi_{\mathbf{t}_N}^{\mathrm{QDA}}\left(\mathbf{x}_{\mathbf{t}_N}^y(\omega_s)\right) = &-\frac{1}{2}\left(\mathbf{x}_{\mathbf{t}_N}^y(\omega_s) - \mathbf{m}_{\mathbf{t}_N}^0\right)^T \mathbf{\Psi}_{\mathbf{t}_N}^{0\,-1}\left(\mathbf{x}_{\mathbf{t}_N}^y(\omega_s) - \mathbf{m}_{\mathbf{t}_N}^0\right)\\ &+ \frac{1}{2}\left(\mathbf{x}_{\mathbf{t}_N}^y(\omega_s) - \mathbf{m}_{\mathbf{t}_N}^1\right)\mathbf{\Psi}_{t_N}^{1-1}\left(\mathbf{x}_{\mathbf{t}_N}^y(\omega_s) - \mathbf{m}_{\mathbf{t}_N}^1\right)\\ &+ \frac{1}{2}\log\frac{|\mathbf{\Psi}_{\mathbf{t}_N}^{1\,-1}|}{|\mathbf{\Psi}_{\mathbf{t}_N}^{0\,-1}|} - \log\frac{\alpha_1}{1-\alpha_1}.\end{aligned} \tag{24}$$

The use of (24) is justified by the fact that the BCGP classification reduces to differentiating independent observations of Np dimension generated from two multivariate Gaussian distributions. Therefore, taking the same set of machinery as in [15] results in (24). We restate that in this case where (19) holds, there is no need for utilizing the sample path measurements (training sample paths) in finding the discriminant (24). This is due to the fact that the statistical properties of a Gaussian process at $\mathbf{t}_N$ are solely determined by $\mathbf{m}^y_{\mathbf{t}_N}$ and $\mathbf{\Psi}^y_{\mathbf{t}_N}$ and, as mentioned before, either closed-form solutions of these are available or they can be approximated element-wise with an arbitrary small error rate by generating a sufficiently large number of sample paths.

The optimal solution proposed in (24) is, in fact, a function of the observation time vector of future sample paths. Therefore, if a future sample point $\mathbf{x}^y_{\mathbf{t}_L}(\omega_s)$ is measured at an arbitrary time vector $\mathbf{t}_L$, which can be partially or totally different from $\mathbf{t}_N$, then the optimal discriminant $\psi_{\mathbf{t}_L}\left(\mathbf{x}^y_{\mathbf{t}_L}(\omega_s)\right)$ is obtained by determining the solution of SDEs at $\mathbf{t}_L$ and replacing $\mathbf{m}^y_{\mathbf{t}_N}$ and $\mathbf{\Psi}^y_{\mathbf{t}_N}$ with $\mathbf{m}^y_{\mathbf{t}_L}$ and $\mathbf{\Psi}^y_{\mathbf{t}_L}$, respectively, in (24).

4.2 SDEs as prior information for the dynamics of class-conditional processes

In practice, the SDEs usually do not provide complete description and are then viewed as prior knowledge concerning the underlying dynamics of the BCGP model. Since we assume that a Gaussian process governs both the dynamics of each class-conditional process (BCGP model in Section 3) and its corresponding set of SDEs (by using model (15)), incompleteness of the SDEs results from the fact that (19) does not necessarily hold. We make the following assumptions on the *nature of the prior information* to which the set of SDEs corresponding to each class give rise: (i) before observing the sample paths at an observation time vector, the SDEs characterize the only information that we have about the system and (ii) the statistical properties of all Gaussian processes that may generate the data are on average (over the parameter space) equivalent to the statistical properties determined from the SDEs. The latter statement will subsequently be formalized.

Assume that the parameters $\boldsymbol{\mu}^y_{\mathbf{t}_N}$ and $\mathbf{\Sigma}^y_{\mathbf{t}_N}$ defining the BCGP model constitute a realization of the random vector $\theta^y_{\mathbf{t}_N} = \left[\boldsymbol{\mu}^y_{\mathbf{t}_N}, \mathbf{\Sigma}^y_{\mathbf{t}_N}\right]$, where $\theta^y_{\mathbf{t}_N}$ has a prior distribution $\pi(\theta^y_{\mathbf{t}_N})$ parameterized by a set $\left\{\breve{\mathbf{m}}^y_{\mathbf{t}_N}, \breve{\mathbf{\Psi}}^y_{\mathbf{t}_N}, \nu^y_{\mathbf{t}_N}, \kappa^y_{\mathbf{t}_N}\right\}$ of hyperparameters. The quantities $\nu^y_{\mathbf{t}_N}$ and $\kappa^y_{\mathbf{t}_N}$ define our certainty about the prior knowledge (here, the set of SDEs presenting the dynamics of the model). If we take the conjugate priors for mean and covariance when the sampling is Gaussian, i.e., a normal-inverse-Wishart distribution (which depends on $\mathbf{t}_N$), then

$$\pi\left(\theta^y_{\mathbf{t}_N}\right) \propto |\mathbf{\Sigma}^y_{\mathbf{t}_N}|^{-(\kappa^y_{\mathbf{t}_N}+Np+1)/2}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\breve{\mathbf{\Psi}}^y_{\mathbf{t}_N}\left(\mathbf{\Sigma}^y_{\mathbf{t}_N}\right)^{-1}\right)\right) \times |\mathbf{\Sigma}^y_{\mathbf{t}_N}|^{-1/2}\exp\left(-\frac{\nu^y_{\mathbf{t}_N}}{2}\left(\boldsymbol{\mu}^y_{\mathbf{t}_N}-\mathbf{m}^y_{\mathbf{t}_N}\right)^T\left(\mathbf{\Sigma}^y_{\mathbf{t}_N}\right)^{-1}\left(\boldsymbol{\mu}^y_{\mathbf{t}_N}-\mathbf{m}^y_{\mathbf{t}_N}\right)\right), \tag{25}$$

with $\boldsymbol{\mu}^y_{\mathbf{t}_N}$ and $\mathbf{\Sigma}^y_{\mathbf{t}_N}$ defined in (6) and (7). Therefore, the above assumption (ii) on the nature of the prior information means that

$$\begin{aligned}\breve{\mathbf{m}}^y_{\mathbf{t}_N} &= \mathbf{m}^y_{\mathbf{t}_N}\\ \breve{\mathbf{\Psi}}^y_{\mathbf{t}_N} &= \left(\kappa^y_{\mathbf{t}_N} - Np - 1\right)\mathbf{\Psi}^y_{\mathbf{t}_N},\end{aligned} \tag{26}$$

with $\mathbf{m}^y_{\mathbf{t}_N}$ defined by (16) and (20) and $\mathbf{\Psi}^y_{\mathbf{t}_N}$ defined by (17) and (21). To see (26), note that from (25) and independence of $\boldsymbol{\mu}^y_{\mathbf{t}_N}$ and $\mathbf{\Sigma}^y_{\mathbf{t}_N}$, we have $E_\pi\left[\boldsymbol{\mu}^y_{\mathbf{t}_N}\right] = \breve{\mathbf{m}}^y_{\mathbf{t}_N}$ and $E_\pi\left[\mathbf{\Sigma}^y_{\mathbf{t}_N}\right] = \frac{\breve{\mathbf{\Psi}}^y_{\mathbf{t}_N}}{\kappa^y_{\mathbf{t}_N}-Np-1}$ (the latter is the mean of an inverse-Wishart distribution). The more confident we are about an a priori set of SDEs that is supposed to represent the underlying stochastic processes at $\mathbf{t}_N{}^y$, the larger we might choose the values of $\nu^y_{\mathbf{t}_N}$ and $\kappa^y_{\mathbf{t}_N}$ and the more concentrated become the priors of the mean and covariance about $\mathbf{m}^y_{\mathbf{t}_N}$ and $\mathbf{\Psi}^y_{\mathbf{t}_N}$, respectively. To ensure a proper prior distribution, we assume $\breve{\mathbf{\Psi}}^y_{\mathbf{t}_N}$ is positive definite, $\kappa^y_{\mathbf{t}_N} > Np-1$, and $\nu^y_{\mathbf{t}_N} > 0$ for all $\mathbf{t}_N$ (cf. p. 126 in [16], p. 178 in [17], and p. 427 in [3]).

Given the preceding framework for uncertainty in the BCGP model, the optimal Bayesian classification theory can be directly adapted. Specifically, the normal-inverse-Wishart distribution prior as defined in (25) and the independence of $\mathbf{x}^y_{\mathbf{t}_N}(\omega_s)$ from training sample paths resemble the same set of conditions as in [6], i.e., having a normal-inverse-Wishart distribution prior and independence of future data points from training data points. As a result, we can follow the same set of machinery to find the *effective class-conditional distributions of the processes* (similar to equation (64) in [6]) and from there obtain the optimal discriminant. Therefore, extending the dimensionality of the problem to Np and using the set of parameters $\left\{\breve{\mathbf{m}}^y_{\mathbf{t}_N}, \breve{\mathbf{\Psi}}^y_{\mathbf{t}_N}, \nu^y_{\mathbf{t}_N}, \kappa^y_{\mathbf{t}_N}\right\}$ in the discriminant presented by Eq. (65) in [6] yields

$$\begin{aligned}\psi^{\mathrm{OBC}}_{\mathbf{t}_N}\left(\mathbf{x}^y_{\mathbf{t}_N}(\omega_s)\right) = K\Bigg(&1+\frac{1}{k^0}\left(\mathbf{x}^y_{\mathbf{t}_N}(\omega_s)-\mathbf{m}^{0\,*}_{\mathbf{t}_N}\right)^T\mathbf{\Pi}^{0\,-1}_{\mathbf{t}_N}\\ &\times\left(\mathbf{x}^y_{\mathbf{t}_N}(\omega_s)-\mathbf{m}^{0\,*}_{\mathbf{t}_N}\right)\Bigg)^{k^0+Np}\\ &-\Bigg(1+\frac{1}{k^1}\left(\mathbf{x}^y_{\mathbf{t}_N}(\omega_s)-\mathbf{m}^{1\,*}_{\mathbf{t}_N}\right)^T\mathbf{\Pi}^{1\,-1}_{\mathbf{t}_N}\\ &\times\left(\mathbf{x}^y_{\mathbf{t}_N}(\omega_s)-\mathbf{m}^{1\,*}_{\mathbf{t}_N}\right)\Bigg)^{k^1+Np},\end{aligned} \tag{27}$$

where

$$K=\left(\frac{\alpha_1}{1-\alpha_0}\right)^2\left(\frac{k^0}{k^1}\right)^{Np}\frac{|\mathbf{\Pi}^0_{\mathbf{t}_N}|}{|\mathbf{\Pi}^1_{\mathbf{t}_N}|}\left(\frac{\Gamma(k^0/2)\Gamma((k^1+pN)/2)}{\Gamma(k^1/2)\Gamma((k^0+pN)/2)}\right)^2, \tag{28}$$

with

$$\begin{aligned}
&\mathbf{\Pi}^y_{\mathbf{t}_N}=\frac{\nu^{y\,*}_{\mathbf{t}_N}+1}{\left(\kappa^{y\,*}_{\mathbf{t}_N}-Np+1\right)\nu^{y\,*}_{\mathbf{t}_N}}\mathbf{\Psi}^{y\,*}_{\mathbf{t}_N},\\
&\mathbf{\Psi}^{y\,*}_{\mathbf{t}_N}=\check{\mathbf{\Psi}}^y_{\mathbf{t}_N}+(n^y-1)\,\hat{\Sigma}^y_{\mathbf{t}_N}+\frac{\nu^y_{\mathbf{t}_N}n^y}{\nu^y_{\mathbf{t}_N}+n^y}\left(\hat{\boldsymbol{\mu}}^y_{\mathbf{t}_N}-\check{\mathbf{m}}^y_{\mathbf{t}_N}\right)\left(\hat{\boldsymbol{\mu}}^y_{\mathbf{t}_N}-\check{\mathbf{m}}^y_{\mathbf{t}_N}\right)^T,\\
&\nu^{y\,*}_{\mathbf{t}_N}=\nu^y_{\mathbf{t}_N}+n^y,\quad \kappa^{y\,*}_{\mathbf{t}_N}=\kappa^y_{\mathbf{t}_N}+n^y,\quad k^y=\kappa^{y\,*}_{\mathbf{t}_N}-Np+1,\\
&\mathbf{m}^{y\,*}_{\mathbf{t}_N}=\frac{\nu^y_{\mathbf{t}_N}\check{\mathbf{m}}^y_{\mathbf{t}_N}+n^y\hat{\boldsymbol{\mu}}^y_{\mathbf{t}_N}}{\nu^y_{\mathbf{t}_N}+n^y},
\end{aligned} \tag{29}$$

where $\check{\mathbf{m}}^y_{\mathbf{t}_N}$ and $\check{\mathbf{\Psi}}^y_{\mathbf{t}_N}$ are determined from (26), and $\hat{\Sigma}^y_{\mathbf{t}_N}$ and $\hat{\boldsymbol{\mu}}^y_{\mathbf{t}_N}$ are the sample mean and sample covariance matrix obtained by using the sample path training sets $\mathbf{S}^0_{\mathbf{t}_N}$ and $\mathbf{S}^1_{\mathbf{t}_N}$ as follows:

$$\begin{aligned}
&\hat{\boldsymbol{\mu}}^y_{\mathbf{t}_N}=\frac{1}{n^y}\sum_{i=1}^{n^y}\mathbf{x}^y_{\mathbf{t}_N}(\omega_i),\\
&\hat{\mathbf{\Sigma}}^y_{\mathbf{t}_N}=\frac{1}{n^y-1}\sum_{i=1}^{n^y}\left(\mathbf{x}^y_{\mathbf{t}_N}(\omega_i)-\hat{\boldsymbol{\mu}}^y_{\mathbf{t}_N}\right)\left(\mathbf{x}^y_{\mathbf{t}_N}(\omega_i)-\hat{\boldsymbol{\mu}}^y_{\mathbf{t}_N}\right)^T.
\end{aligned} \tag{30}$$

As opposed to Section 4.1, where the discriminant can be applied to any future sample path with an arbitrary observation time vector, here, the discriminant depends on both the future and training observation time vectors. Thus, if the future observation time vector $\mathbf{t}_L{}^y$ contains only a set of time points t_i where $t_i \in \mathbf{t}_N{}^y$, one may easily apply the optimal discriminant. This is easily doable by reducing the dimensionality of the problem by considering the training sample paths only at $\mathbf{t}_L{}^y$, i.e., by discarding the training sample points at those $\mathbf{t}_N{}^y$ not in $\mathbf{t}_L{}^y$(denoted by $\mathbf{t}_N{}^y\backslash\mathbf{t}_L{}^y$). However, solving the case where $\mathbf{t}_L{}^y$ includes time points not included in $\mathbf{t}_N{}^y$ is more difficult and requires further study. In this case, although one is able to construct the class of prior knowledge for $\mathbf{t}_L{}^y$ (i.e., constructing $\boldsymbol{\mu}^y_{\mathbf{t}_N}$ and $\mathbf{\Psi}^y_{\mathbf{t}_N}$), the paucity of training sample paths at $\mathbf{t}_L{}^y\backslash\mathbf{t}_N{}^y$ does not permit employing (27).

5 Performance analysis

In this section, we analyze the effect of prior knowledge in the form of stochastic differential equations on the performance of the stochastic discriminant, $\psi^{OBC}_{\mathbf{t}_N}\left(\mathbf{x}^y_{\mathbf{t}_N}(\omega_s)\right)$, defined by (27)–(29). As the metric of performance, we take the true error averaged over the sampling space. The true error of a discriminant trained on an observation time vector $\mathbf{t}_N$, i.e., $\psi_{\mathbf{t}_N}(.)$, is the probability of misclassification, which by considering (11) is defined as

$$\epsilon_{\mathbf{t}_N}=\sum_{y=0}^{1}\alpha^y_{\mathbf{t}_N}P\left((-1)^y\psi_{\mathbf{t}_N}\left(\mathbf{X}^y_{\mathbf{t}_N}(\omega_s)\right)>0\,|\,\mathbf{S}^0_{\mathbf{t}_N},\mathbf{S}^1_{\mathbf{t}_N},\mathbf{X}^y_{\mathbf{t}_N}(\omega_s)\in\mathbf{X}^y\right), \tag{31}$$

where $\mathbf{X}^y$ denotes the class-conditional process that generates the future sample path $\mathbf{X}^y_{\mathbf{t}_N}(\omega_s)$ (we assume independence of future sample paths from training sample paths), $\mathbf{S}^y_{\mathbf{t}_N}$ denotes the set of training sample paths from class y, and $\alpha^y_{\mathbf{t}_N}$ is the mixing probability of the class-conditional process.

Recall that in this work, we consider a separate sampling scheme. With separate sampling in a classical binary classification problem where sample points are generated from two class-conditional densities, there is no sensible estimate of prior probabilities of classes from the sample [15]. In that case, either the ratio of the number of sample points in either class to the total sample size needs to reflect the corresponding prior probability of the class or the prior probabilities need to be known a priori; otherwise, classification rules or error estimation rules suffer performance degradation [15, 18, 19]. The same argument applies to this work in which we consider a binary classification of sample paths that are generated from two class-conditional processes under a separate sampling scheme. In this regard, we assume that the prior probability $\alpha^y_{\mathbf{t}_N}$ is known a priori.

Taking expectation over the sample space, that is over the mixture of Gaussian processes with the means and covariance matrices defined by (16), (20), (17), and (21), yields

$$E[\epsilon_{\mathbf{t}_N}]=\sum_{y=0}^{1}\alpha^y_{\mathbf{t}_N}P\left((-1)^y\psi_{\mathbf{t}_N}\left(\mathbf{X}^y_{\mathbf{t}_N}(\omega_s)\right)>0\,|\,\mathbf{X}^y_{\mathbf{t}_N}(\omega_s)\in\mathbf{X}^y\right). \tag{32}$$

As benchmarks for evaluating the performance of $\psi^{\mathrm{OBC}}_{\mathbf{t}_N}\left(\mathbf{x}^y_{\mathbf{t}_N}(\omega_s)\right)$, we compare its performance to (1) the performance of the stochastic QDA, $\psi^{\mathrm{QDA}}_{\mathbf{t}_N}\left(\mathbf{x}^y_{\mathbf{t}_N}(\omega_s)\right)$, which is defined by (23) and (24), where $l_y = n_y$, with n_y indicating the number of available sample paths, and (2) the performance of a Bayes classifier obtained by plugging (16), (17), (20), and (21), into (24).

5.1 Synthetic experiments

5.1.1 *Experimental set-up*

The following steps are used to set up the experiments:

1. To fix the ground-truth model governing the underlying dynamics of the data, we consider a set of three-dimensional SDEs ($p = 3$) defined by (15) along with the following set of parameters:

$$\begin{aligned}
&\mathbf{A}^0(t)=\mathbf{A}^1(t)=[0.01,0.01,0.01]^T,\\
&\mathbf{a}^0(t)=\mathbf{a}^1(t)=[0,0,0]^T,\\
&\mathbf{X}^0_{t_0}(\omega)=[0,0,0]^T,\quad \mathbf{X}^1_{t_0}(\omega)=[0.25,0.25,0.25]^T,\\
&\mathbf{B}^0(t)=\mathbf{B}^1(t)=0.1\times\begin{cases}\sigma^2=1 & \text{diagonal elements}\\ \rho=0.4 & \text{otherwise}\end{cases}.
\end{aligned} \tag{33}$$

The only difference between the SDEs describing $\mathbf{X}^0$ and $\mathbf{X}^1$ is in the constant initial conditions. Figure 1 presents a single sample path of these two three-dimensional processes for $0 \leq t \leq 100$.

2. Use the ground-truth set of SDEs to generate a set of training sample paths, $\mathbf{S}^y_{t_N}$, of size n^y for class $y = 0, 1$. We let $n^0 = n^1 = n$, where $n \in$, let the length of the observation time vector be $N = 20$, and take $[t_1, t_2, ..., t_N]$ such that $t_i - t_{i-1} = 1$, $i = 2, \ldots, 20$.
3. Use the ground-truth set of SDEs to generate a set of test sample paths, $\mathbf{S}^{y,\text{test}}_{t_N}$, of size $n^{y,\text{test}} = 2,000$ for class $y = 0, 1$, where $n^{0,\text{test}} = n^{1,\text{test}} = n^{\text{test}}$.
4. Use $\mathbf{S}^0_{t_N} \cup \mathbf{S}^1_{t_N}$ to train the stochastic QDA, $\psi^{\text{QDA}}_{\mathbf{t}_N}\left(\mathbf{x}^y_{\mathbf{t}_N}(\omega_s)\right)$, which is defined by (23) and (24) with $l^y = n^y$. Apply the trained classifier to the set of test sample paths, $\mathbf{S}^{0,\text{test}}_{t_N} \cup \mathbf{S}^{1,\text{test}}_{t_N}$, to determine the true error, $\epsilon^{\text{QDA}}_{\mathbf{t}_N}$, which is defined by replacing $\psi_{\mathbf{t}_N}\left(\mathbf{X}^y_{\mathbf{t}_N}(\omega_s)\right)$ with $\psi^{\text{QDA}}_{\mathbf{t}_N}\left(\mathbf{X}^y_{\mathbf{t}_N}(\omega_s)\right)$ in (31). This procedure obtains an accurate estimate of true error.
5. Assume a set of SDEs obtained from prior knowledge (a priori SDEs). Let this a priori set of SDEs be presented by replacing $\mathbf{A}^y(t)$, $\mathbf{B}^y(t)$, $\mathbf{a}^y(t)$, and $\mathbf{X}^y_{t_0}(\omega)$ in (15) with $\tilde{\mathbf{A}}^y(t)$, $\tilde{\mathbf{B}}^y(t)$, $\tilde{\mathbf{a}}^y(t)$, and $\tilde{\mathbf{X}}^y_{t_0}(\omega)$, respectively. To examine the effects of deviations in the drift vector and dispersion matrix in the a priori set of SDEs from the ground-truth model introduced in (33), we assume

   - $\tilde{\mathbf{A}}^0(t) = \mathbf{A}^0(t)$, $\tilde{\mathbf{B}}^0(t) = \mathbf{B}^0(t)$, $\tilde{\mathbf{X}}^0_{t_0}(\omega) = \mathbf{X}^0_{t_0}(\omega)$, $\tilde{\mathbf{X}}^1_{t_0}(\omega) = \mathbf{X}^1_{t_0}(\omega)$, $\tilde{\mathbf{a}}^0(t) = \mathbf{a}^0(t)$, $\tilde{\mathbf{a}}^1(t) = \mathbf{a}^1(t)$.
   - To study the effect of shift in the drift vector, we take $\tilde{\mathbf{A}}^1(t) = \mathbf{A}^1(t) + [\Delta\mu, \Delta\mu, \Delta\mu]^T$, where $\Delta\mu = 0, 0.1, 0.2, 0.3$. Here we assume $\tilde{\mathbf{B}}^1(t) = \mathbf{B}^1(t)$.
   - To study the effect of shift in the dispersion matrix, we assume the off-diagonal elements of $\tilde{\mathbf{B}}^1(t)$ are defined by replacing ρ with ρ_d in (33), where $\rho_d - \rho = \Delta\rho = 0, 0.03, 0.06, 0.1$. Here we assume $\tilde{\mathbf{A}}^1(t) = \mathbf{A}^1(t)$.
   - The hyperparameters defining our uncertainty about the specific choice of a priori SDEs (in fact, about the resultant prior distributions) are $\nu^0_{\mathbf{t}_N} = \nu^1_{\mathbf{t}_N} = \kappa^0_{\mathbf{t}_N} = \kappa^1_{\mathbf{t}_N} = Np + \kappa$. The choice of $Np + \kappa$, $\kappa = 20, 50, 100, 500$, is made to have proper prior distributions (see Section 4.2).
6. Generate 2,000 sample paths from the a priori set of SDEs introduced in Step 5. These sample paths are used to calculate the hyperparameters $\mathbf{m}^y_{\mathbf{t}_N}$ and $\mathbf{\Psi}^y_{\mathbf{t}_N}$ being used in (26) (alternatively, one may solve (16), (17), (20), and (21) directly and use them in (26)).
7. Use $\mathbf{m}^y_{\mathbf{t}_N}$ and $\mathbf{\Psi}^y_{\mathbf{t}_N}$ obtained from Step 6 along with $\mathbf{S}^0_{t_N} \cup \mathbf{S}^1_{t_N}$ to train $\psi^{\text{OBC}}_{\mathbf{t}_N}\left(\mathbf{x}^y_{\mathbf{t}_N}(\omega_s)\right)$, which is defined in (27). Apply the trained classifier to the set of test sample paths, $\mathbf{S}^{0,\text{test}}_{t_N} \cup \mathbf{S}^{1,\text{test}}_{t_N}$, to determine the true error, $\epsilon^{\text{OBC}}_{\mathbf{t}_N}$, which is defined by replacing $\psi_{\mathbf{t}_N}\left(\mathbf{X}^y_{\mathbf{t}_N}(\omega_s)\right)$ with $\psi^{\text{OBC}}_{\mathbf{t}_N}\left(\mathbf{X}^y_{\mathbf{t}_N}(\omega_s)\right)$ in (31).
8. Repeat Steps 2 through 7 a total of $T = 1,000$ times to estimate $E\left[\epsilon^{\text{QDA}}_{\mathbf{t}_N}\right]$ and $E\left[\epsilon^{\text{OBC}}_{\mathbf{t}_N}\right]$.
9. Generate 2,000 sample paths from the ground-truth set of SDEs introduced in (33). Use these sample

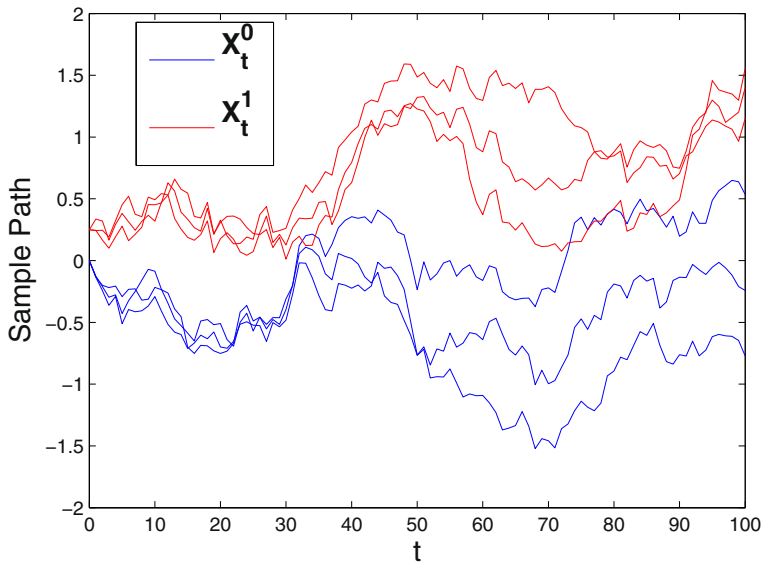

Fig. 1 A single sample path taken from the two three-dimensional processes described by the set of parameters introduced in (33)

paths to train the stochastic QDA, $\psi_{\mathbf{t}_N}^{\mathrm{QDA}}\left(\mathbf{x}_{\mathbf{t}_N}^{y}(\omega_s)\right)$, which is defined by (23) and (24) with $l^y = 2,000$. This provides an accurate estimate of the Bayes (optimal) classifier. Apply this classifier to $\mathbf{S}_{\mathbf{t}_N}^{0,\,\mathrm{test}} \cup \mathbf{S}_{\mathbf{t}_N}^{1,\,\mathrm{test}}$ to obtain the Bayes error, which is a lower bound on the error of any classifier. Note that in our experiments obtaining the Bayes error is possible since we have complete knowledge of the underlying ground-truth models.

5.1.2 Results

Figure 2 shows the effect of a shift in the drift vector from the ground-truth model via plots of the expected true error of $\psi_{\mathbf{t}_N}^{\mathrm{QDA}}(.)$ and $\psi_{\mathbf{t}_N}^{\mathrm{OBC}}(.)$ as functions of the size of training sample paths and κ for $y = 0, 1$, $\tilde{\mathbf{B}}^y(t) = \mathbf{B}^y(t)$, $\tilde{\mathbf{X}}_{t_0}^y(\omega) = \mathbf{X}_{t_0}^y(\omega)$, $\tilde{\mathbf{A}}^0(t) = \mathbf{A}^0(t)$, and $\tilde{\mathbf{A}}^1(t) = \mathbf{A}^1(t) + [\Delta\mu, \Delta\mu, \Delta\mu]^T$, where $\Delta\mu = 0, 0.1, 0.2, 0.3$. If the set of a priori SDEs is equivalent or close to the ground-truth model, e.g., $\Delta\mu = 0$ or $\Delta\mu = 0.1$, then $\psi_{\mathbf{t}_N}^{\mathrm{OBC}}(.)$ outperforms $\psi_{\mathbf{t}_N}^{\mathrm{QDA}}(.)$ for a wide range of training sample sizes and κ. The more the prior distribution generated from the set of a priori SDEs is concentrated about the true underlying parameters of the model and the larger κ, the better is the performance achieved by using $\psi_{\mathbf{t}_N}^{\mathrm{OBC}}(.)$.

Figure 3 presents the effect of the discrepancy between the dispersion matrix of the ground-truth model and that of the a priori set of SDEs. Again, the closer the prior knowledge is to the ground-truth model and the larger κ, the better is the performance achieved by using $\psi_{\mathbf{t}_N}^{\mathrm{OBC}}(.)$.

5.2 An experiment inspired by a model of the evolutionary process

In this section, we use a form of an Ornstein-Uhlenbeck process introduced in [20] for modeling the evolutionary

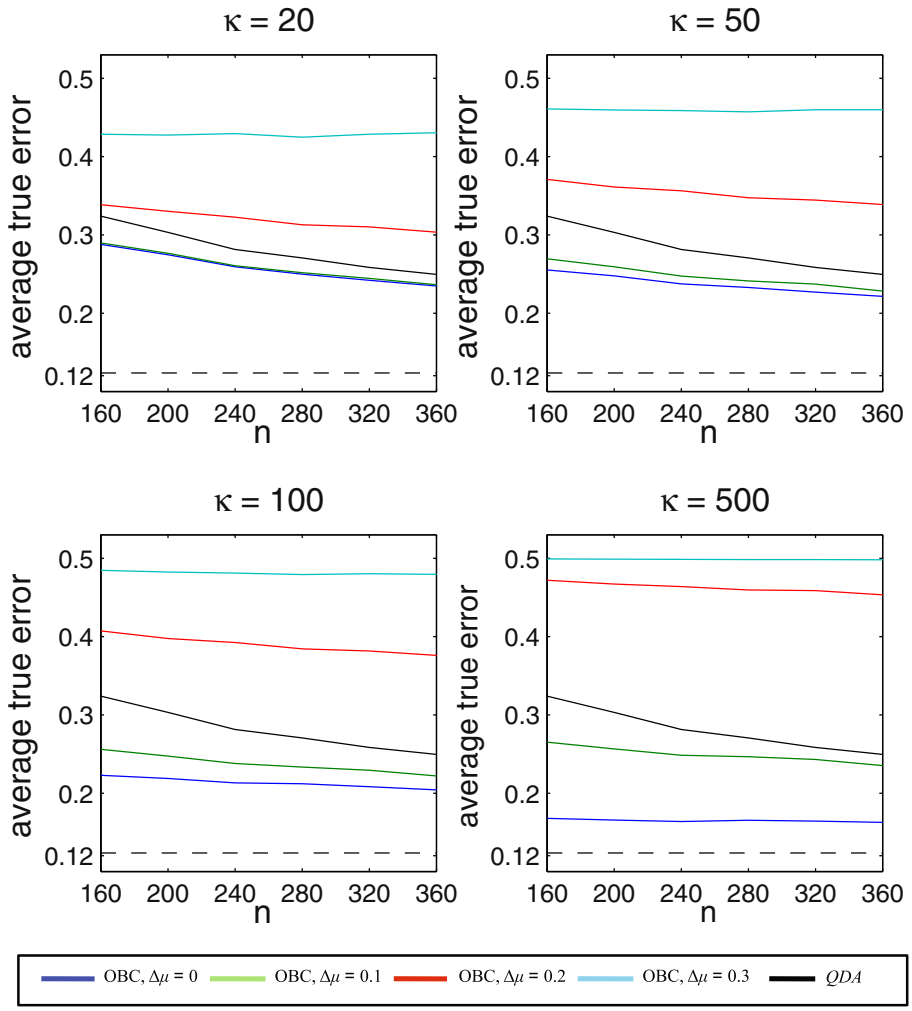

Fig. 2 Expected true error, $E\left[\epsilon_{\mathbf{t}_N}^{\mathrm{QDA}}\right]$ and $E\left[\epsilon_{\mathbf{t}_N}^{\mathrm{OBC}}\right]$, as a function of number of training sample paths in each class and various choices of $\Delta\mu$ and κ. The *dashed line* shows the Bayes error

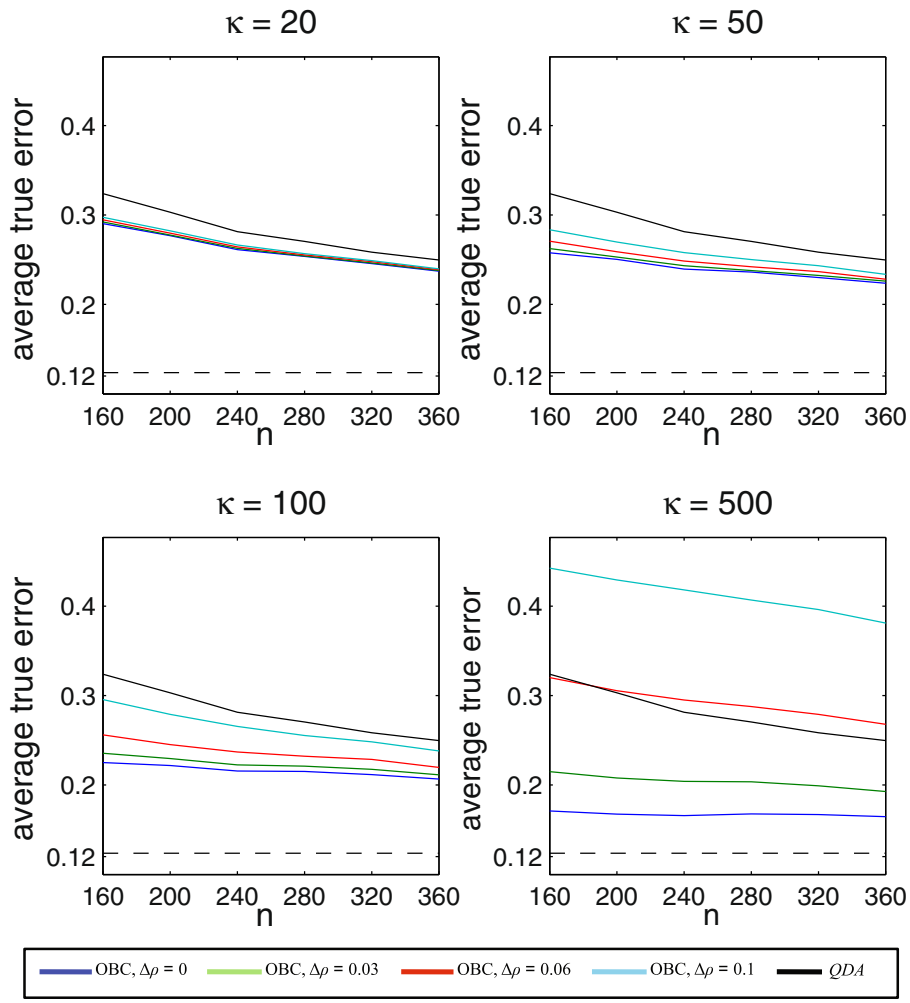

Fig. 3 Expected true error, $E\left[\epsilon_{\mathbf{t}_N}^{\mathrm{QDA}}\right]$ and $E\left[\epsilon_{\mathbf{t}_N}^{\mathrm{OBC}}\right]$, as a function of number of training sample paths in each class and various choices of $\Delta\rho$ and κ. The *dashed line* shows the Bayes error

change of species. This model has been recently employed by [21] to simulate quantitative trait data as a function of single nucleotide polymorphism (SNP) states. The model is presented by the following SDE:

$$\mathrm{d}X_t^y = -\beta^y \left[X_t^y - \theta^y\right] \mathrm{d}t + \sigma^y \mathrm{d}W_t^y, \quad X_0^y = X_a^y, \qquad (34)$$

where X_t^y is the quantitative trait value in a species y, θ^y is the primary target value of the trait, X_a^y is the mean state in an ancestor a, and W_t^y represents Brownian motion. The parameter β^y is the rate of adaptation of species y to the target value—a low rate of adaptation means very slow evolution while a large β^y practically indicates an instantaneous adaptation. The parameter σ^y is an indicator of perturbation due to random selective factors such as random mutations and environmental fluctuations [20]. Similar to [21], we assume the value of the primary target is constant over the history of the species. Nevertheless, the model in (34) can be extended to include situations where the primary target can change over the evolutionary history of the species (see [20]).

Using the model of (34), we generate the evolutionary histories of a quantitative trait of two species, 0 and 1, over a time span of 30 million years with time steps of 1 million years. Similarly to [20, 21], to fix the ground-truth model that generates the data, we vary values of β^y, take $\sigma^y = 1$, and assume $\theta^0 = 80$ and $\theta^1 = 85$. Furthermore, we assume both species have a common ancestor at the state $X_a^y = 1$. Figure 4 presents 20 sample paths from each of these evolutionary processes for the case where $\beta^0 = \beta^1 = \beta$, $\beta = 0.1$ (Fig. 4a) and $\beta = 0.15$ (Fig. 4b). A larger β indicates a faster adaptation of species to the target value. The problem considered here is to use a set of a priori SDEs in constructing a classifier to differentiate the

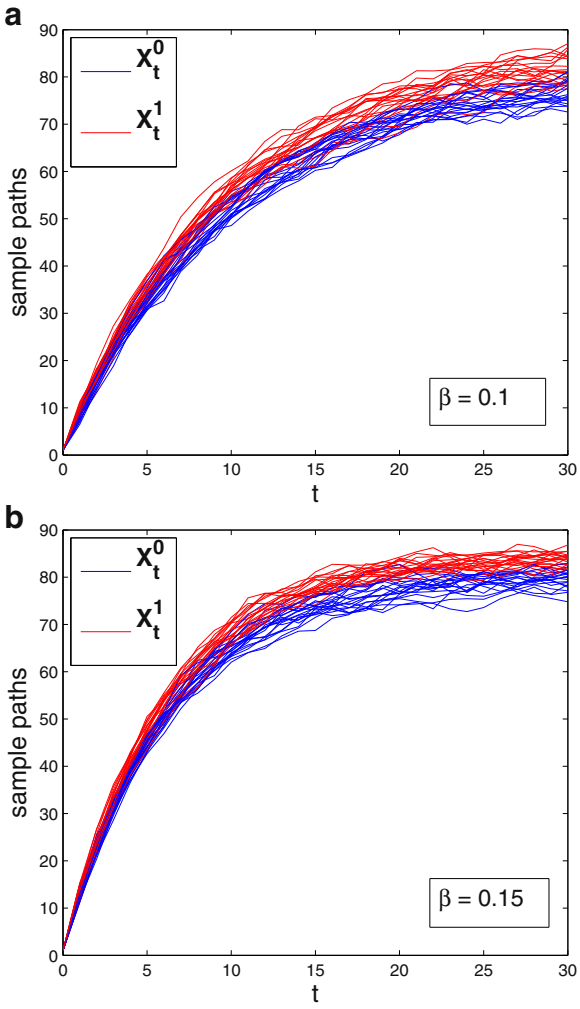

Fig. 4 Multiple sample paths taken from the two one-dimensional evolutionary processes for two values of adaptation rate, (**a**): $\beta = 0.1$; (**b**): $\beta = 0.15$

evolutionary history of an n-size population of species 0 from an n-size population of species 1, where $n \in [60, 140]$.

The general protocol for evaluating the performance of $\psi_{\mathbf{t}_N}^{\mathrm{OBC}}(.)$ is similar to Section 5.1, except for replacing the ground-truth model (33) with (34) and using the following the step instead of Step 5:

- Assume a set of SDEs obtained from prior knowledge (a priori SDEs). Let this a priori set of SDEs be presented by replacing β^y, θ^y, X_a^y, and σ^y by $\tilde{\beta}^y$, $\tilde{\theta}^y$, $\tilde{X}_a^y$, and $\tilde{\sigma}^y$, respectively, in (34). To examine the effect of deviation of the adaptation rate in the a priori set of SDEs from the ground-truth model, we let $\tilde{\theta}^y = \theta^y$, $\tilde{\sigma}^y = \sigma^y$, $\tilde{X}_a^y = X^y$, and $\tilde{\beta}^0 = \beta^0$ and take $\tilde{\beta}^1 = \beta^1 + \Delta\beta$.

5.2.1 Results

Figures 5 and 6 ($\beta = 0.1$ and $\beta = 0.15$, respectively) show the effect of a deviation from the true rate of adaptation to the target value by considering $\tilde{\beta}^1 = \beta^1 + \Delta\beta$, where $\Delta\beta = 0, 0.02, 0.04, 0.06$. They provide plots of the expected true error of $\psi_{\mathbf{t}_N}^{\mathrm{QDA}}(.)$ and $\psi_{\mathbf{t}_N}^{\mathrm{OBC}}(.)$ as functions of the size of training sample paths and κ. In both figures, the closer the prior knowledge is to the ground-truth evolutionary models, the better is the performance achieved by using $\psi_{\mathbf{t}_N}^{\mathrm{OBC}}(.)$. The performance deteriorates and eventually becomes worse than $\psi_{\mathbf{t}_N}^{\mathrm{QDA}}(.)$ as the prior knowledge diverges from the ground-truth model and the certainty about the prior knowledge increases (a bad combination when utilizing prior knowledge). In addition, comparing Figs. 5 and 6 shows that the smaller is the true value of β and the more destructive is a fixed deviation of prior knowledge from the true β.

6 Conclusions

This paper provides the first instance in which prior knowledge in the form of SDEs is used to construct a prior distribution over an uncertainty class of feature-label distributions for the purpose of optimal classification. Given the ubiquity of small samples in biomedicine and other areas where sample data is expensive, time-consuming, limited by regulation, or simply unavailable, we have previously made the point that prior knowledge is the only avenue available. To achieve the mapping of SDE prior knowledge into a prior distribution, we have taken advantage of the form and Gaussianity of (12). This mapping is heavily dependent on the form of the SDEs, and one can expect widely varying mappings for different SDE settings.

In general, all parameters used in the a priori set of SDEs can affect the performance of $\psi_{\mathbf{t}_N}^{\mathrm{OBC}}(.)$. These parameters include every element of the matrices $\tilde{\mathbf{A}}^y(t)$ and $\tilde{\mathbf{B}}^y(t)$ and all the elements of the vectors $\tilde{\mathbf{a}}^y(t)$ and $\tilde{\mathbf{X}}^y{}_{t_0}(\omega)$ used in the SDE's presentation in (15). For example, in the experiment of the evolutionary change of species considered in (34), a deviation from each of the parameters, namely $\tilde{\beta}^y$, $\tilde{\sigma}^y$, $\tilde{\theta}^y$, and $\tilde{X}_a^y$, can affect the performance of $\psi_{\mathbf{t}_N}^{\mathrm{OBC}}(.)$. Although simulation studies can elucidate the effects of deviation of prior knowledge from the ground-truth model (as done herein), it would be beneficial to analytically characterize the performance of $\psi_{\mathbf{t}_N}^{\mathrm{OBC}}(.)$ in terms of all the hyperparameters; however, this may be very difficult to accomplish. One possible approach may be to use an asymptotic Bayesian framework [22] to characterize the performance of $\psi_{\mathbf{t}_N}^{\mathrm{OBC}}(.)$ in terms of sample size, dimensionality, and hyperparameters.

Recognizing that the construction of robust classifiers is simply a special case of optimal Bayesian classification where there are no sample data, so that the "posterior" is identical to the prior [7], the application of SDEs in

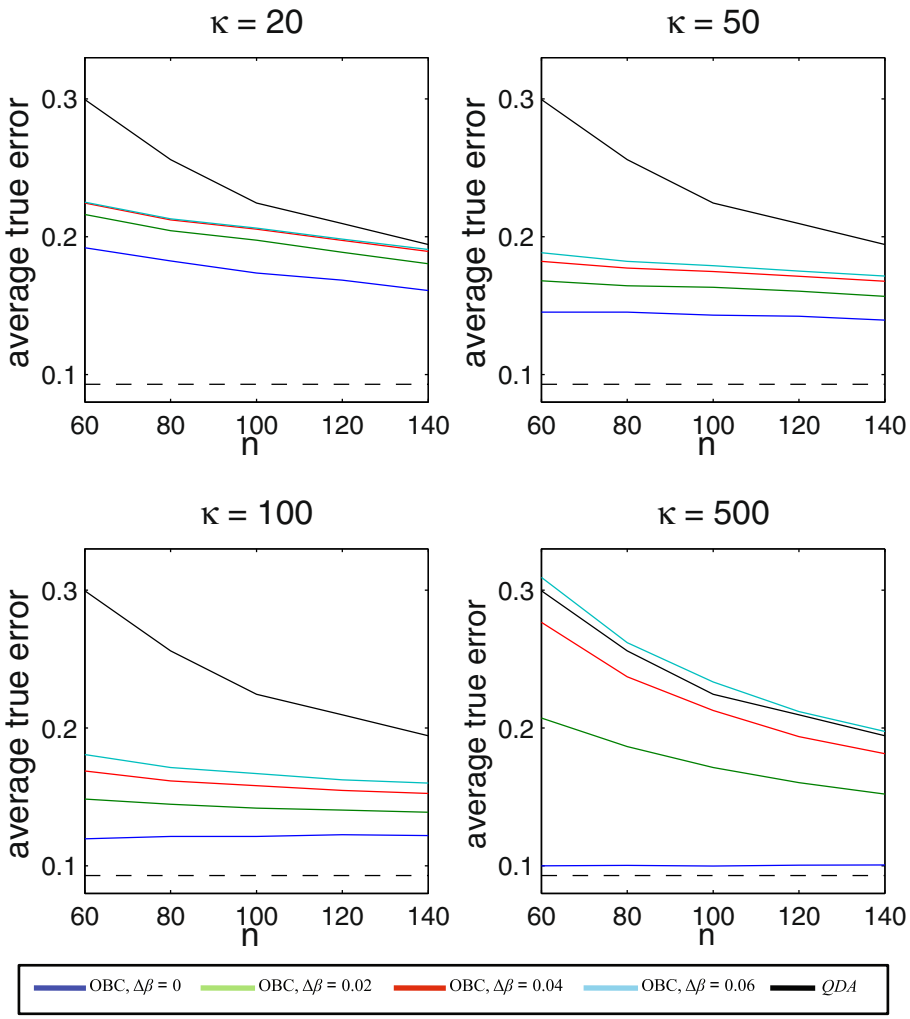

Fig. 5 Expected true error, $E\left[\epsilon_{\mathbf{t}_N}^{\mathrm{QDA}}\right]$ and $E\left[\epsilon_{\mathbf{t}_N}^{\mathrm{OBC}}\right]$, as a function of number of training sample paths in each class and various choices of $\Delta\beta$ and κ and for β=0.1. The *dashed line* shows the Bayes error

this paper is at once applicable to optimal robust classification in a stochastic setting. Beyond that, one can consider the more general setting of optimal Bayesian robust filtering of random processes, where optimization across an uncertainty class of random processes, ideal and observed, is relative to process characteristics such as the auto- and cross-correlation functions [23]. Whereas in this paper we have considered using SDE prior knowledge to construct prior distributions governing uncertainty classes of feature-label distributions, it seems feasible to use SDE knowledge to construct prior distributions governing uncertainty classes of random-process characteristics in the case of optimal filtering. Of course, one must confront the increased abstraction presented by canonical representation of random processes [24, 25]; nevertheless, so long as one remains in the framework of second-order canonical expansions, it should be doable.

Appendix

Definition of q-dimensional Wiener process

A one-dimensional Wiener process over $[0, T]$ is a Gaussian process $W = \{W_t : t \geq 0\}$ satisfying the following properties:

- For $0 \leq t_1 < t_2 < T$, $W_{t_2} - W_{t_1}$ is distributed as $\sqrt{t_2 - t_1}N\left(0, \sigma^2\right)$, where $\sigma > 0$ (for the standard Wiener process, $\sigma = 1$).
- For $0 \leq t_1 < t_2 < t_3 < t_4 < T$, $W_{t_4} - W_{t_3}$ is independent of $W_{t_2} - W_{t_1}$.
- $W_0 = 0$ with probability 1.

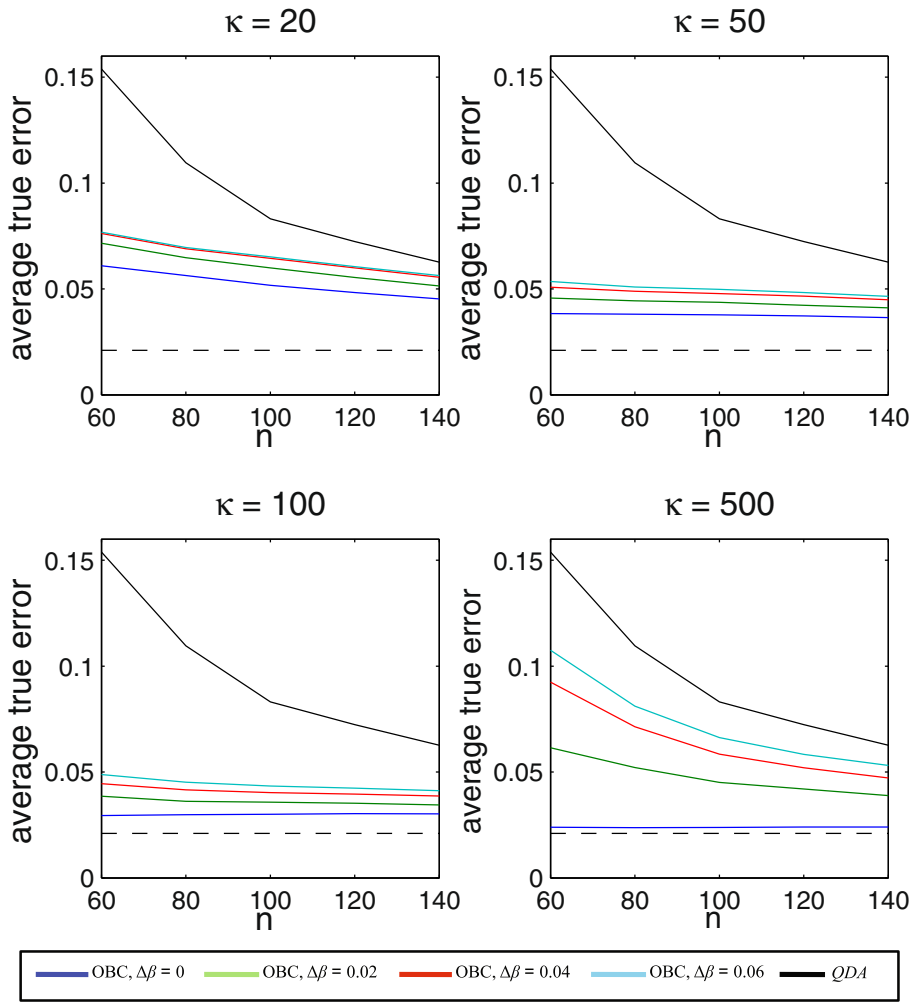

Fig. 6 Expected true error, $E\left[\epsilon_{\mathbf{t}_N}^{\mathrm{QDA}}\right]$ and $E\left[\epsilon_{\mathbf{t}_N}^{\mathrm{OBC}}\right]$, as a function of number of training sample paths in each class and various choices of $\Delta\beta$ and κ and for β =0.15. The *dashed line* shows the Bayes error

- The sample paths of W are almost surely continuous everywhere.

In general, a q-dimensional Wiener process is defined using the homogenous Markov process $\mathbf{X}_t$ for $t \in [t_0, T]$. Let $P(t_1, x; t_2, B) = P(\mathbf{X}_{t_2} \in B | \mathbf{X}_{t_1} = x)$ denote the transition probabilities of a Markov process $\mathbf{X}_t$ for $t_1 < t_2$. For fixed values of t_1, x, and t_2, $P(t_1, x; t_2, .)$ is a probability function (measure) on the σ-algebra $\mathcal{B}$ of Borel subsets of the sample space R^q. Intuitively, $P(t_1, x; t_2, B)$ is the probability that the process be in the set $B \in \mathcal{B}$ at time t_2 given it was in state x at time t_1. A Markov process is homogenous with respect to t if its transition probability $P(t_1, x; t_2, B)$ is stationary. That is, for $t_0 < t_1 < t_2 < T$ and $t_0 < t_1 + u < t_2 + u < T$, it satisfies

$$P(t_1 + u, x; t_2 + u, B) = P(t_1, x; t_2, B). \tag{35}$$

In this case $P(t_1, x; t_2, B)$ is commonly denoted by $P(t_2 - t_1, x; B)$. A q-dimensional Wiener process is a q-dimensional homogenous Markov process defined on $[0, \infty)$ with stationary transition probability defined by a multivariate Gaussian distribution as follows:

$$P(t, x; B) = \int_B \frac{1}{(2\pi t)^{d/2}} e^{-\frac{|y-x|^2}{2t}} dy. \tag{36}$$

Therefore, each dimension of a q-dimensional Wiener process is a one-dimensional Wiener process per se.

Computational complexity

The computational complexity of the algorithm is determined by the computational cost of solving the set of SDEs from the Euler-Maruyama scheme (see Section 4.1) along with the computational cost of evaluating (27). The computational cost of the Euler-Maruyama scheme per sample path is inversely proportional to Δt [26], where

$\Delta t = T/N$, with T and N being defined in Section 3. Thus, for $l = l^0 + l^1$ sample paths, it is $O(l/\Delta t)$. In (27), the computational cost of evaluating $\mathbf{x}_{\mathbf{t}_N}^y(\omega_s) - \mathbf{m}_{\mathbf{t}_N}^{0*}$, with $y = 0, 1$, breaks down to a computation of $\check{\mathbf{m}}_{\mathbf{t}_N}^y$ and $\hat{\boldsymbol{\mu}}_{\mathbf{t}_N}^y$, which are operations with computational costs of $O(l^y Np)$ and $O(n^y Np)$, respectively.

Computation of $\boldsymbol{\Pi}_{\mathbf{t}_N}^{y\,-1}$ in (27) by Gaussian elimination is an $O\left(\max\{n^y, Np\}N^2p^2\right) + O\left(l^y N^2 p^2\right)$ operation (cf. section 3.7.2 in [27]). This will be further simplified because, in order to have a positive definite $\check{\boldsymbol{\Psi}}_{\mathbf{t}_N}^y$, we assume we generate many sample paths by solving the set of SDEs such that $l^y >> Np$ (see Section 4.1), but since $\boldsymbol{\Psi}_{\mathbf{t}_N}^{y\,*}$ and $\boldsymbol{\Pi}_{\mathbf{t}_N}^{y\,-1}$ defined in (29) become positive definite, we do not need to impose the condition of $n^y > Np$. Having a realistic assumption on the number of available sample paths, we can assume $l^y >> n^y$, and therefore, the computation of $\boldsymbol{\Pi}_{\mathbf{t}_N}^{y\,-1}$ becomes an $O\left(l^y N^2 p^2\right)$ calculation. Furthermore, the product of $\mathbf{x}_{\mathbf{t}_N}^y(\omega_s) - \mathbf{m}_{\mathbf{t}_N}^{0*}$ with $\boldsymbol{\Pi}_{\mathbf{t}_N}^{y\,-1}$ is an $O\left(N^2p^2\right)$ calculation. Altogether, by assuming $1/\Delta t < (Np)^2$ and $k^0 + k^1 + Np < (Np)^2$, the overall computational cost of $\psi_{\mathbf{t}_N}^{\mathrm{OBC}}\left(\mathbf{x}_{\mathbf{t}_N}^y(\omega_s)\right)$ is $O\left(\max\{l^0, l^1\}N^2p^2\right)$.

Using a similar approach, we see that the computational cost of QDA, which is solely constructed by using $n^0 + n^1$ training sample paths from classes 0 and 1 (i.e., no prior knowledge) is $O\left(\max\{n^0, n^1\}N^2p^2\right)$. We also note that for computing QDA we need to have $\min\{n^0, n^1\} > Np$ because, otherwise, the sample covariance matrices used in QDA are not invertible.

Additional file

Additional file 1: Supplementary information. I. Definition of QDA in a classical setting. II. Error estimation accuracy. III. Bayesian MMSE error estimator. IV. Review of literature pertaining to classification of stochastic processes.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Electrical and Electronic Engineering, Nazarbayev University, Astana, 010000, Kazakhstan. [2]The Center for Bioinformatics and Genomic Systems Engineering and the Department of Electrical and Computer Engineering, Texas A&M University, 77840, College Station, Texas.

Received: 8 May 2015 Accepted: 12 January 2016
Published online: 20 January 2016

References

1. Braga-Neto, UM, & Dougherty, ER (2015). *Error Estimation for Pattern Recognition*. New York: Wiley-IEEE Press.
2. Kay, S (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory*. New Jersey: Prentice-Hall.
3. Carlin, BP, & Louis, TA (2008). *Bayesian Methods for Data Analysis*. Boca Raton: CRC Press.
4. Dalton, L, & Dougherty, ER (2011). Bayesian minimum mean-square error estimation for classification error–part I: definition and the Bayesian MMSE error estimator for discrete classification. *IEEE Transactions on Signal Processing*, *59*(1), 115–129.
5. Dalton, L, & Dougherty, ER (2011). Bayesian minimum mean-square error estimation for classification error–part II: linear classification of Gaussian models. *IEEE Transactions on Signal Processing*, *59*(1), 130–144.
6. Dalton, L, & Dougherty, ER (2013). Optimal classifiers with minimum expected error within a Bayesian framework – part I: discrete and Gaussian models. *Pattern Recognition*, *46*, 1301–1314.
7. Dalton, L, & Dougherty, ER (2013). Optimal classifiers with minimum expected error within a Bayesian framework – part II: properties and performance analysis. *Pattern Recognition*, *46*, 1288–1300.
8. Knight, J, Ivanov, I, Dougherty, ER (2014). MCMC implementation of the optimal Bayesian classifier for non-gaussian models: model-based RNA-seq classification. *BMC Bioinformatics*, *15*. doi:10.1186/s12859-014-0401-3.
9. Esfahani, MS, & Dougherty, ER (2014). Incorporation of biological pathway knowledge in the construction of priors for optimal Bayesian classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *11*, 202–218.
10. Esfahani, MS, & Dougherty, ER (2015). An optimization-based framework for the transformation of incomplete biological knowledge into a probabilistic structure and its application to the utilization of gene/protein signaling pathways in discrete phenotype classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. doi:10.1109/TCBB.2015.2424407.
11. Jaynes, ET (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, *4*, 227–241.
12. Kloeden, PE, & Platen, E (1995). *Numerical Solution of Stochastic Differential Equations*. New York: Springer.
13. Arnold, L (1974). *Stochastic Differential Equations: Theory and Applications*. New York: Wiley.
14. Higham, D (2001). An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, *43*, 525–546.
15. Anderson, TW (1951). Classification by multivariate analysis. *Psychometrika*, *16*, 31–50.
16. Murphy, KP (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press.
17. DeGroot, MH (1970). *Optimal Statistical Decisions*. New York: McGrawHill.
18. Esfahani, MS, & Dougherty, ER (2014). Effect of separate sampling on classification accuracy. *Bioinformatics*, *30*, 242–250.
19. Braga-Neto, UM, Zollanvari, A, Dougherty, ER (2014). Cross-validation under separate sampling: strong bias and how to correct it. *Bioinformatics*, *30*, 3349–3355.
20. Hansen, TF (1997). Stabilizing selection and the comparative analysis of adaptation. *Evolution*, *51*, 1341–1351.
21. Thompson, K, & Kubatko, LS (2013). Using ancestral information to detect and localize quantitative trait loci in genome-wide association studies. *BMC Bioinformatics*, *14*. doi:10.1186/1471-2105-14-200.
22. Zollanvari, A, & Dougherty, ER (2014). Moments and root-mean-square error of the Bayesian MMSE estimator of classification error in the Gaussian model. *Pattern Recognition*, *47*, 2178–2192.
23. Dalton, L, & Dougherty, ER (2014). Intrinsically optimal Bayesian robust filtering. *IEEE Transactions on Signal Processing*, *62*(3), 657–670.
24. Pugachev, VS (1965). *Theory of Random Functions and Its Applications to Control Problems*. Oxford: Pergamon.
25. Dougherty, ER (1999). *Random Processes for Image and Signal Processing*. New York: SPIE Press and IEEE Presses.
26. Higham, DJ (2015). An introduction to multilevel Monte Carlo for option valuation. *International Journal of Computer Mathematics*, *92*(12).
27. Duda, RO, Hart, PE, Stork, DG (2000). *Pattern Classification*. New York: Wiley.