Journal of Classification (2020) 37:277–297
https://doi.org/10.1007/s00357-019-09316-6

Accurate Bayesian Data Classification Without Hyperparameter Cross-Validation

Mansoor Sheikh[1,2] · A. C. C. Coolen[1,3]

Published online: 2 April 2020

Abstract
We extend the standard Bayesian multivariate Gaussian generative data classifier by considering a generalization of the conjugate, normal-Wishart prior distribution, and by deriving the hyperparameters analytically via evidence maximization. The behaviour of the optimal hyperparameters is explored in the high-dimensional data regime. The classification accuracy of the resulting generalized model is competitive with state-of-the art Bayesian discriminant analysis methods, but without the usual computational burden of cross-validation.

Keywords Hyperparameters · Evidence maximization · Bayesian classification · High-dimensional data

1 Introduction

In the conventional formulation of classification problems, one aims to map data samples $\boldsymbol{x} \in \mathbb{R}^d$ correctly into discrete classes $y \in \{1, \ldots C\}$, by inferring the underlying statistical regularities from a given training set $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ of i.i.d. samples and corresponding classes. The standard approach to this task is to define a suitable parametrization $p(\boldsymbol{x}, y|\boldsymbol{\theta})$ of the multivariate distribution from which the samples in $\mathcal{D}$ were drawn. If the number of samples n is large compared to the data dimensionality d, computing point estimates of the unknown parameters $\boldsymbol{\theta}$ by maximum likelihood (ML) or maximum a posteriori probability (MAP) methods is accurate and usually sufficient. On the other hand, if the ratio d/n is not small, point estimation–based methods are prone to overfitting. This is the ‘curse of dimensionality’. Unfortunately, the regime of finite d/n is quite relevant for medical

✉ Mansoor Sheikh
Mansoor.sheikh@kcl.ac.uk

[1] Institute for Mathematical and Molecular Biomedicine (IMMB), King’s College London, Hodgkin Building 4N/5N (Guy’s’ Campus), London SE1 1UL, UK

[2] Department of Mathematics, King’s College London, The Strand, London WC2R 2LS, UK

[3] Saddle Point Science, London UK

applications, where clinical data-sets often report on relatively few patients but contain many measurements per patient.[1]

In generative classification models, a crucial role is played by the class-specific sample covariance matrices that capture the correlations between the components of $\boldsymbol{x}$. These will have to be inferred, either explicitly or implicitly. While the sample covariance matrix Σ is a consistent estimator for the population covariance matrix Σ_0 in the limit $d/n \to 0$, for finite d/n, the empirical covariance eigenvalue distribution $\varrho(\xi)$ is a poor estimator for its population counterpart $\varrho_0(\lambda)$. This becomes more pronounced as d/n increases.[2] In addition to the clear bias in covariance estimation induced by high ratios of d/n, the geometry of high-dimensional spaces produces further extreme and often counterintuitive values of probability masses and densities (MacKay 1999).

The overfitting problem has been known for many decades, and many strategies have been proposed to combat its impact on multivariate point estimation inferences. Regularization methods add a penalty term to the ML loss function. The penalty strength acts as a hyperparameter, and is to be estimated. The penalty terms punish, e.g. increasing values of the sum of absolute parameter values (LASSO) or squared parameter values (ridge regression), or a linear combination of these (elastic net) (Zou and Hastie 2005). They appear naturally upon introducing prior probabilities in Bayesian inference, followed by MAP point estimation. Feature selection methods seek to identify a subset of 'informative' components of the sample vectors. They range from linear methods such as principal component analysis (Hotelling 1933) to non-linear approaches such as auto-encoders (Hinton and Salakhutdinov 2006). They can guide experimental work, by suggesting which data features to examine. Most of these techniques use heuristics to determine the number of features to select. Early work by Stein and et al. (1956) introduced the concept of shrinking a traditional estimator toward a 'grand average'. In the univariate case, this was an average of averages (Efron and Morris 1977). For the multivariate case, the James-Stein estimator is an admissible estimator of the population covariance matrix (James and Stein 1961). This idea was further developed by Ledoit and Wolf (2004) and Haff (1980). More recent approaches use mathematical tools from theoretical physics to predict (and correct for) the overfitting bias in ML regression analytically (Coolen et al. 2017).

Any sensible generative model for classifying vectors in $\mathbb{R}^d$ will have at least $\mathcal{O}(d)$ parameters. The fundamental cause of overfitting is the fact that in high-dimensional spaces, where d/n is finite even if n is large, the posterior parameter distribution $p(\boldsymbol{\theta}|\mathcal{D})$ (in a Bayesian sense) will be extremely sparse. Replacing this posterior by a delta-peak, which is what point estimation implies, is always a very poor approximation, irrespective of which protocol is used for estimating the location of this peak. It follows that by avoiding point estimation altogether, i.e. by retaining the full posterior distribution and doing all integrations over model parameters *analytically*, one should reduce overfitting effects, potentially allowing for high-dimensional data-sets to be classified reliably. Moreover, only hyperparameters will then have to be estimated (whose dimensionality is normally small, and independent of d), so one avoids the prohibitive computational demands of sampling high-dimensional spaces. The need to do all parameter integrals analytically limits us in practice to parametric generative models with class-specific multivariate Gaussian distributions. Here, the model

[1] This is the case for rare diseases, or when obtaining tissue material is nontrivial or expensive, but measuring extensive numbers of features in such material (e.g. gene expression data) is relatively simple and cheap.

[2] While $\varrho(\lambda)$ is not a good estimator for $\varrho_0(\lambda)$, Jonsson (1982) showed that in contrast $\int \mathrm{d}\lambda\varrho(\lambda)\lambda$ is a good estimate of $\int \mathrm{d}\lambda\varrho_0(\lambda)\lambda$; the bulk spectrum becomes more biased as d/n increases, but the sample eigenvalue *average* does not.

parameters to be integrated over are the class means in $\mathbb{R}^d$ and class-specific $d \times d$ covariance matrices, and with carefully chosen priors, one can indeed obtain analytical results. The Wishart distribution is the canonical prior for the covariance matrices. Analytically tractable choices for the class means are the conjugate (Keehn 1965; Geisser 1964) or the non-informative priors (Brown et al. 1999; Srivastava and Gupta 2006).

As the data dimensionality increases, so does the role of Bayesian priors and their associated hyperparameters, and the method used for computing hyperparameters impacts more on the performance of otherwise identical models. The most commonly used route for hyperparameter estimation appears to be cross-validation. This requires re-training one's model k times for k-fold cross-validation; for leave-one-out cross-validation, the model will need to be re-trained n times.

In this paper, we generalize the family of prior distributions for parametric generative models with class-specific multivariate Gaussian distributions, without loss of analytical tractability, and we compute hyperparameters via evidence maximization, rather than cross-validation. This allows us to derive closed-form expressions for the predictive probabilities of two special model instances. The numerical complexity of our approach does not increase significantly with d since all integrals whose dimensions scale with d are solved analytically.

In Section 2, we first define our generative Bayesian classifier and derive the relevant integrals. Special analytically solvable cases of these integrals, leading to two models (A and B), are described in Section 3 along with the evidence maximization estimation of hyperparameters. Closed-form expressions for the predictive probabilities corresponding to these two models are obtained in Section 3.4. We then examine the behaviour of the hyperparameters in Section 4.1 and carry out comparative classification performance tests on synthetic and real data-sets in Section 5. We conclude our paper with a discussion of the main results.

2 Definitions

2.1 Model and Objectives

We have data $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ consisting of n samples of pairs $(\boldsymbol{x}, y)$, where $\boldsymbol{x} \in \mathbb{R}^d$ is a vector of covariates, and $y \in \{1, \ldots, C\}$ a discrete outcome label. We seek to predict the outcome y_0 associated with a *new* covariate vector $\boldsymbol{x}_0$, given the data $\mathcal{D}$. So we want to compute

$$p(y_0|\boldsymbol{x}_0, \mathcal{D}) = \frac{p(y_0, \boldsymbol{x}_0|\mathcal{D})}{\sum_{y=1}^{C} p(y, \boldsymbol{x}_0|\mathcal{D})} = \frac{p(y_0, \boldsymbol{x}_0|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; y_1, \ldots, y_n)}{\sum_{y=1}^{C} p(y, \boldsymbol{x}_0|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; y_1, \ldots, y_n)}$$

$$= \frac{p(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n; y_0, \ldots, y_n)}{\sum_{y=1}^{C} p(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n; y, y_1, \ldots, y_n)} \qquad (1)$$

We next need an expression for the joint distribution $p(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n; y_0, \ldots, y_n)$. We assume that all pairs $(\boldsymbol{x}_i, y_i)$ are drawn independently from a parametrized distribution $p(\boldsymbol{x}, y|\boldsymbol{\theta})$ whose parameters $\boldsymbol{\theta}$ we do not know. Using de Finetti's representation theorem and the fact that exchangeability is a weaker condition than i.i.d, we can write the joint distribution of $\{(\boldsymbol{x}_i, y_i)\}_{i=0}^{n}$ as

$$p(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n; y_0, \ldots, y_n) = \int \mathrm{d}\boldsymbol{\theta}\, p(\boldsymbol{\theta}) \prod_{i=0}^{n} p(\boldsymbol{x}_i, y_i|\boldsymbol{\theta}) \qquad (2)$$

It now follows that

$$p(y_0|\boldsymbol{x}_0,\mathcal{D}) \;=\; \frac{\int \mathrm{d}\boldsymbol{\theta}\, p(\boldsymbol{\theta}) \prod_{i=0}^{n} p(\boldsymbol{x}_i, y_i|\boldsymbol{\theta})}{\sum_{y=1}^{C} \int \mathrm{d}\boldsymbol{\theta}\, p(\boldsymbol{\theta}) p(\boldsymbol{x}_0, y|\boldsymbol{\theta}) \prod_{i=1}^{n} p(\boldsymbol{x}_i, y_i|\boldsymbol{\theta})} \tag{3}$$

We regard all model parameters with dimensionality that scales with the covariate dimension d as *micro-parameters*, over which we need to integrate (in the sense of $\boldsymbol{\theta}$ above). Parameters with d-independent dimensionality are regarded as *hyperparameters*. The hyperparameter values will be called a 'model' H. Our equations will now acquire a label H:

$$p(y_0|\boldsymbol{x}_0,\mathcal{D},H) \;=\; \frac{\int \mathrm{d}\boldsymbol{\theta}\, p(\boldsymbol{\theta}|H) \prod_{i=0}^{n} p(\boldsymbol{x}_i, y_i|\boldsymbol{\theta},H)}{\sum_{y=1}^{C} \int \mathrm{d}\boldsymbol{\theta}\, p(\boldsymbol{\theta}|H) p(\boldsymbol{x}_0, y|\boldsymbol{\theta},H) \prod_{i=1}^{n} p(\boldsymbol{x}_i, y_i|\boldsymbol{\theta},H)} \tag{4}$$

The Bayes-optimal hyperparameters H are those that maximize the evidence, i.e.

$$\begin{aligned}\hat{H} \;&=\; \mathrm{argmax}_H\, p(H|\mathcal{D}) = \mathrm{argmax}_H \log\left\{\frac{p(\mathcal{D}|H)p(H)}{\sum_{H'} p(\mathcal{D}|H')p(H')}\right\}\\ &=\; \mathrm{argmax}_H \left\{\log \int \mathrm{d}\boldsymbol{\theta}\, p(\boldsymbol{\theta}|H) \prod_{i=1}^{n} p(\boldsymbol{x}_i, y_i|\boldsymbol{\theta},H) + \log p(H)\right\}\end{aligned} \tag{5}$$

What is left is to specify the parametrization $p(\boldsymbol{x}, y|\boldsymbol{\theta})$ of the joint statistics of covariates $\boldsymbol{x}$ and y in the population from which our samples are drawn. This choice is constrained by our desire to do all integrations over $\boldsymbol{\theta}$ analytically, to avoid approximations and overfitting problems caused by point estimation. One is then naturally led to class-specific Gaussian covariate distributions:

$$p(\boldsymbol{x}, y|\boldsymbol{\theta}) = p(y)p(\boldsymbol{x}|y,\boldsymbol{\theta}),\; p(\boldsymbol{x}|y,\boldsymbol{\theta}) = \frac{\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_y)^T \boldsymbol{\Lambda}_y (\boldsymbol{x}-\boldsymbol{\mu}_y)}}{\sqrt{(2\pi)^d/\mathrm{Det}\boldsymbol{\Lambda}_y}} \tag{6}$$

Thus, the parameters to be integrated over are $\boldsymbol{\theta} = \{\boldsymbol{\mu}_y, \boldsymbol{\Lambda}_y, y = 1, \ldots, C\}$, i.e. the class-specific means and precision matrices.

2.2 Integrals to be Computed

In both the inference formula $p(y_0|\boldsymbol{x}_0,\mathcal{D},H)$ (4) and in the expression for $\hat{H}$ (5), the relevant integral we need to do analytically is the one in

$$\Omega(H,n,\mathcal{D}) \;=\; -\log \int \mathrm{d}\boldsymbol{\theta}\, p(\boldsymbol{\theta}|H) \prod_{i=1}^{n} p(\boldsymbol{x}_i, y_i|\boldsymbol{\theta},H) \tag{7}$$

In the case where we require $\Omega(H,n+1,\mathcal{D})$, when evaluating the numerator and the denominator of Eq. 4, we simply replace $\prod_{i=1}^{n}$ by $\prod_{i=0}^{n}$, so that

$$p(y_0|\boldsymbol{x}_0,\mathcal{D}) = \frac{\mathrm{e}^{-\Omega(H,n+1,\mathcal{D})}}{\sum_{z=1}^{C} \mathrm{e}^{-\Omega(H,n+1,\mathcal{D})}|_{y_0=z}} \qquad \hat{H} = \mathrm{argmin}_H \Omega(H,n,\mathcal{D}) \tag{8}$$

Working out $\Omega(H, n, \mathcal{D})$ for the parametrization (6) gives:

$$\Omega(H,n,\mathcal{D}) = \frac{1}{2}nd\log(2\pi) - \sum_{i=1}^{n}\log p_{y_i} \\ -\log\int\left[\prod_{z=1}^{C}\mathrm{d}\boldsymbol{\mu}_z\mathrm{d}\boldsymbol{\Lambda}_z p_z(\boldsymbol{\mu}_z,\boldsymbol{\Lambda}_z)\right]\left[\prod_{i=1}^{n}(\mathrm{Det}\boldsymbol{\Lambda}_{y_i})^{\frac{1}{2}}\right]\mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu}_{y_i})^T\boldsymbol{\Lambda}_{y_i}(\boldsymbol{x}_i-\boldsymbol{\mu}_{y_i})} \tag{9}$$

where $p(y_i) = p_{y_i}$ is the prior probability of a sample belonging to class y_i. For generative models, this is typically equal to the empirical proportion of samples in that specific class. To simplify this expression, we define the data-dependent index sets $I_z = \{i|y_i = z\}$, each of size $n_z = |I_z| = \sum_{i=1}^{n}\delta_{z,y_i}$. We also introduce empirical covariate averages and correlations, with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$:

$$\hat{X}^z_\mu = \frac{1}{n_z}\sum_{i\in I_z}x_{i\mu}, \qquad \hat{C}^z_{\mu\nu} = \frac{1}{n_z}\sum_{i\in I_z}(x_{i\mu}-\hat{X}^z_\mu)(x_{i\nu}-\hat{X}^z_\nu) \tag{10}$$

Upon defining the vector $\hat{\boldsymbol{X}}_z = (\hat{X}^z_1, \ldots, \hat{X}^z_d)$, and the $d \times d$ matrix $\hat{\boldsymbol{C}}_z = \{\hat{C}^z_{\mu\nu}\}$, we can then write the relevant integrals after some simple rearrangements in the form

$$\begin{aligned}\Omega(H,n,\mathcal{D}) &= \frac{1}{2}nd\log(2\pi) - \sum_{z=1}^{C}n_z\log p_z \\ &\quad -\log\int\left[\prod_{z=1}^{C}\mathrm{d}\boldsymbol{\mu}_z\mathrm{d}\boldsymbol{\Lambda}_z p_z(\boldsymbol{\mu}_z,\boldsymbol{\Lambda}_z)(\mathrm{Det}\boldsymbol{\Lambda}_z)^{\frac{n_z}{2}}\mathrm{e}^{-\frac{1}{2}n_z\boldsymbol{\mu}_z^T\boldsymbol{\Lambda}_z\boldsymbol{\mu}_z}\right] \\ &\quad \times\mathrm{e}^{\sum_{z=1}^{C}\boldsymbol{\mu}_z\cdot\boldsymbol{\Lambda}_z\sum_{i\in I_z}\boldsymbol{x}_i-\frac{1}{2}\sum_{z=1}^{C}\sum_{i\in I_z}\boldsymbol{x}_i^T\boldsymbol{\Lambda}_z\boldsymbol{x}_i} \\ &= \frac{1}{2}nd\log(2\pi) - \sum_{z=1}^{C}n_z\log p_z \\ &\quad -\sum_{z=1}^{C}\log\int\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\Lambda}\,p_z(\boldsymbol{\mu}+\hat{\boldsymbol{X}}_z,\boldsymbol{\Lambda})(\mathrm{Det}\boldsymbol{\Lambda})^{\frac{1}{2}n_z}\mathrm{e}^{-\frac{1}{2}n_z\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu}-\frac{1}{2}n_z\mathrm{Tr}(\hat{\boldsymbol{C}}_z\boldsymbol{\Lambda})}\end{aligned} \tag{11}$$

To proceed, it is essential that we compute $\Omega(H, n, \mathcal{D})$ analytically, for arbitrary $\hat{\boldsymbol{X}} \in \mathbb{R}^d$ and arbitrary positive definite symmetric matrices $\hat{\boldsymbol{C}}$. This will constrain the choice of our priors $p_z(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ for the covariate averages and correlations in outcome class z. All required integrals are of the following form, with $\boldsymbol{\Lambda}$ limited to the subset Ξ^d of symmetric positive definite matrices:

$$\Psi_z(H,n,\mathcal{D}) = \int_{\mathbb{R}^d}\mathrm{d}\boldsymbol{\mu}\int_{\Xi^d}\mathrm{d}\boldsymbol{\Lambda}\,p_z(\boldsymbol{\mu}+\hat{\boldsymbol{X}}_z|\boldsymbol{\Lambda})p_z(\boldsymbol{\Lambda})(\mathrm{Det}\boldsymbol{\Lambda})^{\frac{1}{2}n_z}\mathrm{e}^{-\frac{1}{2}n_z\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu}-\frac{1}{2}n_z\mathrm{Tr}(\hat{\boldsymbol{C}}_z\boldsymbol{\Lambda})} \tag{12}$$

We will drop the indications of the sets over which the integrals are done, when these are clear from the context. The tricky integral is that over the inverse covariance matrices $\boldsymbol{\Lambda}$. The choice in Shalabi et al. (2016) corresponded to $p_z(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto \mathrm{e}^{-\frac{1}{2}\boldsymbol{\mu}^2/\beta_z^2}\delta[\boldsymbol{\Lambda} - \mathbf{I}/\alpha_z^2]$, which implied assuming uncorrelated covariates within each class. Here we want to allow for arbitrary class-specific covariate correlations.

2.3 Priors for Class-Specific Means and Covariance Matrices

The integrals over $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ can be done in either order. We start with the integral over $\boldsymbol{\mu}$. In contrast to most studies, we replace the conjugate prior for the unknown mean vector by a multivariate Gaussian with an as yet arbitrary precision matrix $\boldsymbol{A}$. This should allow us to cover a larger parameter space than the conjugate prior (which has $\boldsymbol{\Lambda_z}^{-1}$ as its covariance matrix):

$$p_z(\boldsymbol{\mu}|\boldsymbol{A}) = (2\pi)^{-\frac{d}{2}}\sqrt{\mathrm{Det}\boldsymbol{A}_z}\mathrm{e}^{-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{A}_z\boldsymbol{\mu}} \tag{13}$$

Insertion into Eq. 12 gives

$$\begin{aligned}\Psi_z &= (2\pi)^{-\frac{d}{2}}\int \mathrm{d}\boldsymbol{\Lambda}\, p_z(\boldsymbol{\Lambda})\mathrm{e}^{-\frac{1}{2}n_z\mathrm{Tr}(\hat{\boldsymbol{C}}_z\boldsymbol{\Lambda})-\frac{1}{2}\hat{\boldsymbol{X}}_z^T\boldsymbol{A}_z\hat{\boldsymbol{X}}_z}\left[\mathrm{Det}(\boldsymbol{\Lambda}^{n_z})\mathrm{Det}\boldsymbol{A}_z\right]^{\frac{1}{2}}\\ &\quad\times\int \mathrm{d}\boldsymbol{\mu}\mathrm{e}^{-\frac{1}{2}\boldsymbol{\mu}^T(n_z\boldsymbol{\Lambda}+\boldsymbol{A}_z)\boldsymbol{\mu}-\boldsymbol{\mu}^T\boldsymbol{A}_z\hat{\boldsymbol{X}}_z}\\ &= \int \mathrm{d}\boldsymbol{\Lambda}\, p_z(\boldsymbol{\Lambda})\mathrm{e}^{-\frac{1}{2}n_z\mathrm{Tr}(\hat{\boldsymbol{C}}_z\boldsymbol{\Lambda})}\left[\frac{\mathrm{Det}(\boldsymbol{\Lambda}^{n_z})\mathrm{Det}\boldsymbol{A}_z}{\mathrm{Det}(n_z\boldsymbol{\Lambda}+\boldsymbol{A}_z)}\right]^{\frac{1}{2}}\mathrm{e}^{\frac{1}{2}\boldsymbol{X}^T\boldsymbol{A}_z(n_z\boldsymbol{\Lambda}+\boldsymbol{A}_z)^{-1}\boldsymbol{A}_z\hat{\boldsymbol{X}}_z-\frac{1}{2}\hat{\boldsymbol{X}}_z^T\boldsymbol{A}_z\hat{\boldsymbol{X}}_z}\\ &= \int \mathrm{d}\boldsymbol{\Lambda}\, p_z(\boldsymbol{\Lambda})\mathrm{e}^{-\frac{1}{2}n_z\mathrm{Tr}(\hat{\boldsymbol{C}}_z\boldsymbol{\Lambda})}\left[\mathrm{Det}(n_z\boldsymbol{\Lambda}^{1-n_z}\boldsymbol{A}_z^{-1}+\boldsymbol{\Lambda}^{-n_z})\right]^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}\hat{\boldsymbol{X}}^T[(n_z\boldsymbol{\Lambda})^{-1}+(\boldsymbol{A}_z)^{-1}]^{-1}\hat{\boldsymbol{X}}_z}\end{aligned} \tag{14}$$

Our present more general assumptions lead to calculations that differ from the earlier work of, e.g. Keehn (1965), Brown et al. (1999), and Srivastava et al. (2007). Alternative analytically tractable priors are the transformation-invariant Jeffrey's or Reference priors, which are derived from information-theoretic arguments (Berger et al. 1992). There, the calculation of the predictive probability is simpler, but the sample covariance matrix is not regularized. This causes problems when $n < d$, where the sample covariance matrices would become singular and the predictive probability would cease to be well defined. Our next question is for which choice(s) of $\boldsymbol{A}_z$ we can do also the integrals over $\boldsymbol{\Lambda}$ in Eq. 14 analytically. Expression (14), in line with Keehn (1965), Brown et al. (1999), and Srivastava et al. (2007), suggests using for the measure $p_z(\boldsymbol{\Lambda})$ over all positive definite matrices $\boldsymbol{\Lambda} \in \Xi^d$ a Wishart distribution, which is of the form

$$p(\boldsymbol{\Lambda}) = \frac{(\mathrm{Det}\boldsymbol{\Lambda})^{(r-d-1)/2}}{2^{rd/2}\Gamma_d(\frac{r}{2})(\mathrm{Det}\boldsymbol{S})^{r/2}}\mathrm{e}^{-\frac{1}{2}\mathrm{Tr}(\boldsymbol{S}^{-1}\boldsymbol{\Lambda})} \tag{15}$$

Here, $r > d-1$, $\boldsymbol{S}$ is a positive definite and symmetric $d \times d$ matrix, and $\Gamma_p(x)$ is the multivariate gamma function which is expressed in terms of the ordinary gamma function via:

$$\Gamma_p\left(\frac{r}{2}\right) = \pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma\left(\frac{r}{2}-\frac{j-1}{2}\right) \tag{16}$$

The choice Eq. 15 is motivated solely by analytic tractability. However, since the prior domain is the space of all positive definite matrices, we are assured that upon using Eq. 15, our posterior will be consistent. Distribution (15) implies stating that $\boldsymbol{\Lambda}$ is the empirical precision matrix of a set of r i.i.d. random zero-average Gaussian vectors, with covariance matrix $\boldsymbol{S}$. Since Eq. 15 is normalized, for any $\boldsymbol{S}$, we can use it to do all integrals of the following form analytically:

$$\int_{\Xi^d} \mathrm{d}\boldsymbol{\Lambda}\,(\mathrm{Det}\boldsymbol{\Lambda})^{(r-d-1)/2}\mathrm{e}^{-\frac{1}{2}\mathrm{Tr}(\boldsymbol{S}^{-1}\boldsymbol{\Lambda})} \;=\; 2^{rd/2}\Gamma_d\left(\frac{r}{2}\right)(\mathrm{Det}\boldsymbol{S})^{r/2} \tag{17}$$

In order for Eq. 14 to acquire the form (17), we need a choice for $\boldsymbol{A}_z$ such that the following holds, for some $\gamma_0, \gamma_1 \in \mathbb{R}$: $[(n_z\boldsymbol{\Lambda})^{-1}+(\boldsymbol{A}_z)^{-1}]^{-1} = \gamma_{1z}\boldsymbol{\Lambda}+\gamma_{0z}\mathbf{I}$. Rewriting this condition gives:

$$\boldsymbol{A}_z(\gamma_{0z}, \gamma_1) \;=\; [(\gamma_{1z}\boldsymbol{\Lambda} + \gamma_{0z}\mathbf{I})^{-1} - (n_z\boldsymbol{\Lambda})^{-1}]^{-1} \tag{18}$$

Conditions to ensure $\boldsymbol{A}_z$ is positive definite are considered in the supplementary material. Upon making the choice (18) and using Eq. 17, we obtain for the integral (14):

$$\Psi_z \;=\; \mathrm{e}^{-\frac{1}{2}\gamma_{0z}\hat{\boldsymbol{X}}_z^2}\int \mathrm{d}\boldsymbol{\Lambda}\,p_z(\boldsymbol{\Lambda})\frac{\mathrm{e}^{-\frac{1}{2}n_z\mathrm{Tr}(\hat{\boldsymbol{C}}_z\boldsymbol{\Lambda})-\frac{1}{2}\gamma_{1z}\hat{\boldsymbol{X}}_z^T\boldsymbol{\Lambda}\hat{\boldsymbol{X}}_z}}{\sqrt{\mathrm{Det}[n_z\boldsymbol{\Lambda}^{1-n_z}(\gamma_{1z}\boldsymbol{\Lambda}+\gamma_{0z}\mathbf{I})^{-1}]}} \tag{19}$$

We conclude that we can evaluate (19) analytically, using Eq. 17, provided we choose for $p_z(\boldsymbol{\Lambda})$ the Wishart measure, and with either $\gamma_{0z} \to 0$ and $\gamma_{1z} \in (0, n_z)$ or with $\gamma_{1z} \to 0$ and $\gamma_{0z} \in (0, n_z\lambda_{\min})$. Alternative choices for $(\gamma_{0z}, \gamma_{1z})$ would lead to more complicated integrals than the Wishart one.

The two remaining analytically integrable candidate model branches imply the following choices for the inverse correlation matrix $\boldsymbol{A}_z$ of the prior $p_z(\mu|\boldsymbol{A}_z)$ for the class centres:

$$\gamma_{0z}=0: \boldsymbol{A}_z = \frac{n_z\gamma_{1z}}{n_z-\gamma_{1z}}\boldsymbol{\Lambda}, \qquad \gamma_{1z}=0: \boldsymbol{A}_z = \left[\gamma_{0z}^{-1}\mathbf{I}-(n_z\boldsymbol{\Lambda})^{-1}\right]^{-1} \tag{20}$$

Note that the case $\boldsymbol{A}_z \to 0$, a non-informative prior for class means as in Brown et al. (1999), corresponds to $(\gamma_{0z}, \gamma_{1z}) = (0, 0)$. However, the two limits $\gamma_{0z} \to 0$ and $\gamma_{1z} \to 0$ will generally not commute, which can be inferred from working out (19) for the two special cases $\gamma_{0z} = 0$ and $\gamma_{1z} = 0$:

$$\gamma_{0z}=0: \quad \Psi_z = \left(\frac{\gamma_{1z}}{n_z}\right)^{\frac{d}{2}}\int \mathrm{d}\boldsymbol{\Lambda}\,p_z(\boldsymbol{\Lambda})[\mathrm{Det}(\boldsymbol{\Lambda})]^{\frac{n_z}{2}}\mathrm{e}^{-\frac{1}{2}n_z\mathrm{Tr}(\hat{\boldsymbol{C}}_z\boldsymbol{\Lambda})-\frac{1}{2}\gamma_{1z}\hat{\boldsymbol{X}}_z^T\boldsymbol{\Lambda}\hat{\boldsymbol{X}}_z} \tag{21}$$

$$\gamma_{1z}=0: \quad \Psi_z = \left(\frac{\gamma_{0z}}{n_z}\right)^{\frac{d}{2}}\mathrm{e}^{-\frac{1}{2}\gamma_{0z}\hat{\boldsymbol{X}}_z^2}\int \mathrm{d}\boldsymbol{\Lambda}\,p_z(\boldsymbol{\Lambda})[\mathrm{Det}(\boldsymbol{\Lambda})]^{\frac{n_z-1}{2}}\mathrm{e}^{-\frac{1}{2}n_z\mathrm{Tr}(\hat{\boldsymbol{C}}_z\boldsymbol{\Lambda})} \tag{22}$$

This non-uniqueness of the limit $\boldsymbol{A}_z \to 0$ is a consequence of having done the integral over $\boldsymbol{\Lambda}$ first.

3 The Integrable Model Branches

3.1 The Case $\gamma_{0z} = 0$: Model A

We now choose $\gamma_{0z} = 0$, and substitute for each $z = 1\ldots C$ the Wishart distribution Eq. 15 into Eq. 19, with seed matrix $\boldsymbol{S} = k_z\mathbf{I}$. This choice is named Quadratic Bayes in Brown et al. (1999). We also define the $p \times p$ matrix $\hat{\boldsymbol{M}}_z$ with entries $\hat{M}^z_{\mu\nu} = X^z_\mu X^z_\nu$. The result of working out (19) is, using Eq. 17:

$$\Psi_z \;=\; \left(\frac{2^{n_z}\gamma_{1z}}{n_z k_z^{r_z}}\right)^{\frac{d}{2}}\frac{\Gamma_d\left(\frac{r_z+n_z}{2}\right)}{\Gamma_d\left(\frac{r_z}{2}\right)}[\mathrm{Det}(n_z\hat{\boldsymbol{C}}_z+\gamma_{1z}\hat{\boldsymbol{M}}_z+k_z^{-1}\mathbf{I})]^{-(r_z+n_z)/2} \tag{23}$$

This, in turn, allows us to evaluate (11):

$$\begin{aligned}\Omega(H,n,\mathcal{D}) &= \frac{1}{2}nd\log(\pi) - \sum_{z=1}^{C} n_z \log p_z - \frac{1}{2}d\sum_{z=1}^{C}\left[\log(\gamma_{1z}/n_z) - r_z\log k_z\right] \\ &\quad - \sum_{z=1}^{C}\log\left[\frac{\Gamma_d\left(\frac{r_z+n_z}{2}\right)}{\Gamma_d\left(\frac{r_z}{2}\right)}\right] \\ &\quad + \frac{1}{2}\sum_{z=1}^{C}(r_z+n_z)\log\mathrm{Det}(n_z\hat{\boldsymbol{C}}_z + \gamma_{1z}\hat{\boldsymbol{M}}_z + k_z^{-1}\mathbf{I})\end{aligned} \tag{24}$$

The hyperparameters of our problem are $\{p_z, \gamma_{1z}, r_z, k_z\}$, for $z = 1\ldots C$. If we choose flat hyper-priors, to close the Bayesian inference hierarchy, their optimal values are obtained by minimizing (24), subject to the constraints $\sum_{z=1}^{C} p_z = 1$, $p_z \geq 0$, $r_z \geq d$, $\gamma_{1z} \in [0, n_z]$, and $k_z > 0$. We now work out the relevant extremization equations, using the general identity $\boldsymbol{\partial_x \log \mathrm{Det}\, Q = \mathrm{Tr}(Q^{-1}\partial_x Q)}$:

- Minimization over p_z: $p_z = n_z/n$.
- Minimization over k_z:

$$k_z = 0 \ \text{ or } \ r_z = n_z\left[\frac{dk_z}{\mathrm{Tr}[(n_z\hat{\boldsymbol{C}}_z + \gamma_{1z}\hat{\boldsymbol{M}}_z + k_z^{-1}\mathbf{I})^{-1}} - 1\right]^{-1} \tag{25}$$

- Minimization over r_z, using the digamma function $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log\Gamma(x)$:

$$\begin{aligned} r_z = d \ \text{ or } \ \log k_z = \frac{1}{d}\sum_{j=1}^{d}\Bigg[\psi\left(\frac{r_z+n_z-j+1}{2}\right) - \psi\left(\frac{r_z-j+1}{2}\right)\Bigg] \\ -\frac{1}{d}\log\mathrm{Det}(n_z\hat{\boldsymbol{C}}_z + \gamma_{1z}\hat{\boldsymbol{M}}_z + k_z^{-1}\mathbf{I})\Bigg]\end{aligned} \tag{26}$$

- Minimization over γ_{1z}:

$$\gamma_{1z} \in \{0, n_z\} \ \text{ or } \ \gamma_{1z} = \frac{1}{r_z+n_z}\left[\frac{1}{d}\mathrm{Tr}[(n_z\hat{\boldsymbol{C}}_z + \gamma_{1z}\hat{\boldsymbol{M}}_z + k_z^{-1}\mathbf{I})^{-1}\hat{\boldsymbol{M}}_z]\right]^{-1} \tag{27}$$

In addition, we still need to satisfy the inequalities $r_z \geq d$, $\gamma_{1z} \in [0, n_z]$, and $k_z > 0$.

We observe in the above results that, unless we choose $\gamma_{1z} \in \{0, n_z\}$, i.e. $\boldsymbol{A} = 0$ or $\boldsymbol{A}^{-1} = 0$, we would during any iterative algorithmic solution of our order parameter equations have to diagonalize a $d \times d$ matrix at each iteration step. This would be prohibitively slow, even with the most efficient numerical diagonalization methods. Since $\gamma_{1z} = n_z$ implies that the prior $p_z(\boldsymbol{\mu}|\boldsymbol{A})$ forces all class centres to be in the origin, we will be left for the current model branch only with the option $\gamma_{1z} \to 0$, corresponding to a flat prior for the class centres. We thereby arrive at the Quadratic Bayes model of Brown et al. (1999), with hyperparameter formulae based on evidence maximization.

3.2 The Case $\gamma_{1z} = 0$: Model B

We next inspect the alternative model branch by choosing $\gamma_{1z} = 0$, again substituting for each $z = 1\ldots C$ the Wishart distribution (15) into Eq. 19 with seed matrix $\boldsymbol{S} = k_z\mathbf{I}$. The result is:

$$\Psi_z = \left(\frac{2^{n_z-1}\gamma_{0z}}{n_z k_z^{r_z}}\right)^{\frac{d}{2}} \frac{\Gamma_d\left(\frac{r_z+n_z-1}{2}\right)}{\Gamma_d(\frac{r_z}{2})} [\text{Det}(n_z\hat{\boldsymbol{C}}_z + k_z^{-1}\mathbf{I})]^{-(r_z+n_z-1)/2} \text{e}^{-\frac{1}{2}\gamma_{0z}\hat{\boldsymbol{X}}_z^2} \tag{28}$$

For the quantity (11), we thereby find:

$$\begin{aligned}\Omega(H, n, \mathcal{D}) = {} & \frac{1}{2}nd\log(\pi) + \frac{1}{2}dC\log 2 - \sum_{z=1}^{C} n_z \log p_z - \frac{1}{2}d\sum_{z=1}^{C}\left[\log\left(\frac{\gamma_{0z}}{n_z}\right) - r_z\log k_z\right] \\ & - \sum_{z=1}^{C}\log\left[\frac{\Gamma_d\left(\frac{r_z+n_z-1}{2}\right)}{\Gamma_d\left(\frac{r_z}{2}\right)}\right] + \frac{1}{2}\sum_{z=1}^{C}\gamma_{0z}\hat{\boldsymbol{X}}_z^2 \\ & + \frac{1}{2}\sum_{z=1}^{C}(r_z+n_z-1)\log\text{Det}(n_z\hat{\boldsymbol{C}}_z + k_z^{-1}\mathbf{I})\end{aligned} \tag{29}$$

If as before we choose flat hyper-priors, the Bayes-optimal hyperparameters $\{p_z, \gamma_{1z}, r_z, k_z\}$, for $z = 1\ldots C$ are found by maximizing the evidence (29), subject to the constraints $\sum_{z=1}^{C} p_z = 1$, $p_z \geq 0$, $r_z \geq d$, $\gamma_{0z} \geq 0$, and $k_z > 0$. For the present model branch B, differentiation gives:

- Minimization over p_z: $p_z = n_z/n$.
- Minimization over k_z:

$$k_z = 0 \text{ or } r_z = (n_z - 1)\left[\frac{dk_z}{\text{Tr}[(n_z\hat{\boldsymbol{C}}_z + k_z^{-1}\mathbf{I})^{-1}} - 1\right]^{-1} \tag{30}$$

- Minimization over r_z:

$$r_z = d \text{ or } \log k_z = \frac{1}{d}\sum_{j=1}^{d}\left[\psi\left(\frac{r_z+n_z-j}{2}\right) - \psi\left(\frac{r_z-j+1}{2}\right)\right] - \frac{1}{d}\log\text{Det}(n_z\hat{\boldsymbol{C}}_z + k_z^{-1}\mathbf{I}) \tag{31}$$

- Minimization over γ_{0z}: $\gamma_{0z} = d/\hat{\boldsymbol{X}}_z^2$.

In addition, we still need to satisfy the inequalities $r_z \geq d$ and $k_z > 0$. In contrast to the first integrable model branch A, here we are able to optimize over γ_{0z} without problems, and the resulting model B is distinct from the Quadratic Bayes classifier of Brown et al. (1999).

3.3 Comparison of the Two Integrable Model Branches

Our initial family of models was parametrized by $(\gamma_{0z}, \gamma_{1z})$. We then found that the following two branches are analytically integrable, using Wishart priors for class-specific precision matrices:

$$\text{A}: \quad (\gamma_{0z}, \gamma_{1z}) = (0, \hat{\gamma}_{1z}) \text{ with } \hat{\gamma}_{1z} \to 0 \tag{32}$$

$$\text{B}: \quad (\gamma_{0z}, \gamma_{1z}) = (\hat{\gamma}_{0z}, 0) \text{ with } \hat{\gamma}_{0z} \to d/\hat{\boldsymbol{X}}_z^2 \tag{33}$$

Where conventional methods tend to determine hyperparameters via cross-validation, which is computationally expensive, here we optimize hyperparameters via evidence maximization. As expected, both models give $p_z = n_z/n$. The hyperparameters (k_z, r_z) are to be solved from the following equations, in which $\varrho_z(\xi)$ denotes the eigenvalue distribution of $\hat{\boldsymbol{C}}_z$:

$$\text{A}: \quad k_z = 0 \text{ or } r_z = n_z \left[\frac{1}{\int \mathrm{d}\xi \, \varrho_z(\xi)(n_z k_z \xi + 1)^{-1}} - 1 \right]^{-1} \tag{34}$$

$$r_z = d \text{ or } \frac{1}{d} \sum_{j=1}^{d} \left[\psi\left(\frac{r_z + n_z - j + 1}{2} \right) - \psi\left(\frac{r_z - j + 1}{2} \right) \right] = \int \mathrm{d}\xi \, \varrho_z(\xi) \log(n_z k_z \xi + 1) \tag{35}$$

$$\text{B}: \quad k_z = 0 \text{ or } r_z = (n_z - 1) \left[\frac{1}{\int \mathrm{d}\xi \, \varrho_z(\xi)(n_z k_z \xi + 1)^{-1}} - 1 \right]^{-1} \tag{36}$$

$$r_z = d \text{ or } \frac{1}{d} \sum_{j=1}^{d} \left[\psi\left(\frac{r_z + n_z - j}{2} \right) - \psi\left(\frac{r_z - j + 1}{2} \right) \right] = \int \mathrm{d}\xi \, \varrho_z(\xi) \log(n_z k_z \xi + 1) \tag{37}$$

We see that the equations for (k_z, r_z) of models A and B differ only in having the replacement $n_z \to n_z - 1$ in certain places. Hence, we will have $\left(k_z^A, r_z^A\right) = \left(k_z^B, r_z^B\right) + \mathcal{O}(n_z^{-1})$.

3.4 Expressions for the Predictive Probability

Starting from Eq. 8, we derive an expression for the predictive probability for both models (see supplementary material). The predictive probability for model B:

$$p(y_0|\boldsymbol{x}_0, \mathcal{D}) = \frac{W_{y_0} \mathrm{e}^{-\frac{\gamma_{0y_0}}{2(n_{y_0}+1)}\left[2\hat{\boldsymbol{X}}_{y_0} \cdot (\boldsymbol{x}_0 - \hat{\boldsymbol{X}}_{y_0}) + \frac{1}{n_{y_0}+1}(\boldsymbol{x}_0 - \hat{\boldsymbol{X}}_{y_0})^2\right]} \left(1 + \frac{n_{y_0}}{n_{y_0}+1}(\boldsymbol{x}_0 - \hat{\boldsymbol{X}}_{y_0}) \cdot \boldsymbol{\Xi}_{y_0}^{-1}(\boldsymbol{x}_0 - \hat{\boldsymbol{X}}_{y_0})\right)^{-\frac{1}{2}(r_{y_0}+n_{y_0})}}{\sum_{z=1}^{C} W_z \mathrm{e}^{-\frac{\gamma_{0z}}{2(n_z+1)}\left[2\hat{\boldsymbol{X}}_z \cdot (\boldsymbol{x}_0 - \hat{\boldsymbol{X}}_z) + \frac{1}{n_z+1}(\boldsymbol{x}_0 - \hat{\boldsymbol{X}}_z)^2\right]} \left(1 + \frac{n_z}{n_z+1}(\boldsymbol{x}_0 - \hat{\boldsymbol{X}}_z) \cdot \boldsymbol{\Xi}_{y_0}^{-1}(\boldsymbol{x}_0 - \hat{\boldsymbol{X}}_z)\right)^{-\frac{1}{2}(r_z+n_z)}} \tag{38}$$

with $p_z = n_z/n$, and

$$W_z = p_z \left(\frac{n_z}{n_z + 1} \right)^{\frac{d}{2}} \frac{\Gamma\left(\frac{r_z + n_z}{2} \right)}{\Gamma\left(\frac{r_z + n_z - d}{2} \right)} [\mathrm{Det} \boldsymbol{\Xi}_z]^{-\frac{1}{2}}, \qquad \gamma_{0z} = d/\hat{\boldsymbol{X}}_z^2, \qquad \boldsymbol{\Xi}_z = n_z \hat{\boldsymbol{C}}_z + k_z^{-1} \mathbf{I} \tag{39}$$

Upon repeating the same calculations for model A, one finds that its predictive probability is obtained from expression (38) simply by setting γ_{0y_0} to zero (keeping in mind that for model A, we would also insert into this formula distinct values for the optimal hyperparameters k_z and r_z).

4 Phenomenology of the Classifiers

4.1 Hyperparameters: LOOCV Versus Evidence Maximization

The most commonly used measure for classification performance is the percentage of samples correctly predicted on unseen data (equivalently, the trace of the confusion matrix), and most Bayesian classification methods also use this measure as the optimization target for hyperparameters, via cross-validation. Instead, our method of hyperparameter optimization maximizes the evidence term in the Bayesian inference. In k-fold cross-validation, one needs to diagonalize for each outcome class a $d \times d$ matrix k times, whereas using the evidence maximization route, one needs to diagonalize such matrices only once, giving a factor k reduction in what for large d is the dominant contribution to the numerical demands. Moreover, cross-validation introduces fluctuations into the hyperparameter computation (via the random separations into training and validation sets), whereas evidence maximization is strictly deterministic.

The two routes, cross-validation versus evidence maximization, need not necessarily lead to coincident hyperparameter estimates. In order to investigate such possible differences, we generated synthetic data-sets with equal class sizes $n_1 = n_2 = 50$, and with input vectors of dimension $d = 50$. Using a 100×100 grid of values for the hyperparameters k_1 and k_2, with $k_z \in [0, k_{\mathrm{max},z}]$, we calculated the leave-one-out cross-validation (LOOCV) estimator of the classification accuracy for unseen cases, for a single data realization. The values of (r_1, r_2) were determined via evidence maximization, using formula (35) (i.e. following model branch A, with the non-informative prior for the class means). The value $k_{\mathrm{max},z}$ is either the upper limit defined by the condition $r_z > d - 1$ (if such a limit exists, dependent on the data realization), otherwise set numerically to a fixed large value. The location of the maximum of the resulting surface determines the LOOCV estimate of the optimal hyperparameters (k_1, k_2), which can be compared to the optimized hyperparameters (k_1, k_2) of the evidence maximization method.

Figure 1 shows the resulting surface for uncorrelated data, i.e. $\Sigma_1 = \Sigma_2 = \mathbb{I}_d$. The comparison points from our evidence-based optimal hyperparameters (k_1, k_2) are shown in Table 1. The small values for (k_1, k_2) imply that the model correctly infers that the

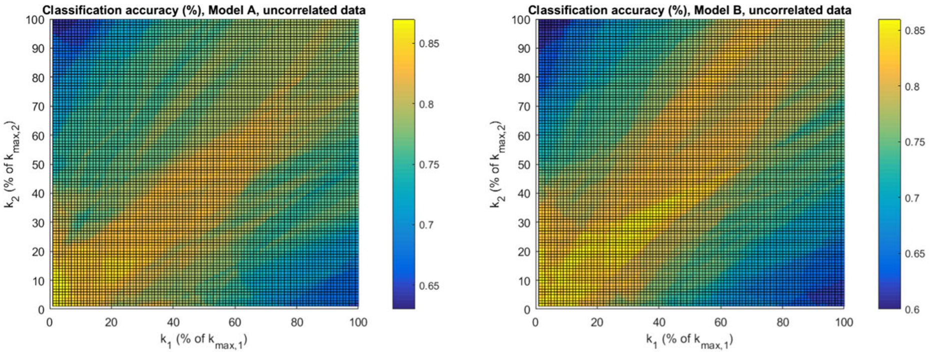


Fig. 1 LOOCV classification accuracy in (k_1, k_2) space for uncorrelated synthetic data, with class means $\boldsymbol{\mu}_1 = (0, 0, \ldots, 0)$ and $\boldsymbol{\mu}_2 = (2.5, 0, \ldots, 0)$, population covariance matrices $\Sigma_1 = \Sigma_2 = \mathbb{I}_d$, and covariate dimension $d = 50$. The hyperparameters (r_1, r_2) for models A and B were determined via Eq. 35. The results are based on a single data realization

Table 1 Comparison of hyperparameter estimation using cross-validation and evidence maximization for correlated and uncorrelated data

(k_1, k_2)	Method	Model A	Model B
Uncorrelated data	Cross-validation	(1–5%, 1–13%)	(1%, 1%)
	Evidence maximization	(3%, 2%)	(2%, 1%)
Correlated data	Cross-validation	(82–87%, 55–61%)	(96–100%, 54–92%)
	Evidence maximization	(94%, 95%)	(94%, 94%)

Entries are the values of (k_1, k_2), given as a percentage of each class k_{max}, corresponding to the maximum classification accuracy (within the granularity of our numerical experiments). A range of values is given when they all share the same classification accuracy

components of $\boldsymbol{x}$ in each class are most likely uncorrelated. The same protocol was subsequently repeated for correlated data, using a Toeplitz covariance matrix, the results of which are shown in Fig. 2 and Table 1. The larger values for (k_1, k_2) imply that here the model correctly infers that the components of $\boldsymbol{x}$ in each class are correlated. In both cases, the differences between optimal hyperparameter values defined via LOOCV as opposed to evidence maximization are seen to be minor (Table 2).

4.2 Overfitting

Next, we illustrate the degree of overfitting for models A and B, using examples of both correlated and uncorrelated synthetic data-sets. We sampled from the data described in Table 3, using case 1 (uncorrelated) and case 8 (correlated). In all cases, we chose $C = 3$ classes of $n_z = 13$ samples each, for a broad range of data dimensions d. See the caption of Table 3 for a full description of the statistical features of these synthetic data-sets. Measuring training and validation classification performance via LOOCV on these data resulted in Fig. 3, where each data-point is an average over 250 simulation experiments. The degree of divergence between the training and validation curves (solid versus dashed) is a direct measure

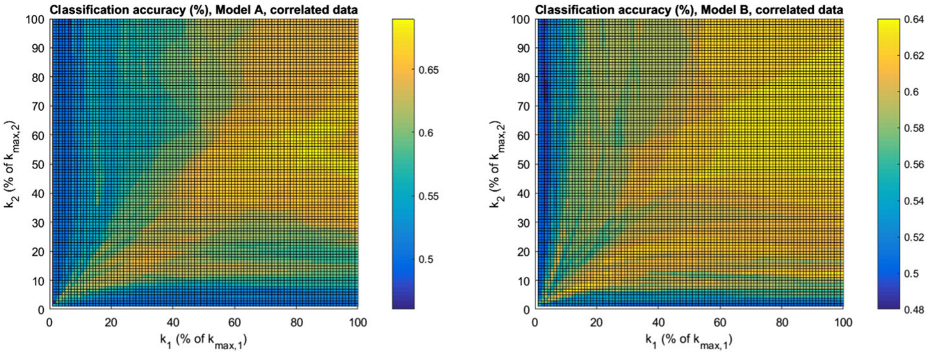


Fig. 2 LOOCV classification accuracy in (k_1, k_2) space for correlated synthetic data, with class means $\boldsymbol{\mu}_1 = (0, 0, \ldots, 0)$ and $\boldsymbol{\mu}_2 = (2.5, 0, \ldots, 0)$, population covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma$ of symmetric Toeplitz form with first row $(d, d-1, \ldots, 2, 1)$, and covariate dimension $d = 50$. The hyperparameters (r_1, r_2) for models A and B were determined via Eq. 35. The results are based on a single data realization

Table 2 Comparison of classification accuracy using cross-validation and evidence maximization methods for estimating hyperparameters using the same data as Figs. 1 and 2

Classification accuracy (%)	Method	Model A	Model B
Uncorrelated data	Cross-validation	87%	86%
	Evidence maximization	86%	83%
Correlated data	Cross-validation	69%	64%
	Evidence maximization	64%	62%

of the degree of overfitting. We observe that model B overfits less for uncorrelated data, and model A overfits less for correlated data. This pattern is also seen more generally in Table 4, for a broader range of synthetic data-sets. However, we note that all models still perform significantly above the random guess level on unseen data, even when $d \gg n_z$. For instance, for $d = 150$ (corresponding to $d/n_z \approx 11.5$), the Bayesian models can still classify some 80% of the unseen data correctly.

We thank the reviewers for pointing out other measurements of agreement between true and predicted values in particular the adjusted Rand Index (Morey and Agresti 1984; Hubert and Arabie 1985) which neatly corrects for chance prediction results.

5 Classification Accuracy

We compare the classification accuracy of our Bayesian models A and B, with hyperparameters optimized by evidence maximization, to other successful state-of-the-art generative classifiers from Srivastava et al. (2007). These include the distribution-based Bayesian classifier (BDA7), the Quadratic Bayes (QB) classifier (Brown et al. 1999), and a non-Bayesian method, the so-called eigenvalue decomposition discriminant analysis (EDDA) as described in Bensmail and Celeux (1996). All three use cross-validation for model selection and hyperparameter estimation. The classifiers (our models A and B and the three benchmark methods from Srivastava et al. (2007)) are all tested on the same synthetic and real data-sets, and following the definitions and protocols described in Srivastava et al. (2007), for a fair comparison. Model A differs from Quadratic Bayes (Brown et al. 1999) only in that our hyperparameters have been estimated using evidence maximization, as described in Section 3, rather than via cross-validation. Model A is seen in Table 4 to have lower error rates than Quadratic Bayes in the majority of the synthetic data-sets. In contrast, model B is mathematically different from both model A and Quadratric Bayes.

5.1 Implementation

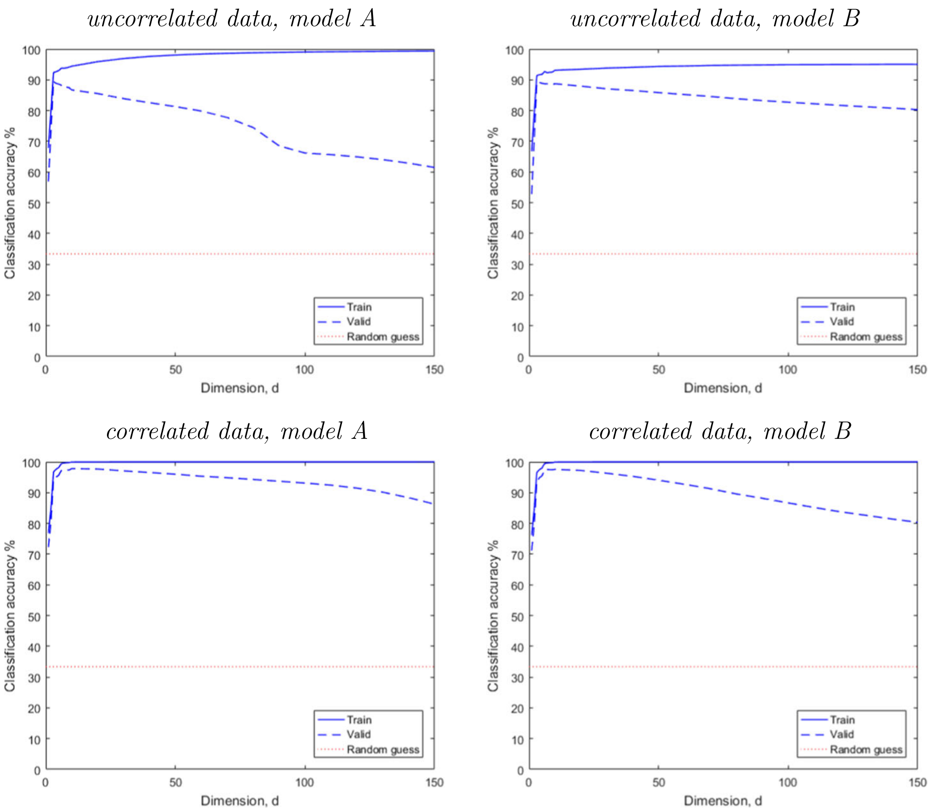
The classifier was implemented in MATLAB.[3] The leave-one-out cross-validation pseudocode is displayed below.

[3] MATLAB 8.0, The MathWorks, Inc., Natick, Massachusetts, United States.

Table 3 Description of synthetic data-sets

	Σ_1	Σ_2	Σ_3	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_3$
Case 1	$\mathbb{I}_d$	$\mathbb{I}_d$	$\mathbb{I}_d$	$(0, 0, \ldots, 0)$	$(3, 0, \ldots, 0, 0)$	$(0, 0, \ldots, 0, 3)$
Case 2	$\mathbb{I}_d$	$2\mathbb{I}_d$	$3\mathbb{I}_d$	$(0, 0, \ldots, 0)$	$(3, 0, \ldots, 0, 0)$	$(0, 0, \ldots, 0, 4)$
Case 3	$\Sigma_{ii} = \left(\frac{9(i-1)}{d-1} + 1\right)^2$	$\Sigma_{ii} = \left(\frac{9(i-1)}{d-1} + 1\right)^2$	$\Sigma_{ii} = \left(\frac{9(i-1)}{d-1} + 1\right)^2$	$(0, 0, \ldots, 0)$	$\mu_{2i} = 2.5\sqrt{\frac{e_i}{d}}\left(\frac{d-i}{\frac{d}{2}-1}\right)$	$\mu_{3i} = (-1)^i \mu_{2i}$
Case 4	$\Sigma_{ii} = \left(\frac{9(i-1)}{d-1} + 1\right)^2$	$\Sigma_{ii} = \left(\frac{9(i-1)}{d-1} + 1\right)^2$	$\Sigma_{ii} = \left(\frac{9(i-1)}{d-1} + 1\right)^2$	$(0, 0, \ldots, 0)$	$\mu_{2i} = 2.5\sqrt{\frac{e_i}{d}}\left(\frac{i-1}{\frac{d}{2}-1}\right)$	$\mu_{3i} = (-1)^i \mu_{2i}$
Case 5	$\Sigma_{ii} = \left(\frac{9(i-1)}{d-1} + 1\right)^2$	$\Sigma_{ii} = \left(\frac{9(d-i)}{d-1} + 1\right)^2$	$\Sigma_{ii} = \left(\frac{9(i-(\frac{d-1}{2}))}{d-1} + 1\right)^2$	$(0, 0, \ldots, 0)$	$(0, 0, \ldots, 0)$	$(0, 0, \ldots, 0)$
Case 6	$\Sigma_{ii} = \left(\frac{9(i-1)}{d-1} + 1\right)^2$	$\Sigma_{ii} = \left(\frac{9(d-i)}{d-1} + 1\right)^2$	$\Sigma_{ii} = \left(\frac{9(i-(\frac{d-1}{2}))}{d-1} + 1\right)^2$	$(0, 0, \ldots, 0)$	$(\frac{14}{\sqrt{d}}, \ldots, \frac{14}{\sqrt{d}})$	$\mu_{3i} = (-1)^i \mu_{2i}$
Case 7	$\mathbf{R}_1^T\mathbf{R}_1$	$\mathbf{R}_2^T\mathbf{R}_2$	$\mathbf{R}_3^T\mathbf{R}_3$	$(0, 0, \ldots, 0)$	$(0, 0, \ldots, 0, 0)$	$(0, 0, \ldots, 0, 0)$
Case 8	$\mathbf{R}_1^T\mathbf{R}_1$	$\mathbf{R}_2^T\mathbf{R}_2$	$\mathbf{R}_3^T\mathbf{R}_3$	$\mathcal{N}_d(0, 1)$	$\mathcal{N}_d(0, 1)$	$\mathcal{N}_d(0, 1)$
Case 9	$\mathbf{R}_1^T\mathbf{R}_1\mathbf{R}_1^T\mathbf{R}_1$	$\mathbf{R}_2^T\mathbf{R}_2\mathbf{R}_2^T\mathbf{R}_2$	$\mathbf{R}_3^T\mathbf{R}_3\mathbf{R}_3^T\mathbf{R}_3$	$(0, 0, \ldots, 0)$	$(0, 0, \ldots, 0, 0)$	$(0, 0, \ldots, 0, 0)$
Case 10	$\mathbf{R}_1^T\mathbf{R}_1\mathbf{R}_1^T\mathbf{R}_1$	$\mathbf{R}_2^T\mathbf{R}_2\mathbf{R}_2^T\mathbf{R}_2$	$\mathbf{R}_3^T\mathbf{R}_3\mathbf{R}_3^T\mathbf{R}_3$	$\mathcal{N}_d(0, 1)$	$\mathcal{N}_d(0, 1)$	$\mathcal{N}_d(0, 1)$

The above class-specific means and covariance matrices were used to sample from a multivariate Gaussian distribution. These are the same data characteristics as those used in Srivastava et al. (2007), reflecting varying degrees of correlation between variables. Note that there is no model mismatch between these data-sets and what is assumed by our generative models


uncorrelated data, model A

uncorrelated data, model B

correlated data, model A

correlated data, model B

Fig. 3 Overfitting in models A and B as measured via LOOCV. Top—uncorrelated data (case 1 in Table 3). Bottom—correlated data (case 8 in Table 3). In all cases, $n_z = 13$ for each of the three classes. Solid lines—classification accuracy on training samples; dashed lines—classification accuracy on validation samples. Horizontal dotted line—baseline performance of a random guess classifier

Algorithm 1 LOOCV classification.

```
procedure LOOCV
    $\mathbf{X} \leftarrow$ design matrix, $\mathbf{y} \leftarrow$ class labels        ▷ Import or generate data, $\mathbf{X} \in \mathbb{R}^{N \times d}$
    specify model A or B                                                    ▷ User input
    loop i = 1:N                                                            ▷ N = number of samples/rows
        $\underline{\mathbf{x}}_0 = \mathbf{X}(i, :) \Rightarrow \mathcal{D} = \mathbf{X} \setminus \underline{\mathbf{x}}_0$        ▷ $\mathcal{D}$ = training set
        calculate sample covariance matrix and eigenvalues from $\mathcal{D}$
        calculate hyperparameters                                           ▷ Eqs. 35 37
        predict $\leftarrow \underline{\mathbf{x}}_0$ class prediction      ▷ Eq. 38
    end loop
    Create confusion matrix from y and predict
end procedure
```

Table 4 Classification performance for synthetic data-sets

Error rate (%)	*d*	BDA7	QB	EDDA	Model A	Model B
Case 1	10	13.2	19.2	11.2	12.0 ± 3.2	11.0 ± 2.8
	50	27.9	33.3	21.7	19.9 ± 4.6	15.6 ± 3.4
	100	35.8	31.1	24.8	32.6 ± 6.0	19.9 ± 4.3
Case 2	10	21.3	27.4	16.1	11.9 ± 3.4	11.4 ± 3.6
	50	26.8	42.6	12.5	9.3 ± 3.2	5.8 ± 2.2
	100	20.8	41.9	9.0	26.5 ± 5.6	3.6 ± 2.1
Case 3	10	10.4	35.0	9.1	27.2 ± 4.9	27.2 ± 5.5
	50	27.2	55.7	21.2	48.6 ± 5.0	49.2 ± 5.2
	100	46.9	56.4	27.7	55.4 ± 5.2	55.1 ± 4.9
Case 4	10	12.6	32.8	11.6	11.3 ± 3.5	11.1 ± 4.1
	50	22.5	30.9	17.0	22.5 ± 4.4	17.8 ± 4.0
	100	37.6	32.1	21.1	30.8 ± 5.2	21.9 ± 4.3
Case 5	10	4.1	15.0	4.4	12.8 ± 4.1	12.8 ± 3.5
	50	1.2	30.6	0.0	9.2 ± 3.4	5.6 ± 2.7
	100	0.2	38.3	0.1	10.9 ± 3.8	5.4 ± 3.4
Case 6	10	5.2	7.9	1.7	4.6 ± 2.3	4.4 ± 2.3
	50	0.5	26.5	0.0	3.9 ± 2.3	3.5 ± 2.4
	100	0.1	29.4	0.0	4.8 ± 2.5	4.5 ± 2.6
Case 7	10	19.5	22.8	19.7	20.0 ± 6.0	27.3 ± 7.4
	50	34.7	30.9	63.9	30.2 ± 5.0	44.7 ± 7.8
	100	40.0	35.2	64.8	35.2 ± 5.1	51.7 ± 7.8
Case 8	10	3.7	2.7	5.1	1.6 ± 1.9	1.5 ± 1.5
	50	9.2	3.5	25.5	4.4 ± 3.2	9.5 ± 5.0
	100	17.3	8.1	55.2	8.7 ± 4.4	23.9 ± 9.0
Case 9	10	1.5	0.9	1.0	0.9 ±1.1	5.4 ± 6.8
	50	1.3	0.9	32.5	1.3 ± 1.2	16.9 ± 14.6
	100	2.9	2.8	67.0	1.5 ± 1.5	22.4 ± 15.3
Case 10	10	0.4	0.1	3.4	0.1 ± 0.6	0.2 ± 0.6
	50	1.7	0.9	32.4	0.8 ± 1.0	15.9 ± 13.6
	100	2.2	2.4	64.0	1.4 ± 1.2	23.4 ± 16.0

Three generative Bayesian models, BDA7, QB, and EDDA (results taken from Srivastava et al. 2007) are used as comparison with our models A and B. Error rates are the percentages of misclassified samples from the test data-set. The error bars for models A and B represent one standard deviation in the error rates, calculated over the 100 data realizations

The rate-limiting step in the algorithm is calculation of sample eigenvalues (approximately $\mathcal{O}(d^3)$). Figure 4 shows the relationship between algorithm time and data dimension for binary classification with 100 data samples.[4]

[4] Leave-one-out cross-validation using an Intel i5-4690 x64-based processor, CPU speed of 3.50GHz, 32GB RAM. As the data dimension increases above 30,000, RAM storage considerations become an issue on typical PCs.

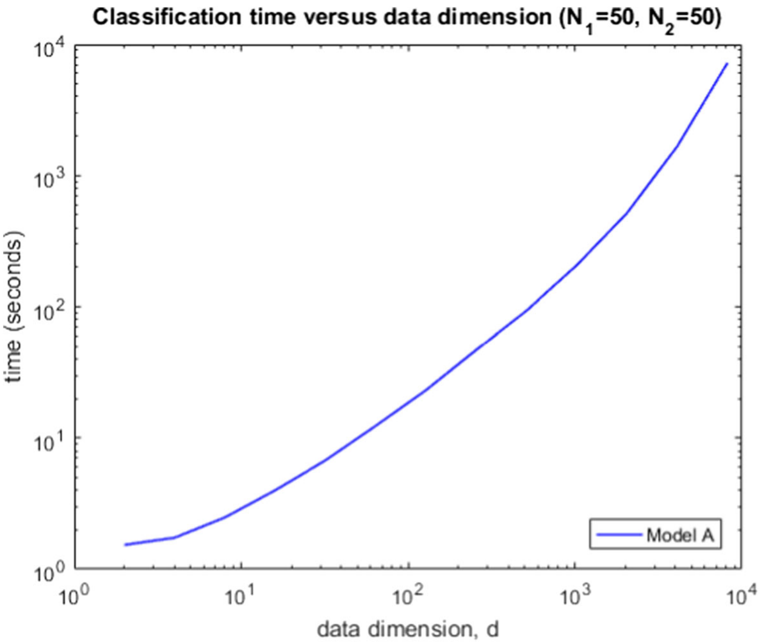


Fig. 4 Processing time in seconds for binary classification on synthetic data (50 samples in each class). Both models had similar timings so only model A is plotted. Note the logarithmic scale

5.2 Synthetic Data

The study of Srivastava et al. (2007) used a set of ten synthetic data-sets, all with Gaussian multivariate covariate distributions and a range of choices for class-specific means and covariance matrices. In the present study, we generated data with exactly the same statistical features. The first six of these choices were also used in Friedman (1989) and involve diagonal covariance matrices. The remaining four represent correlated data. Each data-set has $C = 3$ outcome classes and is separated into a training set, with $n_z = 13$ samples in each class, and a validation set, with $n_z = 33$ samples in each class. In terms of the balance n_z/d, this allows for a direct comparison with the dimensions used in Srivastava et al. (2007). The results shown in Table 4 are all averages over 100 synthetic data runs. The data dimensions are chosen from $d \in \{10, 50, 100\}$. Since all these synthetic data-sets involve multivariate Gaussian covariate distributions, there is no model mismatch with any of the models being compared.

The means and covariance matrices of the synthetic data-sets are given in Table 3. The covariance matrices for the correlated data-sets are defined in terms of auxiliary random $d \times d$ matrices $\mathbf{R}_z$, with i.i.d. entries sampled from the uniform distribution on the interval $[0, 1]$, according to either $\Sigma_z = \mathbf{R}_z^T\mathbf{R}_z$ or $\Sigma_z = \mathbf{R}_z^T\mathbf{R}_z\mathbf{R}_z^T\mathbf{R}_z$. These covariance matrices have a single dominant eigenvalue, and further non-dominant eigenvalues that are closer to zero for data-sets 9-10. Data-sets 7 and 9 have all class means at the origin, whereas each element of the class mean vectors from data-sets 8 and 10 is independently sampled from a standard normal distribution.

Table 4 shows the classification error rates, as percentages of misclassified samples over the validation set. The variability of these for results for the models BDA7, QB, and EDDA, i.e. the error bars in the classification scores, is not reported in Srivastava et al. (2007) (where only the best classifier was determined using a signed ranked test). For completeness, we have included in this study the standard deviation of the error rate over the 100 synthetic data runs for our models A and B. Given that all experiments involved the same dimensions of data-sets and similar average error rates, the error bars for the Srivastava et al. (2007) results are expected to be similar to those of models A and B. We conclude from Table 4 that our models A and B perform on average quite similarly to the benchmark classifiers BDA7, QB, and EDDA. On some data-sets, model A and/or B outperform the benchmarks, on others they are outperformed. However, models A and B achieve this competitive level of classification accuracy without cross-validation, i.e. at a much lower computational cost.

5.3 Real Data

Next, we test the classification accuracy of our models against the real data-sets used in Srivastava et al. (2007), which are publicly available from the UCI machine learning repository.[5] Three data-sets were left out due to problems with matching the formats: *Image segmentation* (different number of samples than Srivastava et al. (2007)), *Cover type* (different format of training/validation/test), and *Pen digits* (different format of training/validation/test). Before classification, we looked for identifying characteristics which could allow for retrospective interpretation of the results, e.g. occurrence of discrete covariate values, covariance matrix entropies, or class imbalances. None were found to be informative. No scaling or pre-processing was done to the data before classification.

We duplicated exactly the protocol of Srivastava et al. (2007), whereby only a randomly chosen 5% or 10% of the samples from each class of each data-set are used for training, leaving the bulk of the data (95% or 90%) to serve as validation (or test) set. The resulting small training sample sizes lead to $n_z \ll d$ for a number of data-sets, providing a rigorous test for classifiers in overfitting-prone conditions. For example, the set *Ionosphere*, with $d = 34$, has original class sizes of 225 and 126 samples leading in the 5% training scenario to training sets with $n_1 = 12$ and $n_2 = 7$. We have used the convention of rounding up any non-integer number of training samples (rounding down the number of samples had only a minimal effect on most error rates). The *baseline* column gives the classification error that would be obtained if the majority class is predicted every time.

We conclude from the classification results shown in Tables 5 and 6 (which are to be interpreted as having non-negligible error bars) that also for the real data, models A and B are competitive with the other Bayesian classifiers. The exceptions are *Ionosphere* (where models A and B outperform the benchmark methods, in both tables) and the data-sets *Thyroid* and *Wine* (where in Table 6, our model A is being outperformed). Note that in Table 6, *Thyroid* and *Wine* have only 2 or 3 data samples in some classes of the training set. This results in nearly degenerate class-specific covariance matrices, which hampers the optimization of hyperparameters via evidence maximization. Model B behaves well even in these tricky cases, presumably due to the impact of its additional hyperparameter $\gamma_{0z} = d/\hat{\boldsymbol{X}}_z^2$. As expected, upon testing classification performance using leave-one-out cross-validation (details not shown here) rather than the 5% or 10% training set methods above, all error rates are significantly lower.

Examining the results from Sections 5.2 and 5.3 does not lead us to conclusions on when one specific model outperforms the other. We are currently pursuing two approaches to this

[5] http://archive.ics.uci.edu/ml/index.php

Table 5 Average error rate using randomly selected 10% of training samples in each class

Error rate (%)	n	Class sizes	d	Baseline	BDA7	QB	EDDA	Model A	Model B
Heart	270	150, 120	10	44.4	27.4	32.0	28.3	30.3	30.1
Ionosphere	351	225, 126	34	35.9	12.5	11.1	23.3	8.3	7.5
Iris	150	50, 50, 50	4	66.6	6.2	5.9	7.4	7.5	6.6
Pima	768	500, 268	8	34.9	28.4	29.7	29.0	28.8	28.9
Sonar	208	97, 111	60	46.6	31.2	33.7	34.8	34.9	33.8
Thyroid	215	150, 35, 30	5	30.2	7.9	9.1	8.6	7.6	7.9
Wine	178	59, 71, 48	13	60.1	7.9	16.9	8.2	15.6	16.0

The remaining 90% of samples were used as a validation set. Error rates are the percentage of misclassified samples over this validation set

problem: (1) finding analytical expressions for the probability of misclassification similar to Raudys and Young (2004) but with the true data-generating distribution different from model assumptions and (2) numerical work generating synthetic data from a multivariate t-distribution with varying degrees of freedom.

6 Discussion

In this paper, we considered generative models for supervised Bayesian classification in high-dimensional spaces. Our aim was to derive expressions for the optimal hyperparameters and predictive probabilities in closed form. Since the dominant cause of overfitting in the classification of high-dimensional data is using point estimates for high-dimensional parameter vectors, we believe that by carefully choosing Bayesian models for which parameter integrals are analytically tractable, we will need point estimates only at hyperparameter level, reducing overfitting.

We showed that the standard priors of Bayesian classifiers that are based on class-specific multivariate Gaussian covariate distributions can be generalized, from which we derive two special model cases (A and B) for which predictive probabilities can be derived analytically in fully explicit form. Model A is known in the literature as Quadratic Bayes (Brown et al. 1999), whereas model B is novel and has not yet appeared in the literature. In contrast to

Table 6 Average error rate using randomly selected 5% of training samples in each class

Error Rate (%)	n	Class sizes	d	Baseline	BDA7	QB	EDDA	Model A	Model B
Heart	270	150, 120	10	44.4	30.6	38.5	33.9	38.8	39.6
Ionosphere	351	225, 126	34	35.9	16.9	16.1	26.0	10.3	8.8
Iris	150	50, 50, 50	4	66.6	6.9	7.6	9.40	12.8	11.4
Pima	768	500, 268	8	34.9	29.7	32.7	30.7	30.3	30.8
Sonar	208	97, 111	60	46.6	36.8	40.4	39.8	45.6	39.0
Thyroid	215	150, 35, 30	5	30.2	11.7	14.8	14.7	34.5	14.6
Wine	178	59, 71, 48	13	60.1	9.6	33.1	11.2	54.4	33.0

The remaining 95% of samples were used as a validation set. Error rates are the percentage of misclassified samples over this validation set common practice for most Bayesian classifiers, we use evidence maximization (MacKay 1999) to find analytical expressions for our hyperparameters in both models. This allows us to find their optimal values without needing to resort to computationally expensive cross-validation protocols.

We found that the alternative (but significantly faster) hyperparameter determination by evidence maximization leads to hyperparameters that are generally very similar to those obtained via cross-validation, and that the classification performance of our models A and B degrades only gracefully in the ‘dangerous’ regime $n \ll d$ where we would expect extreme overfitting. We compared the classification performance of our models on the extensive synthetic and real data-sets that have been used earlier as performance benchmarks in Srivastava and Gupta (2006) and Srivastava et al. (2007). Interestingly, the performance of our models A and B turned out to be competitive with state-of-the-art Bayesian models that use cross-validation, despite the large reduction in computational expense. This will enable users in practice to classify high-dimensional data-sets quicker, without compromising on accuracy.

This paper shows that the analytical approach merits further investigation. Calculating the predictive probability for arbitrary γ_{0z}, γ_{1z} values remains to be done. The main obstacle being the resulting symbolic integration. We believe this could lead to interesting analytically tractable classification models.

Acknowledgements This work was supported by the Biotechnology and Biological Sciences Research Council (UK) and by GlaxoSmithKline Research and Development Ltd. Many thanks to James Barrett for his support.

References

Bensmail, H., & Celeux, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, *91*(436), 1743–1748.

Berger, J.O., Bernardo, J.M., et al. (1992). On the development of reference priors. *Bayesian Statistics*, *4*(4), 35–60.

Brown, P.J., Fearn, T., Haque, M. (1999). Discrimination with many variables. *Journal of the American Statistical Association*, *94*(448), 1320–1329.

Coolen, A.C.C., Barrett, J.E., Paga, P., Perez-Vicente, C.J. (2017). Replica analysis of overfitting in regression models for time-to-event data. *Journal of Physics A: Mathematical and Theoretical*, *50*, 375001.

Efron, B., & Morris, C.N. (1977). *Stein’s paradox in statistics*. New York: WH Freeman.

Friedman, J.H. (1989). Regularized discriminant analysis. *Journal of the American statistical Association*, *84*(405), 165–175.

Geisser, S. (1964). Posterior odds for multivariate normal classifications. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(1), 69–76.

Haff, L. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics*, *8*(3), 586–597.

Hinton, G.E., & Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 361–379).

Jonsson, D. (1982). Some limit theorems for the eigenvalues of a sample covariance matrix. *Journal of Multivariate Analysis*, *12*(1), 1–38.

Keehn, D.G. (1965). A note on learning for Gaussian properties. *IEEE Transactions on Information Theory*, *11*(1), 126–132.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411.

MacKay, D.J. (1999). Comparison of approximate methods for handling hyperparameters. *Neural Computation*, *11*(5), 1035–1068.

Morey, L.C., & Agresti, A. (1984). The measurement of classification agreement: an adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement*, *44*(1), 33–7.

Raudys, S., & Young, D.M. (2004). Results in statistical discriminant analysis: a review of the former Soviet Union literature. *Journal of Multivariate Analysis*, *89*(1), 1–35.

Shalabi, A., Inoue, M., Watkins, J., De Rinaldis, E., Coolen, A.C. (2016). Bayesian clinical classification from high-dimensional data: signatures versus variability. Statistical Methods in Medical Research, 0962280216628901.

Srivastava, S., & Gupta, M.R. (2006). Distribution-based Bayesian minimum expected risk for discriminant analysis. In *2006 IEEE international symposium on information theory* (pp. 2294–2298): IEEE.

Srivastava, S., Gupta, M.R., Frigyik, B.A. (2007). Bayesian quadratic discriminant analysis. *Journal of Machine Learning Research*, *8*(6), 1277–1305.

Stein, C., et al. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 197–206).

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.