Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–20

DOI: 10.1177/0962280216628901
smm.sagepub.com
SAGE

Bayesian clinical classification from high-dimensional data: Signatures versus variability

Akram Shalabi,[1] Masato Inoue,[2] Johnathan Watkins,[3] Emanuele De Rinaldis[4] and Anthony CC Coolen[1]

Abstract

When data exhibit imbalance between a large number d of covariates and a small number n of samples, clinical outcome prediction is impaired by overfitting and prohibitive computation demands. Here we study two simple Bayesian prediction protocols that can be applied to data of any dimension and any number of outcome classes. Calculating Bayesian integrals and optimal hyperparameters analytically leaves only a small number of numerical integrations, and CPU demands scale as $O(nd)$. We compare their performance on synthetic and genomic data to the *mclustDA* method of Fraley and Raftery. For small d they perform as well as *mclustDA* or better. For $d = 10{,}000$ or more *mclustDA* breaks down computationally, while the Bayesian methods remain efficient. This allows us to explore phenomena typical of classification in high-dimensional spaces, such as overfitting and the reduced discriminative effectiveness of signatures compared to intra-class variability.

Keywords

Discriminant analysis, Bayesian classification, overfitting, curse of dimensionality, outcome prediction

1 Introduction

Discriminant analysis[1] is the use of known classifications to find rules that link observations to their classes, which are then used to predict the classes of new observations. It is applied in many settings.[2–4] Its methods are usually probabilistic and model based: observations are assumed to have been generated from a class-specific distribution that must be estimated from the data.[1] Many methods are accurate for data with low covariate dimension d, but problems arise when d is large. The main ones are overfitting,[5,6] i.e. the tendency of models with many parameters to

[1]Institute for Mathematical and Molecular Biomedicine, King's College London, London, UK
[2]Department of Electrical Engineering and Bioscience, School of Advanced Science and Engineering, Waseda University, Tokyo, Japan
[3]Breakthrough Breast Cancer Research Unit, Department of Research Oncology, Guy's Hospital, London, UK
[4]NIHR Biomedical Research Centre – R&D Department, Guy's Hospital, London, UK

Corresponding author:
Anthony CC Coolen, Institute for Mathematical and Molecular Biomedicine, Hodgkin Building, Guy's Campus, King's College London, London, UK.
Email: ton.coolen@kcl.ac.uk

capture the noise in the data as opposed to the signal when the number n of samples is small and inability to do the required computations within practical timescales. With the increasing availability of genomic covariates in medical data sets, where often $d \gg n$,[5–11] these problems can cause outcome predictions to have limited reliability, if they can be generated at all. Hence, for high-dimensional data one often resorts to 'class (prognostic) signatures',[12,13] i.e. selections of covariates with specific combined value patterns that are characteristic of a class, with ad hoc rules for converting signature similarity into classification. To combat overfitting one would prefer to use Bayesian methods, but these involve more computations since each covariate contributes at least one integral to the posterior parameter distribution. Several dimension reduction methods have been suggested, such as principal component analysis,[14–16] subspace clustering,[17,18] or the use of constrained and parsimonious models.[19] The first runs the risk of reduced accuracy through loss of information.[20,21] The second assumes that high-dimensional data live in subspaces with dimensionality less than d[17,18,22] but relies on finding a classifier in high dimensions.[21,22] The third is a compromise between precise modelling and what can be estimated in practice. The latter two have been combined to reduce the number of parameters and dimensions for EM.[17,21,23]

Probabilistic discriminant analysis methods assume that the observations $\mathbf{x}$ (d-dimensional vectors) in each class $y \in \{1, \ldots, c\}$ are described by a conditional distribution $p(\mathbf{x}|\, y)$ specific to y. The number c of classes is usually known. A typical data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ consists of n observations $\mathbf{x}_i$ and their class labels y_i. Each pair $(\mathbf{x}_i, y_i)$ is assumed to be drawn independently from a joint distribution $p(\mathbf{x}, y) = p(\mathbf{x}|\, y)\, p(\, y)$ that describes the population, which is to be estimated from the data. Many methods assume that the observations in each class are multivariate Gaussian[24,25] or Gaussian mixture Models (GMM).[1,14,19,21] A popular model-based discriminative analysis approach was proposed by Fraley and Raftery and implemented in the R software environment[26] (reference package *mclust*[14,16]). It is widely used in medical outcome prediction,[27–31] and we therefore use it as our benchmark. After splitting $\mathcal{D}$ into a training and validation set, the authors of R Core Team[26] fit a GMM to each class in the training set via model-based clustering. The number of classes c and the covariance matrix for each class are inferred from the training set. For each number of GMM components and each assumed covariance complexity level, hierarchical clustering techniques[24,25,32] are applied to the training set to yield trial partitions, whose parameters and class-conditional probabilities are determined using the Expectation-Maximisation algorithm (EM).[33] The R implementation of the programme allows for different choices of priors and various parametrisations for the covariance matrices $\mathbf{C}^r$ of each Gaussian component r, ranging from spherically symmetric ($C^r_{k\ell} = c\delta_{k\ell}$) via nonuniform diagonal ($C^r_{k\ell} = c_k\delta_{k\ell}$) to completely general positive definite matrices. The model options available in the R package *mclust* are summarised in Fraley and Raftery.[28] In this paper, we report explicitly on results from *mclustDA* when used in its default setting 'MclustDA' (where each class is allowed to be a Gaussian mixture), and with default (i.e. conjugate) priors. In addition, we have repeated most of our experiments with the alternative setting 'EDDA' (i.e. Eigenvalue Decomposition Discriminant Analysis, where each class has one Gaussian component). The resulting performance differences are small and are discussed in the relevant sections. The Bayesian information criterion[24,34] is computed for each GMM and used to identify the optimal model, which is then applied to the validation set. Class-conditional probabilities are calculated using Bayes' theorem, and each observation in the validation set is assigned to the class with the highest posterior probability. The possible impact of degenerate solutions is alleviated by using Maximum A Posteriori estimates.[28] For large d, *mclustDA* is known to suffer from overfitting and prohibitive computation demands.[14,16,19,35,36] For data sets of modest

sizes, such as $n = 100$ with $d \leq 3000$, *mclustDA* produces predictions relatively swiftly, but for $d = 10{,}000$ or more *mclustDA* can no longer be used in practice on conventional (multi-core) machines. This rules out its application to large-scale genomic data.

In this paper we follow[37–41] and show how solving integrals and optimal hyperparameters *analytically* reduces the detrimental impact of high data dimensionality in Bayesian class prediction. Our formulae can be applied to data with arbitrary covariate dimension, without approximations at parameter level. In Section 2 we describe and analyse two Bayesian outcome prediction methods: one models the joint distribution of covariates and classes, and one models the class-conditioned covariate densities. We show how class signatures can lose their discriminative power in high-dimensional spaces, in contrast to intra-class covariate variability, which becomes increasingly effective. In Sections 3 and 4 we apply our methods to high-dimensional data. We first show with synthetic data how the curse of dimensionality[42] is lifted in terms of CPU demands, in comparison to *mclustDA*.[14] We then analyse gene expression data from breast cancer and ovarian cancer patient cohorts, with either clinical outcome classes or biologically defined classes, and breast tissue imaging data for tumour detection. We close with a summary.

2 Bayesian multi-class outcome prediction for high covariate dimensions

We focus on Bayesian class prediction that is computationally feasible for data with high covariate dimensions. This requires that those integrals whose dimension scales with d are solved analytically. To simplify formulas we first introduce the empirical frequency f_y and the empirical averages $\langle \mathbf{x} \rangle_y$ and $\langle \mathbf{x}^2 \rangle_y$ over each class y in $\mathcal{D}$ (with $\mathbf{x}_i^2 = \mathbf{x}_i \cdot \mathbf{x}_i$)

$$f_y = \frac{1}{n} \sum_{i=1}^{n} \delta_{yy_i}, \qquad \langle \mathbf{x} \rangle_y = \frac{\sum_{i=1}^{n} \delta_{yy_i} \mathbf{x}_i}{\sum_{i=1}^{n} \delta_{yy_i}}, \qquad \langle \mathbf{x}^2 \rangle_y = \frac{\sum_{i=1}^{n} \delta_{yy_i} \mathbf{x}_i^2}{\sum_{i=1}^{n} \delta_{yy_i}} \tag{1}$$

In these definitions we used the Kronecker delta-symbol, defined as $\delta_{ab} = 1$ when $a = b$ and $\delta_{ab} = 0$ if $a \neq b$. We assume that all classes are represented in $\mathcal{D}$. We will see later that we need $nf_y \geq 2$ for all y when using a training and validation set. We also define the empirical average signal strength X_y and noise strength Σ_y in the covariates, for each class

$$X_y^2 = \langle \mathbf{x} \rangle_y^2 / \mathrm{d}, \qquad \Sigma_y^2 = (\langle \mathbf{x}^2 \rangle_y - \langle \mathbf{x} \rangle_y^2)/d \tag{2}$$

2.1 Model parametrisation

We use $\boldsymbol{\theta}$ to denote all model parameters with d-dependent dimension, and H for all hyperparameters, with d-independent dimensions. We need a sensible parameterisation of the joint distributions $p(\mathbf{x}, y|\boldsymbol{\theta}, H)$ for the covariate observations $\mathbf{x} \in \mathbb{R}^d$ and the classes $y \in \{1, \ldots, c\}$. When overfitting is a real danger, only simple models with a small number of parameters are acceptable. In the spirit of the literature[37–41] and in line with the concept of outcome 'class signatures', we choose in this paper each $p(\mathbf{x}, y|\boldsymbol{\theta}, H)$ to be a homogeneous Gaussian distribution

$$p(\mathbf{x}, y|\boldsymbol{\theta}, H) = p(\mathbf{x}|\, y, \boldsymbol{\theta}, H)\, p_y, \qquad p(\mathbf{x}|\, y, \boldsymbol{\theta}, H) = (\alpha_y \sqrt{2\pi})^{-d} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_y)^2/\alpha_y^2}, \quad \alpha_y \geq 0 \tag{3}$$

with c class probabilities $p_y \in [0, 1]$, subject to $\sum_{y=1}^{c} p_y = 1$, and with c true but unknown 'signature' vectors $\boldsymbol{\mu}_y \in \mathbb{R}^d$. Each $\boldsymbol{\mu}_y$ is given a simple independent Gaussian Bayesian prior

$$p(\boldsymbol{\mu}_y|\beta_y) = (\beta_y\sqrt{2\pi})^{-d}\, \mathrm{e}^{-\frac{1}{2}\boldsymbol{\mu}_y^2/\beta_y^2}, \qquad \beta_y \geq 0 \tag{4}$$

Hence our choice of equations (3) and (4) involves

$$cd \quad \text{parameters} : \boldsymbol{\theta} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c\}, \qquad \boldsymbol{\mu}_y \in \mathbb{R}^d \tag{5}$$

$$3c - 1 \text{ hyperparameters} : \qquad H = \{(\alpha_1, \beta_1, p_1), \ldots, (\alpha_c, \beta_c, p_c)\},\ \alpha_y, \beta_y, p_y \geq 0, \qquad \sum_{y=1}^{c} p_y = 1 \tag{6}$$

We assume in equation (3) that the variances of different components of the covariate vector $\mathbf{x}$ are identical within a class. This will be appropriate if all covariates are of the same type, e.g. gene expression levels or when they have been normalised. However, it is expected to give poor results if covariates refer to distinct modalities, e.g. if gene expression levels are concatenated with other clinical or imaging data. Adding nontrivial covariance matrices or Gaussian mixtures to equations (3) and (4) would increase the number of parameters and hence the risk of overfitting. To enable application of equations (3) and (4) to high-dimensional data, i.e. to covariate vectors with $d \gg 1$, we must marginalise the vectors $\{\boldsymbol{\mu}_y\}$ analytically.

2.2 Method I: Generative Bayesian classification

In the generative framework one regards all the data in $\mathcal{D}$ as informative, including the empirical class frequencies f_y. Given equations (3) and (4), one can then write the joint likelihood of the data $\mathcal{D}$ and any new pair $(\mathbf{x}_0, y_0)$, given parameters $\boldsymbol{\theta}$ and hyperparameters H, following[22] as

$$p(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n, y_0, y_1, \ldots, y_n|\boldsymbol{\theta}, H) = \prod_{i=0}^{n} p(\mathbf{x}_i, y_i|\boldsymbol{\theta}, H) \tag{7}$$

From this expression follows the posterior class prediction for a new covariate observation $\mathbf{x}_0$

$$\begin{aligned} p(y_0|\mathbf{x}_0, \mathcal{D}, H) &= \frac{\int \mathrm{d}\boldsymbol{\theta}\; p(\boldsymbol{\theta}|H)\, p(\mathbf{x}_0, y_0|\boldsymbol{\theta}, H) \prod_{i=1}^{n} p(\mathbf{x}_i, y_i|\boldsymbol{\theta}, H)}{\int \mathrm{d}\boldsymbol{\theta}\; p(\boldsymbol{\theta}|H)\, p(\mathbf{x}_0|\boldsymbol{\theta}, H) \prod_{i=1}^{n} p(\mathbf{x}_i, y_i|\boldsymbol{\theta}, H)} \\ &= \frac{(p_{y_0}/\alpha_{y_0}^d) \int \left(\prod_{z=1}^{c} \mathrm{d}\boldsymbol{\mu}_z\right) \mathrm{e}^{-\frac{1}{2}\sum_{z=1}^{c}\left[\boldsymbol{\mu}_z^2/\beta_z^2 + \delta_{zy_0}(\mathbf{x}_0-\boldsymbol{\mu}_z)^2/\alpha_z^2 + \sum_{i=1}^{n}\delta_{zy_i}(\mathbf{x}_i-\boldsymbol{\mu}_z)^2/\alpha_z^2\right]}}{\sum_{y'=1}^{c}(p_{y'}/\alpha_{y'}^d) \int \left(\prod_{z=1}^{c} \mathrm{d}\boldsymbol{\mu}_z\right) \mathrm{e}^{-\frac{1}{2}\sum_{z=1}^{c}\left[\boldsymbol{\mu}_z^2/\beta_z^2 + \delta_{zy'}(\mathbf{x}_0-\boldsymbol{\mu}_z)^2/\alpha_z^2 + \sum_{i=1}^{n}\delta_{zy_i}(\mathbf{x}_i-\boldsymbol{\mu}_z)^2/\alpha_z^2\right]}} \end{aligned} \tag{8}$$

Since the above integrals are of a Gaussian form, they can all be solved analytically for any d. See e.g. the literature[38–41] for details (where a more general calculation with arbitrary covariance matrices is given). For the present model the result reads

$$p(y_0|\mathbf{x}_0, \mathcal{D}, H) = \frac{(p_{y_0}/S_{y_0}^d)\, \mathrm{e}^{-\frac{1}{2}(\mathbf{x}_0-\mathbf{m}_{y_0})^2/S_{y_0}^2}}{\sum_{z=1}^{c}(p_z/S_z^d)\, \mathrm{e}^{-\frac{1}{2}(\mathbf{x}_0-\mathbf{m}_z)^2/S_z^2}} \tag{9}$$

with the short-hands

$$\mathbf{m}_y = \langle \mathbf{x} \rangle_y \, \frac{nf_y \beta_y^2}{nf_y \beta_y^2 + \alpha_y^2}, \qquad S_y^2 = \alpha_y^2 \, \frac{\alpha_y^2 + (nf_y + 1)\beta_y^2}{\alpha_y^2 + nf_y \beta_y^2} \tag{10}$$

Comparing equations (9) to (3) shows that the Bayesian sample mean $\mathbf{m}_y$ can be interpreted as our estimated 'class signature'. If next we choose noninformative hyperparameter priors, the Bayes-optimal hyperparameters $\hat{H}$ become Type II Maximum Likelihood estimators

$$\begin{aligned}
\hat{H} &= \text{argmax}_H \log p(\mathcal{D}|H) \\
&= \text{argmax}_{\{\alpha_y, \beta_y, p_y\}} \left\{ \log \int \left(\prod_{z=1}^{c} \frac{\mathrm{d}\boldsymbol{\mu}_z \mathrm{e}^{-\frac{1}{2}\boldsymbol{\mu}_z^2/\beta_z^2}}{(\beta_z \sqrt{2\pi})^d} \right) \left(\prod_{i=1}^{n} \frac{p_{y_i} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_{y_i})^2/\alpha_{y_i}^2}}{(\alpha_{y_i} \sqrt{2\pi})^d} \right) \right\} \\
&= \text{argmax}_{\{\alpha_y, \beta_y, p_y\}} \sum_{z=1}^{c} \left\{ \frac{nf_z}{d} \log p_z - (nf_z - 1) \log \alpha_z - \frac{1}{2} \log \left(\alpha_z^2 + nf_z \beta_z^2 \right) - \frac{1}{2} nf_z \left(\frac{\Sigma_z^2}{\alpha_z^2} + \frac{X_z^2}{\alpha_z^2 + nf_z \beta_z^2} \right) \right\}
\end{aligned} \tag{11}$$

The maximum over $\{\alpha_y, \beta_y, p_y\}$ in equation (11) is found by straightforward differentiation, resulting in

$$\forall y: \quad \hat{p}_y = f_y, \quad \hat{\alpha}_y^2 = \Sigma_y^2 + X_y^2 - \hat{\beta}_y^2, \quad \hat{\beta}_y^2 = \left(X_y^2 - \frac{\Sigma_y^2}{nf_y - 1} \right) \theta \left[X_y^2 - \frac{\Sigma_y^2}{nf_y - 1} \right] \tag{12}$$

In this expression we used the step function, defined by $\theta[u < 0] = 0$ and $\theta[u > 0] = 1$. Inserting equation (12) into equation (9) gives the explicit prediction of the generative Bayesian model. For classes y with weak signals one finds $\hat{\beta}_y = 0$, giving $\mathbf{m}_y = \mathbf{0}$ in equation (9). Here the 'Occam's razor' action of the Bayesian method is apparently to decide that for such classes there is insufficient evidence for 'class signatures'.

2.3 Method II: Discriminative Bayesian classification

If we are unsure whether the classes were sampled faithfully from the population, i.e. whether the empirical frequencies f_y estimate the true frequencies p_y, we may wish to extract from $\mathcal{D}$ only information regarding the link between $\mathbf{x}$ and y. Following this discriminative route, one regards the class labels $\{y_1, \ldots, y_n\}$ as conditions (as opposed to potentially informative data) and replaces equation (7) by

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n, y_0 | \mathbf{x}_0, y_1, \ldots, y_n, \boldsymbol{\theta}, H) = p(y_0 | \mathbf{x}_0, \boldsymbol{\theta}, H) \prod_{i=1}^{n} p(\mathbf{x}_i | y_i, \boldsymbol{\theta}, H) \tag{13}$$

The posterior class prediction formula now becomes

$$p(y_0 | \mathbf{x}_0, \mathcal{D}, H) = \frac{\int \mathrm{d}\boldsymbol{\theta} \, p(\boldsymbol{\theta}|H) (p(\mathbf{x}_0, y_0 | \boldsymbol{\theta}, H) / p(\mathbf{x}_0 | \boldsymbol{\theta}, H)) \prod_{i=1}^{n} p(\mathbf{x}_i | y_i, \boldsymbol{\theta}, H)}{\int \mathrm{d}\boldsymbol{\theta} \, p(\boldsymbol{\theta}|H) \prod_{i=1}^{n} p(\mathbf{x}_i | y_i, \boldsymbol{\theta}, H)}$$

$$= \int \left(\prod_{z=1}^{c} \frac{\mathrm{d}\mathbf{u}_z \, \mathrm{e}^{-\frac{1}{2}\mathbf{u}_z^2}}{(2\pi)^{d/2}} \right) \left\{ \frac{(p_{y_0}/\alpha_{y_0}^d)\mathrm{e}^{-\frac{1}{2\alpha_{y_0}^2}\left(\mathbf{x}_0 - \mathbf{m}_{y_0} - \frac{\alpha_{y_0}\beta_{y_0}\mathbf{u}_{y_0}}{\sqrt{\alpha_{y_0}^2 + nf_{y_0}\beta_{y_0}^2}}\right)^2}}{\sum_{z=1}^{c}(p_z/\alpha_z^d)\mathrm{e}^{-\frac{1}{2\alpha_z^2}\left(\mathbf{x}_0 - \mathbf{m}_z - \frac{\alpha_z\beta_z\mathbf{u}_z}{\sqrt{\alpha_z + nf_z\beta_{z2}}}\right)^2}} \right\} \tag{14}$$

with the short-hands (10) (see Appendix 1). The integrals above are no longer strictly Gaussian, unlike those in the literature,[38–41] but most can still be done analytically. We choose for each class z a convenient basis for the integration over $\mathbf{u}_z$, such that the first basis vector points in the direction of $\mathbf{x}_0 - \mathbf{m}_z$. This allows us to write, with $|\mathbf{x}| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$

$$p(y_0|\mathbf{x}_0, \mathcal{D}, H) = \int \left(\prod_{z=1}^{c} \frac{\mathrm{d}u_z \mathrm{d}v_z}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}u_z^2} \, \mathcal{P}(v_z) \right) \left\{ \frac{\frac{p_{y_0}}{\alpha_{y_0}^d} \, \mathrm{e}^{-\frac{1}{2\alpha_{y_0}^2}\left(|\mathbf{x}_0 - \mathbf{m}_{y_0}| - \frac{\alpha_{y_0}\beta_{y_0}u_{y_0}}{\sqrt{\alpha_{y_0}^2 + nf_{y_0}\beta_{y_0}^2}}\right)^2 - \frac{1}{2}\frac{\beta_{y_0}^2 v_{y_0}}{\alpha_{y_0}^2 + nf_{y_0}\beta_{y_0}^2}}}{\sum_{z=1}^{c} \frac{p_z}{\alpha_z^d} \, \mathrm{e}^{-\frac{1}{2\alpha_z^2}\left(|\mathbf{x}_0 - \mathbf{m}_z| - \frac{\alpha_z\beta_z u_z}{\sqrt{\alpha_z^2 + nf_z\beta_z^2}}\right)^2 - \frac{1}{2}\frac{\beta_z^2 v_z}{\alpha_z^2 + nf_z\beta_z^2}}} \right\} \tag{15}$$

$$\mathcal{P}(v) = \frac{\left(\frac{1}{2}v\right)^{\frac{1}{2}(d-3)}\mathrm{e}^{-\frac{1}{2}v}}{2\Gamma\left(\frac{1}{2}(d-1)\right)} \tag{16}$$

(see Appendix 1). Formula (15) describing discriminative classification still requires only $2c$ integrals to be done numerically, so also this route is computationally feasible for arbitrary d. The Type II Maximum Likelihood hyperparameter estimators are almost identical to those of the previous case, since $\log p(\mathcal{D}|H)$ differs from its counterpart equation (11) only through the absence of the term $\sum_{z=1}^{c} nf_z \log p_z$. Hence we obtain the same values for $\hat{\alpha}_y$ and $\hat{\beta}_y$. If we believe that the frequencies f_y mirror those of the population, we choose $\hat{p}_y = f_y$ as in equation (12), otherwise we must choose the uniform prior $\hat{p}_y = 1/c$.

2.4 Asymptotic formulae for large covariate dimension

Both X_y and Σ_y were defined such that they scale as $\mathcal{O}(d^0)$ for large d. Hence the same is true for the optimal hyperparameters $\{\hat{\alpha}_y, \hat{\beta}_y, \hat{p}_y\}$ of both our classification models. We may therefore focus for large d and finite n and c on the asymptotic behaviour of the prediction formulae (9) and (15). In both cases one finds these reducing to deterministic class assignment for $d \to \infty$

$$\lim_{d\to\infty} p(y_0|\mathbf{x}_0, \mathcal{D}, H) = \delta_{y_0, y(\mathbf{x}_0|\mathcal{D}, H)} \tag{17}$$

but with different formulae for the assigned classes $y(\mathbf{x}|\mathcal{D}, H)$. For the generative model we find

$$\textit{generative} : y(\mathbf{x}|\mathcal{D}, H) = \mathrm{argmin}_{y \le c} \left\{ \log \hat{S}_y + \frac{(\mathbf{x} - \widehat{\mathbf{m}}_y)^2}{2\mathrm{d}\hat{S}_y^2} \right\} \tag{18}$$

$$\widehat{\mathbf{m}}_y = \langle \mathbf{x} \rangle_y \, \frac{nf_y\hat{\beta}_y^2}{nf_y\hat{\beta}_y^2 + \hat{\alpha}_y^2}, \quad \hat{S}_y^2 = \hat{\alpha}_y^2 \, \frac{\hat{\alpha}_y^2 + (nf_y + 1)\hat{\beta}_y^2}{\hat{\alpha}_y^2 + nf_y\hat{\beta}_y^2} \tag{19}$$

In contrast, for the more complicated discriminative case one may transform $v = \mathrm{d}\tilde{v}$ and use $\lim_{d\to\infty} \tilde{\mathcal{P}}(\tilde{v}) = \delta(\tilde{v} - 1)$, which follows from the law of large numbers, to establish that

$$\textit{discriminative} : y(\mathbf{x}|\mathcal{D}, H) = \mathrm{argmin}_{y\leq c}\left\{\log\hat{\alpha}_y + \frac{(\mathbf{x} - \widehat{\mathbf{m}}_y)^2}{2\mathrm{d}\hat{\alpha}_y^2} + \frac{1}{2}\frac{\hat{\beta}_y^2}{\hat{\alpha}_y^2 + nf_y\hat{\beta}_y^2}\right\} \tag{20}$$

Since $\lim_{n\to\infty}\hat{\beta}_y = X_y$ and $\lim_{n\to\infty}\hat{S}_y = \lim_{n\to\infty}\hat{\alpha}_y = \Sigma_y$, we observe that if both the covariate dimension d and the sample size n become large simultaneously, the generative and discriminative formulae for $p(y|\mathbf{x}, \mathcal{D}, H)$ become identical.

2.5 Overfitting and combined signature-based and variability-based class separation

When overfitting occurs, the estimated 'class signatures' $\widehat{\mathbf{m}}_y$ move away from the true centres $\boldsymbol{\mu}_y$ towards the sample means $\langle\mathbf{x}\rangle_y$. This is alleviated by the influence of the factor $nf_y\hat{\beta}_y^2/(nf_y\hat{\beta}_y^2 + \hat{\alpha}_y^2)$ in equation (10). If the signal in a class is strong, the signal-to-noise ratio $nf_y\hat{\beta}_y^2/\hat{\alpha}_y^2$ becomes large, and $\widehat{\mathbf{m}}_y$ and $\langle\mathbf{x}\rangle_y$ become approximately equal. However, the signal-to-noise ratio can become very small for classes y with weak signals. This drives $\widehat{\mathbf{m}}_y$ away from $\langle\mathbf{x}\rangle_y$ and towards $\boldsymbol{\mu}_y \approx \mathbf{0}$ for large d and fixed n. A typical example of this movement is the 'Occam's razor' effect which makes $\widehat{\mathbf{m}}_y$ exactly zero by estimating $\hat{\beta}_y = 0$. We calculate the relationship between $\widehat{\mathbf{m}}_y$ and $\langle\mathbf{x}\rangle_y$ in Appendix 2. We can use the result of equation (33) of this calculation to monitor how the location of $\widehat{\mathbf{m}}_y$ changes with increasing dimension d, for fixed $\alpha_y = 1$ and different β_y values, by evaluating the signature shrinkage ratio

$$\lambda = \sqrt{\langle\widehat{\mathbf{m}}_y^2\rangle_{p(X_y^2,\Sigma_y^2)}}\Big/\sqrt{\langle\langle\mathbf{x}\rangle_y^2\rangle_{p(X_y^2,\Sigma_y^2)}} \tag{21}$$

If the signal-to-noise ratio in the data is high, we will have $nf_y\beta_y^2/\alpha_y^2 \gg 1$ and $\lambda \approx 1$. Here the impact of the prior is negligible and our protocol approaches maximum likelihood regression. If the signal-to-noise ratio is low, or we have too few samples, we will find $nf_y\beta_y^2/\alpha_y^2 \ll 1$, and the ratio (21) generally shrinks to zero, especially for large d (see Figure 1). Thus, λ tells us exactly when the present model employs the Bayesian mechanism of balancing evidence against model complexity to combat overfitting.

Another typical phenomenon is observed upon choosing $c = 2$, $f_1 = f_2 = 1/2$, $\alpha_1 < \alpha_2$, and $\beta_1 = \beta_2 = 0$. Now it follows from equation (4) that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$, so the centres of the covariate distributions of the two classes are forced to be identical, and hence *any* classification strategy that is based on finding class signatures is fundamentally ruled out. For large d, the ratio (21) shrinks and can reduce to zero under the 'Occam's razor' effect (see Figure 1). One might naively expect that a vanishing estimated signature $\widehat{\mathbf{m}}_y$ would automatically cause inaccurate prediction. However, the squared length $\mathbf{x}^2$ of a new observation is distributed according to $\mathrm{Ga}(d/2, 2\alpha_y^2)$ (see Appendix 2), from which we deduce that the data are still easily separable for large d, but now on the basis of observed class covariate variability differences (see Figure 2(c)). The typical value of $\mathbf{x}^2$ differs between the two classes according to $d(\alpha_2^2 - \alpha_1^2)$ which is progressively greater than $\widehat{\mathbf{m}}_y^2$ as d becomes larger. For small d, the difference remains similar to $\widehat{\mathbf{m}}_y^2$, and we observe considerable overlap between the class distributions (see Figure 2(a)). For large d the Bayesian method will increasingly rely on variability-based class separation as opposed to separation based on covariate averages.

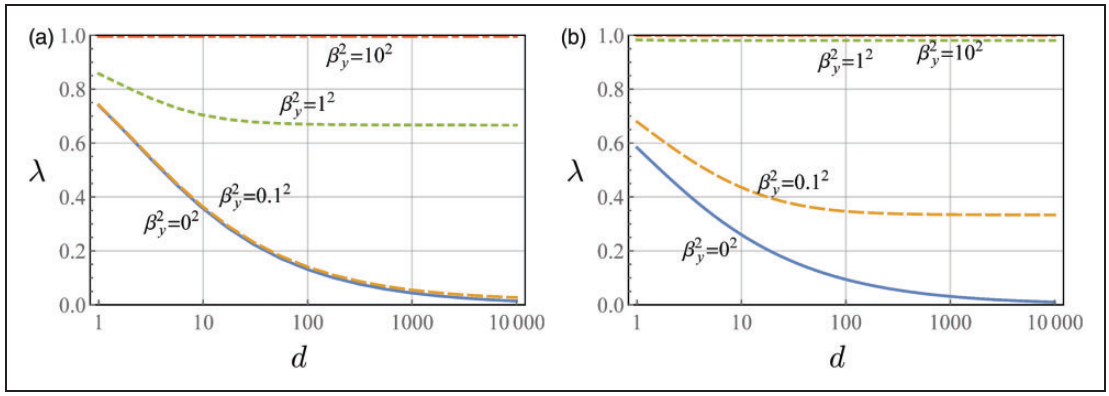

Figure 1. Dependence of the signature shrinkage ratio λ (21) on the covariate dimension d. The lines in each panel denote $\beta_y = 10$, 1, 0.1, and 0, from top to bottom, whereas α_y is fixed to unity. When $\beta_y = 0$ one has $\lambda = \sqrt{R}$. Left: $nf_y = 2$ (so there are only two samples in class y). Right: $nf_y = 50$ (so there are 50 samples in class y), here the signature shrinkage is less severe for nonzero values of β_y.

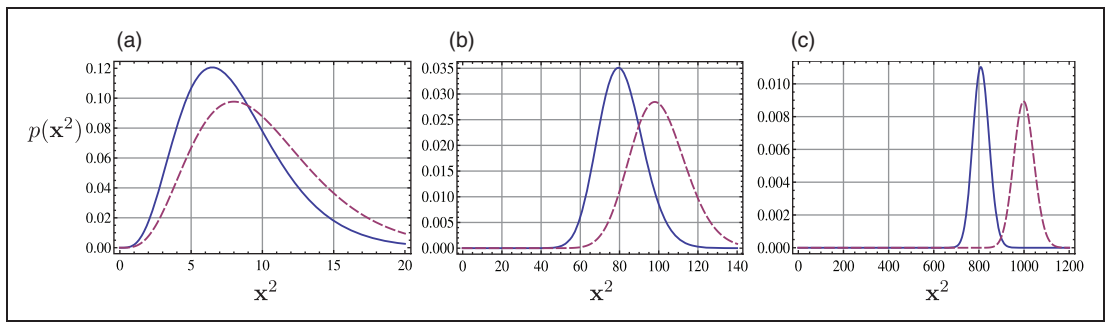

Figure 2. Statistics of covariate variability $\mathbf{x}^2$ in binary classification with $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$, but $\alpha_1^2 < \alpha_2^2$. (a) $d = 10$, where class-specific variability differences are still modest. (b) $d = 100$. (c) $d = 1000$, where the classes have become nearly perfectly separable on the basis of the observed value of $\mathbf{x}^2$. As the dimension d increases, class separation on the basis of variability becomes progressively more effective than separation on the basis of signatures alone (which in this example would be impossible, since the classes have identical signatures).

The solvability of our models allows us to understand the overfitting mechanism in detail. Each of the $\widehat{\mathbf{m}}_y$ and $\langle \mathbf{x} \rangle_y$ are random vectors, whose locations are difficult to predict for small d. However, for larger d each is found at its typical location. The typical location of the estimated signature vectors $\widehat{\mathbf{m}}_y$ can then be used to assess the border-hyperplane behaviour (see Figure 3(a)). We inspect, for instance, the case $c = 2, f_1 = f_2 = \frac{1}{2}$, $\alpha_1 = \alpha_2 = 1$, and $\beta_1 = \beta_2 > 0$. We assume for simplicity that shrinkage does not occur, i.e. we consider $nf_y\beta_y^2/\alpha_y^2 \gg 1$, so that $\widehat{\mathbf{m}}_y \approx \langle \mathbf{x} \rangle_y$. We also specify $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_1^2 = \boldsymbol{\mu}_2^2 = l^2(d)$. After a suitable rotation of the coordinate system in the space of $\mathbf{x}$, the direction of the first three dimensions is given by $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, $\widehat{\mathbf{m}}_1 - \widehat{\mathbf{m}}_2$, and $\widehat{\mathbf{m}}_1 + \widehat{\mathbf{m}}_2$. The first two dimensions are shown in Figure 3(a). The training observations are typically distributed with average

$$\widehat{\mathbf{m}}_y = (\pm\sqrt{l^2(d)},\ \pm\sqrt{(d-1)/n},\ \sqrt{(d-1)/n},\ 0,\ 0, \ldots) \tag{22}$$

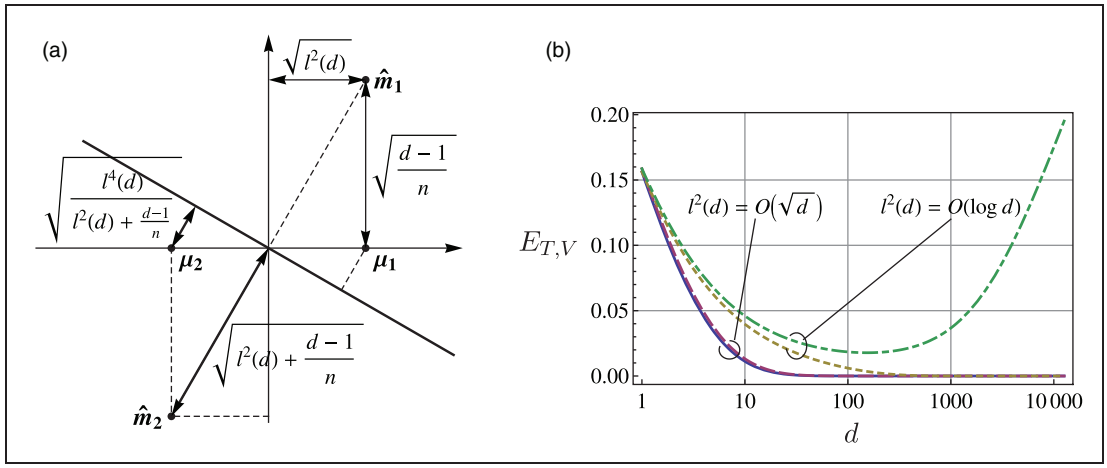

Figure 3. (a) location of the class means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and their Bayesian estimators $\widehat{\mathbf{m}}_1$ and $\widehat{\mathbf{m}}_2$. The thick line denotes the border-hyperplane which is perpendicular to the line $\widehat{\mathbf{m}}_2 - \widehat{\mathbf{m}}_1$. The formulae denote the average distances from the border-hyperplane. (b) classification errors E_T and E_V (fractions of misclassified samples) calculated from training and validation sets, for $l^2(d) = \mathcal{O}(\sqrt{d})$ (leftmost curves, specifically $\boldsymbol{\mu}_1 = (1^{-1/4}, 2^{-1/4}, 3^{-1/4}, \ldots, d^{-1/4})$, where we observe no overfitting) and for $l^2(d) = \mathcal{O}(\log d)$ (rightmost curves, specifically $\boldsymbol{\mu}_1 = (1^{-1/2}, 2^{-1/2}, 3^{-1/2}, \ldots, d^{-1/2})$, where overfitting occurs: the validation error is significantly higher than the training error).

and covariance matrix $\frac{n/2-1}{n/2}\,\mathbb{I}_{\mathrm{d}\times\mathrm{d}}$. According to equation (4), a new observation in the validation set is distributed with average $\pm\boldsymbol{\mu}_y$ and covariance matrix $\mathbb{I}_{\mathrm{d}\times\mathrm{d}}$. By considering the *Mahalanobis* distance[22] between an observation and the prediction border-hyperplane (see Figure 3(a)), and using the short-hand $\epsilon(a) = \frac{1}{2} - \frac{1}{2}\mathrm{erf}(a/\sqrt{2})$, the training and validation errors can then be written as

$$E_T = \epsilon\Big(\sqrt{[l^2(d) + (d-1)/n]/(1-2/n)}\Big), \quad E_V = \epsilon\Big(\sqrt{l^4(d)/[l^2(d) + (d-1)/n]}\Big) \tag{23}$$

When d is large and β_y of $\mathcal{O}(d^0)$ we have $\langle\, \boldsymbol{\mu}_y^2\rangle = \mathrm{d}\beta_y^2$, according to equation (4). It then follows that $l^2(d) \approx \mathrm{d}\beta_y^2 = \mathcal{O}(d)$. Hence both errors in equation (23) will be small, and the observations will be classified correctly in both the training and the validation set. If β_y^2 is of $\mathcal{O}(1/\sqrt{d})$, then $l^2(d) = \mathcal{O}(\sqrt{d})$. Now the observations in training and validation sets will still be separable (see the two curves on the left in Figure 3(b)), despite the equivalent β_y value converging to zero, $\lim_{d\to\infty}\beta_y = 0$. If β_y^2 is of $\mathcal{O}(d^{-1}\log(d))$, however, then $l^2(d) = \mathcal{O}(\log(d))$. Here the training observations are still separable but the observations in the validation set will no longer be so for large d, see the right two curves in Figure 3(b), in spite of the fact that the distance between the two true class centres $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ diverges. The border-hyperplane behaviour in the case where signature shrinkage occurs, i.e. for $nf_y\beta_y^2/\alpha_y^2 \ll 1$ and $\widehat{\mathbf{m}}_y \approx 0$, warrants further investigation, which we consider to be beyond the scope of this paper.

3 Application to synthetic data

We have applied our methods and *mclustDA* first to synthetic data sets. Each set consisted of $n = 100$ covariate vectors and their classifications, generated according to equation (3). We increased d on a logarithmic scale up to $d = 10{,}000$. The largest d value, although typical of biomedical data,[5–11]

cannot be handled by *mclustDA* and underlines the merit of analytically integrable models. The characteristics of our data are shown in Table 1. Classification of sets A1 and A2 was tested using leave-one-out cross-validation (LOOCV), and of sets B1, B2, and C1 using a training set (T, $n = 100$) and a validation set (V, $n = 100$). The curves E_T and E_V in Figure 4 give the fractions of incorrect classifications, all averaged over 100 randomly generated data sets.

The performance of Methods I and II observed in Figure 4(a) and (b) can be understood using Figure 2. The overlap between the covariate distributions of the classes results for small d in high error rates, but for large d separation on intra-class covariate variability differences become increasingly effective. Both Bayesian methods also suffer less from overfitting (marked by a gap between E_T and E_V) than *mclustDA*. The unsupervised clustering in *mclustDA* struggles to separate the observations in sets A1 and A3, since the true class centres are identical. Method I would have struggled too if the sample means $\langle \mathbf{x} \rangle_y$ had been used instead of the estimators $\widehat{\mathbf{m}}_y$ in equation (7). The difference between data sets B1 and B2, whose classification results are shown in Figure 4(c) and (d), is in the distance between the true class centres $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, which in B1 remains finite, but diverges as $d \to \infty$ in B2. The curves in Figure 4(d) are reminiscent of the $I^2(d) = \mathcal{O}(\log d)$ curves in Figure 3(b). Figure 4(c) is similar, except that here the validation error E_V approaches the random guessing level 0.5 sooner. This is because $I^2(d)$ in set B1 is finite, i.e. more difficult to separate: $\lim_{d\to\infty} I^2(d) = \pi^2/6$, whereas $\lim_{d\to\infty} I^2(d) = \infty$ in B2. *mclustDA* performs similarly to Method I for data sets B1 and B2 for $d \leq 3,000$. Its unsupervised hierarchical clustering can learn the characteristics of the observations in the training set without using the class labels y, because the class centres are well separated. Replacing the standard setting 'MclustDA' of the R-implementation of *mclustDA* by 'EDDA' leads for data sets A1 and A3 to further performance deterioration, whereas there is no significant change for data sets B1 and B2.

In Figure 4(e) and (f) we show results on data where the training and validation sets have different class frequencies. Here we expect any method that assumes the training data to be representative of the population in terms of class frequencies to perform badly. This is indeed the case for Method I and *mclustDA*, which both use the 10%/90% class imbalance of the training set to predict labels in

Table 1. Characteristics of our synthetic data.

Data set	n	f_1	f_2	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	α_1	α_2
A1	100	0.5	0.5	$(0,\ldots,0)$	$(0,\ldots,0)$	0.24	0.28
A2	100	0.5	0.5	$(1,\ldots,1,\frac{1}{2},\ldots,\frac{1}{2},0,\ldots,0)$	$(1,\ldots,1,\frac{1}{2},\ldots,\frac{1}{2},0,\ldots,0)$	0.24	0.28
B1 (T,V)	100	0.5	0.5	$(-1,-\frac{1}{2},-\frac{1}{3},\ldots,-\frac{1}{d})$	$(1,\frac{1}{2},\frac{1}{3},\ldots,\frac{1}{d})$	1.00	1.00
B2 (T,V)	100	0.5	0.5	$(-1,-\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{3}},\ldots,-\frac{1}{\sqrt{d}})$	$(1,\frac{1}{\sqrt{2}},\frac{1}{\sqrt{3}},\ldots,\frac{1}{\sqrt{d}})$	1.00	1.00
C1 (T)	100	0.1	0.9	$(0,\ldots,0)$	$(0,\ldots,0)$	0.24	0.28
C1 (V)	100	0.9	0.1	$(0,\ldots,0)$	$(0,\ldots,0)$	0.24	0.28

Data set	n	f_1	f_2	f_3	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_3$	α_1	α_2	α_3
A3	100	0.33	0.33	0.34	$(0,\ldots,0)$	$(0,\ldots,0)$	$(0,\ldots,0)$	0.24	0.26	0.28

Top table: sets with two classes. Data for which training and validation sets have identical characteristics are indicated with (T,V). A1: classes have identical covariate averages, centred in the origin, but distinct intra-class variability. A2: classes have identical covariate averages, with a mixture of zero and nonzero components (values 0, $\frac{1}{2}$ and 1, equally distributed over the d entries), but distinct intra-class variability. B1 and B2: distinct class averages but identical covariate variability. C1: identical class averages and distinct variability, but now the training and validation sets differ in the imbalance of class membership. Bottom table, A3: data with three classes. Here the class averages are identical, but the covariate distributions of the classes have different widths.

the validation set, where the class imbalance is in fact 90%/10%. In contrast, Method II, which disregards class frequency information altogether, does not suffer from this mismatch. In large dimensions, both Method I and Method II will rely increasingly on intra-class variability to classify samples (see Figure 2), whereas *mclustDA* continues to suffer from inappropriate extrapolation of the class imbalance of the training set to the validation set. To test our interpretation of the above performance curves we applied modifications of our methods to data C1 and show the results in Figure 4(f). In Method I we replaced the estimators $\widehat{\mathbf{m}}_y$ by the sample means $\langle \mathbf{x} \rangle_y$, to inhibit the Bayesian switch from signature-based to variability-based classification (and now overfitting indeed persists for large d), we replaced Methods IIa and IIb by their large n asymptotic form (Method n) to suppress the beneficial regularising effect of the hyperparameters (and now overfitting sets in), and we imposed a uniform class balance $f_y = 1/c$ in running *mclustDA* on the validation set (see *mclustDA*u in Figure 4(f)), which reduces overfitting for small d, although performance remains poor for large d. Replacing the standard setting 'MclustDA' of the R-implementation of *mclustDA* by 'EDDA' (results not presented here) leads for data set C1 to further performance deterioration, whereas it improves performance for small dimensions $d \leq 100$ when we impose uniform class balance (as was previously done in Figure 4(e)). Figure 5 shows the result of analysing synthetic data A2 in which the class centres combine zero and nonzero components. Here all three methods considered perform very similarly, with *mclustDA* making slightly more validation errors for larger covariate dimensions (with the usual proviso that *mclustDA* cannot be used for the largest d values).

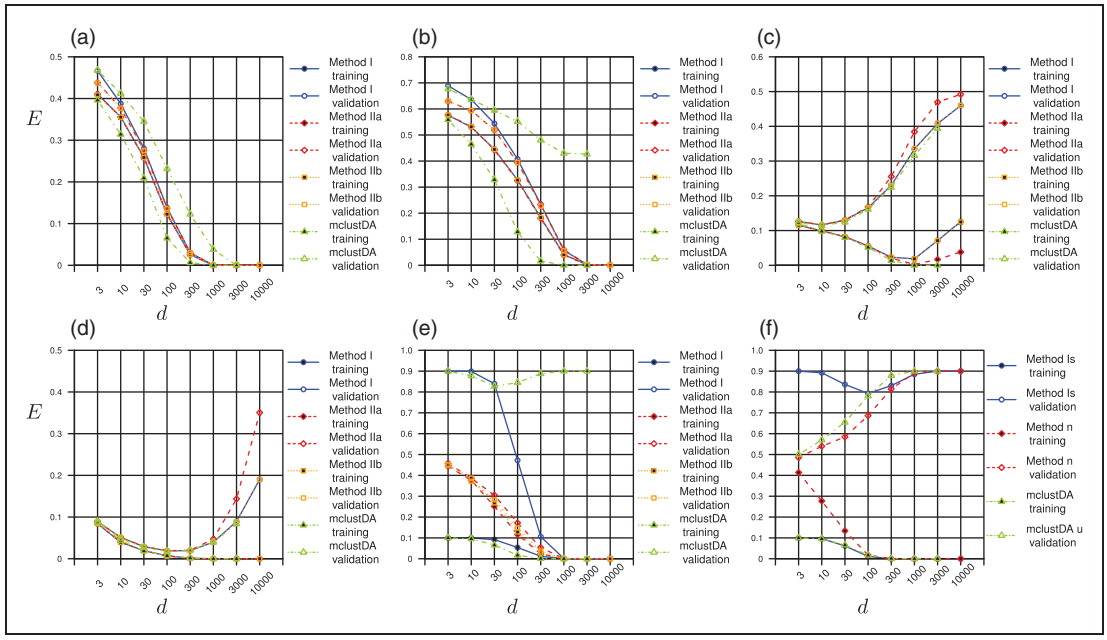

Figure 4. Training and validation errors, E_T and E_V, for the synthetic data described in Table 1, for different covariate dimensions d. All data sets were analysed via our generative method, Method I (9,12), our discriminative method, Method II (15,12) (with $f_y = 1/c$, called Method IIa), the large d approximation (20) of Method II (called Method IIb), and with *mclustDA*. Note that (f) refers to the same data as (e), but in (f) we applied modified versions of some of our methods (see the main text and Table 1 for motivation and details).

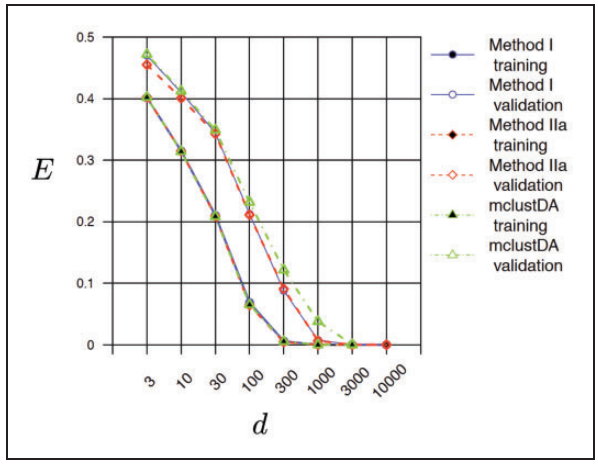

Figure 5. Training errors E_T (filled markers) and LOOCV validation errors E_V (open markers), for data A2 (see Table 1) with different dimensions d, upon analysis via method, Method I (9,12) (blue curves), the large d approximation (20) of Method II (red curves), and *mclustDA* (green curves).

Table 2. Dependence on the covariate dimension d of typical CPU demands (in minutes) required for LOOCV analysis of the data in Figure 4(a), measured for single-processor runs on a standard UNIX workstation. For $d = 10{,}000$ *mclustDA* can no longer be run due to its excessive computation requirements. For smaller values of $d < 300$ all methods use negligible computation time.

d	Method I	Method IIb	mclustDA
300	0.026	0.029	1.136
1000	0.044	0.047	8.498
3000	0.092	0.074	96.200
10000	0.213	0.182	NA

Table 2 shows the typical computation demands of some of the methods tested in Figure 4. Each value is an average over 100 runs. Due to their simplicity and analytical integrability, the Bayesian methods I and IIb considered here have no problems in processing high-dimensional data sets, unlike *mclustDA*. To have a meaningful comparison of CPU demands we provided *mclustDA* for each class with the correct type of covariance matrix (isotropic) and the correct number of components (one).

4 Applications to cancer data with genomic covariates

We applied the various classification methods to the data set *exprset2* of De Rinaldis et al.,[11] which contains gene expression profiles of triple-negative breast cancer patients who were treated at Guy's and St Thomas's Hospitals (London) between 1979 and 2001, and who had at least 5.5 years of follow-up. Expression levels were recorded for $d = 22{,}035$ genes, and patients with missing data or who were lost to follow up were excluded, leaving $n = 165$ for this study. Patients who survived for at

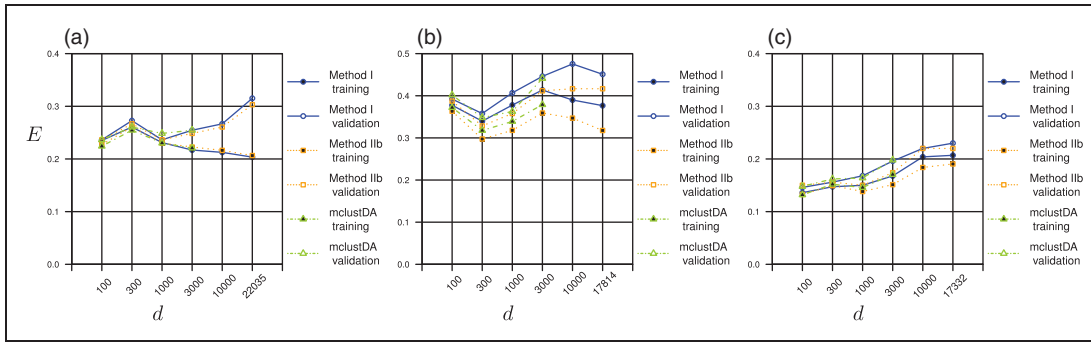

Figure 6. Training and validation errors, E_T and E_V, for three high-dimensional gene expression data sets of breast and ovarian cancer patient cohorts. Validation errors were measured via leave-one-out cross-validation (LOOCV). (a) Triple-negative Breast cancer ($n = 165$ samples), (b) TCGA Ovarian cancer ($n = 204$ samples), and (c) TCGA Breast cancer ($n = 500$ samples).

least five years from initial diagnosis are designated as class $y = 2$ ('good' prognosis, a fraction $f_2 = 0.75$), and patients who died from breast cancer within five years are as class $y = 1$ ('poor' prognosis, a fraction $f_1 = 0.25$). There is no censoring due to competing events.[43] Our aim is to predict a patient's prognosis class from their gene expressions. To apply *mclustDA* to *exprset2*, we needed to reduce the dimensionality, since *mclustDA* cannot handle $d = 22{,}035$. Assuming that genes with greater correlation with outcome are more likely to predict clinical outcome, we followed[12,13] and ranked the genes according to their Pearson correlation with outcome. The highest ranked 100, 300, 1000, 3000, and 10,000 genes were chosen, and to these sets the various classification methods were applied. Validation performance was measured via LOOCV. Figure 6(a) shows that the optimal predictive information resides in the first 1000 ranked genes, and that all methods give very similar results. However, the Bayesian methods can handle much larger d values and confirm in doing so that no relevant information is lost by limiting oneself to the top 1000 genes. Moreover, Method IIb (20) suffers much less from overfitting than *mclustDA*. The optimal validation error E_V is approximately 0.24. Since one can already achieve $E_V = 0.25$ by assigning any new sample simply by default to the largest class $y = 2$, we conclude that the gene expression measurements of *exprset2* either confer only a modest amount of predictive information on five-year survival from triple-negative breast cancer after treatment, or all methods considered fail to match the structure of the data.

Next we use level 3 gene expression data.[a] from The Cancer Genome Atlas (TCGA) ovarian carcinoma cohort.[8] This set consists of $n = 204$ patients and expression levels for 17,814 genes. As outcome variable we use the mutation status for the *TP53*, *BRCA1*, and *BRCA2* genes. Patients are classified according to whether germline or somatic mutations are present in *TP53* and one or more of *BRCA1*, and *BRCA2* ($y = 3$), just *TP53* ($y = 2$), or in none of these three genes ($y = 1$). The relative class sizes are $f_1 = 0.04$, $f_2 = 0.73$, and $f_3 = 0.23$. The question is whether the gene expression profiles of ovarian carcinoma patients can predict the mutation status of three common and known drivers of ovarian tumour genesis.[8] Due to various reasons, including data availability, it is not always possible to directly assess whether a specific gene is mutated. It would therefore be useful to have a surrogate means of predicting mutations, for instance by deriving a gene expression signature that can be used as a substitute. This has indeed been the rationale of several studies, such as Miller et al.[44] and Bernardini et al.[45] (for TP53) or van't Veer et al.,[12] Konstantinopoulos et al.,[46] Press et al.[47] (for BRCA genes). Here the covariate genes were ranked as described for the

previous data set. Figure 6(b) shows that all methods agree that for this data set the optimal predictive information is in the first 300 ranked genes, with Method IIb giving the best validation error rate $E_V \approx 0.33$. However, the error rates are again not impressive, given that $E_V = 0.27$ can be achieved by simply assigning all samples to the largest class $y = 2$. The situation here is in fact worse than with the previous data set, which suggests again that all approaches considered suffer from model mismatch or there is simply no information in these gene expression data to predict the chosen outcome variable.

As a third medical application we use level 3 gene expression data from TCGA breast carcinoma cohort.[9] It consists of $n = 500$ patients with their expression levels for 17,332 genes. As our outcome variable y we now use the dichotomised status for estrogen receptor (ER) and HER2 immunohistochemistry-derived expression. Patients are either ER negative and HER2 negative ($y = 1$), ER positive and HER2 negative ($y = 2$), ER negative and HER2 positive ($y = 3$), or positive for both ER and HER2 ($y = 4$), with relative class sizes $f_1 = 0.19$, $f_2 = 0.66$, $f_3 = 0.04$, and $f_4 = 0.11$. The question is to ascertain whether the gene expression profiles of breast cancer patients can predict the immunohistochemical status of their tumours. The gene expression covariates were ranked as before. Figure 6(c) shows that all methods agree that the optimal predictive information is in the first 100 ranked genes, and all produce an optimal validation error rate $E_V \approx 0.14$. In contrast to the previous examples, this result is significant, since assigning all samples to the largest class would here have given an average validation error of $E_V = 0.34$. We conclude that gene expression profiles of breast cancer patients are reliable predictors of their ER and HER2 status.

In Figure 7 we show the Pearson correlation coefficients between the gene expressions and the clinical outcome variable y, for the above three cancer data sets. The absolute values of these correlations (i.e. the distances from zero) were used to rank the genes.

5 Applications to cancer data with imaging covariates

Finally, we illustrate the application of our method to the problem of how to classify tissue types from imaging data. The task is to differentiate benign from malignant breast tissue, with a view to

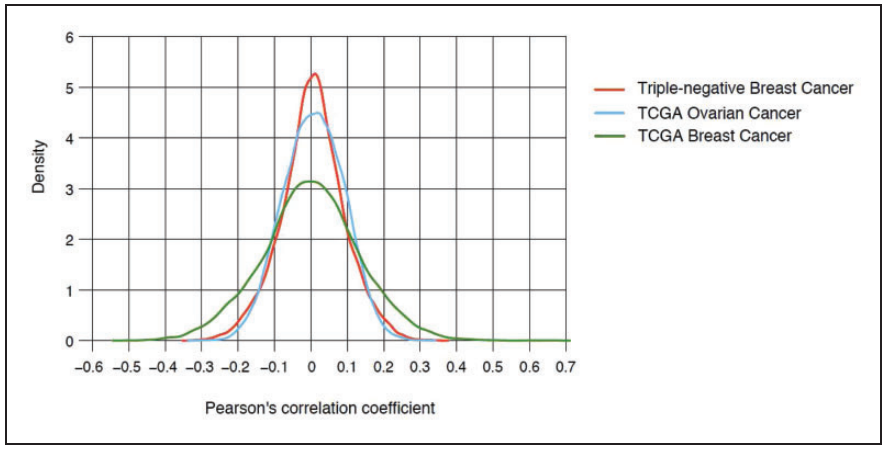

Figure 7. Densities of Pearson correlation coefficients between individual gene expressions and the outcome variable *y*, for the three genomic cancer data sets analysed in this paper.

reducing re-operation rates in breast conserving surgery, based on waveform data obtained from a handheld Terahertz pulsed imaging device.[48,b] Existing methods for intra-operative tumour margin assessment are either accurate but slow (e.g. off-line cytological analysis) or fast but inaccurate (e.g. specimen radiography, with 40–60% sensitivity and 70–90% specificity). Our data set consisted of $n = 100$ breast tissue samples from three classes (see Table 3 for details). The wavelet expansion representation replaces the original waveforms $f(x)$, sampled at 301 equidistant points, with the following convolutions.[c]

$$F_r(\sigma_i, x_j) = \sigma_i^r \int \frac{\mathrm{d}y}{\sigma_i\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(x_j-y)^2/\sigma_i^2} \frac{\mathrm{d}^r}{\mathrm{d}y^r} f(y), \qquad r = 0, 1, 2, 3, 4 \tag{24}$$

The wavelet centres x_j are the middle 100 sampled values of the x-axis in the waveforms (vertical axes in heat maps). The wavelet scales are $\sigma_i = 0.3i$ for $i = 1 \ldots 100$ (horizontal axes in heat maps). Each heat map represents 10,000 values, so using all five heat maps gives a combined covariate dimension $d = 50{,}000$. One would normally already hesitate, in view of overfitting, to attempt tissue classification from the Terahertz waveforms, given the unfavourable ratio $n/d \approx 0.332$. One would advise strongly against using the wavelet expansion, where $n/d \approx 0.002$. However, we see in Table 3 that the Bayesian discriminant analysis protocol (9) exhibits a highly significant classification performance, which in fact improves upon switching to the high-dimensional wavelet representation (further reduction of training and validation errors by roughly a factor two). When focused on tumour detection, the method gives 85% sensitivity and 82% specificity. Since computation time is negligible and performance is promising, this method would appear suitable for intra-operative tumour margin assessment.

6 Discussion

Following the literature[37–41] we have studied Bayesian clinical outcome prediction methods in which the parameter integrals can be done analytically, so that they can handle data with arbitrary numbers of classes and large covariate dimensions. We thereby avoid the computational hurdles normally associated with Bayesian methods in large dimensions (e.g. $d = 20{,}000$ or more as in genomic data sets[11]), without sacrificing possibly valuable information by prior dimension reduction, and we can explore important aspects of discriminant analysis in high dimensions.

Table 3. Classification of tissue types based on Terahertz imaging data. The data set consisted of n = 100 breast tissue samples, all scored via histological analysis and assigned to one of three classes: tumour (class $y = 1$, fraction $f_1 = 0.283$), fibrous (class $y = 2$, fraction $f_2 = 0.457$), and adipose (class $y = 3$, fraction $f_3 = 0.261$). Our method (9) was applied to the original Terahertz waveform data (left column in Figure 8), and to a high-dimensional multi-scale wavelet expansion[49] of the waveforms (heat maps in Figure 8). Validation errors were calculated via LOOCV. Naive assignment to the largest class would give the baseline performance $E = 0.543$, so the observed classification performance is highly significant.

Data ($n = 100$)	d	E_T	E_V
Original Terahertz wave forms	301	0.153	0.174
Multi-scale wavelet expansion	50,000	0.056	0.087

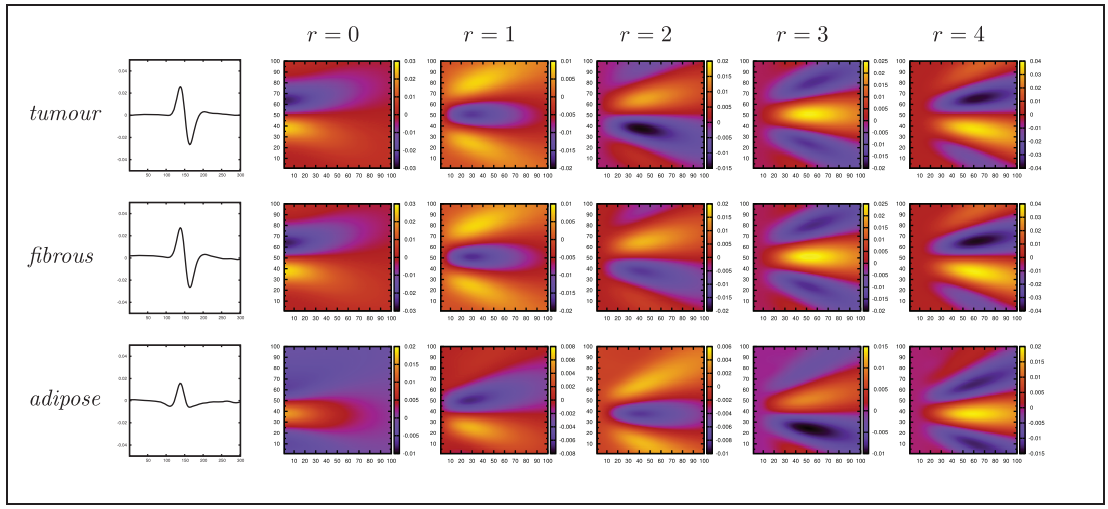

Figure 8. Typical examples of imaging data of three breast tissue types. Left: waveforms obtained from a handheld pulsed Terahertz imaging device, giving $d = 301$ covariate values. Heat maps: multi-scale wavelet expansion[49] of the waveforms, of orders $r = 0, 1, 2, 3, 4$ (see main text). The horizontal axis in each heat map gives the wavelet scale (100 equidistant values), and the vertical coordinate gives the wavelet centre (100 equidistant values). The resulting wavelet representation of the tissue data, where $d = 50{,}000$, rules out the use of *mclustDA* due to prohibitive CPU demands. Clearly, the main difficulty is to distinguish between tumour samples and fibrous samples.

We showed how for large d values Bayesian discriminative methods can exploit intra-class variability, as opposed to differences between the average signals of the different classes as captured by 'signatures'. We compared the integrable Bayesian methods to *mclustDA*,[14,16] which is also built on Bayesian principles and widely applied to medical data.[27–31] The analytically integrable Bayesian methods outperform *mclustDA* significantly in CPU demands, since *mclustDA* estimates more parameters through the use of EM, and *mclustDA* becomes computationally infeasible for $d \approx 10{,}000$ or more. Application to synthetic data showed that for modest dimensions d our methods perform in similar manner to *mclustDA*, but they generally outperform the latter in giving lower validation errors and less overfitting for larger d. Moreover, version II of our methods is significantly more robust against mismatch in relative class sizes between training and validation sets. We have also tested our methods on synthetic data sets with alternative covariate statistics (e.g. discrete covariates, these experiments are not described in this paper), and we have found no significant deterioration in performance.

We next used the various outcome prediction methods to analyse three genomic cancer data sets. In the triple-negative breast cancer data set *exprset2* (where we sought to predict five-year survival) and the TCGA ovarian cancer data set (where we sought to predict oncogene mutations), none of the methods succeeded in achieving nontrivial prediction performance. In contrast, for the TCGA breast cancer data set (where we seek to predict ER and HER2 receptor status), all methods did exhibit statistically significant prediction performance. We also applied our methods to the task of identifying breast tissue types (specifically cancer tissue) from Terahertz imaging data, sufficiently accurately and fast to support intra-operative tumour margin assessment. Despite a very poor ratio $n/d = 0.002$, the most informative multi scale waveform expansion of the data, with $d = 50{,}000$

(so *mclustDA* cannot be used), is found to allow for very precise classification, with hardly any overfitting.

The advantage of *mclustDA* over our analytically integrable models is that it can handle data sets with multimodal class-specific covariate distributions. While there will certainly be situations where this allows *mclustDA* to outperform unimodal discriminant analysis models, in the present clinical applications we did not observe this computationally costly added flexibility of *mclustDA* translating into significant outcome prediction benefit. Many other model-based methods exist that try to overcome the computational limitations of discriminant analysis in high dimensions.[19,21] These include variable selection steps[30,50] and combining subspace clustering with constrained and parsimonious models.[17,21] However, all these methods are expected to suffer from loss of information and/or underestimated uncertainty at parameter level.

The Bayesian routes proposed in this paper can be extended in several ways, without sacrificing the key analytical integrability. The obvious one is to allow for more complicated covariance matrices in the class-specific distributions $p(\mathbf{x}|\, y)$, as in the literature[37–41] (using Wishart priors). Another direction is to investigate the limit $c \to \infty$, where we may be able to predict real-valued outcomes (such as time to relapse) from high-dimensional data.

Acknowledgements

We are grateful to Maarten Grootendorst and Aida Santaolalla for discussions on tissue classification for intra-operative tumour margin assessment and for providing us with Terahertz tissue imaging data.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors gratefully acknowledge support from the Engineering and Physical Sciences Research Council (UK), IDBS, and the Ana Leaf Foundation.

Notes

a. These are defined as the aggregate of processed data from single sample or in some cases grouped by probed loci to form larger contiguous regions (https://wiki.nci.nih.gov/display/TCGA/Data + level).
b. This device has a spatial resolution of around 15 mm and requires 20 s for generating a sample.
c. Waveform derivatives are discretised as usual: $2f^{(1)}(x_i) = f(x_{i+1}) - f(x_{i-1})$, $4f^{(2)}(x_i) = f(x_{i+2}) + f(x_{i-2}) - 2f(x_i)$, $8f^{(3)}(x_i) = f(x_{i+3}) - f(x_{i-3}) - 3f(x_{i+1}) + 3f(x_{i-1})$, and $16f^{(4)}(x_i) = f(x_{i+4}) + f(x_{i-4}) - 4f(x_{i+2}) - 4f(x_{i-2}) + 6f(x_i)$.

References

1. Hastie T and Tibshirani R. Discriminant analysis by Gaussian mixtures. *J R Stat Soc Ser B* 1996; **58**: 155–176.
2. Ripley BD. *Pattern recognition and neural networks*. Cambridge: Cambridge University Press, 1996.
3. Duda RO, Hart PE and Stork D. *Pattern classification*, 2nd ed. New York: Wiley, 2001.
4. McLachlan GJ, Peel D and Bean RW. Modelling high-dimensional data by mixtures of factor analyzers. *Comput Stat Data Anal* 2003; **41**: 379–388.
5. Clarke R, Ressom HW, Wang A, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* 2008; **8**: 37–49.

6. Michiels S, Kramar A and Koscielny S. Multidimensionality of microarrays: statistical challenges and (im)possible solutions. *Mol Oncol* 2011; **5**: 190–196.
7. Wang Y, Miller DJ and Clarke R. Approaches to working in high-dimensional data spaces: gene expression microarrays. *Br J Cancer* 2008; **98**: 1023–1028.
8. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011; **474**: 609–615.
9. The Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012; **490**: 61–70.
10. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012; **483**: 603–307.
11. de Rinaldis E, Gazinska P, Mera A, et al. Integrated genomic analysis of triple-negative breast cancers reveals novel microRNAs associated with clinical and molecular phenotypes and sheds light on the pathways they control. *BMC Genom* 2013; **14**: 643.
12. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**: 530–536.
13. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002; **347**: 1999–2009.
14. Fraley C and Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc* 2002; **97**: 611–631.
15. Jolliffe IT. *Principal component analysis*, 2nd ed. New York: Springer, 2002.
16. Fraley C, Raftery AE, Murphy TB, et al. *mclust Version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation*. University of Washington (USA): Technical Report No 597, 2012.
17. Bouveyron C, Girard S and Schmid C. High-dimensional discriminant analysis. *Commun Stat Theory Methods* 2007; **36**: 2607–2623.
18. Scott DW and Thompson JR. Probability density estimation in higher dimensions. In: *Proceedings of the fifteenth symposium on the interface on computer science and statistics*. Vol. 528. Amsterdam: North-Holland, 1983, pp.173–179.
19. McNicholas PD. On model-based clustering, classification, and discriminant analysis. *J Iran Stat Soc* 2011; **10**: 181–199.
20. Chang WC. On using principal components before separating a mixture of two multivariate normal distributions. *J R Stat Soc Ser C (Appl Stat)* 1983; **32**: 267–275.
21. Bouveyron C and Brunet-Saumard C. Model-based clustering of high-dimensional data: a review. *Comput Stat Data Anal* 2014; **71**: 52–78.
22. Bishop CM. *Pattern recognition and machine learning*. New York: Springer, 2006.
23. Berge L, Bouveyron C and Girard S. HDclassif: an R package for model-based clustering and discriminant analysis of high-dimensional data. *J Stat Softw* 2012; **46**: 1–29.
24. Dasgupta A and Raftery AE. Detecting features in spatial point processes with clutter via model-based clustering. *J Am Stat Assoc* 1998; **93**: 294–302.
25. Fraley C and Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput J* 1998; **41**: 578–588.
26. R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2013.
27. Dean N, Murphy TB and Downey G. Using unlabelled data to update classification rules with applications in food authenticity studies. *J R Stat Soc Ser C (Appl Stat)* 2006; **55**: 1–14.
28. Fraley C and Raftery AE. Model-based methods of classification: using the mclust software in chemometrics. *J Stat Softw* 2007; **18**: 1–13.
29. Iverson AA, Gillett C, Cane P, et al. A single-tube quantitative assay for mRNA levels of hormonal and growth factor receptors in breast cancer specimens. *J Mol Diagn* 2009; **11**: 117–130.
30. Murphy TB, Dean N and Raftery AE. Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *Ann Appl Stat* 2010; **4**: 396–421.
31. Andrews JL and McNicholas PD. Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Stat Comput* 2012; **22**: 1021–1029.
32. Banfield JD and Raftery AE. Model-based Gaussian and non-Gaussian clustering. *Biometrics* 1993; **49**: 803–821.
33. Dempster AP, Laird NM and Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B* 1977; **39**: 1–38.
34. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978; **6**: 461–464.
35. Wehrens R, Buydens LMC, Fraley C, et al. Model-based clustering for image segmentation and large datasets via sampling. *J Classif* 2004; **21**: 231–253.
36. Fraley C, Raftery AE and Wehrens R. Incremental model-based clustering for large datasets with small clusters. *J Comput Graph Stat* 2005; **14**: 529–546.
37. Geisser S and Cornfield J. Posterior distributions for multivariate normal parameters. *J R Stat Soc B* 1963; **25**: 368–376.
38. Keehn DG. A note on learning for Gaussian properties. *IEEE Trans Inform Theory* 1965; **11**: 126–132.
39. Han X, Wakabayashi T and Kimura F. The optimum classifier and the performance evaluation by bayesian approach. In: Ferri FJ, et al. (eds) *Advances in Pattern Recognition*. Heidelberg: Springer, 2000, pp.591–600.
40. Srivastava S and Gupta MR. Distribution-based Bayesian minimum expected risk for discriminant analysis. In: *Proceedings of the IEEE International Symposium on Information Theory*, 2006, pp.2294–2298.
41. Srivastava S, Gupta MR and Frigyik BA. Bayesian quadratic discriminant analysis. *J Mach Learn Res* 2007; **8**: 1277–1305.
42. Bellman R. *Dynamic programming*. Princeton: Princeton University Press, 1957.
43. Klein JP and Moeschberger ML. *Survival analysis techniques for censored and truncated data*. New York: Springer, 2005.
44. Miller LD, Smeds J, George J, et al. An expression signature for p53 status in human breast cabcer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci USA* 2005; **102**: 13550–13555.
45. Bernardini MQ, Baba T, Lee PS, et al. Expression signatures of TP53 mutations in serious ovarian cancers. *BMC Cancer* 2010; **10**: 237.
46. Konstantinopoulos PA, Spentzos D, Karlan BY, et al. Gene expression profile of BRCAness that correlates with responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer. *J Clin Oncol* 2010; **28**: 3555–3561.
47. Press J, Wurz K, Norquist BM, et al. Identification of a preneoplastic gene expression profile in tubal epithelium of BRCA1 mutation carriers. *Neoplasia* 2010; **12**: 993–1002.

48. George DK, Charkhesht A and Vinh NQ. New terahertz dielectric spectroscopy for the study of aqueous solutions. Epub ahead of print 2015.
49. Koenderink JJ. *Solid shape*. Boston: MIT Press, 1990.
50. Raftery AE and Dean N. Variable selection for model-based clustering. *J Am Stat Assoc* 2006; **101**: 168–178.
51. Gradshteyn IS and Ryzhik IM. *Table of integrals, series, and products*, 7th ed. Burlington MA: Academic Press, 2007.

Appendix 1: Calculation of $P(v)$

Here we calculate the following integrals for $d \geq 2$, required in the evaluation of equation (16)

$$\mathcal{P}(v) = \int \left(\prod_{j=1}^{d-1} \frac{\mathrm{d}u_j \mathrm{e}^{-\frac{1}{2}u_j^2}}{\sqrt{2\pi}} \right) \delta \left[v - \sum_{j=1}^{d-1} u_j^2 \right] \tag{25}$$

For $d=2$ and $d=3$ they are easy

$$d = 2: \qquad \mathcal{P}(v) = \int_{-\infty}^{\infty} \frac{\mathrm{d}u}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}u^2} \, \delta(v - u^2) = \frac{\mathrm{e}^{-\frac{1}{2}v}}{(2\pi)^{3/2}\sqrt{v}} \tag{26}$$

$$d = 3: \qquad \mathcal{P}(v) = \int_{0}^{\infty} \mathrm{d}r \, r\mathrm{e}^{-\frac{1}{2}r^2} \, \delta(v - r^2) = \frac{1}{2}\mathrm{e}^{-\frac{1}{2}v} \tag{27}$$

For $d>3$ we first write the delta distribution in integral form, and find after some simple manipulations

$$\mathcal{P}(v) = \int_{-\infty}^{\infty} \frac{\mathrm{d}\hat{v}}{2\pi} \mathrm{e}^{\mathrm{i}\hat{v}v} \left(\int_{-\infty}^{\infty} \frac{\mathrm{d}u}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}u^2(1-2\mathrm{i}\hat{v})} \right)^{d-1} = \int_{-\infty}^{\infty} \frac{\mathrm{d}\hat{v}}{2\pi} \frac{\mathrm{e}^{\mathrm{i}\hat{v}v}}{(1+2\mathrm{i}\hat{v})^{\frac{d-1}{2}}} \tag{28}$$

With the substitution $\hat{v} = \frac{1}{2}\tan\phi$ this can be simplified to

$$\mathcal{P}(v) = \int_{0}^{\pi/2} \frac{\mathrm{d}\phi}{2\pi} \, (\cos\phi)^{\frac{d-5}{2}} \cos\left(\frac{1}{2}v\tan\phi - \frac{1}{2}(d-1)\phi \right) \tag{29}$$

The latter integral can be found in Gradshteyn and Ryzhik[51] (on page 423), which leads us to the simple formula of equation (16)

$$\mathcal{P}(v) = \frac{\left(\frac{1}{2}v\right)^{\frac{1}{2}(d-3)} \mathrm{e}^{-\frac{1}{2}v}}{2\Gamma\left(\frac{1}{2}(d-1)\right)} \tag{30}$$

This is a chi-squared distribution with $d-1$ degrees of freedom. It includes the above cases $d=2$, 3. For numerical evaluation it is convenient to exploit the properties $\int \mathrm{d}v \, \mathcal{P}(v)v = d-1$ and $\int \mathrm{d}v \, \mathcal{P}(v)v^2 = (d-1)(d+1)$, and introduce a zero-average and unit-variance integration variable $\tilde{v}$ via $v = d-1+\tilde{v}\sqrt{2(d-1)}$, giving

$$\mathcal{P}(\tilde{v}) = \frac{\sqrt{\frac{1}{2}(d-1)}}{\Gamma(\frac{1}{2}(d-1)} \left(\frac{1}{2}(d-1) + \tilde{v}\sqrt{\frac{1}{2}(d-1)} \right)^{\frac{1}{2}(d-3)} \mathrm{e}^{-\frac{1}{2}(d-1)-\tilde{v}\sqrt{\frac{1}{2}(d-1)}} \tag{31}$$

Note that $\lim_{d\to\infty} \mathcal{P}(\tilde{v}) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}\tilde{v}^2}$. This follows from the definition of $\mathcal{P}(v)$ in equation (16) via the law of large numbers and can alternatively be derived from equation (31) using the asymptotic properties of the gamma function.

Appendix 2: Effect of overfitting on class centre estimators

Here we quantify the effect of overfitting on the relation between the class centre estimators $\mathbf{m}_y$ and the true class centres $\boldsymbol{\mu}_y$. We recall that the observations $\mathbf{x}$ and the centres $\boldsymbol{\mu}_y$ are assumed to obey the relations defined in equations (3) and (4). From these it follows that X_y^2 and Σ_y^2 are distributed according to

$$p(X_y^2, \Sigma_y^2) = \mathrm{Ga}\left(X_y^2; \frac{d}{2}, \frac{2}{d}\left(\frac{\alpha_y^2}{nf_y} + \beta_y^2\right)\right)\mathrm{Ga}\left(\Sigma_y^2; \frac{d}{2}(nf_y - 1), \frac{2\alpha_y^2}{dnf_y}\right) \tag{32}$$

where $\mathrm{Ga}(x; a, b) = x^{a-1}e^{-x/b}/\Gamma(a)b^a$ denotes the gamma distribution. This expression can be used to calculate the average lengths of $\langle \mathbf{x} \rangle_y$ and $\widehat{\mathbf{m}}_y$ (the squared length is used to simplify the integration)

$$\sqrt{\langle \langle \mathbf{x} \rangle_y^2 \rangle_{p(X_y^2, \Sigma_y^2)}} = \sqrt{\mathrm{d}\left(1 + \frac{nf_y\beta_y^2}{\alpha_y^2}\right)\frac{\alpha_y^2}{nf_y}}, \qquad \sqrt{\langle \widehat{\mathbf{m}}_y^2 \rangle_{p(X_y^2, \Sigma_y^2)}} = \sqrt{\mathrm{d}\left(R + \frac{nf_y\beta_y^2}{\alpha_y^2}\right)\frac{\alpha_y^2}{nf_y}} \tag{33}$$

with

$$R = \frac{2\,nf_y - \frac{nf_y\beta_y^2}{\alpha_y^2}(d-2)(nf_y - 1)}{(d-2)(nf_y - 1)\left(1 + \frac{nf_y\beta_y^2}{\alpha_y^2}\right)} - \frac{4\,\Gamma(\frac{dnf_y}{2} + 1)\;{}_2F_1(\frac{d}{2} - 1, \frac{dnf_y}{2} + 1; \frac{d}{2} + 2; -[(nf_y - 1)(1 + \frac{nf_y\beta_y^2}{\alpha_y^2})]^{-1})}{\Gamma(\frac{d}{2} + 2)\,\Gamma(\frac{d}{2}(nf_y - 1) + 1)\,(d-2)[(nf_y - 1)(1 + \frac{nf_y\beta_y^2}{\alpha_y^2})]^{d/2}} \tag{34}$$

Here ${}_2F_1(a, b; c; z) = \sum_{k=0}^{\infty} \frac{\Gamma(a+k)}{\Gamma(a)} \frac{\Gamma(b+k)}{\Gamma(b)} \frac{\Gamma(c)}{\Gamma(c+k)} \frac{z^k}{k!}$ is the hypergeometric function. From the asymptotic properties of the hypergeometric function it follows that $R \ll 1$ for large d.