City Research Online

City St George's, University of London

Citation: Soch, J. & Allefeld, C. (2026). Multivariate Bayesian inversion for classification and regression. International Journal of Data Science and Analytics, 22(1), 208. doi: 10.1007/s41060-026-01168-9

This is the published version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/38064/

Link to published version: https://doi.org/10.1007/s41060-026-01168-9

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

International Journal of Data Science and Analytics (2026) 22:208
https://doi.org/10.1007/s41060-026-01168-9

REGULAR PAPER

Multivariate Bayesian inversion for classification and regression

Joram Soch[1,2] · Carsten Allefeld[3]

Received: 27 August 2025 / Accepted: 17 May 2026
© The Author(s) 2026

Abstract

We propose the statistical modeling approach to supervised learning (i.e., predicting labels from features) as an alternative to algorithmic machine learning (ML). The approach is demonstrated by employing a multivariate general linear model (MGLM) describing the effects of labels on features, possibly accounting for covariates of no interest, in combination with prior distributions on the model parameters. ML "training" is translated into estimating the MGLM parameters via Bayesian inference and ML "testing" or application is translated into Bayesian model comparison—a reciprocal relationship we refer to as multivariate Bayesian inversion (MBI). We devise MBI algorithms for the standard cases of supervised learning, discrete classification and continuous regression, derive novel classification rules and regression predictions, and use practical examples (simulated and real data) to illustrate benefits of the statistical modeling approach: interpretability, incorporation of prior knowledge, probabilistic predictions. We close by discussing further advantages, disadvantages, and the future potential of MBI.

Keywords Supervised learning · Multivariate analysis · General linear models · Bayesian inference · Marginal likelihood · Model comparison

1 Introduction

The core aims of machine learning (ML) are the identification of patterns in data and the use of these patterns for the prediction of variables of interest [1], allowing for the automatization of complex problems such as optical character recognition [2], human speech recognition [3], or strategic game play [4]. Typically, ML is based on constructing a mapping from measured variables, or *features*, to some variables of interest, or *labels*. For example, ML may proceed by separating data points in a high-dimensional feature space to classify them into discretely labeled categories, as in support vector machines (SVMs) [5, 6], or by transforming features using complex nonlinear mappings to predict continuous target labels, as in deep convolutional neural networks (DNNs/CNNs) [7, 8].

In the ML approach, labels are algorithmically extracted from features. From a statistical perspective, an alternative is to model the features in terms of the labels (see Table 1)—and in many contexts where ML is applied, this is the more plausible approach, because the statistical model is often interpretable in terms of an underlying data-generating process. A statistical model entails the specification of probability distributions for the feature variables given the label variables, a so-called *generative model* [9, 10]. This typically invokes unknown parameters that can be estimated from the data and, in combination with prior distributions on those parameters, it becomes a *full probability model* [11]. Once the model parameters have been estimated, the model can be inverted for probabilistic inference of labels from features [12, 13]. Fundamental advantages of the statistical over the algorithmic approach are the possibility to incorporate explicit knowledge about the relationships between labels and features [14], both through the model structure and prior distributions, and its ability to qualify predictions by associating them with probabilities. For further technical advantages, see below and Sect. 5.

✉ Joram Soch
joram.soch@ovgu.de

Carsten Allefeld
Carsten.Allefeld@city.ac.uk

1 Institute of Psychology, Otto von Guericke University, Magdeburg, Germany

2 Bernstein Center for Computational Neuroscience, Charité – Universitätsmedizin Berlin, Berlin, Germany

3 Department of Psychology and Neuroscience, City St George's, University of London, London, UK

Published online: 25 June 2026

The aim of this paper is to demonstrate this statistical approach by adapting a standard model from multivariate statistics to a predictive modeling context, showing its application in simulations and analyses of existing data sets. Multivariate features are assumed to be normally distributed and influenced by some label variable of interest as well as possibly covariates of no interest (see Table 1). Specifically, we use a multivariate general linear model (MGLM) [15, 16] to describe linear relationships between features, labels, and covariates (see Table 2). Via a fully Bayesian treatment of the MGLM, we calculate conditional probabilities for the label variables, given the multivariate features. We refer to this as *multivariate Bayesian inversion* (MBI) and show how it can be used for both classification and regression.

In the process, we provide a systematic review of Bayesian parameter estimation and Bayesian model comparison for the MGLM. While these techniques are very well known [12, 17, 18] and have been applied in many different contexts before (e.g., [19, 20]), we here combine them to develop a comprehensive Bayes classifier (and Bayes regressor) for the MGLM.

Taken together, the contributions of this work are threefold:

(1) We rewrite the ML task of predicting from multivariate continuous data in terms of an interpretable probabilistic model which, rather than extracting labels from features, describes the effects that labels have on features.
(2) We show how inversion of this model by means of Bayesian model selection can be used to perform ML tasks such as classification and regression and, in addition, provides posterior probabilities for the predicted labels.
(3) We familiarize users of standard ML methods (e.g., SVMs) with statistical model-based predictive analyses and demonstrate how common problems in ML—e.g., inclusion of variables of no interest, integration of prior knowledge, classification analyses for unbalanced categories, regression analyses for bounded variables, and correlated observations—can be handled naturally and elegantly by the MBI framework.

We are aware of the fact that the simple model which we describe here lacks several capabilities of currently leading ML methods (i.e., DNNs)—e.g., facilitating prediction in high-dimensional spaces, coping with non-normal and heteroskedastic error distributions, or allowing for nonlinear relationships between features and labels via kernel-based methods. Our goal is not to outperform these approaches in terms of predictive accuracy. Rather, we want to provide a basic framework for statistical model-based Bayesian machine learning which may be further extended to address these issues. Some possible extensions are discussed in Sect. 5.3.

The present paper is structured as follows. First, we outline the statistical theory and derive the mathematical equations underlying MBI (see Sect. 2). Next, we present a series of MBI applications based on the analysis of simulated and empirical data (see Sects. 3 and 4). In these sections, we also compare results based on MBI with results obtained using SVMs, representing a comparably flexible and currently popular ML approach [1, 21, 22]. Finally, we review various aspects illustrated by these simulations and discuss advantages, disadvantages, and possible extensions of MBI (see Sect. 5).

Terminology: Note that, following the statistical modeling approach, we conceptualize an ML problem as learning a relationship of the form $Y = f(X) + \varepsilon$—in which Y are features and X are labels—as opposed to the typical ML approach of extracting estimates of labels from observations of features, i.e., $\hat{X} = \hat{g}(Y)$. Consequently, we refer to labels as "independent variables" and features as "dependent variables," not the other way around. This is because the statistical model describes Y as a function of X, but when inverted, can be used to deduce X from Y.

Notation: In the following, we denote n-dimensional zero vectors as 0_n and n-dimensional vectors of all ones as 1_n. Similarly, we denote $n \times m$ zero matrices as 0_{nm} and $n \times m$ matrices of all ones as 1_{nm}. We write the $n \times n$ identity matrix as I_n and write its i-th column, i.e., the i-th n-dimensional unit vector, as e_i for $i = 1, \ldots, n$. We use $\mathcal{N}(\mu, \Sigma)$ to denote the multivariate normal distribution with expected value $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}_+^{n \times n}$ where $\mathbb{R}_+^{n \times n}$ is the set of all positive-definite symmetric matrices of size $n \times n$. Finally, we use $\mathcal{MN}(M, U, V)$ to denote the matrix-normal distribution with expected value $M \in \mathbb{R}^{n \times m}$, covariance matrix across rows $U \in \mathbb{R}_+^{n \times n}$, and covariance matrix across columns $V \in \mathbb{R}_+^{m \times m}$.

2 Theory

In this section, we introduce the foundations of multivariate Bayesian inversion (MBI). We first review the notions of predictive algorithms and generative models and then introduce the multivariate general linear model as a generative model (see Sect. 2.1). We go on to describe how, using Bayesian data analysis, one can either estimate the parameters of such a model or perform comparison between the models themselves (see Sect. 2.2). Using the MBI framework for predictive analyses rests on the reciprocal estimation of parameters and labels, i.e., either label values are known and model parameters have to be estimated or model parameters are known and label values have to be predicted (see Sect. 2.3). Finally, upon partitioning available data into train-

Table 1 Terminology. Analogous terms commonly used in the ML and statistics literatures, accompanied by an explanation of our use of them

Machine learning	Statistics	Explanation
features	dependent variables	measured variables which are the input to a predictive algorithm in machine learning or the output of a generative model in statistics
labels	independent variables	controlled (sometimes observed) variables which are thought to have some relationship with the features
classes	groups, categories	discrete labels
targets	covariate, predictor	continuous labels
feature matrix	data matrix	an observation-by-variable matrix containing dependent variables
label information	design matrix	an observation-by-regressor matrix containing independent variables
class labels	categorical design matrix	discrete categories represented by indicator regressors taking the values 0 or 1
target values	parametric design matrix	continuous values represented by regressors taking real numbers as values

ing and test data, MBI can be used for discrete classification or continuous regression, respectively (see Sect. 2.4).

2.1 Linear models for multivariate data

2.1.1 Predictive algorithms versus generative models

Consider some features $Y \in \mathcal{Y}$ and some labels $X \in \mathcal{X}$, i.e., random variables (dependent and independent variables, see Table 1) with sample spaces $\mathcal{Y}$ and $\mathcal{X}$, respectively. Then, a function $g : \mathcal{Y} \to \mathcal{X}$ is called a *predictive algorithm*. In supervised learning, one typically calibrates g based on a training set of features and labels (Y_1, X_1) and then evaluates the performance of g by predicting the labels X_2 of test features Y_2,

$$\hat{g} \leftarrow (Y_1, X_1) \quad \text{and} \quad \hat{X}_2 = \hat{g}(Y_2) \,, \tag{1}$$

where $\hat{g}$ denotes the calibrated predictive algorithm and $\hat{X}_2$ are the predicted labels of the test features [23, 24]. A common example for this is the *multivariate continuous prediction task* in which $Y \in \mathbb{R}^v$ is a random vector of v continuous features and either $X \in \{1, \ldots, C\}$ is a discrete class index or $X \in \mathbb{R}$ is a real number. In the former case, the multivariate continuous prediction task is referred to as a *classification task* and is typically solved using algorithms such as multinomial logistic regression or support vector classification [5, 25]. In the latter case, the multivariate continuous prediction task is referred to as a *regression problem* and can be solved using, e.g., multiple linear regression or support vector regression [6].

Alternatively, the relationship between features Y and labels X can be furnished by taking a statistical perspective. Here, a statement about the probability distribution of Y, given X and some model parameters $\theta \in \Theta$, is referred to as a *generative model*. The term implies that, given fixed parameters θ, features Y can be generated by sampling the data-generating mechanism,

$$m : \; Y \sim \mathcal{D}(X, \theta) \,, \tag{2}$$

where m denotes the generative model and $\mathcal{D}(X, \theta)$ signifies an arbitrary probability distribution parametrized in terms of X and θ. This distribution implies a probability function $p(Y|X, \theta, m) : \mathcal{Y} \to \mathbb{R}_{\geq 0}$, which—when viewed as a function θ for fixed Y—is referred to as the *likelihood function* of the generative model.

The idea behind this generative approach is to model the effects that labels have on features ($X \to Y$), with the intention to later invert the model in order to estimate labels from features ($Y \to X$). When labels are latent variables causally influencing observable features—which is often, but not always the case in real-world applications (see Sect. 4)—the generative model captures the data-generating mechanism by describing $p(Y|X)$ rather than $p(X|Y)$ (see Sect. 2.7).

A typical example for this is a Gaussian model where $\mathcal{D}$ is a (potentially multivariate) Gaussian distribution and θ are weights or coefficients characterizing linear or nonlinear effects of X on Y. Importantly, the fact that generative models are based on explicit probabilistic assumptions about features, labels, and parameters makes them compatible to principled forms of model estimation and model selection such as maximum likelihood, Bayesian posterior inference,

classical information criteria, or Bayesian model comparison [11, 12, 18, 26, 27].

For example, in Bayesian inference, a *prior distribution* over model parameters $p(\theta|m)$ is specified which allows to infer the *posterior probabilities* of the model parameters, given the features $p(\theta|Y, m)$, or to calculate the *marginal likelihood* of the measured features, given the model $p(Y|m)$.

In the following, we show how the multivariate general linear model can be written as a generative model and made amenable to Bayesian analyses, where the results of these analyses can then be employed as a means to solve multivariate continuous prediction tasks in the style of predictive algorithms.

2.1.2 Multivariate general linear models

Let $Y \in \mathbb{R}^{n\times v}$ denote an observed random matrix of measured data comprising n observations of v features and let $X \in \mathbb{R}^{n\times p}$ denote a known matrix comprising n associated observations of p predictor variables (also called "regressors") that possibly influence the measured data. Then, we define a multivariate general linear model (MGLM) m as

$$\begin{aligned} m:\ & Y = XB + E, \\ & E \sim \mathcal{MN}(0_{nv}, V, \Sigma) \end{aligned} \tag{3}$$

where $B \in \mathbb{R}^{p\times v}$ denotes a matrix of unknown (regression) coefficients, $E \in \mathbb{R}^{n\times v}$ denotes an unobserved error matrix, and $\mathcal{MN}(0_{nv}, V, \Sigma)$ denotes a matrix-normal distribution with zero mean, known observation covariance matrix $V \in \mathbb{R}_+^{n\times n}$ and unknown feature covariance matrix $\Sigma \in \mathbb{R}_+^{v\times v}$.

In applications to data sets comprising multiple features for multiple experimental units (e.g., participants in a study), V models the error covariation between different experimental units and Σ models the error covariation between of features within each experimental unit. In applications to spatiotemporal data sets, where observations correspond to different time points and features correspond to different locations in space, V may be referred to as a *temporal covariance matrix* and Σ may be referred to as a *spatial covariance matrix* [16].

Note that, in equation (3), X and V are assumed to be known, whereas B and Σ are assumed to be unknown (see Table 2). This is compatible with the fact that observations, i.e., rows of Y, typically have some known autocorrelation structure or are assumed independent (hence, V is prespecified), whereas features, i.e., columns of Y, typically have some unknown covariation which needs to be inferred (hence, Σ is estimated). The assumption of matrix-normally distributed errors also means that all observations have the same covariance across columns Σ and that all features have the same covariance across rows V. If observations are independent and identically distributed (i.i.d.), as is often assumed in ML applications, then $V = I_n$ and the model simplifies to (see Appendix A)

$$\begin{aligned} m:\ & y_i = x_i B + \varepsilon_i, \\ & \varepsilon_i^T \sim \mathcal{N}(0_v, \Sigma),\ i = 1, \ldots, n \end{aligned} \tag{4}$$

where $y_i \in \mathbb{R}^{1\times v}$, $x_i \in \mathbb{R}^{1\times p}$, $\varepsilon_i \in \mathbb{R}^{1\times v}$ are row vectors, and $\mathcal{N}(0_v, \Sigma)$ denotes a multivariate normal distribution with zero mean and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$.

Since E follows a matrix-normal distribution and XB is constant, the MGLM formulated in (3) is equivalent to the following generative model (see Appendix A):

$$m:\ Y \sim \mathcal{MN}(XB, V, \Sigma)\ . \tag{5}$$

With the matrix-normal probability density function $\mathcal{MN} : \mathbb{R}^{n\times v} \to \mathbb{R}_{>0}$, equation (5) then implies the following likelihood function for m (see Appendix A):

$$\begin{aligned} p(Y|X, B, V, \Sigma, m) &= \mathcal{MN}(Y; XB, V, \Sigma) \\ &= \sqrt{\frac{1}{(2\pi)^{nv}|\Sigma|^n|V|^v}}\cdot \\ &\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}(Y - XB)^{\mathrm{T}}V^{-1}(Y - XB)\right)\right]. \end{aligned} \tag{6}$$

For reasons of mathematical convenience, we write this in terms of the precision matrices $P = V^{-1} \in \mathbb{R}_+^{n\times n}$ and $T = \Sigma^{-1} \in \mathbb{R}_+^{v\times v}$ (see Table 2):

$$p(Y|X, B, P, T, m) = \sqrt{\frac{|T|^n|P|^v}{(2\pi)^{nv}}}\cdot \&\exp\left[-\frac{1}{2}\operatorname{tr}\left(T(Y - XB)^{\mathrm{T}}P(YXB)\right)\right]. \tag{7}$$

Note that, among the conditioned variables, V (or equivalently, P) and m are constant which is why, in the following sections, we will drop them for clarity of notation. In contrast with that, B and T are random entities described by some probability distribution. X is considered fixed when estimating B and T (see Sect. 2.3.1) and random when using a distribution over B and T to predict it (see Sect. 2.3.2). Taken together, we will therefore refer to the likelihood function of the MGLM as $p(Y|X, B, T)$.

2.2 Bayesian inference for MGLMs

2.2.1 Bayesian parameter estimation

In Bayesian inference, one usually specifies a prior density $p(\theta|m)$, i.e., a probability distribution for θ unconditional on Y, and the goal is to derive the posterior density $p(\theta|Y, m) : \Theta \to \mathbb{R}_{\geq 0}$, i.e., the probability distribution for θ conditional

Table 2 *Notation*. Symbols used in the mathematical description of MBI. Quantities are either measured (empirically observed), known (experimentally controlled), unknown (statistically estimated), or predicted (reconstructed from the features)

Symbol	Dimension	Category	Meaning
n, v, p	Scalar	known	number of observations, variables and regressors
Y	$n \times v$	measured	data matrix
X	$n \times p$	known	design matrix
B	$p \times v$	unknown	regression coefficients
E	$n \times v$	unknown	error matrix
V	$n \times n$	known	observation covariance matrix
Σ	$v \times v$	unknown	feature covariance matrix
P	$n \times n$	known	inverse of V
T	$v \times v$	unknown	inverse of Σ
x	$n \times 1$	predicted	classes or targets
Y_1, X_1	$n_1 \times \cdot$	see above	Y and X in the training data
Y_2, X_2	$n_2 \times \cdot$	see above	Y and X in the test data
y_{2i}	$1 \times v$	measured	a single row of Y in the test data
x_{2i}	$1 \times p$	predicted	a single row of X in the test data

on Y, given the model m [11, 18, 28–30]. Bayes' theorem implies that this posterior density is proportional to the product of the likelihood function $p(Y|\theta, m)$ and the prior density:

$$p(\theta|Y, m) \propto p(Y|X, \theta, m)\, p(\theta|m) \tag{8}$$

Following (7), the MGLM model parameters are $\theta = (B, T) \in \mathbb{R}^{p\times v} \times \mathbb{R}_+^{v\times v} = \Theta$ where $\mathbb{R}_+^{n\times n}$ is the set of all positive-definite symmetric matrices of size $n \times n$. Thus, Bayesian MGLM e B and T. In order to obtain a tractable posterior, we will use a conjugate prior, i.e., a prior distribution that, when combined with the likelihood function, leads to a posterior distribution that belongs to the same family of probability distributions. For the MGLM, this conjugate prior is a normal-Wishart distribution on the coefficient matrix B and the precision matrix T ([11, ch. 3]; [31, sect. 8])

$$\begin{aligned} p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \\ \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) \end{aligned} \tag{9}$$

where $M_0 \in \mathbb{R}^{p\times v}$ is the prior mean, $\Lambda_0 \in \mathbb{R}_+^{p\times p}$ is the prior precision matrix, $\Omega_0 \in \mathbb{R}_+^{v\times v}$ is the prior inverse scale matrix, $\nu_0 \in \mathbb{R}$ are the prior degrees of freedom, and $\mathcal{MN} : \mathbb{R}^{p\times v} \to \mathbb{R}_{>0}$ and $\mathcal{W} : \mathbb{R}_+^{v\times v} \to \mathbb{R}_{>0}$ are the respective probability density functions.

Because this distribution is a conjugate prior, the application of (8) results in a posterior distribution that is also a normal-Wishart distribution (see Appendix B)

$$\begin{aligned} p(B, T|Y) = \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \\ \cdot \mathcal{W}(T; \Omega_n^{-1}, \nu_n) \end{aligned} \tag{10}$$

where the posterior parameters M_n, Λ_n, Ω_n, and ν_n turn out to be functions of the prior parameters, the feature variables Y and the predictor variables X (see Appendix B):

$$\begin{aligned} M_n &= \Lambda_n^{-1}(X^{\mathrm{T}}PY + \Lambda_0 M_0) \\ \Lambda_n &= X^{\mathrm{T}}PX + \Lambda_0 \\ \Omega_n &= \Omega_0 + Y^{\mathrm{T}}PY + M_0^{\mathrm{T}}\Lambda_0 M_0 - M_n^{\mathrm{T}}\Lambda_n M_n \\ \nu_n &= \nu_0 + n\ . \end{aligned} \tag{11}$$

Note that, if observations are assumed to be i.i.d., we have $P = V^{-1} = I_n^{-1} = I_n$ and the matrix P simply disappears from those equations. This is the case for most applications of linear regression models and will also be assumed throughout most parts of this paper (but see Simulation 4 in Sect. 3).

2.2.2 Bayesian model comparison

In addition to obtaining posterior distributions for the model parameters θ, Bayesian inference also allows to perform statistical inference on the generative model m itself. This is achieved by calculating the marginal likelihood $p(Y|m) : \mathcal{Y} \to \mathbb{R}_{\geq 0}$, i.e., the probability of the data Y, independent of particular parameter values θ, given only the model m [12, 32–34]. The law of marginal probability implies that this marginal likelihood is an integral of the product of likelihood function and prior distribution:

$$p(Y|m) = \int_\Theta p(Y|\theta, m)\, p(\theta|m)\, \mathrm{d}\theta\ . \tag{12}$$

Application of (12) to the MGLM with $\theta = (B, T) \in \mathbb{R}^{p\times v} \times \mathbb{R}_+^{v\times v} = \Theta$ yields

$$\begin{aligned} p(Y|m) &= \int_{\mathbb{R}^{p\times v}\times\mathbb{R}_+^{v\times v}} p(Y|X, B, T)\, p(B, T)\, \mathrm{d}(B, T) \\ &= \int_{\mathbb{R}_+^{v\times v}} p(T) \left[\int_{\mathbb{R}^{p\times v}} p(B|T)\, p(Y|X, B, T)\, \mathrm{d}B \right] \mathrm{d}T \end{aligned} \tag{13}$$

and the marginal likelihood with (11) becomes (see Appendix C)

$$\begin{aligned} p(Y|m) &= \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}} \sqrt{\frac{\left|\frac{1}{2}\Omega_0\right|^{\nu_0}}{\left|\frac{1}{2}\Omega_n\right|^{\nu_n}}} \frac{\Gamma_v\left(\frac{\nu_n}{2}\right)}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \end{aligned} \tag{14}$$

where $\Gamma_p : \mathbb{R} \to \mathbb{R}$ is the multivariate Gamma function of order $p \in \mathbb{N}$. Once again, if observations are assumed i.i.d., such that rows of the data matrix Y are uncorrelated—as it is the case in most linear modeling contexts—we have $P = I_n$, such that the term $|P|^v$ in (14) would be replaced by 1.

If two or more models are differing by the predictor variables X they utilize (giving rise to different likelihood functions $p(Y|X, \theta, m)$) or by the prior parameters $(M_0, \Lambda_0, \Omega_0, \nu_0)$ they impose (giving rise to different prior distributions $p(\theta|m)$), this constitutes a case for model comparison. Consider a model space $\mathcal{M} = \{m_1, \ldots, m_M\}$ where M is the number of models. Then, posterior model probabilities can be calculated from marginal likelihoods (14) using Bayes' theorem

$$p(m_i|Y) = \frac{p(Y|m_i)\, p(m_i)}{\sum_{m\in\mathcal{M}} p(Y|m)\, p(m)} \tag{15}$$

for $i = 1, \ldots, M$ where $p(m_i)$ are prior model probabilities, i.e., prior beliefs about the relative likelihoods of models $m \in \mathcal{M}$ before seeing the data Y.

2.3 Multivariate Bayesian inversion

2.3.1 Estimation of parameters, given the labels

In Sect. 2.2.1, we have introduced how Bayesian inference can be used for MGLM parameter estimation. In this section, we explain how this method constitutes the first step of a multivariate Bayesian inversion analysis.

The fact that, according to equation (7), the probability of the measured data Y is conditional on both the model's regressors X and the model parameters $\theta = (B, T)$, means that both of these quantities can be considered either estimated and known or unknown and to-be-predicted. In some situation, e.g., in a designed experiment, it could be that regressor values are known and model parameters have to be learned. In other contexts, e.g., in consumer choice prediction, it could be that model parameters are known and regressor values have to be inferred.

Consider a data set $(Y, X, P) \in \mathbb{R}^{n\times v} \times \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times n}$ specified by data matrix Y, design matrix X, and precision matrix P (see Sect. 2.1.2). In this case, X would be considered known and B and T have to be estimated following (8):

$$p(B, T|Y) \propto p(Y|X, B, T)\, p(B, T)\,. \tag{16}$$

To do so, a prior density $p(B, T)$ as in (9) has to be specified by fixing the prior parameters M_0, Λ_0, Ω_0 and ν_0 (see Sect. 2.2.1). If there is no prior knowledge, i.e., one does not have concrete prior beliefs about the model parameters B and T, one possibility is to use non-informative prior parameters [30] given by

$$M_0 = 0_{pv},\ \ \Lambda_0 = 0_{pp},\ \ \Omega_0 = 0_{vv},\ \ \nu_0 = 0\,. \tag{17}$$

Using (10) and (11), the posterior distribution becomes

$$\begin{aligned} p(B, T|Y) = \mathcal{MN}(B; M_1, \Lambda_1^{-1}, T^{-1}) \\ \cdot \mathcal{W}(T; \Omega_1^{-1}, \nu_1) \end{aligned} \tag{18}$$

where the posterior parameters after seeing the data are (see Appendix D):

$$\begin{aligned} M_1 &= \Lambda_1^{-1} X^{\mathrm{T}} P Y \\ \Lambda_1 &= X^{\mathrm{T}} P X \\ \Omega_1 &= Y^{\mathrm{T}} P Y - M_1^{\mathrm{T}} \Lambda_1 M_1 \\ \nu_1 &= n\,. \end{aligned} \tag{19}$$

Note that the prior parameters (17) are specified in such a way that the posterior parameters (19) are only influenced by the observed data Y and the known regressors X in this example (cf. eq. 11). Moreover, as we only estimate one $T = \Sigma^{-1}$, the feature covariance matrix Σ is assumed to be constant across observations, i.e., the same for all data points. For multivariate Bayesian classification (see Sect. 2.4.1), this means that MBI pools covariance across classes. For multivariate Bayesian regression (see Sect. 2.4.1), this means that the covariance is not dependent on the target variable.

2.3.2 Identification of labels, given the parameters

In Sect. 2.2.2, we have introduced how Bayesian inference can be used for MGLM model comparison. In this section,

we explain how this method constitutes the second step of a multivariate Bayesian inversion analysis.

Other than before, it could be that a distribution $p(B, T)$ is available, for example (but not necessarily) because it has been estimated from independent data. In this case, the model parameters B and T can be assumed known—or at least partially known, or known with some uncertainty—when approaching new data.

Consider a single data point $(y_i, x_i) \in \mathbb{R}^{1\times v} \times \mathbb{R}^{1\times p}$ specified by a new observation y_i and the associated regressor values x_i, where i is an index over previously unseen data. Let us assume that the goal is to identify the x_i belonging to a single new y_i. For this one observation, the multivariate model reads (cf. eq. 4):

$$y_i = x_i B + \varepsilon_i, \;\; \varepsilon_i^{\mathrm{T}} \sim \mathcal{N}(0, T^{-1}) \; . \tag{20}$$

Now, B and T can be considered known in the sense that the distribution $p(B, T)$ acts as the prior distribution for the new data point, such that the prior predictive distribution of the observed y_i, given the unknown x_i follows from (12). For example, assume that the distribution $p(B, T)$ is given by the posterior distribution $p(B, T|Y)$ in (18) specified by the parameters M_1, Λ_1, Ω_1 and ν_1 from (19). Then, assuming this as our prior knowledge and using (14), the marginal likelihood $p(y_i|x_i)$ becomes

$$\begin{aligned} &p(y_i|x_i) \\ &= \sqrt{\frac{1}{(2\pi)^v}} \sqrt{\frac{|\Lambda_1|^v}{|\Lambda_2|^v}} \sqrt{\frac{\left|\frac{1}{2}\Omega_1\right|^{\nu_1}}{\left|\frac{1}{2}\Omega_2\right|^{\nu_2}}} \frac{\Gamma_v\left(\frac{\nu_2}{2}\right)}{\Gamma_v\left(\frac{\nu_1}{2}\right)} \end{aligned} \tag{21}$$

where the posterior parameters after seeing the new data are (see Appendix D):

$$\begin{aligned} M_2 &= \Lambda_2^{-1}(X^{\mathrm{T}} P Y + x_i^{\mathrm{T}} y_i) \\ \Lambda_2 &= X^{\mathrm{T}} P X + x_i^{\mathrm{T}} x_i \\ \Omega_2 &= Y^{\mathrm{T}} P Y + y_i^{\mathrm{T}} y_i - M_2^{\mathrm{T}} \Lambda_2 M_2 \\ \nu_2 &= n + 1 \; . \end{aligned} \tag{22}$$

Crucially, different regressor values can be plugged in for x_i in equation (22), enabling to treat the identification of labels as a model selection problem. More precisely, we are testing competing models for the observations y_i by submitting different possibilities for the regressors x_i, to see which of these possibilities maximizes the marginal likelihood $p(y_i|x_i)$ and is thus the most likely value of x_i (for a similar approach, see [20]). Given that we can specify some prior knowledge about the predictor variables x_i, the marginal likelihoods from (21) can be used to obtain posterior probabilities for label values x_i, given the measured data y_i:

$$p(x_i|y_i) \propto p(y_i|x_i)\, p(x_i) \; . \tag{23}$$

Note that, other than in equation (16), X is not considered fixed anymore, but x_i is now a random variable which has prior distribution $p(x_i)$ and can be estimated in the form of a posterior distribution $p(x_i|y_i)$.

Multivariate Bayesian inversion (MBI) is the alternating procedure in which first, MGLM parameters are estimated, given known labels, and in which second, labels are inferred on, given known MGLM parameters. In Sects. 2.4.1 and 2.4.2, we will outline an application of this MBI framework for cross-validated predictive analyses which are based on splitting the data into training and test set in order to predict labels for left-out observations using either classification or regression.

2.4 Cross-validated MBI for prediction

2.4.1 Multivariate Bayesian classification

Let $x \in \mathbb{N}_{>0}^{n\times 1}$ be a vector of nonzero class indices $(1, 2, 3, \ldots)$ corresponding to the observations in the data matrix $Y \in \mathbb{R}^{n\times v}$ (see Fig. 1) and let $(Y_1, x_1) \in \mathbb{R}^{n_1\times v} \times \mathbb{N}_{>0}^{n_1\times 1}$ and $(Y_2, x_2) \in \mathbb{R}^{n_2\times v} \times \mathbb{N}_{>0}^{n_2\times 1}$ denote training and test set, respectively. The number of data points is n_1 and n_2, such that $n_1 + n_2 = n$.

First, we define a "regressor generator function" $X_C : \mathbb{N}_{>0}^{n\times 1} \to \mathbb{R}^{n\times C}$ which sets the j-th column in the i-th row to 1, if the i-th observation belongs to the j-th class, where $C = \max(x)$ is the number of classes (see Fig. 1A):

$$(X_C(x))_{ij} = \begin{cases} 1 \; , \text{ if } x_i = j \\ 0 \; , \text{ otherwise} \; . \end{cases} \tag{24}$$

The result is a "categorical design matrix" in which each column vector is a binary indicator, i.e., only taking values 0 or 1, representing one class.

Since x_2 is unknown, i.e., we do not know the classes in the test set, we have to specify a set of candidate models. Let m_{ij} be the model asserting that the single observation y_{2i} in the test set belongs to the j-th class (see Fig. 1B):

$$\begin{aligned} m_{ij} : (X_2)_i &= X_C(j) \\ &= [0, \ldots, 1, \ldots, 0] = e_j^{\mathrm{T}} \; , \end{aligned} \tag{25}$$

i.e., the i-th row of X_2 is a $1 \times C$ vector of zeros with a one at the j-th position.

Then, we can calculate the marginal likelihood of the test data point y_{2i} based on (21):

$$p(y_{2i}|m_{ij}) = \sqrt{\frac{1}{(2\pi)^v}} \sqrt{\frac{|\Lambda_1|^v}{|\Lambda_{2i}(j)|^v}} \sqrt{\frac{\left|\frac{1}{2}\Omega_1\right|^{\nu_1}}{\left|\frac{1}{2}\Omega_{2i}(j)\right|^{\nu_2}}} \frac{\Gamma_v\left(\frac{\nu_2}{2}\right)}{\Gamma_v\left(\frac{\nu_1}{2}\right)} . \tag{26}$$

This allows to quantify posterior class probabilities for each observation based on (15):

$$p(x_{2i} = j|y_{2i}) = \frac{p(y_{2i}|m_{ij})\, p(m_{ij})}{\sum_{k=1}^{C} p(y_{2i}|m_{ik})\, p(m_{ik})} \tag{27}$$

where $p(m_{ij})$ is the prior probability that test set observation i belongs to class j.

Note that the only terms in (26) depending on the candidate model m_{ij} are

$$\begin{aligned} \Lambda_{2i}(j) &= X_C(x_1)^{\mathrm{T}} P_1 X_C(x_1) + X_C(j)^{\mathrm{T}} X_C(j) \\ \Omega_{2i}(j) &= Y_1^{\mathrm{T}} P_1 Y_1 + y_{2i}^{\mathrm{T}} y_{2i} - M_{2i}(j)^{\mathrm{T}} \Lambda_{2i}(j) M_{2i}(j) , \end{aligned} \tag{28}$$

such that the posterior class probabilities from (27) can be written as

$$p(x_{2i} = j|y_{2i}) = \frac{\sqrt{\frac{|\Lambda_{2i}(j)|^{-v}}{|\Omega_{2i}(j)|^{n_1+1}}} \cdot p(m_{ij})}{\sum_{k=1}^{C} \sqrt{\frac{|\Lambda_{2i}(k)|^{-v}}{|\Omega_{2i}(k)|^{n_1+1}}} \cdot p(m_{ik})} , \tag{29}$$

for all test data points $i = 1, \ldots, n_2$. If one were to predict the classes in the test set, picking the one which maximizes the posterior probability would be a reasonable choice:

$$\hat{x}_{2i} = \arg\max_{j=1}^{C} p(x_{2i} = j|y_{2i}) . \tag{30}$$

2.4.2 Multivariate Bayesian regression

Let $x \in \mathbb{R}^{n\times 1}$ be a vector of real number target values ($x_i \in \mathbb{R}$, $i = 1, \ldots, n$) corresponding to the observations in the data matrix $Y \in \mathbb{R}^{n\times v}$ (see Fig. 2) and let $(Y_1, x_1) \in \mathbb{R}^{n_1\times v} \times \mathbb{R}^{n_1\times 1}$ and $(Y_2, x_2) \in \mathbb{R}^{n_2\times v} \times \mathbb{R}^{n_2\times 1}$ denote training and test set, respectively. The number of data points are n_1 and n_2, such that $n_1 + n_2 = n$.

First, we define a "regressor generator function" $X_R : \mathbb{R}^{n\times 1} \to \mathbb{R}^{n\times 2}$ which horizontally concatenates the target vector and a vector of ones (see Fig. 2A):

$$X_R(x) = [x,\ 1_n] . \tag{31}$$

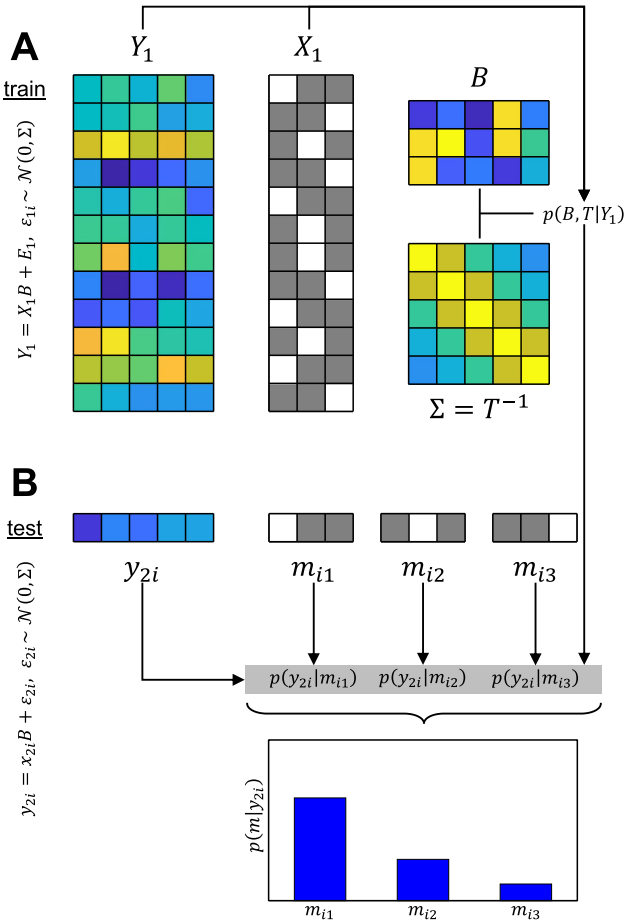

Fig. 1 *Multivariate Bayesian inversion for classification.* **A** In the training set, a categorical design matrix X_1 is formed and the posterior distribution over regression coefficients B and inverse feature covariance $T = \Sigma^{-1}$ is calculated from the measured data Y_1. **B** In the test set, each category j is successively tested against each data point y_{2i} as a possible model m_{ij}, resulting in posterior class probabilities $p(x_{2i} = j|y_{2i})$. Multivariate Bayesian classification is a two-step procedure, separately estimating parameters (top) and inferring on classes (bottom). Note that between-observation covariance, denoted in the text via V and P, has been omitted (i.e., set to $V = P = I_n$) in this example for simplicity

This results in a "regression design matrix" in which the first column represents the target variable and the second column is a constant regressor.

Since x_2 is unknown, i.e., we do not know the targets in the test set, we have to specify a set of candidate models. Let $m_i(\lambda)$ be the model asserting that the target value belonging to the single observation y_{2i} is equal to λ (see Fig. 2B):

$$m_i(\lambda) : (X_2)_i = X_R(\lambda) = [\lambda,\ 1] . \tag{32}$$

i.e., the i-th row of X_2 is a 1×2 vector with λ as first entry and one as second entry.

Then, the marginal likelihood of the test data point y_{2i} is a function of λ based on (21):

$$p(y_{2i}|m_i(\lambda)) = \sqrt{\frac{(p_{2i})^v}{(2\pi)^v}} \sqrt{\frac{|\Lambda_1|^v}{|\Lambda_{2i}(\lambda)|^v}} \sqrt{\frac{\left|\frac{1}{2}\Omega_1\right|^{v_1}}{\left|\frac{1}{2}\Omega_{2i}(\lambda)\right|^{v_2}}} \frac{\Gamma_v\left(\frac{v_2}{2}\right)}{\Gamma_v\left(\frac{v_1}{2}\right)} . \tag{33}$$

This allows to derive a posterior target density over λ for each observation based on (15):

$$p(x_{2i} = \lambda|y_{2i}) = \frac{p(y_{2i}|m_i(\lambda))\, p(\lambda)}{\int_{\mathbb{R}} p(y_{2i}|m_i(\vartheta))\, p(\vartheta)\, \mathrm{d}\vartheta} \tag{34}$$

where $p(\lambda)$ and $p(\vartheta)$ are the prior probability density over the i-th test set label x_{2i}.

Note that the only terms in (33) depending on the candidate value λ are

$$\begin{aligned} \Lambda_{2i}(\lambda) &= X_R(x_1)^{\mathrm{T}} P_1 X_R(x_1) + X_R(\lambda)^{\mathrm{T}} X_R(\lambda) \\ \Omega_{2i}(\lambda) &= Y_1^{\mathrm{T}} P_1 Y_1 + y_{2i}^{\mathrm{T}} y_{2i} - M_{2i}(\lambda)^{\mathrm{T}} \Lambda_{2i}(\lambda) M_{2i}(\lambda) , \end{aligned} \tag{35}$$

such that the posterior target density from (34) can be written as

$$p(x_{2i} = \lambda|y_{2i}) = \frac{\sqrt{\frac{|\Lambda_{2i}(\lambda)|^{-v}}{|\Omega_{2i}(\lambda)|^{n_1+1}}} \cdot p(\lambda)}{\int_{\mathbb{R}} \sqrt{\frac{|\Lambda_{2i}(\vartheta)|^{-v}}{|\Omega_{2i}(\vartheta)|^{n_1+1}}} \cdot p(\vartheta)\, \mathrm{d}\vartheta} , \tag{36}$$

for all test data points $i = 1, \ldots, n_2$. If one was to predict the target value for each data point in the test set, reporting either the posterior expected value or the maximum-a-posteriori estimate would be reasonable choices:

$$\begin{aligned} \hat{x}_{\text{mean}} &= \int_{\mathbb{R}} \lambda \cdot p(x_{2i} = \lambda|y_{2i})\, \mathrm{d}\lambda \\ \hat{x}_{\text{MAP}} &= \arg\max_{\lambda \in \mathbb{R}} p(x_{2i} = \lambda|y_{2i}) . \end{aligned} \tag{37}$$

2.5 Inclusion of covariates

In Sects. 2.4.1 and 2.4.2, MBC and MBR were formulated without covariates, i.e., with only class indices x or target values x influencing the measured data Y. We will now write down the design matrices for MBI with inclusion of covariates, i.e., incorporating confound variables Z which are not themselves the target of prediction, but may have influences on the available features Y.

Let $Z \in \mathbb{R}^{n \times c}$ be a matrix of covariates corresponding to the observations in the data matrix $Y \in \mathbb{R}^{n \times v}$ and let $Z_1 \in \mathbb{R}^{n_1 \times c}$ and $Z_2 \in \mathbb{R}^{n_2 \times c}$ denote covariate values from training and test set, respectively. Let $(Z_2)_i$ be the $1 \times c$ vector of covariate values for the i-th observation in the test set.

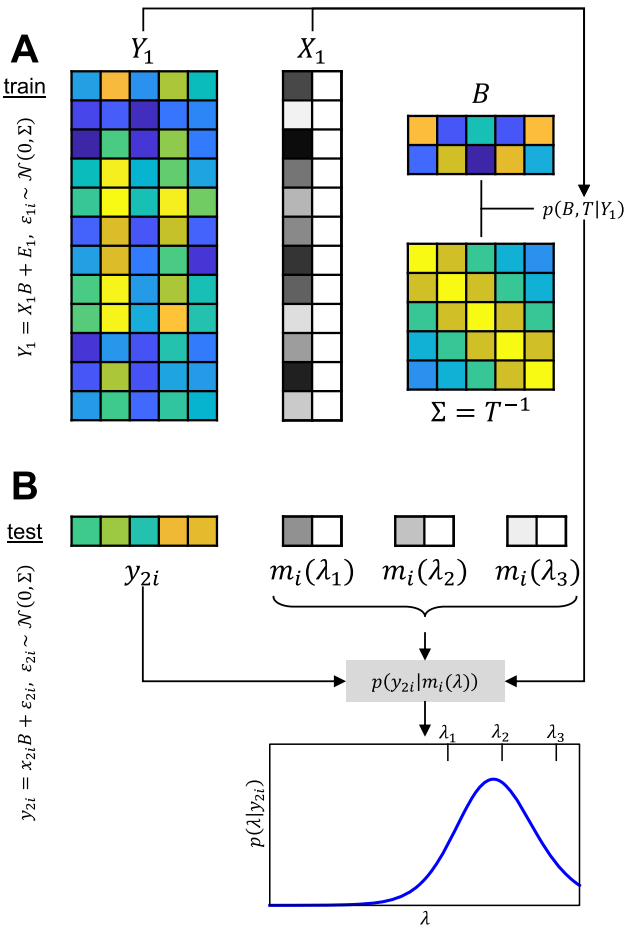

Fig. 2 *Multivariate Bayesian inversion for regression.* **A** In the training set, a regression design matrix X_1 is formed and the posterior distribution over regression coefficients B and inverse feature covariance $T = \Sigma^{-1}$ is calculated from the measured data Y_1. **B** In the test set, the marginal likelihood for each data point y_{2i} is a continuous function of the possible target values λ, resulting in a posterior target density $p(x_{2i} = \lambda|y_{2i})$. Target values corresponding to the three models shown are marked on top of the posterior density plot. Multivariate Bayesian regression is a two-step procedure, separately estimating parameters (top) and inferring on targets (bottom). Note that between-observation covariance, denoted in the text via V and P, has been omitted (i.e., set to $V = P = I_n$) in this example for simplicity

Then, the design matrix for multivariate Bayesian classification in the training set is

$$X_1 = \left[X_C(x_1)\ Z_1\right] \tag{38}$$

and the model assuming the j-th class for the i-th observation in the test set is

$$m_{ij} : (X_2)_i = \left[X_C(j)\ (Z_2)_i\right] . \tag{39}$$

Similarly, the design matrix for multivariate Bayesian regression in the training set is

$$X_1 = \left[X_R(x_1)\ Z_1\right] \tag{40}$$

and the model assuming target value λ for the i-th observation in the test set is

$$m_i(\lambda) : (X_2)_i = \left[X_R(\lambda)\ (Z_2)_i\right] . \tag{41}$$

In other words, different than class indices and target values which are known in the training data and considered unknown in the test data, covariates will be considered known in both parts of the data set. Given these modified design matrices X_1 and $(X_2)_i$, the process of calculating posterior probabilities is essentially the same, e.g., given by equation (29) for MBC and (36) for MBR.

2.6 Cross-validation

Multivariate Bayesian inversion, as outlined in Sects. 2.4.1 and 2.4.2, relies on cross-validation to predict labels based on features. Cross-validation (CV) successively splits a given data set into training and test sets and evaluates predictive performance on the test set, which is repeated until each data point has once been part of the test set [35, 36].

Unless otherwise stated, the following simulations (see Sect. 3) and analyses (see Sect. 4) use *k-fold cross-validation* with $k = 10$. This means that the data set is split into 10 equally large subsets (or sets as equally large as possible, if the number of CV folds k does not divide the number of data points n[1]) and in each CV fold, 9 subsets are used for training, i.e., estimation of the parameters (see Sect. 2.3.1), and the remaining subset is used for testing, i.e., identification of the labels (see Sect. 2.3.2). Splits into CV folds are non-random, but strictly sequential by index of data point.

For multivariate Bayesian classification (Simulations 1–3, Analyses 1–3), we use 10-fold CV *on points per class*. This means that all classes are split into 10 equally large subsets separately, such that in each CV fold, the distribution of classes is the same in training and test set (or as close as possible, if the number of CV folds does not divide the number of data points per class). This ensures that class imbalances are preserved in all subsets and thus do not have a spurious influence on classification.

For Analysis 2 (MNIST digit recognition; see Sect. 4.2) and Analysis 4 (brain age prediction; see Sect. 4.4), we did not use cross-validation, as the data sets were already partitioned in training and validation sets. In these cases, the posterior distribution over model parameters was estimated from the training set, the class labels (Analysis 2) or target values (Analysis 4) were predicted in the validation set, and predictive performance measures are solely based on the validation set.

2.7 Comparison with other Bayesian methods

There exist a number of Bayesian approaches to machine learning [12] some of which are closely related to the general statistical approach presented here.

MBI differs from *Bayesian discriminative methods* such as Bayesian logistic regression [11, ch. 16] and Bayesian linear regression [11, ch. 14] by the fact that the latter typically model the predictive distribution of the labels given the features, i.e., the probability distribution $p(X|Y)$ (cf. Section 2.1.1). MBI in turn is more closely related to *Bayesian generative methods* such as Gaussian naive Bayes classifiers [37] or Variational Gaussian mixture models [38] which describe the generative distribution of the features given the labels, i.e., the probability distribution $p(Y|X)$, and later inverts this distribution to infer on labels given the features.

MBC and MBR are *multi-stage procedures*, meaning that data are splitted into training and test sets, in order to first estimate parameters of the generative model then extract values of the label variable (see Figs. 1 and 2). This is in contrast with expectation maximization (EM) [39] and variational Bayesian (VB) [40] techniques which treat labels as latent random variables and jointly estimate the posterior distribution over label variables and model parameters.

The generative model underlying MBC is equivalent to the model behind (Bayesian) *linear discriminant analysis* (LDA) [41] when the design matrix only includes categorical regressors ($X = X_c(x)$) and when all observations are regarded as independent ($V = I_n$). In this case, all data points in the training set are assumed to follow multivariate normal distributions with class-specific means and class-independent covariance. By estimating B and T (cf. eq. 10), MBI implicitly estimates those class means and covariance matrix, such that new data points in the test set can be assigned to one class. This "LDA scenario" is exemplified in Simulation 3 (see Sect. 3.3) and Analyses 1/2 (see Sects. 4.1/4.2).

The *novel contribution* here is that MBI generalizes (Bayesian) LDA (i) by using the same generative model for classification and regression, (ii) by allowing for correlations between data points in this model, (iii) by systematically integrating covariates into predictive analysis, and (iv) by incorporating prior knowledge about class frequencies.

[1] For details, see the MATLAB implementation (https://github.com/JoramSoch/MBI/blob/main/MATLAB/ML_CV.m) or the Python implementation cf. method "crossval" for class "cvMBI" in (https://github.com/JoramSoch/MBI/blob/main/Python/MBI.py) of cross-validation in the MBI GitHub repository.

3 Simulations

In this section, we apply cross-validated multivariate Bayesian inversion to several synthetic data sets necessitating either classification (Simulations 1–3) or regression (Simulation 4). For each of these examples, we describe (i) the purpose behind the simulation study, (ii) the generative model underlying the data set, (iii) possible scientific interest in data sets of this form, (iv) steps of data analysis, we give a (v) summary of the results and provide (vi) a conclusion.

When of interest (Simulations 2–4), we compare MBI's performance with that of support vector machines (SVM), as a representative easy-to-implement and currently popular alternative approach [42, 43].

3.1 Simulation 1: two-class classification

Purpose: The purpose of this simulation is to show how between-class variance (quantified as the Euclidean distance between class means) as well as within-class variance (quantified as the standard deviation around class means) influence classification accuracy in multivariate Bayesian classification (MBC).

Generative model: We simulate two classes of equal size in two dimensions: Each observation $y_i \in \mathbb{R}^{1\times 2}$, $i = 1, \ldots, n$ is generated as

$$\begin{aligned} y_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i, \\ \varepsilon_i &\sim \mathcal{N}\left([0,0], \sigma^2\Sigma\right) \end{aligned} \tag{42}$$

where $n = 250$ (with 125 per class) and some correlation between the two dimensions is induced via the between-feature covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}. \tag{43}$$

The terms $x_{1i}, x_{2i} \in \{0, 1\}$ are indicating the presence of the two classes, and the class means are given by (see Fig. 3)

$$\beta_1 = [-\mu, +\mu] \quad \text{and} \quad \beta_2 = [+\mu, -\mu]. \tag{44}$$

Taken together, this is equivalent to the model behind linear discriminant analysis (LDA) [41], with class-dependent multivariate means and class-independent between-feature covariance. MBC generalizes LDA to additional variables influencing the features (see Simulation 2), an arbitrary number of classes (see Simulation 3), and potential between-observation correlation (see Simulation 4).

The parameters μ and σ are controlling (i) the distance between the centers of the two distributions and (ii) the amount of spreading (standard deviation) for the two distributions. In our simulations, they are specified as

$$\begin{aligned} \sigma = 1 \quad &\text{with} \quad \mu \in \{0, 0.25, 0.5, 0.75, 1\} \text{ or} \\ mu = 1 \quad &\text{with} \quad \sigma \in \{1, 2, 3, 4, 5\}. \end{aligned} \tag{45}$$

Scientific interest: Note that μ/σ is proportional to the Mahalanobis distance between the two distributions [44] which is a measure for class separability: With large μ and small σ, observations from the two classes are very distinct, whereas with small μ and large σ, the two classes are practically overlapping (see Fig. 3A/B), which makes classification either a very easy or a very hard task.

Thus, a scientific question related to this simulated data set—or a real data set of the same form—could be how well the two classes can be separated, given some difference between class means and some variation within classes.

Data analysis: MBC is performed according to the methodology described in Sects. 2.3.1, 2.3.2 and 2.4.1 based on the multivariate general linear model

$$\begin{aligned} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + E, \\ E &\sim \mathcal{MN}(0_{n2}, \sigma^2 I_n, \Sigma) \end{aligned} \tag{46}$$

and using k-fold cross-validation (CV) with $k = 10$ CV folds.

Results: We observe that, trivially, (i) with increasing distance between the distributions, classification accuracy increases, and (ii) with increasing variance of the distributions, classification accuracy decreases (see Fig. 3). Moreover, (iii) when the distance between distributions is zero, all points receive a posterior probability close to 1/2 (see Fig. 3A), and (ii) when variance of distributions is large, many close to the decision boundary points receive such an indecisive posterior probability (see Fig. 3B).

Conclusion: We have seen that classification accuracy in binary classification increases with distance of class means and decreases with variability of classes in multivariate space. Moreover, posterior class probabilities reflect the certainty of predictions made by MBC, with posterior probability becoming practically 1/2, if the two classes overlap.

3.2 Simulation 2: two classes with confound

Purpose: The purpose of this simulation is to show that including a confound (i.e., a covariate that systematically varies with class membership) into the classification model (MBC) results in higher classification accuracy than removing the covariate effect from the feature variables before classification (SVC).

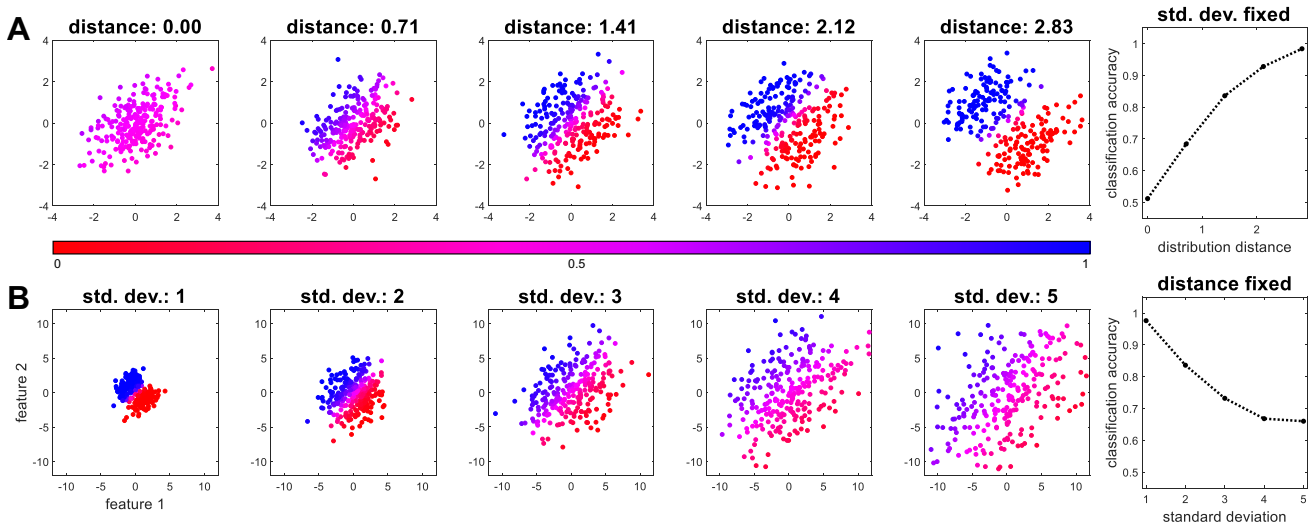

Fig. 3 *Simulation 1: two-class classification.* Simulated data ($n = 250$) from two classes (blue, red) in two dimensions (x, y) when varying **A** the distance between the distributions (measured as Euclidean distance between the multivariate means) or **B** the spread of the distributions (measured as standard deviation of the noise). The color of each dot indicates the posterior probability assigned to this data point in multivariate Bayesian classification (see colorbar). Assigning classes by maximum posterior probability results in a classification accuracy for each of the scenarios (right panels)

Generative model: We again simulate two-dimensional data by drawing from two classes, but this time adding a continuous covariate to the outcome variables: More precisely, each observation $y_i \in \mathbb{R}^{1\times 2},\ i = 1, \ldots, n$ is generated as

$$\begin{aligned} y_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + z_i\beta_3 + \varepsilon_i, \\ \varepsilon_i &\sim \mathcal{N}\left([0,0], \sigma^2\Sigma\right) \end{aligned} \tag{47}$$

where $n = 250$, the class means β_1 and β_2 are as before (see Sect. 3.1) with $\mu = 1$, $\sigma = 2$ and Σ given in (43). The continuous variable $z_i \in [-1, +1]$ is sampled as

$$\begin{aligned} z_i &= u_i + 0.25\, x_{1i} - 0.25\, x_{2i} \quad \text{with} \\ u_i &\sim \mathcal{U}(-0.75, +0.75) \end{aligned} \tag{48}$$

where $\mathcal{U}(a, b)$ is a uniform distribution with minimum a and maximum b. This sampling causes $z_i \in [-0.5, +1]$ for class 1 and $z_i \in [-1, +0.5]$ for class 2 (see Fig. 4B), such that the covariate z is mildly correlated with class membership and tends to be higher for class 1. The effect of the covariate on the two features is specified as

$$\beta_3 = [0, 2] \tag{49}$$

which means that the covariate has no effect on y_{i1} and exclusively acts on the second feature y_{i2} (see Fig. 4B).

Scientific interest: When a data set contains a continuous variable that has an effect on the features, but also varies with class membership (such as z_i), accounting for the effect of this confounder is crucial, such that patterns observed in the features may not be mistaken as being due to class, but as (partly) due to the covariate.

Thus, scientific questions related to this simulated data set—or a real data set of similar form—could be whether the two classes can be separated based on the features when accounting for the effect of the covariate associated with class membership, and whether separability differs depending on how the covariate is accounted for.

Data analysis: MBC is performed using the MGLM framework as before (see Sect. 3.1). In order to account for the correlated covariate (see Sect. 2.5), the multivariate linear model is extended to

$$\begin{aligned} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} x_1 & x_2 & z \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + E, \\ E &\sim \mathcal{MN}(0_{n2}, \sigma^2 I_n, \Sigma)\ . \end{aligned} \tag{50}$$

For comparison, SVC is performed by training a linear support vector machine with soft-margin parameter $C = 1$ on the data points from the two classes. In order to account for the confound in SVC, a linear regression model is fit to the features, removing the effect associated with z and keeping the residuals for classification. This is here referred to as the "prior regression" approach (see Fig. 4F) and is known to be problematic sometimes [45, 46]. Both MBC and SVC are performed with 10-fold cross-validation.

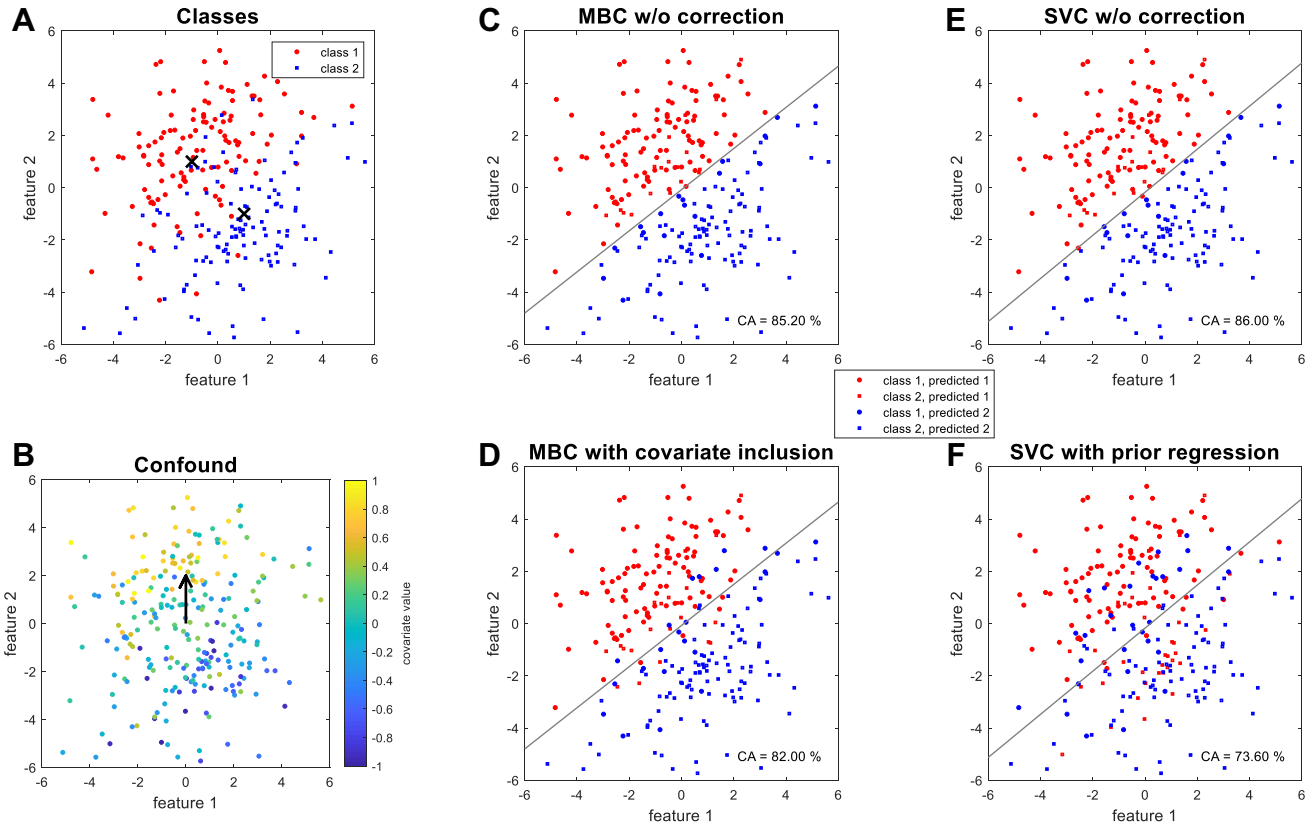

Fig. 4 *Simulation 2: two classes with confound.* **A** Two-dimensional data ($n = 250$) are simulated as coming from two classes (class means indicated by black crosses) and **B** there is an additional effect of a confounding variable in one of the dimensions (covariate effect indicated by black arrow). **C** When performing MBC, classes can be separated with high accuracy (decision boundary shown as gray line). **D** Because this effect is in part based on the covariate, accounting for this confound reduces classification accuracy. **E** When performing SVC, a similar linear decision boundary also results in high accuracy. **F** Because accounting for the covariate is done via prior regression in SVC, the confounding nature of the covariate is not properly accounted for and classification accuracy decreases by much more. Also note that MBC as well as SVC with correction achieve classification beyond linear boundaries (see panels **D** and **F**), but for different reasons: MBC, because it includes an additional covariate into the model; SVC, because prior regression transforms the data into another space (for comparison purposes, data are plotted in the original space)

Results: We observe that, the way the data are generated, values of the covariate are in fact correlated with the second feature and related to class membership (see Fig. 4A/B). Both MBC and SVC result in linear decision boundaries separating the data points into predicted classes (see Fig. 4C/E). Plausibly, because separation into classes is in part based on the covariate effect, accounting for this covariate results in a reduction in classification accuracy in MBC (see Fig. 4D). However, because the SVM approach removes the covariate effect separately from classification and thus fails to account for the correlation of class membership and covariate, the classification accuracy reduces by much more in SVC (see Fig. 4F).

In addition to this prior regression approach, we also tested a second correction method for SVC. This method includes the covariate as a feature, instead of removing its effect from the other features (see Sect. 4.4). We find that this preserves the classification accuracy of SVC without correction (CA = 86.00%; results not shown). However, this only shows that both, features and covariate, are correlated with class membership and it does not show how informative the features are about class membership when correcting for the effect of the covariate.

Conclusion: We have seen that covariate inclusion captures a confound's effect on the features while accounting for the correlation of covariate and class which leads to a more realistic estimation of class separability than removing the confound's effect via prior regression before classification.

3.3 Simulation 3: three-class classification

Purpose: The purpose of this simulation is to show that (i) MBC, unlike SVC, facilitates posterior class probabilities resulting in gradual classification boundaries and that (ii)

prior class probabilities can be used to bias decisions into direction of particular classes by incorporating prior information.

Generative model: We simulate three classes of equal size in two dimensions: Each observation $y_i \in \mathbb{R}^{1\times 2}$, $i = 1, \ldots, n$ is generated as

$$\begin{aligned} y_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + \varepsilon_i, \\ \varepsilon_i &\sim \mathcal{N}\left([0,0], \sigma^2\Sigma\right) \end{aligned} \tag{51}$$

where $n = 300$ (with 100 per class), $\sigma = 2$, Σ is as given in (43) and $x_{1i}, x_{2i}, x_{3i} \in \{0, 1\}$ where $\sum_{j=1}^{3} x_{ji} = 1$. The class means are specified as (see Fig. 5A)

$$\begin{aligned} \beta_1 &= [-1, +1] \\ \beta_2 &= [+2, +2] \\ \beta_3 &= [+1, -1]\ . \end{aligned} \tag{52}$$

Scientific interest: Multi-class or n-ary classification occurs frequently in empirical research problems such as optical character recognition, object classification, or language detection. Thus, scientific questions related to this simulated data set—or a real data set of the same form—could be to what extent classes differ in multivariate space and whether some pairs of classes can be better distinguished than other pairs.

Data analysis: MBC is performed using the MGLM framework as before (see Sect. 3.1) based on the multivariate general linear model

$$\begin{aligned} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + E, \\ E &\sim \mathcal{MN}(0_{n2}, \sigma^2 I_n, \Sigma)\ . \end{aligned} \tag{53}$$

For exploratory reasons, we also perform MBC under manipulation of the prior class probabilities, i.e., by letting $p(m_{ij})$ in (29) to be equal across classes, i.e., $p(m_{ij}) = 1/3$ for all classes (see Fig. 5B) or by setting $p(m_{ij}) = 2/3$ for one class and $p(m_{ij}) = 1/6$ for the other two classes (see Fig. 5D/E/F).

For comparison, SVC is performed by training a linear support vector machine with soft-margin parameter $C = 1$ on the data points from the three classes. Both MBC and SVC are performed with 10-fold cross-validation.

Results: We observe that, when classification was performed by maximizing posterior probability, MBC and SVC would give rise to nearly identical decision boundaries (see Fig. 5B/C). However, close to the decision boundaries, MBC provides graded information in the form of posterior probabilities rather than discrete predictions. Moreover, when incorporating prior knowledge by, e.g., increasing the prior probability for one of the classes, the decision boundaries move accordingly, allocating more posterior probability to the respective class (see Fig. 5D/E/F).

Conclusion: We have seen that MBC provides probabilistic information in multi-class decision problems and that it allows to incorporate prior information about class frequency, according to possible knowledge about the prevalence of categories.

3.4 Simulation 4: continuous regression

Purpose: The purpose of this simulation is to show that MBR, in the presence of between-observation covariance, achieves predictive correlation comparable with SVR and results in predicted targets whose distribution is more similar to the training data than with SVR.

Generative model: We simulate multi-dimensional data with continuous effects: Each observation $y_i \in \mathbb{R}^{1\times v}$, $i = 1, \ldots, n$ is generated as

$$\begin{aligned} y_i &= x_i\beta_1 + \beta_0 + \varepsilon_i, \\ \varepsilon_i &\sim \mathcal{N}\left(0_{1v}, \sigma^2\Sigma\right) \end{aligned} \tag{54}$$

where $x_i \in [-1, +1]$ is a continuous regressor sampled from a uniform distribution, β_1 and β_0 are $1 \times v$ row vectors with slopes and intercepts for the features y_i and Σ is a $v \times v$ matrix inducing between-feature covariance[2]:

$$\Sigma = \begin{bmatrix} \nu^0 & \nu^1 & \cdots & \nu^{v-1} \\ \nu^1 & \nu^0 & \cdots & \nu^{v-2} \\ \vdots & \vdots & \ddots & \vdots \\ \nu^{v-1} & \nu^{v-2} & \cdots & \nu^0 \end{bmatrix},\ 0 < \nu < 1\ . \tag{55}$$

For demonstration purposes, we here also modulate the covariance between data points by applying an $n \times n$ matrix V inducing between-observation covariance:

$$V = \begin{bmatrix} \tau^0 & \tau^1 & \cdots & \tau^{n-1} \\ \tau^1 & \tau^0 & \cdots & \tau^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \tau^{n-1} & \tau^{n-2} & \cdots & \tau^0 \end{bmatrix},\ 0 < \tau < 1\ , \tag{56}$$

[2] For simplicity, this and the next matrix are specified as Toeplitz matrices with constant descending diagonals, i.e., the entries only depend on the absolute difference between row and column index $(i - j)$.

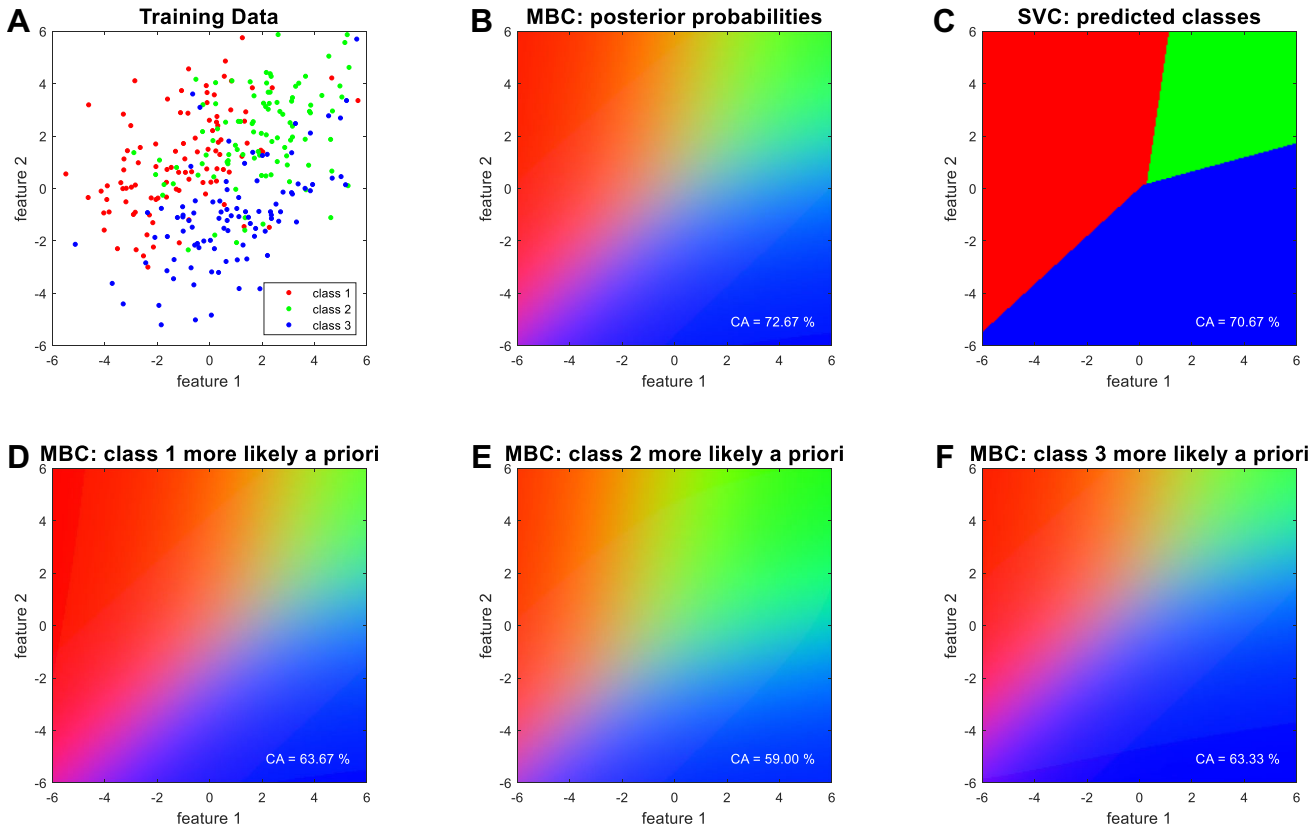

Fig. 5 *Simulation 3: three-class classification.* **A** Simulated data ($n = 300$) from three classes (red, green, blue) in two dimensions (x, y). **B** Posterior class probabilities from MBC as a function of location in feature space are visualized as RGB triplets (e.g., [0.5, 0.0, 0.5] results in purple, which corresponds to equal posterior probability for classes 1 and 3; linear RGB converted to sRGB for visualization purposes). **C** Predicted classes from SVC are shown as a function of location in feature space. **D**,**E**,**F** Varying the prior class probabilities for MBC results in higher posterior probabilities for the class which is more likely a priori. Since classes were equally likely a priori, classification accuracy (CA) is highest when a uniform prior distribution is used (see panel **B**)

such that the simulation model for the entire data set is given by:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x & 1_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_0 \end{bmatrix} + E, \qquad (57)$$
$$E \sim \mathcal{MN}(0_{nv}, \sigma^2 V, \Sigma) \; .$$

In our simulations, we set $n = 200$, $v = 10$, $\sigma = 1$, $\nu = 0.5$, $\tau = 0.25$ and all entries of β_0 and β_1 were sampled from the standard normal distribution $\mathcal{N}(0, 1)$.

Scientific interest: If the rows of Y correspond to units of time (i.e., time series measurements), τ may also be called "time constant" governing the "temporal covariance" V. If the columns of Y correspond to units of space (e.g., measurement locations), then ν may also be called "space constant" governing the "spatial covariance" Σ [16]. In this situation, a scientific question could be whether and to what extent the continuous quantity x_i can be inferred based multivariate data y_i, accounting for (known) temporal covariance and (estimated) spatial covariance. Moreover, values of x_i may fall into a range that may not necessarily preserved by a particular algorithm.

Data analysis: MBR is performed according to the methodology described in Sect. 2.4.2 based on the multivariate general linear model given by (57), with the aim of reconstructing the target values x_i, $i = 1, \ldots, n$. As the prior density $p(x_i) = p(\lambda)$ in (36), we use a continuous uniform distribution, i.e., $p(\lambda) = 1/2$ where $-1 < \lambda < +1$. Resulting posterior densities are calculated via discretization of the interval $[-1, +1]$ in steps of 0.01 (see Fig. 6D).

For comparison, SVR is performed by training a linear support vector machine with soft-margin parameter $C = 1$ on the data matrix Y and the target vector x. Both MBR and SVR are performed with 10-fold cross-validation.

Results: We observe that, trivially, maximum-a-posteriori (MAP) estimates from MBR fall into $[-1, +1]$, because the prior probability is zero outside this interval. Notably, the resulting distribution of predicted target values is closer to

the distribution of actual target values (see Fig. 6A) for MAP estimates from MBR (see Fig. 6E), when compared with SVM predictions (see Fig. 6F). MBR and SVR are otherwise comparable in terms of predictive correlation and mean absolute error (see Fig. 6B/C). In addition to this, it appears that posterior target densities $p(\lambda|y_{2i})$ from MBR are univariate normal distributions when uniform prior densities are applied (see Fig. 6D).

Conclusion: We have seen that MBR is able to reconstruct continuous labels from a multi-dimensional feature space and that it can effectively constrain the reconstructed variable into its natural range by a prior distribution restricted to that range.

3.5 Comparison with alternative approaches

Purpose: In addition to comparison with SVMs (see Figs. 4, 5, 6), we also compare MBI's simulation performance against that of other ML approaches for selected simulations. This includes generative and discriminative as well as Bayesian and non-Bayesian methods for classification and regression problems.

Data analysis: We re-analyze data from Simulation 3 using the following techniques:

- *naive Bayes classification* (GNB; multivariate normal likelihood, uniform prior);
- *linear discriminant analysis* (LDA; linear discriminant function);
- *multi-class logistic regression* (LogReg; regularization parameter $\lambda = 1$);
- *random forest classification* (RFC; bootstrap aggregation)
- *neural network classification* (NNC; two hidden layers with ten neurons each, sigmoid activation function)

Moreover, we re-analyze data from Simulation 4 using the following techniques:

- *naive Bayes regression* (GNB; multivariate normal likelihood with linear effect of label variable on feature variables, uniform prior density over label variable);
- *multiple linear regression* (LinReg; weighted least squares estimation);
- *random forest regression* (RFR; bootstrap aggregation)
- *neural network regression* (NNR; two hidden layers with ten neurons each, linear activation function)

We distinguish between generative approaches, i.e., methods modeling feature variables as a function of the label variable (e.g., LDA, and including MBC), and discriminative approaches, i.e., methods modeling the label variable as a function of the feature variables (e.g., LogReg, and including SVC).

In order to systematically compare diverse approaches, predictive performance of all techniques is measured as rank graduation accuracy (RGA) [47] which always falls between 0 and 1, with a value of 0.5 corresponding to chance-level performance. RGA is equivalent to area under the curve (AUC) in binary classification, but generalizes to n-ary classification and regression problems [48].

Results: Because MBC and LDA are equivalent for multi-class classification with i.i.d. observations and without covariates (see Sect. 2.7), both show identical performance for Simulation 3 (RGA $= 0.8591$). As expected, they perform better than GNB which disregards covariances between features (RGA $= 0.8448$). The highest performance is achieved by SVC (RGA $= 0.8611$); all other discriminative approaches have performance lower than that of MBC (see Table 3).

In Simulation 4, differences between regression techniques are negligible, with MBR (RGA $= 0.9327$) and NNR (RGA $= 0.9318$) performing best among generative and discriminative approaches, respectively; GNB (RGA $= 0.9218$) and RFR (RGA $= 0.9143$) performing worst among generative and discriminative approaches, respectively. All other discriminative approaches have predictive performance between that of GNB and NNR (see Table 3).

Conclusion: MBI-based predictive analysis for classification and regression reaches predictive performance comparable with that of established (generative or discriminative, Bayesian or non-Bayesian) ML techniques in simulation studies.

4 Applications

In this section, we apply cross-validated multivariate Bayesian inversion to several empirical data sets necessitating either classification (Analyses 1–3) or regression (Analysis 4). Again, we describe (i) the purpose behind each empirical example, (ii) specificities, dimension (and possibly, acquisition) of the respective data set, (iii) possible scientific interest in this data set, (iv) steps of data analysis, we give a (v) summary of the results and provide (vi) a conclusion.

For all analyses, we compare MBI's performance with that of support vector machines (SVM), as a popular ML approach that is often used for multivariate continuous data [42, 43].

4.1 Analysis 1: Egyptian skull data

Purpose: The purpose of this analysis is to show that MBC achieves (above-chance) classification accuracy comparable with SVC in different binary and multi-class classification scenarios for a single data set without covariates.

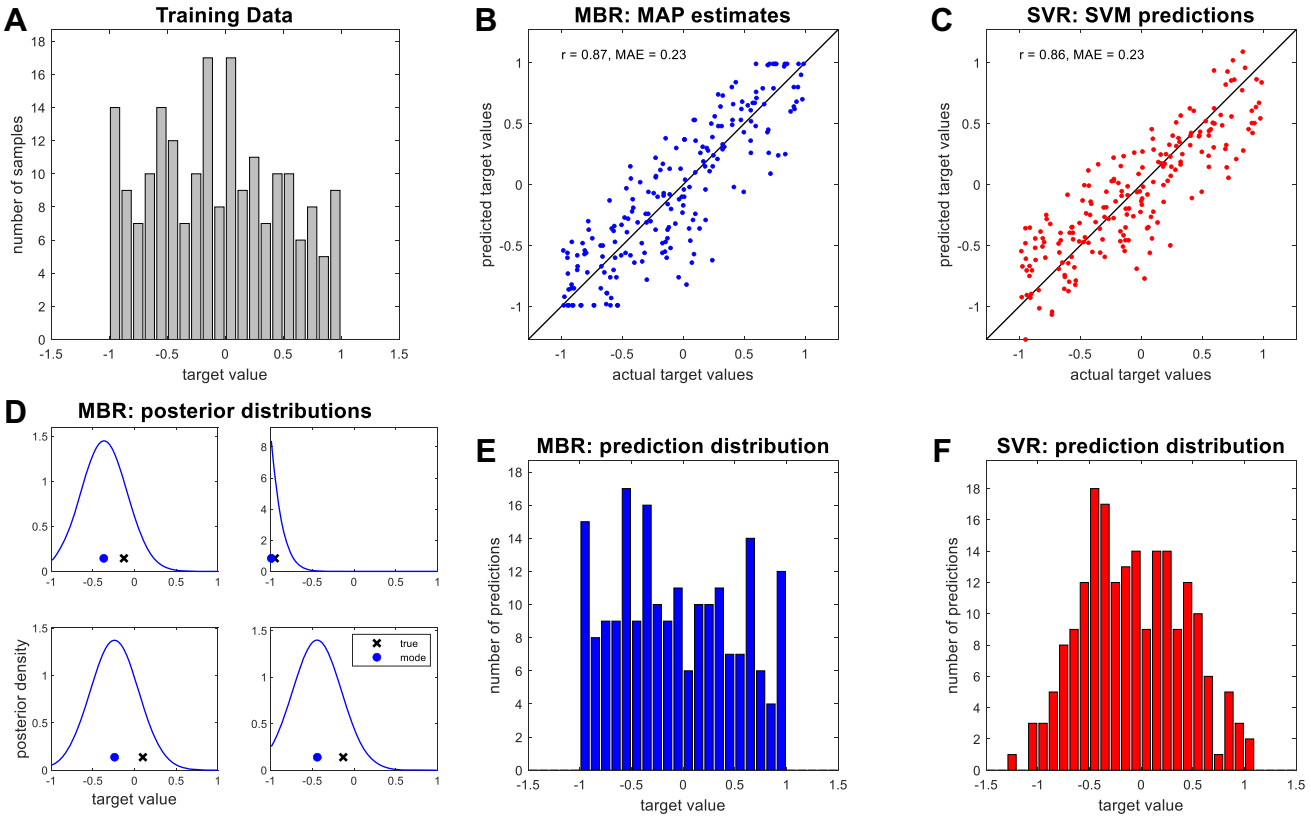

Fig. 6 *Simulation 4: continuous prediction.* **A** A continuous predictor variable x (the histogram of which is shown) is used to sample a 200×10 data matrix Y (for details, see text). **B** Maximum-a-posteriori (MAP) estimates from MBR and **C** support vector machine (SVM) predictions from SVR are plotted against the actual labels. **D** Exemplary posterior target densities for the first four simulated data points, along with MAP estimate (blue dots) and true values (black crosses). **E** MAP estimates and **F** SVM predictions for the reconstructed targets are plotted as histograms

Table 3 *Simulations: comparisons with alternative approaches.* Predictive performance for generative and discriminative techniques when analyzing data from Simulation 3 and Simulation 4. For details regarding MBC and SVC, see Fig. 5, and for details regarding MBR and SVR, see Fig. 6. In all cases, predictive performance is given as rank graduation accuracy (RGA) [48]

Simulation 3	RGA	Simulation 4	RGA
generative methods			
naive Bayes classification	0.8448	naive Bayes regression	0.9218
multivariate Bayesian classification	0.8591	multivariate Bayesian regression	0.9327
linear discriminant analysis	0.8591		
discriminative methods			
multi-class logistic regression	0.8544	multiple linear regression	0.9250
support vector classification	0.8611	support vector regression	0.9309
random forest classification	0.8169	random forest regression	0.9143
neural network classification	0.7818	neural network regression	0.9318

Data set: The Egyptian skull data set ([49]; [26, p. 287]) comprises 4 skull measurements (maximal breadth of the skull, MB; basibregmatic height, BH; basialveolar length, BL; nasal height, NH) from 5×30 human skulls coming from 5 time periods (4000 BCE, 3300 BCE, 1850 BCE, 200 BCE, 150 CE). Thus, we have $v = 4$ features from $n = 150$ observations belonging to $C = 5$ classes. The entire data set is visualized in Fig. 7A.

Scientific interest: Scientific questions related to this data set could be:

- To what extent are skulls from various eras differing in terms of MB, BH, BL, NH?
- Are skulls differing more strongly, if they are from temporally more distant eras?

- Given a random skull, with what accuracy can it be assigned to a time period?

Data analysis: Analyses are performed for classification into five groups (all time periods) or distinction into three categories (−4000 vs. −1850 vs. 150) or binary classification (−4000 vs. 150). MBC is performed as described in Sect. 2.4.1, setting $V = I_n$ for no covariance between observations. SVC is performed, setting the soft-margin parameter $C = 1$. Both MBC and SVC are performed with 10-fold cross-validation.

Results: We observe that, as could be expected, classification accuracy increases with decreasing number of classes (see Fig. 7C). However, MBC as well as SVC both exceed chance level in each classification task (see Fig. 7C). Confusion matrices suggest that the most extreme time periods (−4000, 150) were predicted with highest accuracy (see Fig. 7B) and that the 3rd era (−1850) can be better recognized than the 2nd and the 4th era (−3300, −200), possibly because the latter were closer in time to the 1st and the 5th era (−4000, 150) than to the middle-most time period (see Fig. 7B).

Conclusion: We have seen that a comparably large number of classes (5) can be separated based on a comparably small number of samples per class (30) using a comparably low number of features (4). Moreover, confusion matrices show that misclassifications become less frequent with increasing temporal distance between true and predicted class which highlights empirical plausibility of the results.

4.2 Analysis 2: MNIST digit recognition

Purpose: The purpose of this analysis is to show that MBC (and SVC) accuracy is increasing with growing number of training samples and that MBC-based posterior probabilities approximate true class frequencies of predicted test data points.

Data set: The MNIST database [50, 51] consists of images of handwritten digits 0 to 9. Original images from the US National Institute of Standards and Technology (NIST) were size-normalized into a 20×20 pixel box, converting binary black-and-white images into gray-scale levels in the interval $[0, 255]$, and placed into a 28×28 pixel image by computing the center of mass of the pixels. The data set contains $n_1 = 60{,}000$ training samples, $n_2 = 10{,}000$ test samples, the number of classes is $C = |\{0, \ldots, 9\}| = 10$ and the number of features is $v = 28 \times 28 = 784$.

Scientific interest: Scientific questions related to this data set could be:

- With what accuracy can handwritten digits be recognized from pixelated images?
- Which pairs of digits are more often confused with each other than other pairs?
- Given a digit image Y has been assigned a probability p of being digit X, what is the frequency of this assignment being correct?

Data analysis: During preprocessing, 4 pixels were removed on either side of each image, extracting only the center 20×20 pixels, resulting in a reduced number of features $v = 20 \times 20 = 400$. Moreover, all pixel intensity values were rescaled into the interval $[0, 1]$. Each preprocessed image matrix $\tilde{Y}_i \in \mathbb{R}^{20\times 20}$ was vectorized into a row vector $y_i \in \mathbb{R}^{1\times 20\cdot 20}$, giving an $n_1 \times v$ training data matrix Y_1 and an $n_2 \times v$ test data matrix Y_2.

MBC is performed as described in Sect. 2.4.1, setting $V = I_n$. SVC is performed, setting the soft-margin parameter $C = 1$. Both MBC and SVC are performed for classification into ten categories (digits 0-9). And with 10-fold cross-validation.

Results: We observe that classification accuracy on the test set depends on the number of training samples used for estimating model parameters and that SVC reaches higher classification accuracy than MBC across all numbers of training samples, with both clearly exceeding chance level in the digit classification task (see Fig. 8A). This is also evident from confusion matrices which show higher misclassification rates for MBC than for SVC, as the former especially confuses 3 with 5 and 4 with 9 (see Fig. 8B/C).

Next, we evaluated how accurately the posterior class probabilities calculated for test samples reflect the actual class of those samples. Each test sample is assigned ten posterior class probabilities (for digits 0-9) and can then be predicted as belonging to the class for which posterior class probability is highest. Therefore, we binned test data items by highest posterior probability (25 bins, bin width 0.04) and calculated, for each bin, the frequency that the class with the highest posterior probability was also the true class. There is a linear increasing relationship between posterior probability of the most likely class and the frequency of the most likely class being the true class (see Fig. 8D), indicating that posterior probabilities are somewhat veridical. However, the relationship is slightly below the identity line (see Fig. 8D), suggesting that MBC posterior probabilities based on the MNIST training data are a bit overconfident.

For illustration, we also show the posterior inverse scale matrix Ω_1 (cf. eq. 19c), embodying the covariance between features (see Fig. 8E), and the posterior precision matrix Λ_1 (cf. eq. 19b), embodying the covariance between classes (see Fig. 8F), according to the training data. As nearby image pixels tend to be correlated (digits are contiguous, so neighboring pixels have similar intensity values) and all image matrices were vectorized before training, Ω_1 is characterized by diagonal lines which become weaker, the higher the distance from the main diagonal (and thus, the more two pixels

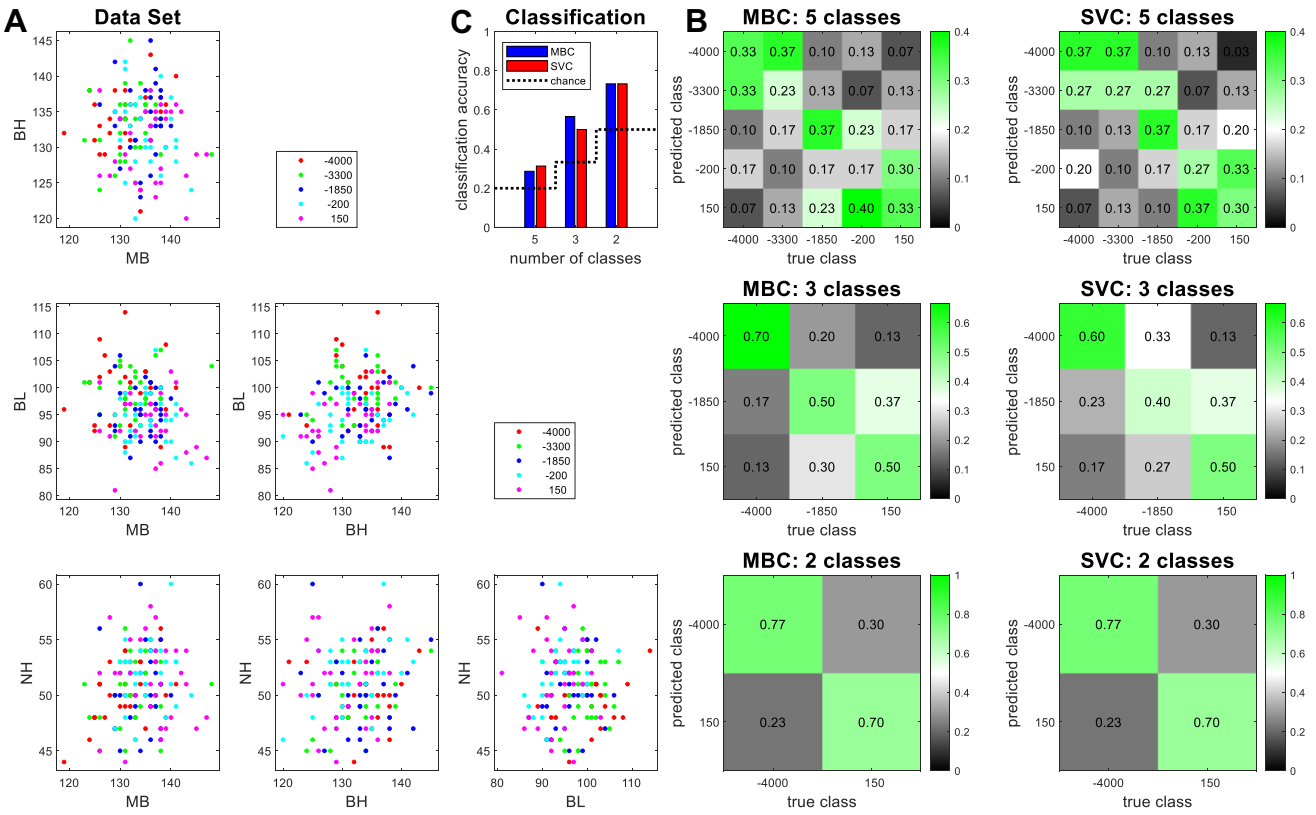

Fig. 7 *Analysis 1: Egyptian skull data.* **A** Maximal breadth of the skull (MB), basibregmatic height (BH), basialveolar length (BL) and nasal height (NH) are shown for 30 skulls each from around 4000 BCE (red), 3300 BCE (green), 1850 BCE (blue), 200 BCE (cyan) and 150 CE (magenta). Note that all measurements are given in integer millimeters (mm), causing an impression of discreteness in the plots. **B** 5-class, 3-class and binary classification were exercised with the goal to detect the historical era based on the four skull measurements. Confusion matrices show the proportion of predicted classes (y-axis), given the true class (x-axis). **C** Classification accuracy, i.e., the proportion of correct predictions across all observations, is shown as a function of classification algorithm and number of classes. Classification accuracy decreases with increasing number of classes, but MBC and SVC achieve comparable performance

are apart from each other in the image). As classes are mutually exclusive (no data point belongs to two or more digits at the same time), Λ_1 is a diagonal matrix with all off-diagonal entries (representing between-class covariances) being equal to zero.

Finally, we present exemplary test set cases which were correctly classified with either highest posterior probability (see Fig. 9, top row) or lowest posterior probability (see Fig. 12, middle row) or which were misclassified, with posterior probability being maximal for the wrong digit (see Fig. 12, bottom row). These examples anecdotally show that images are more likely to be correctly classified, if they are drawn in bold, rather than thin lines (see Fig. 12, top and middle row), and that they are more likely to be misclassified, if they are similar to other digits (see Fig. 12, bottom row).

Conclusion: We have seen that classification accuracy improves with increasing size of the training set, as the posterior distribution over MGLM model parameters (see Sect. 2.2.1) is more accurately estimated when more training data are available. Moreover, posterior probabilities derived from MBC are veridical with respect to the true class in the sense that they approximately reflect the frequency with which a test set sample is the class to which the highest posterior probability was assigned.

4.3 Analysis 3: birth weight data

Purpose: The purpose of this analysis is to show that MBC achieves higher classification accuracy than SVC and less "classification artifacts" (i.e., (near-)exclusive prediction of one class) in situations where one categorical variable (here: smoking) is predicted while other categorical variables (here: ethnicity) and parametric variables (here: age) influence the feature variables.

Data set: The low birth weight data set ([52]; [26, p. 288]) consists of a number of measurements from $n = 189$ US mother–baby pairs which was collected to investigate potential causes of low birth weight. Collected variables include the birth weight (in kg), the mother's weight (in kg), smoking status (non-smoker, smoker), ethnicity (white, black, other),

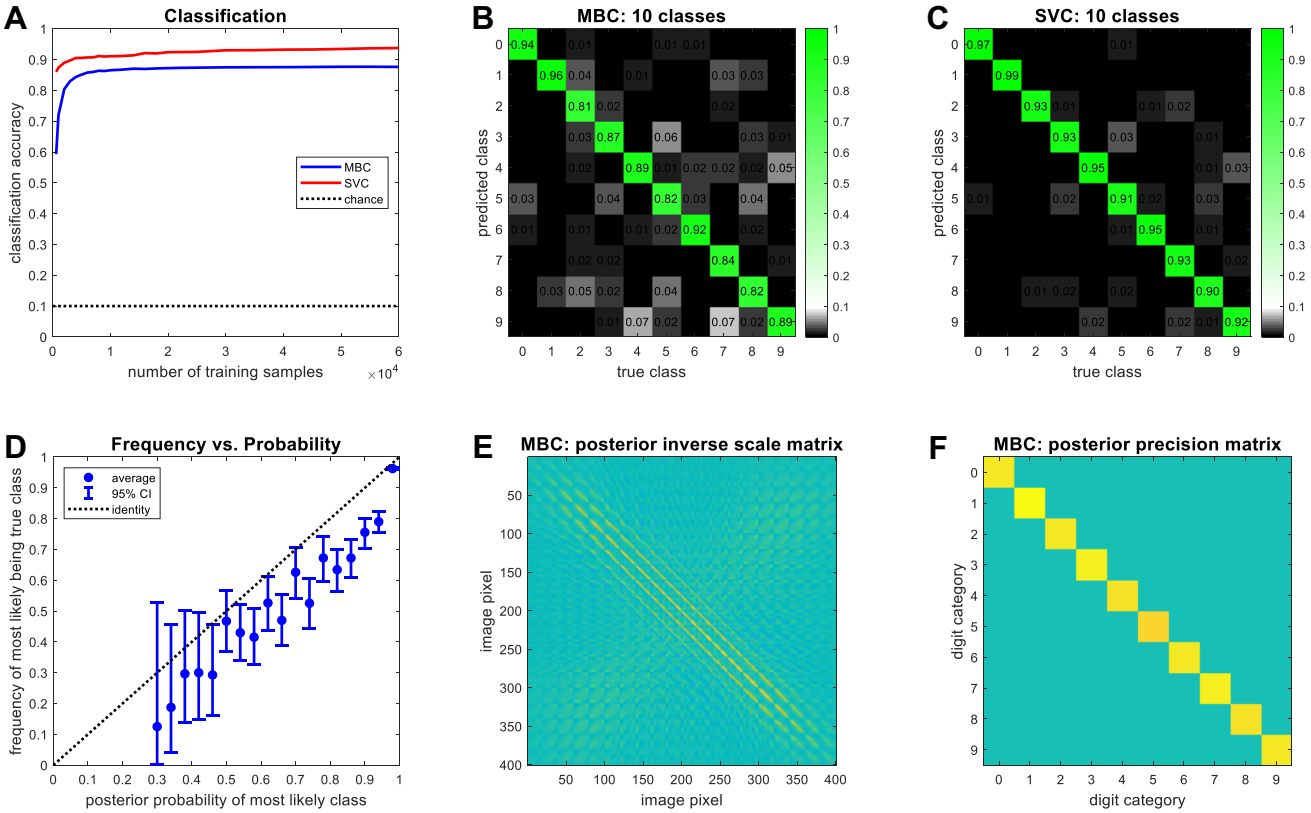

Fig. 8 *Analysis 2: MNIST digit recognition.* **A** Classification accuracy in the test set ($n_2 = 10{,}000$) is shown as a function of the number of training samples (ranging from 600 to 60k) used to estimate parameters for multivariate Bayesian classification (MBC; blue) and support vector classification (SVC; red). Classification analyses were performed to distinguish digits 0-9. For the full training set ($n_1 = 60{,}000$), confusion matrices show the proportion of predicted classes (y-axis) given the true class (x-axis) for **B** MBC and **C** SVC. Misclassification rates are higher for MBC. **D** MBC assigns posterior class probabilities to test set samples and the optimal choice is given by predicting the class with the highest posterior probability. We plot the frequency of the most likely class being the actual class against binned highest posterior probability. Blue dots represent estimated frequencies, and error bars denote 95% binomial confidence intervals. **E** MBC's posterior inverse scale matrix Ω_1 (after training), indicating covariance between image pixels. Nearby pixels show higher covariance than pixels which are far apart. **F** MBC's posterior precision matrix Λ_1 (after training), indicating covariance between classes. Because classes are mutually exclusive, no between-class covariance exists

hypertension (no, yes), the mother's age (in yrs), and the number of visits to the doctor before birth. The entire data set is visualized in Fig. 10A.

Scientific interest: Presumably, the original purpose of acquiring the data set was to identify causes of low birth weight, i.e., to predict birth weight from the other variables. Turning this question around, one can also ask whether the effects of lifestyle, demographics, and health on body weight are so strong that these categorical variables can be detected from the parametric ones. In our analysis, the goal therefore is to predict the 3 discrete variables (smoking, ethnicity, hypertension) from the 2 continuous features (birth weight, mother's weight). Challenges in this prediction task are (i) strongly unequal class sizes (see Fig. 10A), (ii) controlling for the dimensions not currently predicted, and (iii) confounding variables (mother's age, visits to the doctor).

Data analysis: Analyses are performed for classification into two categories (smoking, hypertension) or three categories (ethnicity). MBC is performed as described in Sect. 2.4.1, setting $V = I_n$ as data points were regarded independent. SVC is performed, setting the soft-margin parameter $C = 1$. Both MBC and SVC are performed with 10-fold cross-validation, as described in Sect. 2.6.

In order to account for potentially confounding variables, a design matrix Z was generated by forming difference regressors ($+1$ vs. -1) for the categories not currently classified (e.g., hypertension when predicting smoking) and including continuous covariates (e.g., mother's age) as further regressors. For MBC, this covariate design matrix Z was added to the categorical design matrix X_C used for classification itself, as described in Sect. 2.5. For SVC, a linear regression model is fit to the features, removing the effect associated with Z and keeping the residuals for classification analyses.

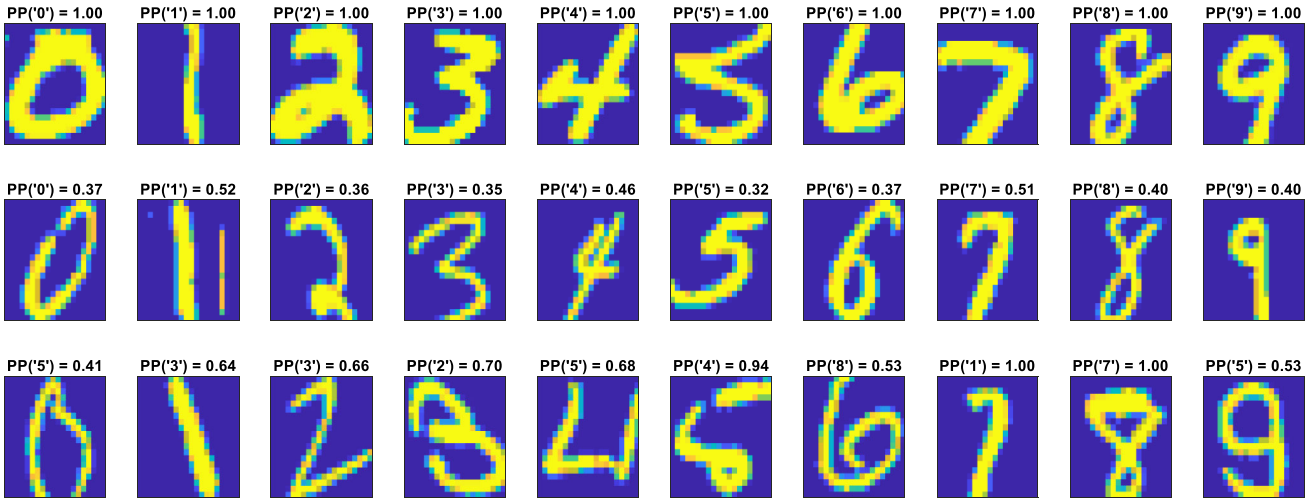

Fig. 9 *Analysis 2: exemplary test set cases.* For each class of digits, three samples from the test set are shown with the highest posterior class probability assigned to these samples ("PP('5') = 0.41" means that this test sample is most likely digit 5, with a posterior probability of around 41%). The first row shows correct classifications with the highest posterior class probability for each digit ("high-confidence hits"). The second row shows correct classifications with the lowest posterior class probability for each digit ("low-confidence hits"). The third row shows one random incorrect classification in which the class with the highest posterior probability differs from the true class ("misses")

To assess classification performance, we used the balanced accuracy (BA), i.e., the average of class-wise accuracies, rather than the classification accuracy (CA), as CA can lead to misleading conclusions in the case of unequal group sizes which are, however, compensated by the BA [53, 54].

Results: We observe that SVC with prior regression fails to accommodate the unequal group sizes in all three classification tasks (smoking, ethnicity, hypertension), frequently predicting the largest class, independent of the true class (see Fig. 10B). Note that, when samples were weighted by class size in the respective training set, predicted class was more correlated with true class for all three label variables, producing above-chance accuracies (results not shown). We here report unweighted results for SVM analyses, as no classification approach was informed with group sizes in training or test set analysis.

In contrast with SVC, MBC with covariate inclusion can seemingly adjust for the unequal numbers of samples, resulting in acceptable class-wise accuracies throughout (see Fig. 10B). As a consequence, balanced accuracy based on SVC is at chance level whereas MBC achieves above-chance prediction accuracy (see Fig. 10C). We hypothesize that MBC with covariate inclusion manages to account for interactions between discrete variables and for the effects of covariates (see Fig. 10A), thereby more faithfully estimating the true pattern underlying the categories currently classified.

We also ran another SVC analysis that includes the covariates as features, instead of removing their effects from the other features (see Section 4.4). We find that this gives above-chance MBC-like balanced accuracies for smoking (BA = 65.29%) and ethnicity (BA = 48.12%), but chance-level performance for hypertension (BA = 50.00%; results not shown). This, however, only demonstrates that features and covariates *together* are predictive of smoking and ethnicity, but it makes no statement about how informative the features are about those variables when *correcting* for the covariates' effects.

Conclusion: We have seen that MBC outperforms SVC in a relatively small data set with few features, potentially confounding continuous variables (here: mother's age, visits to the doctor), potentially interacting discrete variables (here: smoking, ethnicity, hypertension) and highly unequal class sizes.

4.4 Analysis 4: brain age prediction

Purpose: The purpose of this analysis is to show that MBR achieves higher predictive correlation than SVR when variables covarying with the features are included as covariates influencing the features (MBR), rather than as features predicting the labels (SVR).

Data set: In brain age prediction, the goal is to estimate an individual person's chronological age (in years) from structural magnetic resonance imaging (MRI) scans of this person [55, 56]. For the present analysis, we are working with dimensionality-reduced structural MRI information from $n = 3{,}300$ subjects that was obtained by averaging

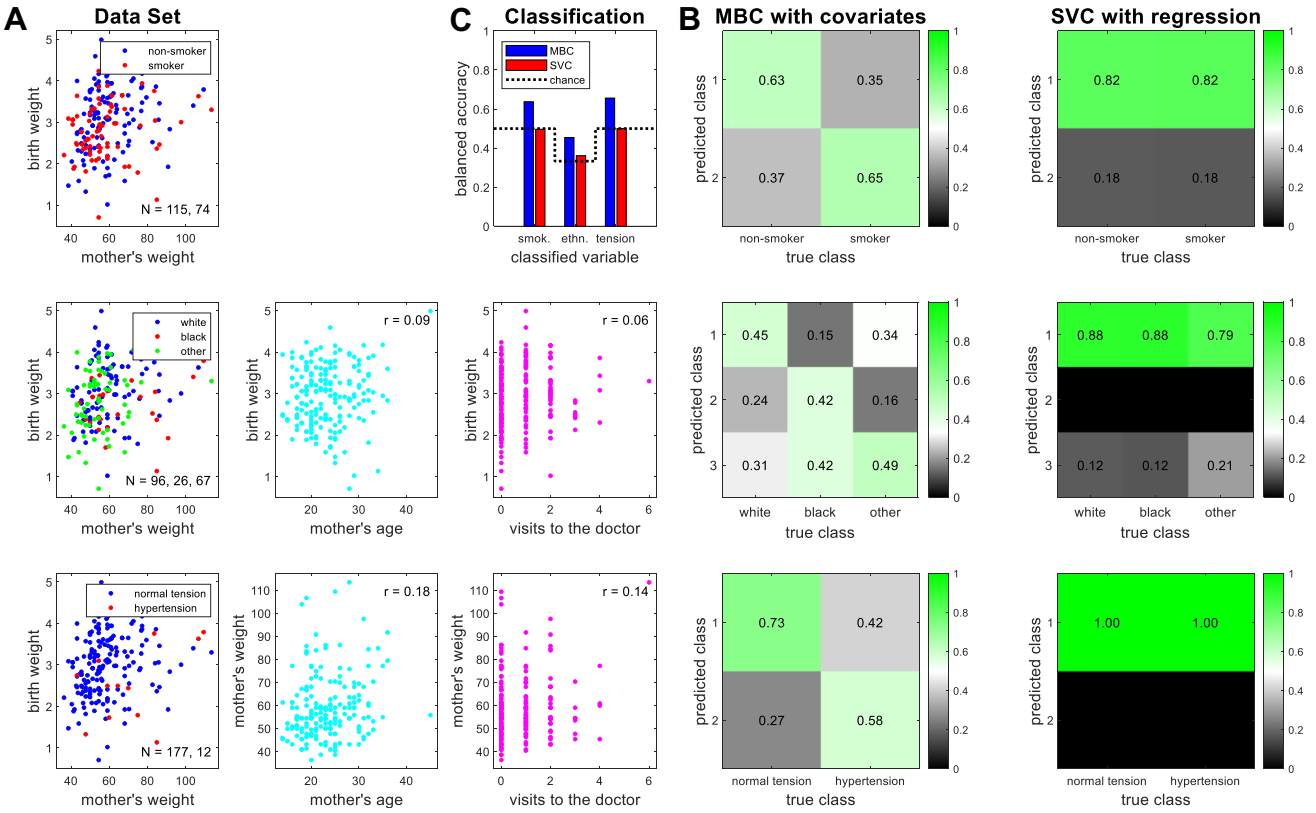

Fig. 10 *Analysis 3: birth weight data.* **A** Birth weight [kg] and mother's weight [kg] as a function of smoking status (top), ethnicity (middle), hypertension (bottom) and as a function of mother's age [yrs] (cyan) and visits to the doctor (magenta). **B** Classification analyses were executed with the goal to distinguish the discrete variables based on the two features, accounting for covariates. Confusion matrices show the proportion of predicted classes (y-axis), given the true class (x-axis). **C** Balanced accuracy, i.e., the average over class-wise accuracies (diagonal entries in confusion matrices), is shown as a function of classification algorithm and classified variable. Whereas SVC remains at chance level, MBC exhibits above-chance prediction accuracy

within the regions from an anatomical atlas of the human brain and already used in an earlier study [57].[3]

As the data were analyzed within a predictive analytics competition,[4] the analysis was not fully cross-validated, but the data set was split into $n_1 = 2{,}640$ training samples (available during the competition) and $n_2 = 660$ validation samples (target values not disclosed, used to decide the competition) with chronological age between 17 and 89 years (see Fig. 11, left column). For each of these subjects, gray matter (GM) and white matter (WM) densities for 116 brain regions were extracted. Additionally, each subject's gender (coded as $+/-1$) and the site at which they were acquired (1 out of 17) were known [57, 58].

Scientific interest: Scientific questions related to this data set could be:

- Which are the brain features that are most strongly varying with chronological age?
- With what precision can chronological age be detected based on brain features?
- How can the non-uniform distribution of target values in the training set be handled?

Data analysis: SVR was performed by calibrating a linear support vector machine with soft-margin parameter $C = 1$ on the training data, using chronological age as the target variable x and the $2 \times 116 + 17 + 1 = 250$ variables (= GM/WM × regions + sites + gender) as feature matrix Y, and then predicting chronological age in the validation data (see [57, sec. 2.2, 2.3.2] for details). Adding site and gender to the feature matrix, although different from the "prior regression" approach used above (see Simulation 2 and Analysis 3), is compatible with common practice in biomedical research in which covariates are added to the feature matrix when predicting class labels [59] or target values [60].

[3] The data set can be obtained from: https://github.com/JoramSoch/PAC_2019.

[4] See: https://github.com/ohbm/OpenScienceRoom2019/issues/10.

By contrast, in a statistical modeling context, it is not really useful to employ covariates such as site and gender as features for prediction of age, since gender was putatively controlled when splitting the data and the whole age range was putatively sampled at each site [61]. And even if this wasn't the case, prediction of a subject's chronological age from gender or site is not really something that would be demanded. Instead, structural MRI features could be varying as a function of gender or site, such that these must be included as influential covariates, not additional features.

MBR was therefore performed using the MGLM framework outlined above (see Sects. 2.4.2 and 2.5) based on the multivariate general linear model

$$\begin{aligned} Y &= \left[x\ 1_n\ Z\right] B + E, \\ E &\sim \mathcal{MN}(0_{nv}, I_n, \Sigma) \end{aligned} \tag{58}$$

where Y is a $n_1 \times (2 \times 116)$ matrix of WM and GM densities, x is an $n_1 \times 1$ vector of chronological age, 1_n is a constant regressor of the same size, and Z is a $n_1 \times 17$ covariate design matrix including indicator regressors for site and gender effects, such that B is a $(2+17) \times (2 \times 116)$ matrix of regression coefficients, describing the influences of x and Z on Y. In this way, effects of site and gender on structural MRI measures are accounted for when reconstructing chronological age from these quantities.

For MBR, we applied three different prior target densities (see Fig. 11, bottom row): (i) a uniform prior, with constant prior density for ages between 0 and 100 years; (ii) a data-driven prior, the probability for each integer age being equal to its frequency of occurrence in the training set; and (iii) a fitted prior, from fitting a shifted gamma distribution to the target values in the training set (see Fig. 11, top left). Like in Simulation 4, posterior densities were obtained via discretization of the interval [0, 100] in steps of 1.

Results: We observe that (i) all MBR variants achieve a higher predictive correlation (uniform: $r = 0.91$; data driven: $r = 0.91$; fitted: $r = 0.92$) than SVR ($r = 0.83$) when predicting chronological age (see Fig. 11, top row); (ii) this is underscored by the fact that the distribution of MBR predictions is closer to the actual distribution of targets than the distribution of SVR predictions (see Fig. 11, middle row); and (iii) using the fitted prior density for age results in the smallest mean absolute error (MAE = 4.72 yrs), outperforming the uniform prior (MAE = 5.77 yrs) by about a year and SVR (MAE = 6.99 yrs) by more than two years (see Fig. 11, top row).

We also present exemplary posterior target densities using the three different priors (see Fig. 12) which anecdotally show how the data-driven and the fitted prior can occasionally draw the MAP estimates closer to the true value (see Fig. 12, right column). Replicating Simulation 4, we also repeat our observation that posterior target densities from MBR seem to approximate univariate normal distributions when uniform prior densities are applied (see Fig. 12, first row).

Conclusion: We have seen that MBR outperforms SVR by including confounding variables as covariates influencing the features, rather than as features allowing to predict the targets. Additionally, different prior distributions can be incorporated, with an ecologically plausible prior distribution resulting in the smallest mean absolute error.

4.5 Comparison with alternative approaches

Purpose: In addition to comparison with SVMs (see Figs. 7-8, 10-11), we also compare MBI's empirical performance against that of other ML approaches for selected analyses. This includes generative and discriminative as well as Bayesian and non-Bayesian methods for classification and regression problems.

Data analysis: We analyze data from Analysis 2 using the same techniques and settings described in Sect. 3.5: *naive Bayes classification* (GNB); *linear discriminant analysis* (LDA); *multi-class logistic regression* (LogReg); *random forest classification* (RFC); *neural network classification* (NNC; one hidden layer with 400 neurons). Moreover, we re-analyze data from Analysis 4 using the same techniques and settings described in Sect. 3.5: *naive Bayes regression* (GNB); *multiple linear regression* (LinReg); *random forest regression* (RFR); *neural network regression* (NNR; one hidden layers with 250 neurons).

As before, we distinguish between generative approaches (e.g., GNB, and including MBC) and discriminative approa ches (e.g., RFC, and including SVC) and measure predictive performance using the rank graduation accuracy (RGA) [47] which always falls between 0 and 1, with a value of 0.5 corresponding to chance-level performance.

Results: Because MBC and LDA are equivalent for multi-class classification with i.i.d. observations and without covariates (see Sect. 2.7), both show identical performance for Analysis 2 (RGA = 0.9396). As expected, they perform better than GNB which disregards covariances between features (RGA = 0.8934). The highest performance is achieved by NNC (RGA = 0.9908) and all discriminative approaches except for LogReg have performance higher than that of MBC (see Table 4).

In Analysis 4, MBR (RGA = 0.9510) and LinReg (RGA = 0.9546) perform best among generative and discriminative approaches, respectively. Except for GNB, which fails to account for covariances between features (RGA = 0.8123), differences between regression techniques are negligible, all performing slightly below the level of MBR and LinReg (see Table 4).

Conclusion: MBI-based predictive analysis for classification and regression reaches predictive performance com-

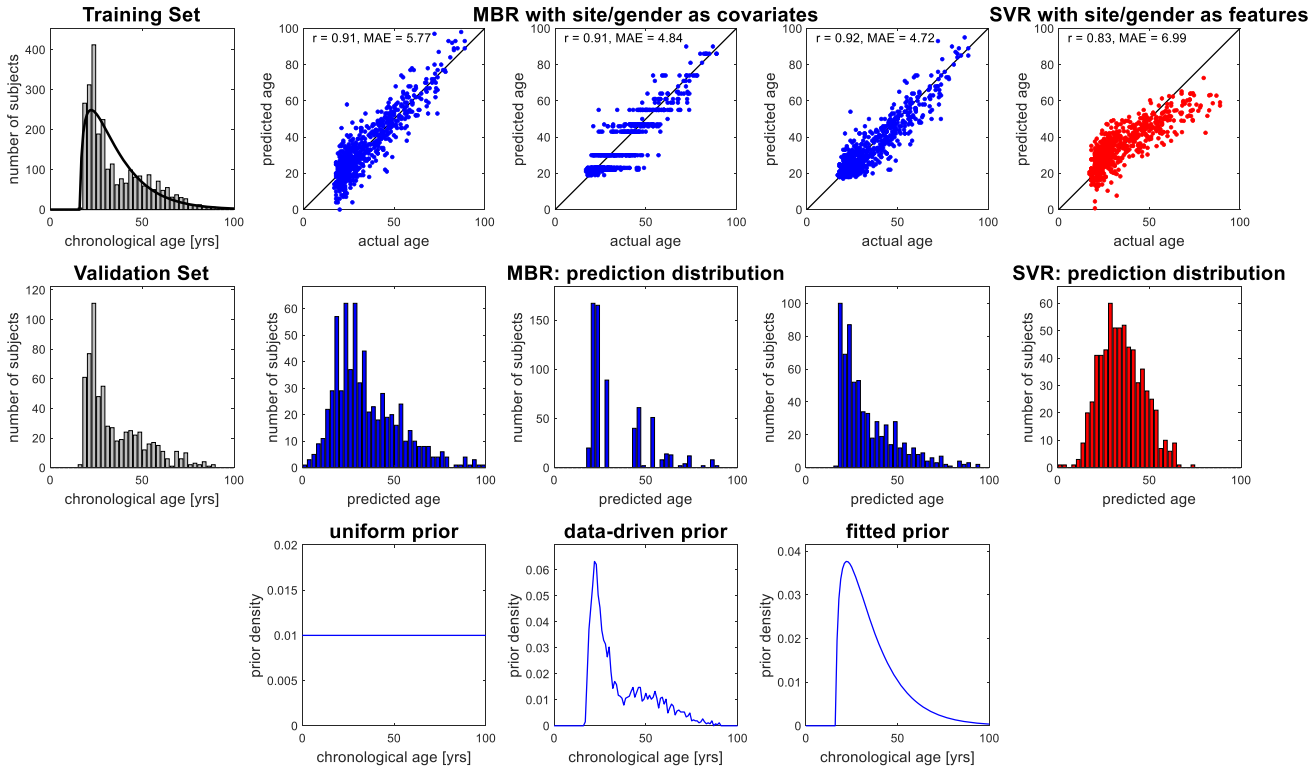

Fig. 11 *Analysis 4: brain age prediction.* Chronological age in training and validation set (left, gray) was predicted from structural MRI features (not shown) using MBR with site and gender as covariates (middle, blue), using different prior densities on age (bottom row), and using SVR with site and gender as features (right, red), resulting in higher predictive correlation and lower absolute errors for MBR (top row) and distributions of predicted values that are closer to the actual distribution (middle row)

parable with that of established (generative or discriminative, Bayesian or non-Bayesian) ML techniques in empirical data sets.

5 Discussion

In this section, we discuss some aspects pertaining to multivariate Bayesian inversion (MBI). We categorize these aspects into "advantages," i.e., benefits of MBI over other approaches, "disadvantages," i.e., shortcomings relative to other approaches, and "extensions," i.e., possibilities for increasing the applicability of MBI.

5.1 Advantages

5.1.1 Explainability and robustness

→ *Examples: Simulations 1–4, Analyses 1–4.*

The proposed MBI framework aligns well with the principles of the SAFE machine learning approach [62]. In contrast with discriminative algorithms, where model weights typically lack a direct substantive interpretation, MBI is based on an explicit generative model in which parameters quantify the effect of label variables on observed features (see Sect. 2.1). This means that the regression coefficients can be interpreted in a straightforward manner as conditional dependencies of features on labels, thereby providing a transparent link between model structure and domain-relevant mechanisms. This property enhances *interpretability* not only at the level of individual parameters, but also at the level of the full model, facilitating scientific insight and supporting informed decision-making.

With respect to *accuracy*, MBI demonstrates competitive performance relative to widely used machine learning approaches. Empirical evaluations show that MBI either outperforms or achieves comparable results to established methods, particularly support vector machines, across multiple real-world data sets (see Sects. 4.1-4.4). This indicates that the generative modeling strategy underlying MBI does not come at the cost of predictive performance. Instead, it combines strong accuracy with the additional benefit of probabilistic outputs, suggesting that high predictive quality and principled statistical modeling are not mutually exclusive, but can be jointly realized within a unified Bayesian framework.

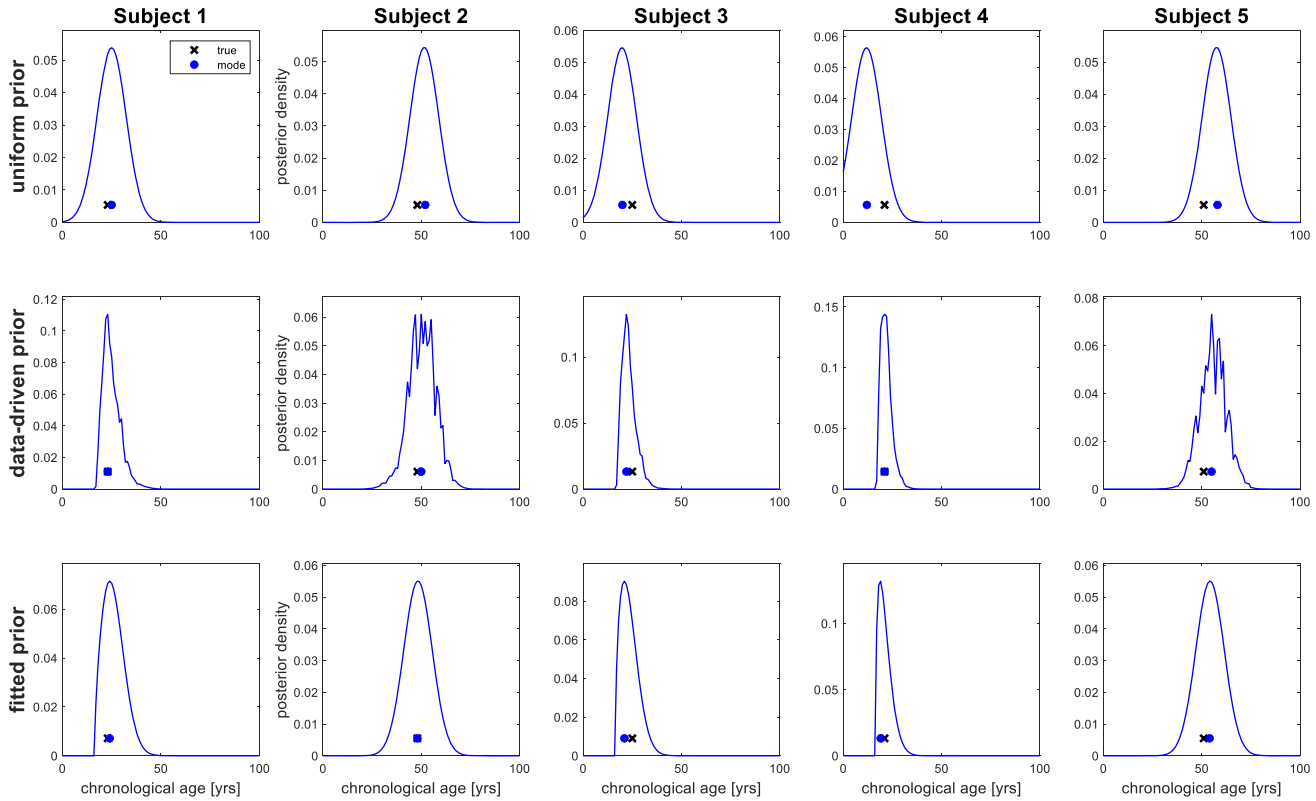

Fig. 12 *Analysis 4: posterior distributions.* Exemplary posterior distributions over chronological age with maximum-a-posteriori estimates (blue dots) and true values (black crosses) for the first five subjects in the validation set (see Fig. 11, left column) when using multivariate Bayesian regression with uniform prior (top row), data-driven prior (middle row) or fitted prior (bottom row) over chronological age (see Fig. 11, bottom row). For explanations, see text

Table 4 *Analyses: comparisons with alternative approaches.* Predictive performance for generative and discriminative techniques when analyzing data from Analysis 2 and Analysis 4. For details regarding MBC and SVC, see Fig. 8, and for details regarding MBR and SVR, see Fig. 11. In all cases, predictive performance is given as rank graduation accuracy (RGA) [48]

Analysis 2	RGA	Analysis 4	RGA
generative methods			
naive Bayes classification	0.8934	naive Bayes regression	0.8123
multivariate Bayesian classification	0.9396	multivariate Bayesian regression	0.9510
linear discriminant analysis	0.9389		
discriminative methods			
multi-class logistic regression	0.9229	multiple linear regression	0.9546
support vector classification	0.9732	support vector regression	0.9325
random forest classification	0.9845	random forest regression	0.9367
neural network classification	0.9908	neural network regression	0.9542

Finally, MBI exhibits favorable properties in terms of *sustainability*, understood here as robustness and reliability of inference. By basing predictions on full posterior probabilities rather than point estimates, MBI enables balanced decision-making that naturally accounts for uncertainty. This probabilistic treatment reduces the risk of overconfident or unstable predictions and supports more robust conclusions, particularly in settings with limited or noisy data. The robustness of this approach is further corroborated by comparisons based on the robustness-focused RGA measure [48] in which MBI shows consistent and stable performance compared to alternative methods (see Sects. 3.5 and 4.5). Together, these characteristics highlight MBI as a sustainable machine learning approach in the SAFE sense, combining robustness with principled uncertainty quantification.

5.1.2 Classification and regression

→ *Examples: Simulations 1–4, Analyses 1–4.*

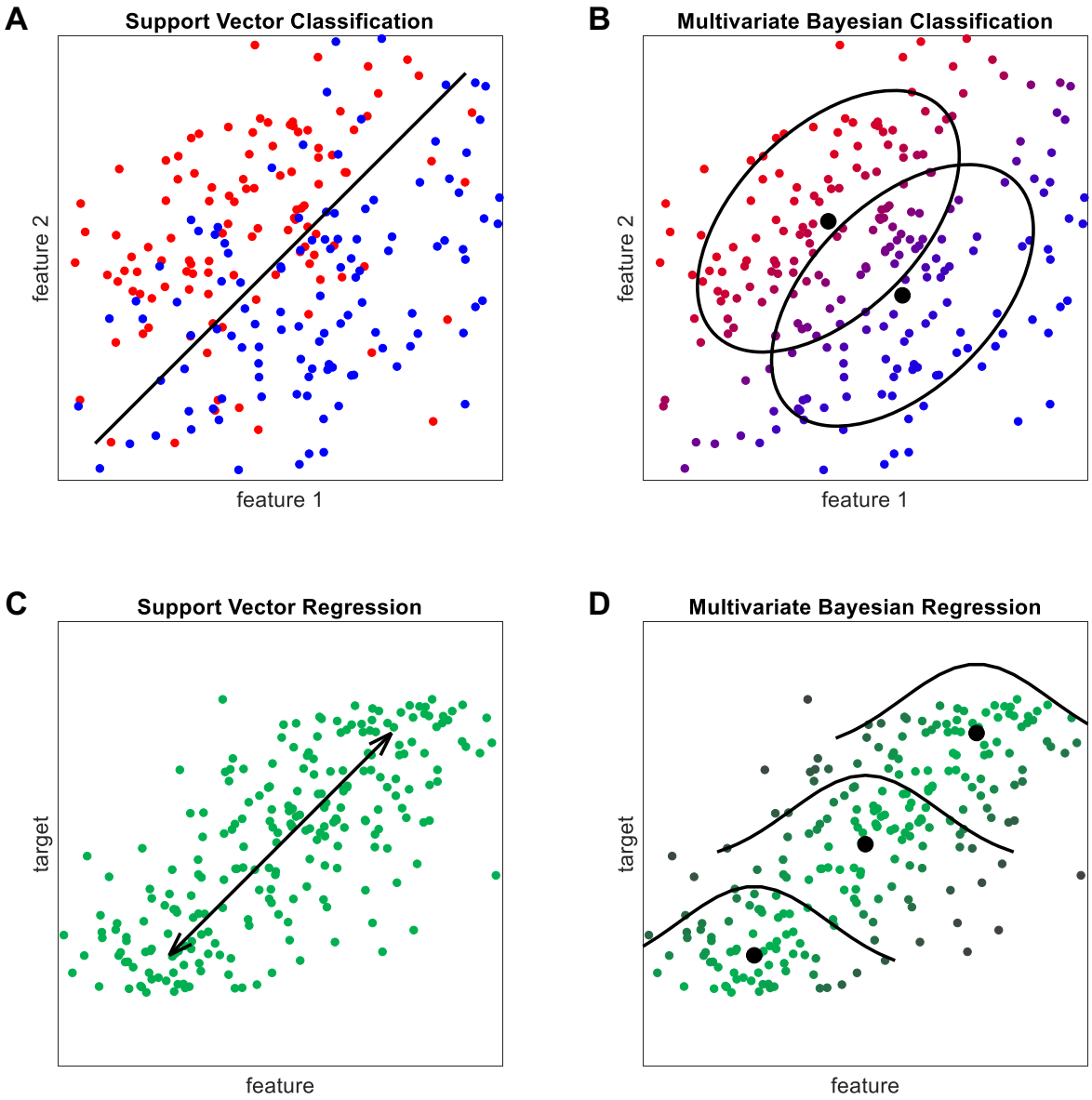

Fig. 13 *Different kinds of predictive analysis.* Conceptual comparison of traditional methods (left; here: SVMs) and probabilistic methods (right; here: MBI) for classification (top) and regression (bottom), reiterating predictive algorithms versus generative models (see Sect. 2.1.1). **A** Support vector classification searches for a hyperplane that optimally separates multivariate patterns from different classes. **B** Multivariate Bayesian classification estimates a probability distributions of the data, conditional on class, and uses them to calculate posterior probabilities for each class, conditional on the data. **C** Support vector regression searches for a projection of the data set that best matches the features onto the targets. **D** Multivariate Bayesian regression estimates a probability distribution of the features, conditional on the targets, and uses them to calculate posterior densities across target values, conditional on the observed features

Unlike other approaches that are solely designed for discrete prediction, e.g., logistic regression ([12, sec. 4.3]), or continuous prediction, e.g., multiple linear regression ([12, sec. 3.1]), the MBI framework can be used for classification into discrete categories as well as for regression of continuous variables (see Fig. 13); it has that in common with convolutional deep neural networks [7, 8], support vector machines [63, 64] or k-nearest neighbor algorithms [65, 66].

In multivariate Bayesian classification (MBC), the quantity of interest is treated as a discrete random variable, resulting in posterior class probabilities (cf. Figure 1B and see, e.g., Figure 3A). In multivariate Bayesian regression (MBR), the quantity of interest is treated as a continuous random variable, resulting in a posterior target density (cf. Figure 2B and see, e.g., Figure 6D).

Importantly, MBI works identically for both cases during the training phase, except for being based on different design matrices (see Fig. 1A vs. Figure 2A). This also means that, in principle, the same model can be used for classification into categories and regression of a continuous variable at the same time, reciprocally accounting for the other, provided that both quantities of interest are included in the design matrix used in the training phase.

5.1.3 Posterior probabilities

→ *Examples: Simulations 1/3/4, Analyses 2/4.*

Unlike many other approaches, e.g., linear discriminant analysis [41], support vector classification [5], or perceptrons [67], MBI provides posterior probabilities for the quantity to be predicted (see Fig. 13); like, e.g., logistic regression [68, 69] or naive Bayes classifiers [37, 70]. These are probabilities for observations belonging to particular classes in MBC and probability densities for targets having particular values in MBR.

Notably, support vector machines can be cast as probabilistic models [71]. Although probabilistic classification can be achieved using methods such as Platt scaling [72], such posterior inferences are not informed by prior distributions, but rather based on fitting a logistic regression model to the classifier's scores [73], making it a methodologically unrelated postprocessing step not part of the algorithm itself. Moreover, the method cannot be easily extended to continuous prediction tasks for which posterior target densities from MBR are a relatively simple solution.

Probabilistic results are particularly useful for some prediction contexts, e.g., disease classification [74, 75] or individual treatment projection [76, 77]. Whereas non-probabilistic approaches for classification always make a distinct decision, a user of MBC can also refrain from making a decision, if the posterior class probability does not exceed a pre-defined threshold. Moreover, these probabilities can be combined with loss or gain functions to identify optimal decisions by application of Bayesian decision theory. Similarly, while non-probabilistic approaches for regression only give point estimates, the posterior density from MBR quantifies remaining uncertainty after seeing the data and informs the user about the precision with which the target variable is known.

5.1.4 Prior knowledge

→ *Examples: Simulations 3/4, Analysis 4.*

The fact that MBI provides posterior probabilities comes hand in hand with the fact that it uses prior probabilities. In this way, it can integrate prior knowledge about the frequency of the distinguished classes or about the distribution of the predicted variable. For example, it might be known beforehand that one class occurs much more often than the other (e.g., healthy controls vs. patients) or that a variable tends to assume values in a specific range of its support (e.g., age or IQ).

This is especially useful in cases where it is known that the quantity of interest is constrained to a certain range of values and values outside this range are not meaningful, e.g., in brain age prediction [56]. While some approaches will inevitably also return negative values for chronological age (see [57, Fig. 2]; also see Fig. 11), specification of a prior distribution can easily prevent this (see Fig. 6E). In such cases, the prior distribution should be motivated from the experimental design by which the variable was controlled (see Fig. 6A) or from world knowledge about the process from which the variable was sampled in case of field studies (see Fig. 11).

Although this was not practiced in the present work, it is also conceivable that not only the prior model probabilities are manipulated, but the prior hyperparameters governing the MGLM parameters themselves are modified (see eqs. 9 and 17). This requires another level of prior knowledge, pertaining not just to the distribution of predicted quantities, but also to the specific mechanism connecting variables X and Y; however, it could also be helpful for regularization in the case of high-dimensional data (see Sect. 5.3.1).

5.1.5 Feature dependencies

→ *Examples: Simulations 1/2, Analyses 1–3, Simulation 4.*

Like many other approaches, MBI accounts for dependencies between measured variables (between-feature covariance; Σ in eq. 3). This is in contrast with, e.g., Gaussian naive Bayes [37] which assumes zero covariance between features (emulated by a diagonal matrix Σ). In principle, MBC implements a Bayes classifier for the multivariate GLM that is aware of non-diagonal feature covariance.

This exploits the way how several features jointly vary as a function of class or target and enables to delineate patterns belonging to classes or targets in multi-dimensional space (see Figs. 3 and 4). In fact, the between-feature covariance plays a key role, because MBI's (unnormalized) posterior distribution is given by the ratio of two determinants (see eqs. 29 and 36), where the two matrices are hyperparameters governing the feature covariance (Ω) and the regressor covariance (Λ).

Unlike many other approaches, MBI, however, also accounts for dependencies between data points (between-observation covariance; V in eq. 3). This is important when the features to predict from are, e.g., time series data, as accounting for the correlation between data points then drastically improves prediction accuracy (cf. [16]).

5.1.6 Covariate integration

→ *Examples: Simulation 2, Analyses 3/4.*

Many classification approaches only accept a feature matrix Y and a label vector x as inputs.[5] This makes it difficult to handle variables which are not predictive with respect to the labels (like the features), but which might influence the features (like covariates) and thus increase predictive performance, if accounted for. MBI naturally integrates such covariates by simply adding them as regressors (see Sect. 2.5): For the training set, they are additional columns in the design matrix, and in the test data, they are also included as regressors, but other than the quantities of interest, their values are fixed in the model inversion process which serves to predict the quantity of interest.

We have seen that including those covariates outperforms a "prior regression" approach [45, 46] in which variation related to them is removed from the features before predictive analysis (see Figs. 4 and 10). We have also seen that including the covariates outperforms employing them as features, i.e., adding them to the feature matrix (see Fig. 11) in the hope that this will explain the variable to be predicted. Note that treating a variable as a covariate rather than as a feature increases the number of parameters from r (one weight per each covariate-as-feature-variable) to $r \times v$ (one weight per each covariate influencing each feature). This suggests that the latter approach can more flexibly account for possible effects of the covariates.

5.1.7 Multi-class classification

→ *Examples: Simulation 3, Analyses 1–3.*

Using several examples, we have shown that MBC is capable of multi-class classification (see Figs. 5 and 7). Unlike other approaches, e.g., support vector classification [78], MBI uses the same algorithm for binary and n-ary classification, i.e., a categorical design matrix (see Fig. 1A and eq. 24). This means that MBC does not require separate 1-vs-1 comparisons for multi-class classification which avoids long computation times when the number of classes is very high (but see Sect. 5.2.5) and there is no need for a decision rule to merge several 1-vs-1 comparisons into a multi-class decision ([79]; [12, p. 339]).

[5] See, for example, the Python implementation of support vector machines (https://github.com/cjlin1/libsvm/blob/master/python/libsvm/svmutil.py) or the MATLAB implementation of logistic regression (https://www.mathworks.com/help/stats/fitmnr.html).

5.2 Disadvantages

5.2.1 Linear mappings

→ *Examples: Analyses 2/4.*

Since MBI is based on multivariate general linear models (see eq. 3), it intrinsically assumes linear relationships between labels and features, and linear effects of covariates on features. As such, it is principally restricted to such linear mappings and not able to represent nonlinear relationships between labels and features in multi-dimensional space. This is different from kernel-based methods [80], e.g., support vector machines, which also allow to exploit nonlinear patterns in the data by replacing a linear kernel with, e.g., radial basis function or polynomial kernels, leading to nonlinear decision boundaries in cases of classification [81]. It is also a disadvantage when compared with convolutional or deep neural networks [7] which allow to represent nonlinear relationships through inclusion of hidden layers or nonlinear activation functions [8].

For example, it is plausible to expect that the mapping from chronological age to structural MRI data is nonlinear in nature. Here, we did not explore this question, as we always applied linear SVMs as alternative methods for comparison with MBI. In an MBI-based approach, one way to faithfully capture those patterns would be to replace the MGLM-derived likelihood function by one allowing for nonlinear relationships between X and Y—which would, however, lead to non-tractable posterior distributions and marginal likelihoods (see Sect. 5.3.2). Another approach would be to apply nonlinear basis functions to either features or targets ([12, sec. 3.1]) or to equip MBI with a kernel-based architecture ([12, sec. 6.1]).

5.2.2 Normal errors

→ *Examples: Analyses 1/2/4.*

In addition to linearity of the mapping from classes or targets to features, MBI assumes random variation in the features to exclusively consist of matrix-normally distributed errors. This also entails that the between-observation covariance V is the same across all columns and that the between-feature covariance Σ is the same across all rows of the data matrix Y. While the latter belief is not strictly problematic, because it corresponds to the commonly made "independent and identically distributed" (i.i.d.) assumption, the normal distribution itself may not always be a valid assumption.

The theoretical justification for normality of the errors is (i) that residual components on top of signals in the features can typically be assumed to consist of a sum of a large number of unmodeled random variables and (ii) the central limit theorem asserts that such a sum will tend to a normal

distribution ([17, sect. 3.3]). Univariate analyses depending on this justification such as multiple linear regression have turned out very robust and reliable, such that the assumption is also legitimate for multivariate approaches ([15, p. 352]). One of our examples suggests that MBI can handle moderate levels of non-normality (cf. Figure 7A). However, there will certainly exist situations in which the assumption is not warranted and in which generalized linear models may be a better choice than the MGLM [82].

5.2.3 Equal covariance

→ *Examples: Simulation 3, Analyses 1–3.*

Through the between-feature covariance Σ, MBI can accommodate arbitrary covariance between measured signals, and because the matrix (or its inverse, $T = \Sigma^{-1}$) is fully estimated, MBI makes no assumptions about the specific structure of this covariance. However, since the MGLM describes a stationary process, multivariate Bayesian classification assumes an identical covariance for all classes.

This means that if, for example, signals belonging to two classes are not differing in terms of their multivariate mean, but only in terms of the amount of variance or the direction of covariance between the features ([83, Fig. 3]), or if classes are differing in terms of multivariate mean *and* feature covariance (cf. Analysis 2), the estimate of the between-feature covariance will be based on the pooled data from both classes (cf. eq. 22c). Thus, MBC will likely not be able to distinguish between the classes, because the measured signals from both classes equally well fit with the pooled estimate. In each analysis, the null hypothesis should therefore precisely specify whether classes are equal only in terms of their multivariate mean (in which case such a result is statistically correct) or in terms of the distribution as such.

5.2.4 Conjugate priors

→ *Examples: Simulation 4, Analyses 2-4.*

Multivariate Bayesian inversion uses conjugate normal-Wishart priors (cf. eq. 9) to estimate parameters of the MGLM underlying classification and regression. When a non-informative prior distribution is used (cf. eq. 17), no hyperparameters need to be specified. However, as this choice is mainly made for mathematical convenience and computational efficiency, it may also be seen as a weakness.

A fixed prior distribution reduces the model's flexibility to accommodate uncertainty about model parameters and its capability to adapt inferences to the specifics of real-world situations. With that in mind, the potential of continuous shrinkage priors [84] or spike-and-slab priors [85, 86] might be investigated. The same holds for alternative prior distributions when inferring on values of the label variable (cf. eq. 34), e.g., heavy-tailed t-distributions for regression problems [87, 88].

As these changes would make the model intractable in closed form, approximate inference techniques such as sampling inference and variational Bayes would have to be employed when using non-conjugate priors (see Sect. 5.3.2).

5.2.5 Unimodal posterior

→ *Examples: Simulation 4, Analysis 4.*

When performing multivariate Bayesian regression and using a uniform prior density (see Sect. 3.4 and 4.4), the posterior target density typically has a global maximum and no other local maxima (see Figs. 6 and 12). In other words, the posterior target density is inherently unimodal (and appears univariate normal).

This means that if, for example, the relationship between targets and features is in a way, such that two distinct values in separated regions of the target space are equally likely a posteriori, the posterior target density in MBR will most probably end up with its mode between these two values. In the worst case, this can lead to misleading conclusions, by reporting the medium value as the reconstruction, although reporting the two extreme values with equal posterior probability would have been more appropriate. Potential remedies for this are different prior densities or nonlinear basis functions.

5.2.6 Computational cost

→ *Examples: Simulations 3/4, Analyses 2/4.*

For each observation in the test set, for which a posterior class probability or the posterior target density is calculated, MBI requires the determinants of two matrices for each possible value of the predicted variable (see eqs. 29 and 36). These two matrices are the posterior precision governing the regressor covariance ($\Lambda_{2i} \in \mathbb{R}^{p \times p}$) and the posterior scale governing the feature covariance ($\Omega_{2i} \in \mathbb{R}^{v \times v}$).

With the number of regressor (p) or the number of features (v) becoming very high, these matrices become very large, such that the calculation of determinants can potentially become slow. This especially holds for discretized MBR when the prior density is intended to be very fine-grained or spanning a large range of values, such that the calculation in equation (36) has to be repeated very often. Systematic comparison of computation times confirms this, showing a 5-fold or 20-fold increase in computation time for MBR, compared with SVR (see Table 5). Computation times of MBC are comparable with those of SVC and sometimes lower, especially when working with a large number of classes, e.g., in Analyses 1 and 2 (see Table 5).

The issue of excessively long computation times for MBR can, if necessary, be mitigated by assuming that the pos-

terior target density is a univariate normal distribution (cf. Figures 2, 6 and 12)—in which case its parameters can be calculated from just three pairs of target value λ and unnormalized posterior density $\tilde{p}(\lambda)$.

5.3 Extensions

5.3.1 High-dimensional data

→ *Example: Analyses 2/4.*

As it can be seen from the equations for posterior class probabilities (see eq. 29) and the posterior target density (see eq. 36), computation of posterior probabilities always requires the calculation of two determinants, the posterior precision Λ_{2i} governing the regressor covariance and the posterior scale Ω_{2i} governing the feature covariance. Note that, when the number of features v is larger than the number of observations n, the matrix Ω_{2i} will be ill-conditioned and not invertible (cf. eq. 22), such that its determinant is zero and the posterior probabilities are not defined.

One reason for this is that the prior scale matrix is set to a $v \times v$ zero matrix $\Omega_0 = 0_{vv}$ (cf. eq. 17) which has the advantage that MBI does not require user-specified hyperparameters, but causes the posterior scale matrix Ω_{2i} to become singular. This can be avoided by setting the prior scale matrix to a scalar multiple of the identity matrix $\Omega_0 = \omega_0 I_v$. The posterior scale matrix will then become a mixture of this prior covariance term and data-based covariance terms (cf. eqs. 11 and 22):

$$\Omega_{2i} = \omega_0 I_v + Y^{\mathrm{T}} P Y + y_i^{\mathrm{T}} y_i - M_2^{\mathrm{T}} \Lambda_2 M_2 \,. \tag{59}$$

The magnitude of ω_0 will determine the extent to which the features are regarded as independent signals, while the remaining terms are incorporating possible dependencies between the features based on the measured signals. Conceptually, imposing this prior knowledge on T via Ω_0 is equivalent to regularizing our estimate of the between-feature covariance, similar to Tikhonov regularization in ridge regression [89, 90]. Provided that the dimensionality of Ω_{2i} $(v \times v)$ allows for calculation of matrix determinants in a software package implementing MBI, this can allow for prediction of classes or target variables from high-dimensional data ($v \gg n$), i.e., data sets with much more features than observations.

5.3.2 Nonlinear mappings

→ *Example: Analyses 2/4.*

As we have noted above (see Sects. 5.2.1 and 5.2.2), MBI is restricted to linear mappings from regressors to features and matrix-normally distributed errors. Sometimes, such linear mappings are not possible or not realistic to assume, but specifying a nonlinear model often renders the joint likelihood intractable and the posterior distribution cannot be expressed in closed form [12, 28, 91]. However, provided that such an alternative likelihood function $p(Y|X, \theta, m)$ and a prior distribution $p(\theta|m)$ are given mathematically, MBI can be finessed to handle nonlinear mappings using one out of two strategies.

The first strategy, *sampling-based inference*, calculates the posterior distribution (in the training data) numerically and approximates the marginal likelihood (in the test data) by drawing from the prior and inserting into the likelihood [92, 93]. The second strategy, *variational Bayesian inference*, assumes the posterior distribution (in the training data) to factorize over sets of parameters [32] and calculates the variational free energy [94] as a lower bound to the marginal likelihood (in the test data).

In conclusion, while MBI does not allow for nonlinear relationships between label variables and observed features, specifying an appropriate likelihood function and prior distribution allows to perform non-standard MBI with nonlinear mappings [39, 95] via sampling or variational approaches.

6 Conclusions

We have systematically reviewed fully Bayesian inference for the multivariate general linear model (MGLM) and have shown that Bayesian MGLM estimation can enable cross-validated prediction of discrete or continuous variables. The resulting techniques of multivariate Bayesian classification (MBC) and regression (MBR) have a number of desirable properties. Further research should investigate how the proposed framework can be extended to other statistical tasks such as clustering, expectation maximization, prediction from high-dimensional data, nonlinear prediction or sampling-based inference.

Appendix A Likelihood function of the MGLM

The model equation of the multivariate general linear model (MGLM) reads:

$$m: \; Y = XB + E, \;\; E \sim \mathcal{MN}(0_{nv}, V, \Sigma) \,. \tag{A.1}$$

The linear transformation theorem for the matrix-normal distribution states that, if $\Xi \in \mathbb{R}^{n \times p}$ is matrix-normally distributed, then any (bi-)linear transformation of Ξ in the form $\Upsilon = A\Xi B + C$ with constant $A \in \mathbb{R}^{r \times n}$, $B \in \mathbb{R}^{p \times s}$ and $C \in \mathbb{R}^{r \times s}$ is also matrix-normally distributed [96, ch. 2-3]:

$$\Xi \sim \mathcal{MN}(M, U, V) \quad \Rightarrow \quad \Upsilon = A\Xi B + C \sim$$

Table 5 *Computational cost.* Computation times for predictive analyses based on multivariate Bayesian inversion (MBI) and support vector machines (SVM) for main analyses reported in this paper. All computations were performed using MATLAB R2025b running on a 64-bit Windows 11 PC with 32 GB RAM and an Intel Core Ultra 5 Processor 225 H working at 1.70 GHz

Analysis	Figure	MBI	SVM
Simulation 1	Fig. 3A	0.07 s	–
	Fig. 3B	0.05 s	–
Simulation 2	Fig. 4C/E	0.02 s	0.01 s
	Fig. 4D/F	0.02 s	0.01 s
Simulation 3	Fig. 5B/C	1.42 s	0.05 s
	Fig. 5D/E/F	4.23 s	–
Simulation 4	Fig. 6B/C	0.29 s	0.06 s
Analysis 1	Fig. 7B	0.03 s	0.07 s
Analysis 2	Fig. 8A	01:42:33 h	32:36 min
	Fig. 8B/C	03:59 min	06:11 min
Analysis 3	Fig. 10B	0.05 s	0.07 s
Analysis 4	Fig. 11, 2nd/5th column	16.14 s	0.94 s
	Fig. 11, 3rd column	13.39 s	–
	Fig. 11, 4th column	15.33 s	–

$$\mathcal{MN}(AMB + C, AUA^{\mathrm{T}}, B^{\mathrm{T}}VB)\ . \tag{A.2}$$

Combining (A.1) and (A.2) leads to the following distribution for Y:

$$m:\ Y \sim \mathcal{MN}(XB, V, \Sigma)\ . \tag{A.3}$$

In the independent and identically distributed (i.i.d.) case, i.e., with $V = I_n$, we can set $A = e_i^{\mathrm{T}}$, $B = I_v$, $C = 0_{1v}$ for $i = 1, \ldots, n$ and apply (A.2) to (A.3) to obtain

$$m:\ y_i = x_i B + \varepsilon_i,\ \varepsilon_i^T \sim \mathcal{N}(0_v, \Sigma),\ i = 1, \ldots, n\ . \tag{A.4}$$

The probability density function of the matrix-normal distribution for a random matrix $\Xi \in \mathbb{R}^{n\times p}$ with mean $M \in \mathbb{R}^{n\times p}$ and covariances $U \in \mathbb{R}_+^{n\times n}$ and $V \in \mathbb{R}_+^{p\times p}$ is given by [97, sect. 2.1]

$$\begin{aligned}&\mathcal{MN}(X; M, U, V)\\ =&\sqrt{\frac{1}{(2\pi)^{np}|U|^n|V|^p}}\\ &\cdot\exp\left[-\frac{1}{2}\,\mathrm{tr}\left(V^{-1}(X-M)^{\mathrm{T}}U^{-1}(X-M)\right)\right]\ .\end{aligned} \tag{A.5}$$

Combining (A.3) and (A.5) leads to the following likelihood function for m:

$$\begin{aligned}&p(Y|X, B, V, \Sigma, m) = \mathcal{MN}(Y; XB, V, \Sigma)\\ =&\sqrt{\frac{1}{(2\pi)^{nv}|\Sigma|^n|V|^v}}\\ &\cdot\exp\left[-\frac{1}{2}\,\mathrm{tr}\left(\Sigma^{-1}(Y-XB)^{\mathrm{T}}V^{-1}(Y-XB)\right)\right]\ .\end{aligned} \tag{A.6}$$

Substituting $P = V^{-1}$ and $T = \Sigma^{-1}$, we finally have:

$$\begin{aligned}&p(Y|X, B, P, T, m)\\ =&\sqrt{\frac{|T|^n|P|^v}{(2\pi)^{nv}}}\\ &\cdot\exp\left[-\frac{1}{2}\mathrm{tr}\left(T(Y-XB)^{\mathrm{T}}P(Y-XB)\right)\right]\ .\end{aligned} \tag{A.7}$$

Appendix B Posterior distribution for the MGLM

From (9), the conjugate prior for the MGLM is given by a normal-Wishart prior distribution over the model parameters B and $T = \Sigma^{-1}$:

$$p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0)\ . \tag{B.1}$$

From (8), the posterior distribution is proportional to the product of likelihood function and prior density, i.e., the joint likelihood function:

$$p(B, T|Y) \propto p(Y|X, B, T)\, p(B, T) = p(Y, B, T)\ . \quad \text{(B.2)}$$

From (7), we have the likelihood function:

$$\begin{aligned} &p(Y|X, B, T) \\ =&\sqrt{\frac{|T|^n|P|^v}{(2\pi)^{nv}}} \\ &\cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T(Y - XB)^{\mathrm{T}}P(Y - XB)\right)\right]\ . \end{aligned} \quad \text{(B.3)}$$

The probability density function of the matrix-normal distribution for a random matrix $\Xi \in \mathbb{R}^{n\times p}$ with mean $M \in \mathbb{R}^{n\times p}$ and covariances $U \in \mathbb{R}^{n\times n}_+$ and $V \in \mathbb{R}^{p\times p}_+$ is:

$$\begin{aligned} &\mathcal{MN}(X; M, U, V) \\ =&\sqrt{\frac{1}{(2\pi)^{np}|U|^n|V|^p}} \\ &\cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(V^{-1}(X - M)^{\mathrm{T}}U^{-1}(X - M)\right)\right]\ . \end{aligned} \quad \text{(B.4)}$$

The probability density function of the Wishart distribution for a random matrix $\Xi \in \mathbb{R}^{p\times p}_+$ with scale matrix $V \in \mathbb{R}^{p\times p}_+$ and degrees of freedom $n \in \mathbb{R},\ n > p$ is:

$$\begin{aligned} \mathcal{W}(X; V, n) = &\frac{1}{\Gamma_p\left(\frac{n}{2}\right)}\sqrt{\frac{1}{2^{np}|V|^n}}|X|^{(n-p-1)/2} \\ &\exp\left[-\frac{1}{2}\mathrm{tr}\left(V^{-1}X\right)\right]\ . \end{aligned} \quad \text{(B.5)}$$

Combining (B.3) with (B.1) and plugging in (B.4) and (B.5), the joint likelihood of the model is given by:

$$\begin{aligned} &p(Y, B, T) \\ &= \sqrt{\frac{|T|^n|P|^v}{(2\pi)^{nv}}}\exp\left[-\frac{1}{2}\mathrm{tr}\left(T(Y-XB)^{\mathrm{T}}P(Y-XB)\right)\right] \\ &\quad\cdot\sqrt{\frac{|T|^p|\Lambda_0|^v}{(2\pi)^{pv}}}\exp\left[-\frac{1}{2}\mathrm{tr}\left(T(B-M_0)^{\mathrm{T}}\Lambda_0(B-M_0)\right)\right] \\ &\quad\cdot\frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)}\sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}}|T|^{(\nu_0-v-1)/2}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega_0 T\right)\right]\ . \end{aligned} \quad \text{(B.6)}$$

Collecting identical variables gives:

$$\begin{aligned} p(Y, B, T) = &\sqrt{\frac{|T|^n|P|^v}{(2\pi)^{nv}}}\sqrt{\frac{|T|^p|\Lambda_0|^v}{(2\pi)^{pv}}}\sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}}\frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \\ &\cdot|T|^{(\nu_0-v-1)/2}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega_0 T\right)\right]\cdot \\ &\exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[(Y - XB)^{\mathrm{T}}P(Y - XB)\right.\right.\right. \\ &\left.\left.\left.+(B - M_0)^{\mathrm{T}}\Lambda_0(B - M_0)\right]\right)\right]\ . \end{aligned} \quad \text{(B.7)}$$

Expanding the products in the exponent gives:

$$\begin{aligned} p(Y, B, T) = &\sqrt{\frac{|T|^n|P|^v}{(2\pi)^{nv}}}\sqrt{\frac{|T|^p|\Lambda_0|^v}{(2\pi)^{pv}}}\sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}}\frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \\ &\cdot|T|^{(\nu_0-v-1)/2}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega_0 T\right)\right] \\ &\cdot\exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[Y^{\mathrm{T}}PY - Y^{\mathrm{T}}PXB\right.\right.\right. \\ &- B^{\mathrm{T}}X^{\mathrm{T}}PY + B^{\mathrm{T}}X^{\mathrm{T}}PXB \\ &+\ B^{\mathrm{T}}\Lambda_0 B - B^{\mathrm{T}}\Lambda_0 M_0 - M_0^{\mathrm{T}}\Lambda_0 B \\ &\left.\left.\left.+M_0^{\mathrm{T}}\Lambda_0\mu_0\right]\right)\right]\ . \end{aligned} \quad \text{(B.8)}$$

Completing the square over B, we finally have

$$\begin{aligned} p(Y, B, T) = &\sqrt{\frac{|T|^n|P|^v}{(2\pi)^{nv}}}\sqrt{\frac{|T|^p|\Lambda_0|^v}{(2\pi)^{pv}}}\sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}}\frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \\ &\cdot|T|^{(\nu_0-v-1)/2}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega_0 T\right)\right]\cdot \\ &\exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[(B - M_n)^{\mathrm{T}}\Lambda_n(B - M_n)\right.\right.\right. \\ &\left.\left.\left.+(Y^{\mathrm{T}}PY + M_0^{\mathrm{T}}\Lambda_0 M_0 - M_n^{\mathrm{T}}\Lambda_n M_n)\right]\right)\right]\ . \end{aligned} \quad \text{(B.9)}$$

with the posterior hyperparameters

$$\begin{aligned} M_n &= \Lambda_n^{-1}(X^{\mathrm{T}}PY + \Lambda_0 M_0) \\ \Lambda_n &= X^{\mathrm{T}}PX + \Lambda_0\ . \end{aligned} \quad \text{(B.10)}$$

Ergo, the joint likelihood is proportional to

$$p(Y, B, T) \propto |T|^{p/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[(B-M_n)^{\mathrm{T}} \Lambda_n (B-M_n)\right]\right)\right] \cdot |T|^{(\nu_n-v-1)/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega_n T\right)\right] \tag{B.11}$$

with the posterior hyperparameters

$$\begin{aligned} \Omega_n &= \Omega_0 + Y^{\mathrm{T}}PY + M_0^{\mathrm{T}}\Lambda_0 M_0 - M_n^{\mathrm{T}}\Lambda_n M_n \\ \nu_n &= \nu_0 + n\ . \end{aligned} \tag{B.12}$$

Taken together, we see that the joint likelihood (B.11) is proportional to a normal-Wishart distribution, such that the posterior distribution is also a normal-Wishart distribution[6]

$$p(B, T|Y) = \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_n^{-1}, \nu_n) \tag{B.13}$$

and the posterior hyperparameters are given by

$$\begin{aligned} M_n &= \Lambda_n^{-1}(X^{\mathrm{T}}PY + \Lambda_0 M_0) \\ \Lambda_n &= X^{\mathrm{T}}PX + \Lambda_0 \\ \Omega_n &= \Omega_0 + Y^{\mathrm{T}}PY + M_0^{\mathrm{T}}\Lambda_0 M_0 - M_n^{\mathrm{T}}\Lambda_n M_n \\ \nu_n &= \nu_0 + n\ . \end{aligned} \tag{B.14}$$

Appendix C Marginal likelihood for the MGLM

Again, we assume a normal-Wishart distribution as conjugate prior distribution over the model parameters B and $T = \Sigma^{-1}$:

$$p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0)\ . \tag{C.1}$$

From (12), the marginal likelihood is obtained by integrating over the product of likelihood function and prior density, i.e., over the joint likelihood function:

$$\begin{aligned} p(Y|m) &= \int_{\mathbb{R}^{p\times v}\times\mathbb{R}_+^{v\times v}} p(Y|X, B, T)\, p(B, T)\, \mathrm{d}(B, T) \\ &= \int_{\mathbb{R}_+^{v\times v}} \int_{\mathbb{R}^{p\times v}} p(Y, B, T)\, \mathrm{d}B\, \mathrm{d}T\ . \end{aligned} \tag{C.2}$$

[6] The derivation of this posterior distribution is included in [98, ch. 3.4] and can also be found at: https://statproofbook.github.io/P/mblr-post.

In Appendix B, the joint likelihood $p(Y, B, T)$ has been obtained as (see eq. B.9)

$$\begin{aligned} p(Y, B, T) = &\sqrt{\frac{|T|^n|P|^v}{(2\pi)^{nv}}}\sqrt{\frac{|T|^p|\Lambda_0|^v}{(2\pi)^{pv}}}\sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}}\frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \\ &\cdot |T|^{(\nu_0-v-1)/2}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega_0 T\right)\right] \\ &\cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[(B-M_n)^{\mathrm{T}}\Lambda_n(B-M_n)\right.\right.\right. \\ &\left.\left.\left.+(Y^{\mathrm{T}}PY + M_0^{\mathrm{T}}\Lambda_0 M_0 - M_n^{\mathrm{T}}\Lambda_n M_n)\right]\right)\right]\ . \end{aligned} \tag{C.3}$$

In the following, we will make use of the fact that probability density integrates to one over the whole domain of a random variable, e.g.:

$$\int_{\mathbb{R}^{n\times p}} \mathcal{MN}(X; M, U, V)\, \mathrm{d}X = 1 \tag{C.4}$$

$$\int_{\mathbb{R}_+^{p\times p}} \mathcal{W}(X; V, n)\, \mathrm{d}X = 1 \tag{C.5}$$

Using the matrix-normal probability density function (B.4), we can rewrite this as

$$\begin{aligned} p(Y, B, T) = &\sqrt{\frac{|T|^n|P|^v}{(2\pi)^{nv}}}\sqrt{\frac{|T|^p|\Lambda_0|^v}{(2\pi)^{pv}}}\sqrt{\frac{(2\pi)^{pv}}{|T|^p|\Lambda_n|^v}} \\ &\sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}}\frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \\ &\cdot |T|^{(\nu_0-v-1)/2}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega_0 T\right)\right] \\ &\cdot \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \\ &\cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[Y^{\mathrm{T}}PY + M_0^{\mathrm{T}}\Lambda_0 M_0\right.\right.\right. \\ &\left.\left.\left.-M_n^{\mathrm{T}}\Lambda_n M_n\right]\right)\right]\ . \end{aligned} \tag{C.6}$$

Now, B can be integrated out easily by applying (C.4):

$$\begin{aligned} &\int_{\mathbb{R}^{p\times v}} p(Y, B, T)\, \mathrm{d}B \\ &= \sqrt{\frac{|T|^n|P|^v}{(2\pi)^{nv}}}\sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}}\sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}}\frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \\ &\cdot |T|^{(\nu_0-v-1)/2} \\ &\cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[\Omega_0 + Y^{\mathrm{T}}PY + M_0^{\mathrm{T}}\Lambda_0 M_0\right.\right.\right. \\ &\left.\left.\left.-M_n^{\mathrm{T}}\Lambda_n M_n\right]\right)\right]\ . \end{aligned} \tag{C.7}$$

Using the Wishart probability density function (B.5), we can rewrite this as

$$\begin{aligned}\int_{\mathbb{R}^{p\times v}} p(Y,B,T)\,\mathrm{d}B &= \sqrt{\frac{|P|^v}{(2\pi)^{nv}}}\sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}}\sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}}\sqrt{\frac{2^{\nu_n v}}{|\Omega_n|^{\nu_n}}}\frac{\Gamma_v\left(\frac{\nu_n}{2}\right)}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \\ &\quad\cdot \mathcal{W}(T;\Omega_n^{-1},\nu_n)\,. \end{aligned} \tag{C.8}$$

Finally, T can also be integrated out by applying (C.5):

$$\begin{aligned}\int_{\mathbb{R}_+^{v\times v}}\int_{\mathbb{R}^{p\times v}} p(Y,B,T)\,\mathrm{d}B\,\mathrm{d}T &= \sqrt{\frac{|P|^v}{(2\pi)^{nv}}}\sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}}\sqrt{\frac{\left|\frac{1}{2}\Omega_0\right|^{\nu_0}}{\left|\frac{1}{2}\Omega_n\right|^{\nu_n}}}\frac{\Gamma_v\left(\frac{\nu_n}{2}\right)}{\Gamma_v\left(\frac{\nu_0}{2}\right)}\,. \end{aligned} \tag{C.9}$$

Combining (C.9) and (C.2), the marginal likelihood of the MGLM is[7]

$$p(Y|m) = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}}\sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}}\sqrt{\frac{\left|\frac{1}{2}\Omega_0\right|^{\nu_0}}{\left|\frac{1}{2}\Omega_n\right|^{\nu_n}}}\frac{\Gamma_v\left(\frac{\nu_n}{2}\right)}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \tag{C.10}$$

where the posterior hyperparameters M_n, Λ_n, Ω_n and ν_n are given by (B.14).

Appendix D Multivariate Bayesian inversion

In this section, we use the superscript $\square^{(1)}$ to denote quantities belonging to the training data and the superscript $\square^{(2)}$ to denote quantities belonging to the test data.

In the first step of an MBI analysis, the data set of features and labels is

$$Y^{(1)} = Y \in \mathbb{R}^{n\times v},\ X^{(1)} = X \in \mathbb{R}^{n\times p},\ P^{(1)} = P \in \mathbb{R}^{n\times n},\ n^{(1)} = n \tag{D.1}$$

and from (17), the prior hyperparameters for the first step are given by:

$$M_0^{(1)} = 0_{pv},\ \Lambda_0^{(1)} = 0_{pp},\ \Omega_0^{(1)} = 0_{vv},\ \nu_0^{(1)} = 0\,. \tag{D.2}$$

[7] The derivation of this marginal likelihood is included in [98, ch. 3.4] and can also be found at: https://statproofbook.github.io/P/mblr-lme.

Plugging this into (B.14), the posterior hyperparameters after the first step are:

$$\begin{aligned} M_1 = M_n^{(1)} &= \Lambda_n^{(1)-1}(X^{(1)\mathrm{T}}P^{(1)}Y^{(1)} + \Lambda_0^{(1)}M_0^{(1)}) \\ &= \Lambda_1^{-1}X^{\mathrm{T}}PY \\ \Lambda_1 = \Lambda_n^{(1)} &= X^{(1)\mathrm{T}}P^{(1)}X^{(1)} + \Lambda_0^{(1)} \\ &= X^{\mathrm{T}}PX \\ \Omega_1 = \Omega_n^{(1)} &= \Omega_0^{(1)} + Y^{(1)\mathrm{T}}P^{(1)}Y^{(1)} + M_0^{(1)\mathrm{T}}\Lambda_0^{(1)}M_0^{(1)} \\ &\quad - M_n^{(1)\mathrm{T}}\Lambda_n^{(1)}M_n^{(1)} \\ &= Y^{\mathrm{T}}PY - M_1^{\mathrm{T}}\Lambda_1 M_1 \\ \nu_1 = \nu_n^{(1)} &= \nu_0^{(1)} + n^{(1)} = n\,. \end{aligned} \tag{D.3}$$

In the second step of an MBI analysis, features and labels of the single data point are

$$Y^{(2)} = y_i \in \mathbb{R}^{1\times v},\ X^{(2)} = x_i \in \mathbb{R}^{1\times p},\ P^{(2)} = p_i = 1,\ n^{(2)} = 1 \tag{D.4}$$

and from (D.3), the prior hyperparameters for the second step are the posterior hyperparameters obtained in the first step:

$$M_0^{(2)} = M_1,\ \Lambda_0^{(2)} = \Lambda_1,\ \Omega_0^{(2)} = \Omega_1,\ \nu_0^{(2)} = \nu_1\,. \tag{D.5}$$

Plugging this into (B.14), the posterior hyperparameters after the second step are:

$$\begin{aligned} M_2 = M_n^{(2)} &= \Lambda_n^{(2)-1}(X^{(2)\mathrm{T}}P^{(2)}Y^{(2)} + \Lambda_0^{(2)}M_0^{(2)}) \\ &= \Lambda_2^{-1}(x_i^{\mathrm{T}}p_i y_i + \Lambda_1\Lambda_1^{-1}X^{\mathrm{T}}PY) \\ &= \Lambda_2^{-1}(X^{\mathrm{T}}PY + x_i^{\mathrm{T}}y_i) \\ \Lambda_2 = \Lambda_n^{(2)} &= X^{(2)\mathrm{T}}P^{(2)}X^{(2)} + \Lambda_0^{(2)} \\ &= x_i^{\mathrm{T}}p_i x_i + X^{\mathrm{T}}PX \\ &= X^{\mathrm{T}}PX + x_i^{\mathrm{T}}x_i \\ \Omega_2 = \Omega_n^{(2)} &= \Omega_0^{(2)} + Y^{(2)\mathrm{T}}P^{(2)}Y^{(2)} + M_0^{(2)\mathrm{T}}\Lambda_0^{(2)}M_0^{(2)} \\ &\quad - M_n^{(2)\mathrm{T}}\Lambda_n^{(2)}M_n^{(2)} \\ &= Y^{\mathrm{T}}PY - M_1^{\mathrm{T}}\Lambda_1 M_1 + y_i^{\mathrm{T}}p_i y_i \\ &\quad + M_1^{\mathrm{T}}\Lambda_1 M_1 - M_2^{\mathrm{T}}\Lambda_2 M_2 \\ &= Y^{\mathrm{T}}PY + y_i^{\mathrm{T}}y_i - M_2^{\mathrm{T}}\Lambda_2 M_2 \\ \nu_2 = \nu_n^{(2)} &= \nu_0^{(2)} + n^{(2)} = \nu_1 + 1 = n + 1\,. \end{aligned} \tag{D.6}$$

Acknowledgements The authors wish to thank Dirk Ostwald (OvGU Magdeburg) for providing helpful comments on an earlier version of this manuscript. Additionally, the authors thank the acquisitors and curators of the Egyptian skull data set [49], the MNIST data set [50], the birth weight data set [52], and the brain age competition data set [61] for providing open access to these valuable resources.

Author Contributions J.S. developed the methods, wrote the code, analyzed the data, and prepared the manuscript. C.A. helped in developing the methods and revised the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability MATLAB and Python implementations of MBI for classification and regression are available from GitHub (https://github.com/JoramSoch/MBI). These tools allow for both, customized train-and-test analyses as well as fully automatic cross-validated prediction. MATLAB and Python code underlying data analysis is provided in the same GitHub repository. Empirical data can be found within this repository (e.g., Analysis 1: https://github.com/JoramSoch/MBI/tree/main/Examples/Analysis_1). Simulated data can be generated using the routines in this repository (e.g., Simulation 1: https://github.com/JoramSoch/MBI/tree/main/Examples/Simulation_1).

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Efron, B., Hastie, T.: Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Institute of Mathematical Statistics monographs. Cambridge University Press, New York, NY (2016)
2. Wei, T.C., Sheikh, U.U., Rahman, A.A.-H.A.: Improved optical character recognition with deep neural network. In: 2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA), pp. 245–249. IEEE, Batu Feringghi (2018). https://doi.org/10.1109/CSPA.2018.8368720. https://ieeexplore.ieee.org/document/8368720/Accessed2022-04-12
3. Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., Zweig, G.: Achieving Human Parity in Conversational Speech Recognition (2016) https://doi.org/10.48550/ARXIV.1610.05256. Accessed 2022-04-12
4. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Driessche, G., Graepel, T., Hassabis, D.: Mastering the game of Go without human knowledge. Nature **550**(7676), 354–359 (2017). https://doi.org/10.1038/nature24270. (**Accessed 2022-04-12**)
5. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995). https://doi.org/10.1007/BF00994018. (**Accessed 2022-04-12**)
6. Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V.: Support Vector Regression Machines. In: Mozer, M.C., Jordan, M., Petsche, T. (eds.) Advances in Neural Information Processing Systems, vol. 9. MIT Press, (1996). https://proceedings.neurips.cc/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf
7. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. **29**(6), 82–97 (2012). https://doi.org/10.1109/MSP.2012.2205597. (**Accessed 2022-04-12**)
8. Deng, L., Hinton, G., Kingsbury, B.: New types of deep neural network learning for speech recognition and related applications: an overview. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8599–8603. IEEE, Vancouver, BC, Canada (2013). https://doi.org/10.1109/ICASSP.2013.6639344. http://ieeexplore.ieee.org/document/6639344/ Accessed 2022-04-12
9. Ng, A., Jordan, M.: On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In: Dietterich, T., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems, vol. 14. MIT Press, (2001). https://proceedings.neurips.cc/paper_files/paper/2001/file/7b7a53e239400a13bd6be6c91c4f6c4e-Paper.pdf
10. Frässle, S., Yao, Y., Schöbi, D., Aponte, E.A., Heinzle, J., Stephan, K.E.: Generative models for clinical applications in computational psychiatry. WIREs Cognitive Science **9**(3) (2018) https://doi.org/10.1002/wcs.1460. Accessed 2022-04-12
11. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis, 3rd edn. Chapman and Hall/CRC, Boca Raton (2013)
12. Bishop, C.M.: Pattern Recognition and Machine Learning, 1st, (ed.): corr. 2nd printing, 2011th edn. Springer, New York (2006). ((2007))
13. Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J.: Bayesian decoding of brain images. Neuroimage **39**(1), 181–205 (2008). https://doi.org/10.1016/j.neuroimage.2007.08.013. (**Accessed 2015-10-30**)
14. Shmueli, G.: To Explain or to Predict? Statistical Science **25**(3), (2010) https://doi.org/10.1214/10-STS330. Accessed 2022-05-30
15. Allefeld, C., Haynes, J.-D.: Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. Neuroimage **89**, 345–357 (2014). https://doi.org/10.1016/j.neuroimage.2013.11.043. (**Accessed 2015-10-30**)
16. Soch, J., Allefeld, C., Haynes, J.-D.: Inverse transformed encoding models—a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding. Neuroimage **209**, 116449 (2020). https://doi.org/10.1016/j.neuroimage.2019.116449. (**Accessed 2021-01-07**)
17. Timm, N.H.: Applied Multivariate Analysis. Springer Texts in Statistics. Springer, New York Berlin Heidelberg (2002)
18. Koch, K.-R.: Introduction to Bayesian Statistics, 2nd, Updated and Enlarged ed. 2007 edition edn. Springer, Berlin; New York (2007)
19. Soch, J., Allefeld, C.: MACS—a new SPM toolbox for model assessment, comparison and selection. J. Neurosci. Methods **306**, 19–31 (2018). https://doi.org/10.1016/j.jneumeth.2018.05.017. (**Accessed 2019-01-23**)
20. Lee, D., Yun, S., Jang, C., Park, H.-J.: Multivariate Bayesian decoding of single-trial event-related fMRI responses for memory retrieval of voluntary actions. PLoS ONE **12**(8), 0182657

(2017). https://doi.org/10.1371/journal.pone.0182657. (**Accessed 2022-09-13**)
21. Sarker, I.H.: Machine learning: algorithms, real-world applications and research directions. SN Comput. Sci. **2**(3), 160 (2021). https://doi.org/10.1007/s42979-021-00592-x. (**Accessed 2025-02-05**)
22. Huang, H.-H., Xu, T., Yang, J.: Comparing logistic regression, support vector machines, and permanental classification methods in predicting hypertension. BMC Proc. **8**(S1), 96 (2014). https://doi.org/10.1186/1753-6561-8-S1-S96. (**Accessed 2025-02-05**)
23. Vapnik, V.N.: Statistical Learning Theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control, Wiley (1998)
24. Valente, G., Castellanos, A.L., Hausfeld, L., De Martino, F., Formisano, E.: Cross-validation and permutations in MVPA: validity of permutation strategies and power of cross-validation schemes. Neuroimage **238**, 118145 (2021). https://doi.org/10.1016/j.neuroimage.2021.118145
25. Böhning, D.: Multinomial logistic regression algorithm. Ann. Inst. Stat. Math. **44**(1), 197–200 (1992). https://doi.org/10.1007/BF00048682. (**Accessed 2022-04-12**)
26. Claeskens, G., Hjort, N.L.: Model Selection and Model Averaging. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press (2008)
27. MacKay, D.J.C.: Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge, UK; New York (2003)
28. Penny, W., Kiebel, S., Friston, K.: Variational Bayesian inference for fMRI time series. Neuroimage **19**(3), 727–741 (2003). https://doi.org/10.1016/S1053-8119(03)00071-5. (**Accessed 2015-10-30**)
29. Penny, W.D., Trujillo-Barreto, N.J., Friston, K.J.: Bayesian fMRI time series analysis with spatial priors. Neuroimage **24**(2), 350–362 (2005). https://doi.org/10.1016/j.neuroimage.2004.08.034. (**Accessed 2015-10-30**)
30. Soch, J., Haynes, J.-D., Allefeld, C.: How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection. Neuroimage **141**, 469–489 (2016). https://doi.org/10.1016/j.neuroimage.2016.07.047. (**Accessed 2019-02-27**)
31. Murphy, K.P.: Conjugate Bayesian Analysis of the Gaussian Distribution. Technical report, University of British Columbia (2007). https://www.cs.ubc.ca/~lowmurphyk/Papers/bayesGauss.pdf
32. Penny, W., Flandin, G., Trujillo-Barreto, N.: Bayesian comparison of spatially regularised general linear models. Hum. Brain Mapp. **28**(4), 275–293 (2007). https://doi.org/10.1002/hbm.20327. (**Accessed 2015-10-30**)
33. Ostwald, D., Starke, L.: Probabilistic delay differential equation modeling of event-related potentials. Neuroimage **136**, 227–257 (2016). https://doi.org/10.1016/j.neuroimage.2016.04.025
34. Soch, J., Meyer, A.P., Haynes, J.-D., Allefeld, C.: How to improve parameter estimates in GLM-based fMRI data analysis: cross-validated Bayesian model averaging. Neuroimage **158**, 186–195 (2017). https://doi.org/10.1016/j.neuroimage.2017.06.056. (**Accessed 2019-08-21**)
35. Molinaro, A.M., Simon, R., Pfeiffer, R.M.: Prediction error estimation: a comparison of resampling methods. Bioinformatics **21**(15), 3301–3307 (2005). https://doi.org/10.1093/bioinformatics/bti499. (**Accessed 2022-04-12**)
36. Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., Salakoski, T.: An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. Comput. Stat. Data Anal. **55**(4), 1828–1844 (2011). https://doi.org/10.1016/j.csda.2010.11.018. (**Accessed 2022-04-12**)
37. Hand, D.J., Yu, K.: Idiot's Bayes: Not So Stupid after All? Int. Stat. Rev. **69**(3), 385 (2001). https://doi.org/10.2307/1403452. (**Accessed 2022-04-12**)
38. Attias, H.: Inferring parameters and structure of latent variable models by variational bayes. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. UAI'99, pp. 21–30. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA (1999)
39. Roweis, S., Ghahramani, Z.: A Unifying Review of Linear Gaussian Models. Neural Comput. **11**(2), 305–345 (1999). https://doi.org/10.1162/089976699300016674. (**Accessed 2022-09-08**)
40. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**(518), 859–877 (2017). https://doi.org/10.1080/01621459.2017.1285773. (**Accessed 2026-02-20**)
41. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**, 179–188 (1936)
42. Koutsouleris, N., Kahn, R.S., Chekroud, A.M., Leucht, S., Falkai, P., Wobrock, T., Derks, E.M., Fleischhacker, W.W., Hasan, A.: Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. The Lancet Psychiatry **3**(10), 935–946 (2016). https://doi.org/10.1016/S2215-0366(16)30171-7. (**Accessed 2025-02-05**)
43. Sanfelici, R., Dwyer, D.B., Antonucci, L.A., Koutsouleris, N.: Individualized diagnostic and prognostic models for patients with psychosis risk syndromes: a meta-analytic view on the state of the art. Biol. Psychiat. **88**(4), 349–360 (2020). https://doi.org/10.1016/j.biopsych.2020.02.009. (**Accessed 2025-02-05**)
44. Mahalanobis, P.C.: On the generalized distance in statistics. Sankhyā: The Indian Journal of Statistics, Series A (2008-) **80**, 1–7 (2018). Accessed 2023-07-05
45. Li, L., Rakitsch, B., Borgwardt, K.: ccSVM: correcting Support Vector Machines for confounding factors in biological data classification. Bioinformatics **27**(13), 342–348 (2011). https://doi.org/10.1093/bioinformatics/btr204. (**Accessed 2025-02-05**)
46. Hamdan, S., Love, B.C., Polier, G.G., Weis, S., Schwender, H., Eickhoff, S.B., Patil, K.R.: Confound-leakage: confound removal in machine learning leads to leakage. GigaScience **12**, 071 (2022). https://doi.org/10.1093/gigascience/giad071. (**Accessed 2025-02-05**)
47. Raffinetti, E.: A rank graduation accuracy measure to mitigate artificial intelligence risks. Quality & Quantity **57**(S2), 131–150 (2023). https://doi.org/10.1007/s11135-023-01613-y. (**Accessed 2026-03-18**)
48. Giudici, P., Raffinetti, E.: RGA: a unified measure of predictive accuracy. Adv. Data Anal. Classif. **19**(1), 67–93 (2025). https://doi.org/10.1007/s11634-023-00574-2. (**Accessed 2026-03-18**)
49. Thomson, A., Randall-Maciver, R.: Ancient Races of the Thebaid. Oxford University Press, Oxford, United Kingdom (1905). http://www.dm.unibo.it/~lowsimoncin/EgyptianSkulls.html
50. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791. (**Accessed 2025-01-21**)
51. Deng, L.: The MNIST database of handwritten digit images for machine learning research [Best of the Web]. IEEE Signal Process. Mag. **29**(6), 141–142 (2012). https://doi.org/10.1109/MSP.2012.2211477. (**Accessed 2025-01-21**)
52. Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression. Wiley Series in Probability and Mathematical Statistics. Wiley, New York (1989). https://www.worldcat.org/title/applied-logistic-regression/oclc/19514573
53. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The Balanced Accuracy and Its Posterior Distribution. In: 2010 20th International Conference on Pattern Recognition, pp. 3121–3124. IEEE, Istanbul, Turkey (2010). https://doi.org/10.1109/ICPR.2010.764. http://ieeexplore.ieee.org/document/5597285/ Accessed 2020-09-08

54. Carrillo, H., Brodersen, K.H., Castellanos, J.A.: Probabilistic Performance Evaluation for Multiclass Classification Using the Posterior Balanced Accuracy. In: Armada, M.A., Sanfeliu, A., Ferre, M. (eds.) ROBOT2013: First Iberian Robotics Conference vol. 252, pp. 347–361. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-03413-3_25. http://link.springer.com/10.1007/978-3-319-03413-3_25 Accessed 2022-04-12
55. Dosenbach, N.U.F., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., Barnes, K.A., Dubis, J.W., Feczko, E., Coalson, R.S., Pruett, J.R., Barch, D.M., Petersen, S.E., Schlaggar, B.L.: Prediction of individual brain maturity using fMRI. Science **329**(5997), 1358–1361 (2010). https://doi.org/10.1126/science.1194144. (**Accessed 2020-10-28**)
56. Cole, J.H., Poudel, R.P.K., Tsagkrasoulis, D., Caan, M.W.A., Steves, C., Spector, T.D., Montana, G.: Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. Neuroimage **163**, 115–124 (2017). https://doi.org/10.1016/j.neuroimage.2017.07.059. (**Accessed 2020-09-08**)
57. Soch, J.: Distributional transformation improves decoding accuracy when predicting chronological age from structural MRI. Front. Psych. **11**, 604268 (2020). https://doi.org/10.3389/fpsyt.2020.604268. (**Accessed 2021-06-02**)
58. Fisch, L., Leenings, R., Winter, N.R., Dannlowski, U., Gaser, C., Cole, J.H., Hahn, T.: Editorial: predicting chronological age from structural neuroimaging: the predictive analytics competition 2019. Front. Psych. **12**, 710932 (2021). https://doi.org/10.3389/fpsyt.2021.710932. (**Accessed 2022-04-12**)
59. Waschkies, K.F., Soch, J., Darna, M., Richter, A., Altenstein, S., Beyle, A., Brosseron, F., Buchholz, F., Butryn, M., Dobisch, L., Ewers, M., Fliessbach, K., Gabelin, T., Glanz, W., Goerss, D., Gref, D., Janowitz, D., Kilimann, I., Lohse, A., Munk, M.H., Rauchmann, B., Rostamzadeh, A., Roy, N., Spruth, E.J., Dechent, P., Heneka, M.T., Hetzer, S., Ramirez, A., Scheffler, K., Buerger, K., Laske, C., Perneczky, R., Peters, O., Priller, J., Schneider, A., Spottke, A., Teipel, S., Düzel, E., Jessen, F., Wiltfang, J., Schott, B.H., Kizilirmak, J.M.: Machine learning-based classification of Alzheimer's disease and its at-risk states using personality traits, anxiety, and depression. Int. J. Geriatr. Psychiatry **38**(10), 6007 (2023). https://doi.org/10.1002/gps.6007. (**Accessed 2024-05-02**)
60. Treder, M.S., Shock, J.P., Stein, D.J., Du Plessis, S., Seedat, S., Tsvetanov, K.A.: Correlation constraints for regression models: controlling bias in brain age prediction. Front. Psych. **12**, 615754 (2021). https://doi.org/10.3389/fpsyt.2021.615754. (**Accessed 2025-02-05**)
61. Cole, J.H., Ritchie, S.J., Bastin, M.E., Valdés Hernández, M.C., Muñoz Maniega, S., Royle, N., Corley, J., Pattie, A., Harris, S.E., Zhang, Q., Wray, N.R., Redmond, P., Marioni, R.E., Starr, J.M., Cox, S.R., Wardlaw, J.M., Sharp, D.J., Deary, I.J.: Brain age predicts mortality. Mol. Psychiatry **23**(5), 1385–1392 (2018). https://doi.org/10.1038/mp.2017.62. (**Accessed 2020-10-28**)
62. Giudici, P.: Safe machine learning. Statistics **58**(3), 473–477 (2024). https://doi.org/10.1080/02331888.2024.2361481. (**Accessed 2026-03-19**)
63. Gunn, S.R.: Support Vector Machines for Classification and Regression. Project Report, s.n., University of Southampton (1998). https://eprints.soton.ac.uk/256459/
64. Brereton, R.G., Lloyd, G.R.: Support vector machines for classification and regression. Analyst **135**(2), 230–267 (2010). https://doi.org/10.1039/B918972F. (**Accessed 2022-04-12**)
65. Laaksonen, J., Oja, E.: Classification with learning k-nearest neighbors. In: Proceedings of International Conference on Neural Networks (ICNN'96), vol. 3, pp. 1480–1483. IEEE, Washington, DC, USA (1996). https://doi.org/10.1109/ICNN.1996.549118. http://ieeexplore.ieee.org/document/549118/ Accessed 2022-04-12
66. Kramer, O.: K-Nearest Neighbors. In: Dimensionality Reduction with Unsupervised Nearest Neighbors vol. 51, pp. 13–23. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38652-7_2. http://link.springer.com/10.1007/978-3-642-38652-7_2 Accessed 2022-04-12
67. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. **65**(6), 386–408 (1958). https://doi.org/10.1037/h0042519. (**Accessed 2022-04-12**)
68. Berkson, J.: Application to the logistic function to bio-assay. J. Am. Stat. Assoc. **39**(227), 357 (1944). https://doi.org/10.2307/2280041. (**Accessed 2022-04-12**)
69. Berkson, J.: Why i prefer logits to probits. Biometrics **7**(4), 327 (1951). https://doi.org/10.2307/3001655. (**Accessed 2022-04-12**)
70. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn. **29**(2–3), 103–130 (1997). https://doi.org/10.1023/A:1007413511361. (**Accessed 2026-02-20**)
71. Franc, V., Zien, A., Schölkopf, B.: Support Vector Machines as Probabilistic Models. In: Proceedings of the 28th International Conference on Machine Learning, pp. 665–672. International Machine Learning Society, Madison, WI, USA (2011)
72. Platt, J.C.: Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: Advances in Large Margin Classifiers, pp. 61–74. MIT Press, (1999)
73. Lin, H.-T., Lin, C.-J., Weng, R.C.: A note on Platt's probabilistic outputs for support vector machines. Mach. Learn. **68**(3), 267–276 (2007). https://doi.org/10.1007/s10994-007-5018-6. (**Accessed 2022-05-30**)
74. Hackmack, K., Paul, F., Weygandt, M., Allefeld, C., Haynes, J.-D.: Multi-scale classification of disease using structural MRI and wavelet transform. Neuroimage **62**(1), 48–58 (2012). https://doi.org/10.1016/j.neuroimage.2012.05.022. (**Accessed 2019-03-01**)
75. Eitel, F., Schulz, M.-A., Seiler, M., Walter, H., Ritter, K.: Promises and pitfalls of deep neural networks in neuroimaging-based psychiatric research. Exp. Neurol. **339**, 113608 (2021). https://doi.org/10.1016/j.expneurol.2021.113608. (**Accessed 2022-02-23**)
76. Huys, Q.J.M., Maia, T.V., Frank, M.J.: Computational psychiatry as a bridge from neuroscience to clinical applications. Nat. Neurosci. **19**(3), 404–413 (2016). https://doi.org/10.1038/nn.4238. (**Accessed 2020-09-08**)
77. Frässle, S., Marquand, A.F., Schmaal, L., Dinga, R., Veltman, D.J., Wee, N.J.A., Tol, M.-J., Schöbi, D., Penninx, B.W.J.H., Stephan, K.E.: Predicting individual clinical trajectories of depression with generative embedding. NeuroImage Clin. **26**, 102213 (2020). https://doi.org/10.1016/j.nicl.2020.102213. (**Accessed 2022-04-13**)
78. Chang, C.-C., Lin, C.-J.: LIBSVM FAQ. www.csie.ntu.edu.tw (2022). https://www.csie.ntu.edu.tw/~lowcjlin/libsvm/faq.html#f419
79. Hsu, C.-W., Lin, C.-J.: A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. **13**(2), 415–425 (2002). https://doi.org/10.1109/72.991427. (**Accessed 2022-04-13**)
80. Hofmann, T., Schölkopf, B., Smola, A.J.: Kernel methods in machine learning. The Annals of Statistics **36**(3), (2008). https://doi.org/10.1214/009053607000000677. Accessed 2022-04-13
81. Scholkopf, B., Kah-Kay Sung, Burges, C.J.C., Girosi, F., Niyogi, P., Poggio, T., Vapnik, V.: Comparing support vector machines with Gaussian kernels to radial basis function classifiers. IEEE Transactions on Signal Processing **45**(11), 2758–2765 (1997) https://doi.org/10.1109/78.650102. Accessed 2022-04-13
82. Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. J. R. Stat. Soc. Ser. A **135**(3), 370 (1972). https://doi.org/10.2307/2344614. (**Accessed 2022-05-30**)

83. Hebart, M.N., Baker, C.I.: Deconstructing multivariate decoding for the study of brain function. Neuroimage **180**, 4–18 (2018). https://doi.org/10.1016/j.neuroimage.2017.08.005. (**Accessed 2022-04-13**)
84. Polson, N.G., Scott, J.G.: Shrink globally, act locally: Sparse Bayesian regularization and prediction. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. (eds.) Bayesian Statistics 9, pp. 501–538. Oxford University Press, (2010). https://doi.org/10.1093/acprof:oso/9780199694587.003.0022
85. Mitchell, T.J., Beauchamp, J.J.: Bayesian variable selection in linear regression. J. Am. Stat. Assoc. **83**(404), 1023–1032 (1988). https://doi.org/10.1080/01621459.1988.10478694
86. George, E.I., McCulloch, R.E.: Variable selection via Gibbs sampling. J. Am. Stat. Assoc. **88**(423), 881–889 (1993). https://doi.org/10.1080/01621459.1993.10476353
87. West, M.: On scale mixtures of normal distributions. Biometrika **74**(3), 646–648 (1987). https://doi.org/10.1093/biomet/74.3.646
88. Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.-S.: A weakly informative default prior distribution for logistic and other regression models. Ann. Appl. Stat. **2**(4), 1360–1383 (2008). https://doi.org/10.1214/08-AOAS191
89. Tikhonov, A.N.: Solution of incorrectly formulated problems and the regularization method. Sov. Math. **4**, 1035–1038 (1963)
90. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970). https://doi.org/10.1080/00401706.1970.10488634. (**Accessed 2022-04-13**)
91. Frässle, S., Lomakina, E.I., Razi, A., Friston, K.J., Buhmann, J.M., Stephan, K.E.: Regression DCM for fMRI. Neuroimage **155**, 406–421 (2017). https://doi.org/10.1016/j.neuroimage.2017.02.090. (**Accessed 2022-04-13**)
92. Penny, W.D., Stephan, K.E., Daunizeau, J., Rosa, M.J., Friston, K.J., Schofield, T.M., Leff, A.P.: Comparing families of dynamic causal models. PLoS Comput. Biol. **6**(3), 1000709 (2010). https://doi.org/10.1371/journal.pcbi.1000709. (**Accessed 2015-10-30**)
93. Aponte, E.A., Yao, Y., Raman, S., Frässle, S., Heinzle, J., Penny, W.D., Stephan, K.E.: An introduction to thermodynamic integration and application to dynamic causal models. Cogn. Neurodyn. **16**(1), 1–15 (2022). https://doi.org/10.1007/s11571-021-09696-9. (**Accessed 2022-04-13**)
94. Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W.: Variational free energy and the Laplace approximation. Neuroimage **34**(1), 220–234 (2007). https://doi.org/10.1016/j.neuroimage.2006.08.035. (**Accessed 2015-10-30**)
95. Särkkä, S.: Bayesian Filtering and Smoothing. Institute of Mathematical Statistics Textbooks, vol. 3. Cambridge University Press, Cambridge, U.K. ; New York (2013)
96. Anderson, T.W.: An Introduction to Multivariate Statistical Analysis, 3rd edn. Wiley, New York (2003)
97. Glanz, H., Carvalho, L.: An expectation-maximization algorithm for the matrix normal distribution with an application in remote sensing. J. Multivar. Anal. **167**, 31–48 (2018). https://doi.org/10.1016/j.jmva.2018.03.010. (**Accessed 2026-01-28**)
98. Ramírez-Hassan, A.: Introduction to Bayesian Econometrics: A Guided Tour Using R. Bookdown, Online (2025). https://bookdown.org/aramir21/IntroductionBayesianEconometricsGuidedTour/

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