Gaussian Process Classification: Singly vs. Doubly Stochastic Models, and New Computational Schemes

Jens Röder
Corporate Sector Research and Advance Engineering
Multimedia, Telematic and Surround Sensing Systems (CR/AEM)
Robert Bosch GmbH, Hildesheim, Germany

Raimon Tolosana-Delgado
Maritime Engineering Laboratory (LIM)
Universitat Politécnica de Catalunya (UPC), Barcelona, Spain

Fred A. Hamprecht
Heidelberg Collaboratory for Image Processing (HCI)
University of Heidelberg, Germany
fred.hamprecht@iwr.uni-heidelberg.de

May 2011

Abstract

The aim of this paper is to compare four different methods for binary classification with an underlying Gaussian process with respect to theoretical consistency and practical performance. Two of the inference schemes, namely classical indicator kriging and simplicial indicator kriging, are analytically tractable and fast. However, these methods rely on simplifying assumptions which are inappropriate for categorical class labels. A consistent and previously described model extension involves a doubly stochastic process. There, the unknown posterior class probability $f(\cdot)$ is considered a realization of a spatially correlated Gaussian process that has been squashed to the unit interval, and a label at position $\mathbf{x}$ is considered an independent Bernoulli realization with success parameter $f(\mathbf{x})$. Unfortunately, inference for this model is not known to be analytically tractable. In this paper, we propose two new computational schemes for the inference in this doubly stochastic model, namely the "Aitchison Maximum Posterior" and the "Doubly Stochastic Gaussian Quadrature". Both methods are analytical up to a final step where optimization or integration must be carried out numerically. For the comparison of practical performance, the methods are applied to storm forecasts for the Spanish coast based on wave heights in the Mediterranean Sea. While the error rate of the doubly stochastic models is slightly lower, their computational cost is much higher.

1 Introduction

Environmental issues sometimes require binary classification: for example, the question whether to raise a dam (Hsu et al., 2010) or to issue a dengue fever warning (Yu et al., 2010) is a yes-or-no decision. Using supervised learning, such questions can be approached using historical precedents collected in a training set of exemplars with categorical labels. When it is fair to assume a relatively simple dependence of the response on the features, methods such as linear discriminant analysis or logistic regression are popular. In more complex settings, more flexible nonparametric methods are required. A successful nonparametric method originally proposed for regression is simple kriging (also called Gaussian process regression) which has found widespread use in geostatistics (Chilès and Delfiner, 1999), machine learning (Williams and Rasmussen, 1996) and signal processing (Wiener, 1949). The popularity of the method is due to its flexibility, mathematical tractability, its natural Bayesian interpretation and success in a wide range of applications (Gibbs, 1997; Rasmussen, 1996; Lim et al., 2002). In simple kriging, the outputs observed at points in feature space are assumed to arise from the realization of a Gaussian process with or without Gaussian noise; an approximation or interpolation (given the observed data and assumptions on the mean and covariance structure of the Gaussian process) is then obtained from the best linear unbiased estimator.

Unfortunately, inference is more complicated in classification, i.e. when the label at each point in feature space comes from a finite set. Whereas a Gaussian prior can be combined with a Gaussian likelihood in the case of regression (resulting in a simple computational scheme revolving around a linear system of equations), a Gaussian likelihood is obviously inappropriate for discrete class labels (Rasmussen and Williams, 2006).

In this paper, we compare four different approximation schemes for *binary* Gaussian process classification: classical indicator kriging (CIK; Journel, 1983), simplicial indicator kriging (SIK; Tolosana-Delgado et al., 2008) and the doubly stochastic Gaussian process, in its maximum posterior (AMP) and predictive quadrature (DSGQ) flavors. The last two are new computational schemes for a model that has previously been studied in environmental applications (Diggle et al., 1998) and in machine learning (Williams and Barber, 1998). This model is motivated and presented in the following paragraphs.

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ be the training set of a supervised learning problem, where $y_i \in \{0, 1\}$ denotes the binary class label of a feature vector $\mathbf{x}_i \in \mathbb{R}^m$. Let $\mathbf{y} \in \{0, 1\}^n$ and $\mathbf{X} \in \mathbb{R}^{n \times m}$ be the vector of labels and matrix of feature vectors, respectively. The goal is to predict the posterior class probability $p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$ of the unknown label y_* at a point $\mathbf{x}_*$ given the training set. In geostatistics, the feature vectors are typically the spatial coordinates where the labels were observed, i.e. $m = 2, 3$ or 4 (when depth, time, an external drift or a combination thereof is taken into consideration), and $y_i = I(\mathbf{x}_i)$ is the indicator of the label of interest at location $\mathbf{x}_i$. For instance, $I(\cdot)$ could indicate that a particular pollutant is above or below a given legal threshold, in a typical success/failure Bernoulli framework.

After Journel (1983), the classical approach is to compute the posterior probabilities by fitting the binary labels directly, without regard to the fact that the y_i cannot be normally distributed. This approximation is called (classical) indicator kriging and has a long record of successful applications. However, there are two main problems with CIK: First, it quite often delivers probabilities that are smaller than 0 or larger than 1, and second, the order relation of probabilities is violated. The latter means, as explained in more detail in Section 2.2, that the difference on the real line is not adequate to express distances between probabilities.

These drawbacks are tackled by SIK (Tolosana-Delgado et al., 2008). There,

1. the probability $p(y = 1|\mathbf{x})$ is considered an unobservable realization $f(\mathbf{x})$ of a Gaussian process "squashed" to the open unit interval $]0, 1[$ as defined in the next section.

But, as will be discussed in Section 2.4, SIK still makes simplifying assumptions that may not be satisfactory in general. In particular, $f_i := f(\mathbf{x}_i)$ can only be either p or $1-p$, $p \in]0, 0.5[$, depending on which of the two classes is observed at the training location $\mathbf{x}_i$, regardless of any other observations in the vicinity. This is contrary to intuition. For example, consider one point in feature space and its immediate neighbors, and two possible scenarios: first, that a "success" has been observed at all these points; and secondly, that "success" has been observed only at the central point and "failure" at all others. According to SIK, the posterior probability at the central point would be the same in both scenarios. The model must hence be extended such that we fully distinguish between an observed label and its estimated probability:

2. the observed labels y_i are *conditionally* independent realizations of Bernoulli distributions with parameters $f_i = p(y=1|\mathbf{x}_i)$, i.e. $y|f_i \sim \mathcal{B}ern(f_i)$.

A graphical representation of this doubly (1. and 2.) stochastic model is shown in Fig. 1.

In summary, CIK and SIK are based on model approximations that may be inconsistent with some or all characteristics of a classification setting, but that are linear in output measurements and thus analytically tractable and fast. In contrast, the doubly stochastic model is consistent, but predictive inference needs to be approximated.

The contribution of this paper is twofold. First, we compare these two antithetic approaches with respect to theoretical consistency and practical performance. Second, for this comparison, two new approximation schemes for the doubly stochastic model are presented, the Aitchison maximum posterior (AMP) and the doubly stochastic Gaussian quadrature (DSGQ), targeting at two slightly different concepts: the most likely value of the probability of success of the Bernoulli process at an unsampled location (AMP), and the numerical integration (or "quadrature") of the predictive probability of obtaining a success (DSGQ). Both methods are analytical up to a final step where optimization or integration must be performed numerically. This extends the insight into the doubly stochastic model and may form the basis for future research.

Many other approximation schemes for the doubly stochastic model have been proposed before, among them Laplace's method (Williams and Barber, 1998), the integrated nested Laplace approximation (Rue et al., 2009), Markov chain Monte Carlo (MCMC) approximations (Neal, 1999; Diggle et al., 1998), expectation propagation (Minka, 2001), the cavity TAP approximation (Opper and Winther, 2000) and a variational approximation (Gibbs and MacKay, 2000). An excellent review is presented by Kuss and Rasmussen (2005). In the field of geostatistics, modifications of CIK are also abundant (Journel and Posa, 1990; Carr and Mao, 1993; Suro-Perez and Journel, 1991; Pardo-Igúzquiza and Dowd, 2005), though most of them target the estimation of the conditional expectation of an underlying continuous characteristic: the exception may be transition probabilities (Carle and Fogg, 1996) and discrete variable maximum entropy treatments (Christakos, 1990; Bogaert, 2002), as both intrinsically model the probabilities of observing a categorical variable in a given "place". Carle and Fogg (1996) model the conditional probabilities to pass from a category to another at a given lag distance and use this information to build the $(n+1)$-dimensional contingency table of n observed and one unobserved label. Bogaert (2002) fits a log-linear model to this table which is hence based on a multi-Poisson distribution instead of relying on an unobservable trans-Gaussian distribution.

The next section reviews the most typically used method in the geostatistical community, CIK, as well as an alternative based on the Aitchison geometry of the unit interval, SIK. The underlying geometry is also presented in Section 2. Section 3 introduces two distributions on the unit interval, compatible with this geometry, that play the role of prior and posterior distributions of the vector of probabilities of interest. Section 4 then uses these distributions to derive the properties of a doubly stochastic Gaussian process for probabilities. Two estimators for the unknown probabilities that are based on these properties, AMP and DSGQ, are presented in the

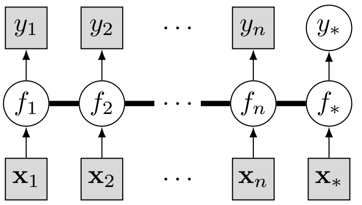

Figure 1: Graphical representation of the doubly stochastic model. Observed variables are shaded squares, circles represent unknowns. The thick lines indicate a fully connected graph. The first stochastic layer is given by the posterior class probabilities $f_i, i = 1, \dots, n$ and f_* that are considered function values of a squashed Gaussian process, the second layer is given by the observed labels that are Bernoulli distributed with parameter f_i.

same section.

An experimental comparison of all methods is presented in Section 5. To that end, a given forecast of wave heights in the Mediterranean Sea is classified in two conditions: *Eastwind-storm* (called *Llevant* in Catalan) and *any other situation* (either calm or any other of the dominant windstorms of this region) are the two possible labels. In this setting, the feature vector $\mathbf{x} \in \mathbb{R}^m$ consists of the values at m predefined pixels of a forecast map.

2 Classical and simplicial indicator kriging

In this section, we briefly review two approximation methods for Gaussian process classification that do not consider a Bernoulli distribution for the observed labels, namely classical and simplicial indicator kriging. We first describe simple kriging which is then applied to predict posterior class probabilities in a classification setting.

2.1 Simple kriging

Let $\{(\mathbf{x}_1, f_1), \dots, (\mathbf{x}_n, f_n)\}$ be n pairs of sampling points $\mathbf{x}_i$ (feature vectors), and outputs $f_i = f(\mathbf{x}_i), i = 1, \dots, n$ (labels). In case of simple kriging it is assumed that $f(\mathbf{x})$ is a realization of a Gaussian process with known mean and covariance structure $\mathrm{C}(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{Cov}[f(\mathbf{x}_i), f(\mathbf{x}_j)]$, i.e. the joint distribution of any subset of observed or unobserved points is a multivariate normal. Assuming a zero mean, the estimate for the function value at an arbitrary point $\mathbf{x}_*$ in feature space is of the form (see e.g. Chilès and Delfiner, 1999; Rasmussen and Williams, 2006)

$$\hat{f}_* = \hat{f}(\mathbf{x}_*) = \sum_{i=1}^{n} \lambda_i(\mathbf{x}_*) f(\mathbf{x}_i) \tag{1}$$

i.e. the simple kriging estimator is a linear combination of the function values at the sampling points. The coefficients $\lambda_i, i = 1, \dots, n$, depend on the position of prediction and are obtained maximizing the well-known normal conditional density (see e.g. Rasmussen and Williams, 2006)

$$p(f_*|\mathbf{f}, \mathbf{X}, \mathbf{x}_*) = \frac{1}{\sqrt{2\pi\left(\sigma_*^2 - \boldsymbol{\sigma}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma}\right)}} \exp\left(-\frac{1}{2} \frac{\left(f_* - \boldsymbol{\sigma}^T \boldsymbol{\Sigma}^{-1} \mathbf{f}\right)^2}{\sigma_*^2 - \boldsymbol{\sigma}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma}}\right)$$

where the covariances $\Sigma_{ij} = \mathrm{C}(\mathbf{x}_i, \mathbf{x}_j)$, $[\boldsymbol{\sigma}]_i = \sigma_i = \mathrm{C}(\mathbf{x}_i, \mathbf{x}_*)$ and $\sigma_* = \mathrm{C}(\mathbf{x}_*, \mathbf{x}_*)$. This gives kriging weight

$$\lambda_i(\mathbf{x}_*) = [\mathbf{\Sigma}^{-1}\boldsymbol{\sigma}]_i$$

The matrix $\mathbf{\Sigma}$ is invertible if the covariance function is strictly positive definite and if all the sampling points are distinct. Note that the coefficients $\lambda(\mathbf{x}_*)$ do not sum up to one in general, and can even be negative. For further details on simple kriging, other kriging "flavors" or examples of covariance functions, see e.g. (Chilès and Delfiner, 1999) and (Rasmussen and Williams, 2006).

2.2 Classical indicator kriging

The easiest possibility to predict posterior class probabilities at a point $\mathbf{x}_*$ in feature space is classical indicator kriging (CIK, Journel, 1983). There, the binary class labels $y_i \in \{0, 1\}$ are treated as function values, i.e. $f_i := y_i, i = 1, \dots, n$, and the probability of success is directly given by the simple kriging estimate $\hat{f}_*$. According to Chilès and Delfiner (1999), simple kriging should be theoretically preferred over ordinary kriging when working with indicators, because here knowledge of the variogram sill (equivalent to the variance of a Bernoulli variable, $\mathrm{var}[y] = p(1-p)$) implies knowledge of the mean of the indicator random function (equivalent to the mean of the same variable, $\mathrm{E}[y] = p$).

However, CIK has two major drawbacks. First, although the data $y_i \in \{0, 1\}$, it is not guaranteed that the interpolation $\hat{f}_* \in]0, 1[$, which is necessary in order to interpret it as a probability. Second, the order relation of probabilities is violated, i.e. distances between probabilities are not represented accurately by their difference on the real line. Consider the following example of two pairs of probabilities: $(0.001, 0.01)$ and $(0.501, 0.51)$. In the first case, the second probability is ten times higher, whereas in the second case, the probabilities are almost equal; but the actual distances on the real line are 0.009 in both cases. This suggests that a change of geometry may be adequate.

2.3 Geometry in the one-dimensional simplex $\mathbb{S}^2$

Consider the line segment $](0, 1), (1, 0)[\subset \mathbb{R}^2$, which is equal to the one-dimensional positive simplex $\mathbb{S}^2$. The simplex $\mathbb{S}^2$ is useful to represent the probability of a certain event together with its complementary probability because the components of an element of $\mathbb{S}^2$ always add up to 1. Moreover, it has a Euclidean vector space structure, called *Aitchison Geometry*, if it is endowed with the following three operations (Pawlowsky-Glahn and Egozcue, 2001; Billheimer et al., 2001). There, $\mathcal{C}(a) := (a_1/(a_1 + a_2), a_2/(a_1 + a_2))$ divides each component of a vector by the sum of its components to ensure the closure under addition and scalar multiplication.

(i) Vector addition: $a \oplus b := \mathcal{C}(a_1 b_1, a_2 b_2)$, representing addition of information following Bayes' Theorem

(ii) Scalar multiplication: $\lambda \odot a := \mathcal{C}(a_1^\lambda, a_2^\lambda), \lambda \in \mathbb{R}$

(iii) Scalar product: $\langle a, b \rangle := 1/c_0^2 \ln(a_1/a_2) \ln(b_1/b_2)$

The constant c_0^2 is a scaling parameter. As explained in detail in Section 3, it is intimately related to the variance of the normal distribution on the hypercube. It follows immediately from the above definitions that the additive neutral element of $\mathbb{S}^2$ is $\mathcal{C}(1, 1) = (1/2, 1/2)$ and the inverse element of $a = (a_1, a_2)$ is (a_2, a_1). Moreover, we automatically obtain an algebraic definition of the distance in $\mathbb{S}^2$:

$$d(a, b) = \|a \ominus b\| = \sqrt{\langle a \ominus b, a \ominus b \rangle} = \frac{1}{c_0} \sqrt{\left(\ln\left(\frac{a_1}{a_2}\right) - \ln\left(\frac{b_1}{b_2}\right) \right)^2} \tag{2}$$

where subtraction is defined by addition with the inverse element. The norm of a vector (a_1, a_2) in this geometry is $\|a\| = \sqrt{\langle a, a\rangle}$.

As in every Euclidean vector space we can choose an orthonormal basis – which, in this case, consists of one vector only: $e_b = \mathcal{C}(\exp(c_0), 1)$. In the *coordinate representation*, each element $a \in \mathbb{S}^2$ is uniquely represented with respect to the chosen basis:

$$\alpha = \langle a, e_b \rangle = \frac{1}{c_0} \ln \left(\frac{a_1}{1 - a_1} \right) \tag{3}$$

Conversely, the element a can be computed from its coordinate representation by scalar multiplication: $a = \alpha \odot e_b = \mathcal{C}(\exp(c_0\alpha), 1) = (a_1, 1 - a_1)$. The mapping from $\mathbb{S}^2$ to $\mathbb{R}$ assigning a coordinate to each point is an isomorphism. Furthermore, all points in $\mathbb{S}^2$ are uniquely determined by their first component, so we can identify a point a_1 on the real interval $]0, 1[$ with the point $(a_1, a_2) = (a_1, 1 - a_1)$ on $\mathbb{S}^2$ and hence the interval $]0, 1[$ with the simplex $\mathbb{S}^2$. This leads to an isomorphism from the interval $]0, 1[$ to $\mathbb{R}$. In the following, we use Latin letters for the elements of $\mathbb{S}^2$ and $]0, 1[$ and the corresponding Greek letters for the respective coordinates in $\mathbb{R}$.

2.4 Simplicial indicator kriging

The main drawbacks of CIK, stated in Section 2.2, are tackled by SIK (Tolosana-Delgado et al., 2008). This method is based on the realization that there is no need to establish an identity between a probability $f \in]0, 1[$ and its representation ϕ on the real line. They are better connected by the logit transformation:

$$\phi = \ln \frac{f}{1 - f} \tag{4}$$

Note the correspondence between Eqs. (3) and (4). The constant $1/c_0$ is omitted here because it cancels out in the final estimate f_* of SIK.

The simplicial kriging estimate is obtained in four steps:

1. estimate $f_i = p(y = 1|\mathbf{x}_i)$, the probabilities of success at each sample (of the training set); many estimation methods are possible (Tolosana-Delgado et al., 2008), e.g. a Bayesian estimate combining a Jeffreys' prior with the observed class likelihood, which would yield $\hat{f}_i = 3/4$ if a success is observed at $\mathbf{x}_i$, and $\hat{f}_i = 1/4$ otherwise;

2. get the logistic transformation of these estimates; being a log-ratio, this implies that extreme values of $p = 1$ or $p = 0$ should be avoided in the preceding step;

3. apply kriging the logistic-transformed estimates $\hat{\phi}_i = \ln(\hat{f}_i/1 - \hat{f}_i)$ to obtain an interpolation ϕ_* at an unclassified sample x_*;

4. undo the logistic transform, to obtain an interpolated probability $\hat{f}_* = \exp(\hat{\phi}_*)/(1 + \exp(\hat{\phi}_*))$.

The rationale behind SIK is to build the linear combination in Eq. (1) using the operations on the simplex explained in the preceding Section 2.3. Thus, SIK is an interpolation or approximation technique for probabilities within the framework of *squashed* Gaussian processes, as will be described next.

Tolosana-Delgado et al. (2008) also show that, if the estimates $\hat{f}_i$ are just $1-p$ or p, $p \in]0, 0.5[$, wherever a success, respectively a failure is observed, results get actually very simple. In this simplified situation, the CIK and SIK variograms are intimately related, and the estimate from the latter can be derived from that of the former, denoted here as f_*^{CIK}, by

$$f_* = \mathrm{logit}^{-1}\left(2\ln\frac{1-p}{p}\cdot\left(f_*^{CIK} - 0.5\right)\right). \tag{5}$$

Though Eq. (5) is an interesting way of "recycling" old, inconsistent CIK results into valid probabilities, estimating f_i by two values only (namely p and $1-p$) may still be a gross simplification.

But the main problem of SIK is its inability to "transfer information" between labeled points in the first step, as detailed in the thought experiment specified in the introduction. SIK is unable to deliver this result, because $f(\mathbf{x}_i)$ is estimated separately at each point $\mathbf{x}_i$, even in the presence of a nugget effect.

3 Distribution models for probabilities

In Section 3.1, we take the Euclidean vector space structure on the interval $]0,1[$ given in Section 2.3, and define the normal distribution on the unit hypercube $]0,1[^n$. This distribution serves as prior distribution for three of the methods for Gaussian process classification considered in this paper, namely SIK (presented in the previous section) and the two new methods (introduced in Section 4). The posterior obtained after updating this prior with a binomial likelihood is derived in Section 3.2.

3.1 The normal distribution on the unit hypercube

The transformation induced by the isomorphism presented in Section 2.3 maps the conventional normal distribution defined on the real line to the interval $]0,1[$:

Definition 1 *A random variable Z is said to be normally distributed on $]0,1[$, denoted $Z \sim \mathcal{N}_{]0,1[}(\mu,\sigma^2)$, if its coordinate representation* (3) *is normally distributed on $\mathbb{R}$ with mean μ and variance σ^2. (Pawlowsky-Glahn, 2003)*

It follows that the random variable Z has Lebesgue density

$$g(z|\mu,\sigma^2) = \frac{1}{c_0 z(1-z)}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\left(\frac{1}{c_0}\ln\left(\frac{z}{1-z}\right)-\mu\right)^2 / \left(2\sigma^2\right)\right) \tag{6}$$

$$= \frac{1}{z(1-z)}\frac{1}{\sqrt{2\pi c_0^2\sigma^2}}\exp\left(-\left(\ln\left(\frac{z}{1-z}\right)-c_0\mu\right)^2 / \left(2c_0^2\sigma^2\right)\right), z \in]0,1[, \mu\in\mathbb{R}, \sigma\in\mathbb{R}^+ \tag{7}$$

where the first factor in (6) comes from measure theory and compensates for the unfamiliar definition (2) of the distance in $\mathbb{S}^2$. Fig. 2 shows the probability density functions for $c_0 = 1$ and varying values of μ and σ^2.

The distribution of the latent variable at a single position in feature space lives on the interval $]0,1[$. The joint distribution of several variables – which will typically be dependent – then lives on the Cartesian product of these line segments, i.e. on the hypercube $]0,1[^n$.

Definition 2 *A random vector $\boldsymbol{Z}$ is normally distributed on $]0,1[^n$, denoted $\boldsymbol{Z} \sim \mathcal{N}^n_{]0,1[}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, if its coordinate representation* (3) *is multivariate normally distributed on $\mathbb{R}^n$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.*

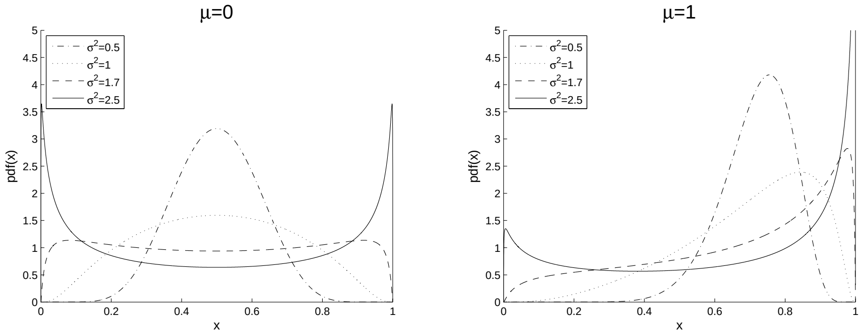

Figure 2: The normal distribution in $\mathbb{S}^2$ for different parameter values. For $\mu = 0$ (left panel), we obtain a symmetric density function around 0.5 ($0.5 \in]0, 1[$ has coordinate representation $0 \in \mathbb{R}$). The bigger the variance σ^2 the more probability mass is concentrated near the boundaries of the interval. In contrast to the usual normal distribution, the density function is apparently not symmetric for $\mu \neq 0$ (right panel). The expectation value of this distribution converges to 1 for $\mu \to +\infty$, and to 0 for $\mu \to -\infty$.

If we squash a Gaussian process to the unit interval according to the inverse of Eq. (3), its finite-dimensional distributions are normally distributed in the unit hypercube.

Remark 3 *Note that σ^2 and c_0 are intimately related and c_0 actually becomes a scaling parameter, or as was already mentioned, the units of the problem. This can be easily inferred from its behavior in term (7) for the one-dimensional distribution and is particularly evident for $\mu = 0$. In this case, σ^2 and c_0 become equivalent parameters. These considerations carry over to the multivariate case in Definition 2, where the multiplication of c_0 by a constant can be compensated by adapting $\mathbf{\Sigma}$ accordingly.*

Finally, note that the multivariate normal in the hypercube is not the only possible choice to model the prior distribution of a probability random field. Another approach not pursued here is using copulas instead (e.g., Kazianka and Pilz, 2010).

3.2 The Aitchison distribution

3.2.1 The one-dimensional case

The "Aitchison (1982) distribution" on the unit interval $\mathcal{A}(\boldsymbol{\theta}, \psi)$ is defined as an exponential family with log-likelihood, cumulant function and density given by

$$\begin{aligned} L_Z(z|\boldsymbol{\theta}, \psi) &= \theta_1 \ln(z) + \theta_2 \ln(1-z) + \psi \ln^2(z/(1-z)) \\ \kappa(\boldsymbol{\theta}, \psi) &= \int_0^1 \exp\left(-L_Z(z|\boldsymbol{\theta}, \psi)\right) dz \\ \ln f_Z(z|\boldsymbol{\theta}, \psi) &= \kappa(\boldsymbol{\theta}, \psi) + L_Z(z|\boldsymbol{\theta}, \psi) \end{aligned} \tag{8}$$

with location vector $\boldsymbol{\theta} = [\theta_1, \theta_2]$ and precision ψ. Eq. (8) integrates to a finite quantity (and is thus a proper density) when

- either $\theta_1 + \theta_2 \geq 0$ and $\psi < 0$,
- or both $\theta_1, \theta_2 > 0$ and $\psi \leq 0$.

This family generalizes both the logistic-normal model (obtained with the first condition when $\theta_1 + \theta_2 = 0$) and the Beta density (obtained in the second condition when $\psi = 0$), at the cost of only one parameter more than those needed for the logistic-normal case.

The Aitchison distribution of the first kind naturally occurs in a Bayesian framework when a logistic-normal prior $\mathcal{N}_{]0,1[}(\mu, \sigma^2)$ is chosen for the probability parameter p of a Binomial distribution $\mathcal{B}i(p, N)$. If our observation was y_1 successes and $y_2 = N - y_1$ failures, the posterior for p has an Aitchison distribution with parameters $\theta_1 = y_1 + \mu/\sigma^2, \theta_2 = y_2 - \mu/\sigma^2$, and precision $\psi^{-1} = -2\sigma^2$, unmodified by the updating.

3.2.2 The n-dimensional case

Extending the Aitchison distribution to the unit hypercube $]0, 1[^n$ requires another set of indices $i, j = 1, \ldots, n$, giving a joint log-likelihood of

$$L_{\mathbf{Z}}(\mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\Psi}) = \sum_{i=1}^{n} \left[\theta_1^i \ln(z_i) + \theta_2^i \ln(1 - z_i) + \sum_{j=1}^{n} \psi_{ij} \ln \frac{z_i}{1 - z_i} \ln \frac{z_j}{1 - z_j} \right] \tag{9}$$

where $\psi_{ij} = \psi_{ji}$ is a measure of mutual influence between the “edges” i and j of the hypercube. The validity conditions are a generalization of the one-dimensional case: the matrix $\boldsymbol{\Psi} = [\psi_{ij}]$ must be negative (semi-) definite, and regarding $\boldsymbol{\theta}$, either $\theta_1^i + \theta_2^i \geq 0$ or else $\theta_1^i, \theta_2^i > 0$ for each i. With respect to ζ_i, the logistic coordinate of z_i (Eq. (3)), this likelihood can be expressed as

$$\begin{aligned} L_{\mathbf{Z}}(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\Psi}) &= \sum_{i=1}^{n} \theta_1^i \ln \left(\frac{e^{c_0 \zeta_i}}{1 + e^{c_0 \zeta_i}} \right) + \theta_2^i \ln \left(\frac{1}{1 + e^{c_0 \zeta_i}} \right) + \sum_{j=1}^{n} \psi_{ij} \zeta_i \zeta_j \\ &= \sum_{i=1}^{n} \theta_1^i c_0 \zeta_i - (\theta_1^i + \theta_2^i) \ln \left(1 + e^{c_0 \zeta_i} \right) + \sum_{j=1}^{n} \psi_{ij} \zeta_i \zeta_j \end{aligned}$$

The n-tuple Aitchison distribution on the unit hypercube $]0, 1[^n$ occurs as the posterior distribution of the probability field of a Gaussian process classification problem, as the next section shows.

4 Doubly stochastic Gaussian process

We introduce the doubly stochastic model in Section 4.1. Both methods presented in the subsequent sections, namely the AMP and the DSGQ, are based on these model assumptions and are different estimators for the unknown posterior class probability $p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$.

4.1 Posing the model

Let us from now on use the coordinate representation $\phi_i \in \mathbb{R}$ for $f_i \in]0, 1[$ as introduced in Section 2.3,

$$\phi_i = \frac{1}{c_0} \log \left(\frac{f_i}{1 - f_i} \right) \Leftrightarrow f_i = \frac{e^{c_0 \phi_i}}{1 + e^{c_0 \phi_i}} \tag{10}$$

In the real coordinate space we can perform the usual Bayesian inference for regression without any restrictions, warranted by the *principle of working on coordinates* (Pawlowsky-Glahn, 2003).

Recall that our goal is to predict the probability distribution of the unknown label y_* at a point $\mathbf{x}_*$, given the training set $\mathcal{D}$. In Section 1, we have introduced the two following model assumptions:

1. the probability $p(y=1|\mathbf{x})$ is considered an unobservable realization $f(\mathbf{x})$ of a Gaussian process squashed to the unit interval,
2. the observed labels y_i are *independent* realizations of Bernoulli distributions with parameters $f_i = p(y=1|\mathbf{x}_i)$, i.e. $y|f_i \sim \mathcal{B}ern(f_i)$.

This two-layer model can be successfully tackled in a Bayesian framework.

The second assumption implies that the likelihood of a sampled label vector $\mathbf{y}$ is

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i) = \prod_{i=1}^{n} f_i^{y_i}(1-f_i)^{1-y_i}$$

where $\mathbf{f} = (f_1, \dots, f_n)^T$. Taking logs, we obtain

$$\ln(p(\mathbf{y}|\mathbf{f})) = \sum_{i=1}^{n} y_i \ln(f_i) + (1-y_i)\ln(1-f_i)$$

or in coordinates

$$\begin{aligned}\ln(p(\mathbf{y}|\mathbf{f})) &= \sum_{i=1}^{n} y_i \ln\frac{e^{c_0\phi_i}}{1+e^{c_0\phi_i}} + (1-y_i)\ln\frac{1}{1+e^{c_0\phi_i}} \\ &= \sum_{i=1}^{n} c_0\phi_i y_i + (y_i + 1 - y_i)\ln\frac{1}{1+e^{c_0\phi_i}} = \sum_{i=1}^{n} c_0\phi_i y_i - \ln(1+e^{c_0\phi_i}) \\ &= c_0\boldsymbol{\phi}^T\mathbf{y} - \sum_{i=1}^{n}\ln(1+e^{c_0\phi_i}) = c_0\boldsymbol{\phi}^T\mathbf{y} - g(\boldsymbol{\phi}) \qquad (11)\end{aligned}$$

where $g(\boldsymbol{\phi}) = \sum_{i=1}^{n}\ln(1+e^{c_0\phi_i})$.

According to our first prior assumption, we may consider the unobserved success probability $p(y=1|\mathbf{x})$ to follow normal distribution on the hypercube, as given by Definition 2. This assumption implies that we must know its mean vector and covariance matrix. If we have no information favoring one predicted class over the other, the mean may be considered zero in coordinate space, corresponding to a probability of 1/2 for each of the possible labels (Fig. 2), such that the prior distribution can be written

$$p(\mathbf{f}, f_*|\mathbf{X}, \mathbf{x}_*) = \mathcal{N}_{]0,1[^{n+1}}(\mathbf{0}, \mathbf{C}), \quad \mathbf{C} = \begin{pmatrix} \boldsymbol{\Sigma}(\mathbf{X}) & \boldsymbol{\sigma}(\mathbf{X}, \mathbf{x}_*) \\ \boldsymbol{\sigma}(\mathbf{X}, \mathbf{x}_*)^T & \sigma_*^2 \end{pmatrix}. \qquad (12)$$

The several covariances $\boldsymbol{\Sigma}(\mathbf{X})$ among sampled locations and $\boldsymbol{\sigma}(\mathbf{X}, \mathbf{x}_*)$ between a sampled location and the unsampled one, are derived from a second-order stationary covariance function, giving smoothness to the hidden random function in feature space. Note that, as a result of the derivations later on, the covariance function enters the final prediction at $\mathbf{x}_*$ only through the prior. Hence, as the multiplication of c_0 by a constant can be compensated by adapting $\mathbf{C}$ according to Remark 3, c_0 can be set to 1 in the prior and thus later on in Eqs. (13) and (22).

4.2 Aitchison maximum posterior (AMP)

4.2.1 The posterior distribution

Let $\mathbf{P}$ be proportional to the precision matrix of the prior distribution in Eq. (12), i.e.

$$\mathbf{P} = -\frac{1}{2}\mathbf{C}^{-1} = \begin{pmatrix} \boldsymbol{\Psi}(\mathbf{X}) & \boldsymbol{\psi}(\mathbf{X}, \mathbf{x}_*) \\ \boldsymbol{\psi}(\mathbf{X}, \mathbf{x}_*)^T & \psi_*^2 \end{pmatrix} \qquad (13)$$

where, for future reference,

$$\boldsymbol{\psi}(\mathbf{X},\mathbf{x}_*) = -\,\boldsymbol{\Sigma}(\mathbf{X})^{-1}\boldsymbol{\sigma}(\mathbf{X},\mathbf{x}_*)\cdot\psi_*^2 \tag{14}$$

$$\boldsymbol{\Sigma}(\mathbf{X}) = -\,\frac{1}{2}\left(\boldsymbol{\Psi}(\mathbf{X}) - \psi_*^{-2}\cdot\boldsymbol{\psi}(\mathbf{X},\mathbf{x}_*)\boldsymbol{\psi}(\mathbf{X},\mathbf{x}_*)^T\right)^{-1} \tag{15}$$

(see e.g. Petersen and Pedersen, 2008, page 45). This allows to express the log-density as

$$\ln(p(\mathbf{f},f_*|\mathbf{X},\mathbf{x}_*)) = \kappa_0 + \boldsymbol{\phi}^T\boldsymbol{\Psi}(\mathbf{X})\boldsymbol{\phi} + 2\boldsymbol{\phi}^T\boldsymbol{\psi}(\mathbf{X},\mathbf{x}_*)\phi_* + \psi_*^2\phi_*^2 \tag{16}$$

Taking logs, Bayes theorem becomes a sum of log-prior and log-likelihood, plus an irrelevant closing constant,

$$\begin{aligned}\ln(p(f_*,\mathbf{f}|\mathbf{X},\mathbf{y},\mathbf{x}_*)) &= \ln(p(\mathbf{f},f_*|\mathbf{X},\mathbf{x}_*)) + \ln(p(\mathbf{y}|\mathbf{f},f_*,\mathbf{X},x_*)) + \kappa_1\\ &= \ln(p(\mathbf{f},f_*|\mathbf{X},\mathbf{x}_*)) + \ln(p(\mathbf{y}|\mathbf{f})) + \kappa_1\\ &= \boldsymbol{\phi}^T\boldsymbol{\Psi}(\mathbf{X})\boldsymbol{\phi} + 2\boldsymbol{\phi}^T\boldsymbol{\psi}(\mathbf{X},\mathbf{x}_*)\phi_* + \psi_*^2\phi_*^2 + c_0\boldsymbol{\phi}^T\mathbf{y} - g(\boldsymbol{\phi}) + \kappa_2\end{aligned} \tag{17}$$

Note that in this expression, the unknown ϕ_* at the unsampled location is involved only in quadratic terms, like in Eq. (16). On the contrary, $\boldsymbol{\phi}$ associated to sampled locations has the quadratic terms plus terms coming from the binomial likelihood of Eq. (11). Thus, the conditional log-likelihoods (17) are either a logistic-normal one for unsampled locations, or an Aitchison one like Eq. (9) for sampled locations.

4.2.2 The maximum posterior estimator

One can obtain a joint estimation for $(\mathbf{f},f_*)$ taking that value that maximizes the posterior density of Eq. (17), i.e. the most likely a posteriori value. To maximize this joint log-likelihood we can take derivatives. Equating these derivatives to zero

$$0 = \frac{d\ln(p(f_*,\mathbf{f}|\mathbf{X},\mathbf{y},\mathbf{x}_*))}{d\phi_*} = 2\boldsymbol{\psi}(\mathbf{X},\mathbf{x}_*)^T\boldsymbol{\phi} + 2\psi_*^2\phi_* \tag{18}$$

$$0 = \frac{d\ln(p(f_*,\mathbf{f}|\mathbf{X},\mathbf{y},\mathbf{x}_*))}{d\boldsymbol{\phi}} = 2\boldsymbol{\Psi}(\mathbf{X})\boldsymbol{\phi} + 2\boldsymbol{\psi}(\mathbf{X},\mathbf{x}_*)\phi_* + c_0(\mathbf{y}-\mathbf{f}) \tag{19}$$

one obtains a non-linear system of equations. From its solution the value ϕ_* is extracted, and then plugged into the coordinate-probability inverse relationship of Eq. (10), to obtain the sought maximum posterior probability estimate. Note that in Eq. (19) we used

$$\frac{dg(\boldsymbol{\phi})}{d\phi_i} = \frac{d}{d\phi_i}\sum_{j=1}^{n}\ln(1+e^{c_0\phi_j}) = c_0\frac{e^{c_0\phi_i}}{1+e^{c_0\phi_i}} = c_0 f_i$$

However, the conditional independence between sampled and unsampled observations displayed in Fig. 1 implies that the estimation can be more efficiently obtained in two steps:

1. first, estimate the logistic coordinates $\boldsymbol{\phi}$ of the probability field at the sampled locations as the mode of an n-tuple of Aitchison density solving an n-dimensional system of non-linear equations;

2. then, estimate the coordinates ϕ_*, using Eq. (18). Because the distribution of ϕ_* conditional to $\boldsymbol{\phi}$ is a normal one, this estimation is just equivalent to regression or simple kriging.

This procedure first extracts all the information from the binomial observations to estimate the success probabilities at the *observed* locations, and then interpolates them to the *unobserved* ones. This behavior is desirable, as non-sampled locations therefore have no influence on the success probability estimates at sampled locations. However, it remains to be proved that we can actually ignore ϕ_* in Eq. (19), and that Eq. (18) is equivalent to simplicial indicator kriging. The rest of this subsection supports these statements.

To prove the first statement, one first isolates ϕ_* in Eq. (18) as a function of $\boldsymbol{\phi}$, and substitutes it in Eq. (19), which gives

$$\begin{aligned} 0 &= 2\boldsymbol{\Psi}(\mathbf{X})\boldsymbol{\phi} + 2\boldsymbol{\psi}(\mathbf{X},\mathbf{x}_*)\left(-\frac{1}{\psi_*^2}\boldsymbol{\psi}(\mathbf{X},\mathbf{x}_*)^T\boldsymbol{\phi}\right) + c_0(\mathbf{y}-\mathbf{f}) \\ &= 2\left(\boldsymbol{\Psi}(\mathbf{X}) - \frac{1}{\psi_*^2}\boldsymbol{\psi}(\mathbf{X},\mathbf{x}_*)\boldsymbol{\psi}(\mathbf{X},\mathbf{x}_*)^T\right)\boldsymbol{\phi} + c_0(\mathbf{y}-\mathbf{f}) \end{aligned}$$

Taking into account then Eq. (15) we obtain

$$0 = -\boldsymbol{\Sigma}(\mathbf{X})^{-1}\boldsymbol{\phi} + c_0(\mathbf{y}-\mathbf{f}) \tag{20}$$

which is what we would obtain as Eq. (19), if we had no ϕ_* at all. Its solution thus maximizes the marginal posterior distribution of the sampled locations *only*, as stated in the first step.

Now assume that $\mathbf{f}$ (and $\boldsymbol{\phi}$) at sampled locations are known, either a priori or by solving Eq. (20). Then, maximization of the joint likelihood (17) is done only with respect to f_*. Thus, after taking derivatives we obtain just Eq. (18). Taking then into account the relations between the block matrices in precision and variance described in Eq. (14), one obtains

$$\begin{aligned} & 0 &=& \frac{1}{2}\frac{d\ln(p(f_*,\mathbf{f}|\mathbf{X},\mathbf{y},\mathbf{x}_*))}{d\phi_*} = \boldsymbol{\phi}^T\boldsymbol{\psi}(\mathbf{X},\mathbf{x}_*) + \psi_*^2\phi_* \\ \Leftrightarrow\quad & 0 &=& \boldsymbol{\phi}^T(-\boldsymbol{\Sigma}(\mathbf{X})^{-1}\boldsymbol{\sigma}(\mathbf{X},\mathbf{x}_*)\cdot\psi_*^2) + \psi_*^2\phi_* \\ \Leftrightarrow\quad & 0 &=& \boldsymbol{\phi}^T(-\boldsymbol{\Sigma}(\mathbf{X})^{-1}\boldsymbol{\sigma}(\mathbf{X},\mathbf{x}_*)) + \phi_* \\ \Leftrightarrow\quad & \phi_* &=& \boldsymbol{\phi}^T(\boldsymbol{\Sigma}(\mathbf{X})^{-1}\boldsymbol{\sigma}(\mathbf{X},\mathbf{x}_*)) \end{aligned}$$

i.e. the interpolated coordinate ϕ_* is a linear combination of the "observed" coordinates $\boldsymbol{\phi}$, with weights $\boldsymbol{\lambda} = \boldsymbol{\Sigma}(\mathbf{X})^{-1}\boldsymbol{\sigma}(\mathbf{X},\mathbf{x}_*)$, equal to the simple kriging weights (see Section 2.1), as stated in the second step.

4.3 Doubly Stochastic Gaussian Quadrature (DSGQ)

4.3.1 Predictive estimation

However, we are actually not interested in the posterior probability of obtaining a success at an unsampled location, given by taking the conditional estimate $p(f_*|\mathbf{f},\mathbf{X},\mathbf{y},\mathbf{x}_*)$ from the obtained maximum posterior estimate $p(f_*,\mathbf{f}|\mathbf{X},\mathbf{y},\mathbf{x}_*)$. We would rather prefer the predictive probability $p(y_*|\mathbf{X},\mathbf{y},\mathbf{x}_*)$, accounting for the fact that the estimate $\hat{f}_*$ is also uncertain.

Following the definition of predictive estimation, we know that

$$\begin{aligned} p(y_*|\mathbf{X},\mathbf{y},\mathbf{x}_*) &= \int p(y_*,f_*|\mathbf{X},\mathbf{y},x_*)df_* \\ &= \int p(y_*|f_*)p(f_*|\mathbf{X},\mathbf{y},\mathbf{x}_*)df_* \end{aligned}$$

which may be computed taking

$$p(y_*|f_*) = \left(\frac{e^{c_0\phi_*}}{1+e^{c_0\phi_*}}\right)^{y_*}\left(\frac{1}{1+e^{c_0\phi_*}}\right)^{1-y_*} = \frac{e^{c_0\phi_* y_*}}{1+e^{c_0\phi_*}}$$

and computing $p(f_*|\mathbf{X},\mathbf{y},\mathbf{x}_*)$ as the exponential of Eq. (17). However, this would require the numerical computation of the closing constant κ_2. Instead, we can go on with the calculations as in (Rasmussen and Williams, 2006) and (Kuss and Rasmussen, 2005), using the conditional independence assumptions reflected by Fig. 1. This gives

$$\begin{aligned} &= \int p(y_*|f_*)\int p(f_*,\mathbf{f}|\mathbf{X},\mathbf{y},\mathbf{x}_*)d\mathbf{f}\,df_* \\ &= \int p(y_*|f_*)\int p(f_*|\mathbf{f},\mathbf{X},\mathbf{x}_*)p(\mathbf{f}|\mathbf{X},\mathbf{y})d\mathbf{f}\,df_* \\ &= \frac{1}{c_1}\int p(y_*|f_*)\int p(f_*|\mathbf{f},\mathbf{X},\mathbf{x}_*)p(\mathbf{y}|\mathbf{f},\mathbf{X})p(\mathbf{f}|\mathbf{X})d\mathbf{f}\,df_* \end{aligned} \tag{21}$$

Using the probability density function of the normal distribution in $]0,1[^n$, plugging in the coordinate representation (10) and repeatedly applying the substitution rule of integration, we obtain explicit expressions for all terms in (21), viz.

$$p(f_*|\mathbf{f},\mathbf{X},\mathbf{x}_*) = \frac{1}{\sqrt{2\pi\left(\sigma_*^2-\boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}\right)}}\exp\left(-\frac{1}{2}\frac{\left(\phi_*-\boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}\right)^2}{\sigma_*^2-\boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}}\right) \tag{22}$$

$$p(\mathbf{y}|\mathbf{f},\mathbf{X}) = p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n}\left[\left(\frac{e^{c_0\phi_i}}{1+e^{c_0\phi_i}}\right)^{y_i}\left(\frac{1}{1+e^{c_0\phi_i}}\right)^{1-y_i}\right] = \prod_{i=1}^{n}\frac{e^{c_0\phi_i y_i}}{1+e^{c_0\phi_i}}$$

$$p(\mathbf{f}|\mathbf{X}) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}\boldsymbol{\phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}\right)$$

where we have used among others the derivation in (Rasmussen and Williams, 2006, chap. 2.2) for Eq. (22). In these equations, we have used the shorter notation $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\mathbf{X})$, as well as $\boldsymbol{\sigma} = \boldsymbol{\sigma}(\mathbf{X},\mathbf{x}_*)$.

The integration in (21) now is with respect to $\boldsymbol{\phi}$ and ϕ_*, logistic coordinates of the unobservable probabilities $\mathbf{f}$ and f_*. Inserting the terms mentioned before, we obtain

$$p(y_*|\mathbf{X},\mathbf{y},\mathbf{x}_*) = c_2\iint\frac{e^{c_0\phi_* y_*}}{1+e^{c_0\phi_*}}\prod_{i=1}^{n}\frac{e^{c_0\phi_i y_i}}{1+e^{c_0\phi_i}}\exp\left(-\frac{\boldsymbol{\phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}}{2}-\frac{(\phi_*-\boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi})^2}{2s_*^2}\right)d\boldsymbol{\phi}\,d\phi_* \tag{23}$$

with $s_*^2 := \sigma_*^2 - \boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}$ and $c_2 := (c_1\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}\sqrt{2\pi s_*^2})^{-1}$. This integral cannot be solved in closed form.

4.3.2 Approximating the integral

In this section, we derive a computational scheme for the calculation of the predictive doubly stochastic mode for Gaussian process classification presented before.

Since the integral in Eq. (23) cannot be solved analytically, we approximate the exact logistic function in (10) by a stretched error function, i.e.

$$\frac{e^{c_0\phi_* y_*}}{1+e^{c_0\phi_*}} \approx \Phi\left((-1)^{y_*+1}k_0\phi_*\right) \tag{24}$$

where Φ denotes the error function. This allows for substantial simplifications leading to the result in equation (26). Choosing

$$k_0 = \arg\min_k \max_{\phi_*} \left| \frac{e^{c_0\phi_*}}{1+e^{c_0\phi_*}} - \Phi(k\phi_*) \right| \approx 0.5876 c_0$$

we obtain a good approximation with a maximum deviation of

$$\max_{\phi_*} \left| \frac{e^{c_0\phi_*}}{1+e^{c_0\phi_*}} - \Phi(k_0\phi_*) \right| < 0.01$$

for every c_0 (see Fig. 3). Of course, the same calculation is valid for ϕ_i and $y_i, i = 1, \ldots, n$, instead of ϕ_* and y_*.

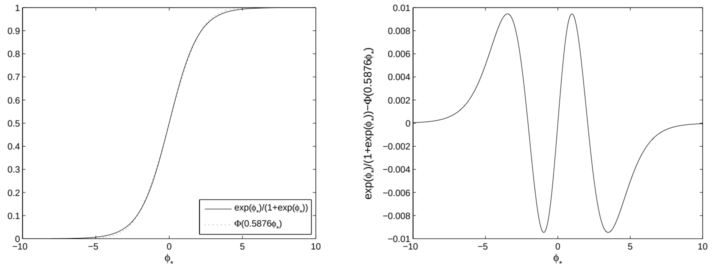

Figure 3: Comparison of the original logistic function and its stretched inverse probit approximation for $c_0 = 1$. The left panel shows the two functions, the right panel their difference.

Working toward the final simplification, we define a multivariate generalization of the Heaviside function $H(\xi)$.

Definition 4 *Let*

$$\boldsymbol{H}_{\boldsymbol{y}}(\boldsymbol{\xi}) := \begin{cases} 0 & \textit{if} \;\; \exists i : (-1)^{y_i+1}\xi_i < 0 \\ \frac{1}{2} & \textit{if} \;\; \forall i : (-1)^{y_i+1}\xi_i \geq 0 \;\textit{and}\; \exists i : \xi_i = 0 \\ 1 & \textit{if} \;\; \forall i : (-1)^{y_i+1}\xi_i > 0 \end{cases}$$

Special cases of $\mathbf{H}_{\mathbf{y}}(\boldsymbol{\xi})$ in one dimension are $\mathbf{H}_1(\xi) = H(\xi)$ and $\mathbf{H}_0(\xi) = H(-\xi) = 1 - H(\xi)$. Summarized in words, the function $\mathbf{H}_{\mathbf{y}}$ is – up to a null set with respect to the Lebesgue measure – equal to 1 in exactly one orthant of $\mathbb{R}^n$ and equal to 0 elsewhere, where the orthant is specified by the components of $\mathbf{y}$.

One can verify that $\Phi(k_0\phi_*) = (\mathbf{H}_1 * \mathcal{N}_{0,\frac{1}{k_0^2}})(\phi_*)$ and $\Phi(-k_0\phi_*) = (\mathbf{H}_0 * \mathcal{N}_{0,\frac{1}{k_0^2}})(\phi_*)$, and hence

$$\prod_{i=1}^{n} \Phi\left((-1)^{y_i+1}k_0\phi_i\right) = \prod_{i=1}^{n} \left(\mathbf{H}_{y_i} * \mathcal{N}_{0,\frac{1}{k_0^2}}\right)(\phi_i) = \left(\mathbf{H}_{\mathbf{y}} * \mathcal{N}_{\mathbf{0},\frac{1}{k_0^2}\mathbf{I}}\right)(\boldsymbol{\phi}). \tag{25}$$

Inserting the approximation in Eq. (24), Definition 4 and Eq. (25) in Eq. (23), we can continue the main calculation so that

$$\begin{aligned} p(y_*|\mathbf{X},\mathbf{y},\mathbf{x}_*) \approx c_2 & \int \left(\mathbf{H}_{\mathbf{y}} * \mathcal{N}_{\mathbf{0},\frac{1}{k_0^2}\mathbf{I}}\right)(\boldsymbol{\phi}) \;\exp\left(-\frac{1}{2}\boldsymbol{\phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}\right) \\ & \times \int \left(\mathbf{H}_{y_*} * \mathcal{N}_{0,\frac{1}{k_0^2}}\right)(\phi_*) \;\exp\left(-\frac{1}{2s_*^2}(\phi_* - \boldsymbol{\sigma}(\mathbf{x}_*)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi})^2\right) d\phi_*\, d\boldsymbol{\phi} \end{aligned}$$

Considering only the inner integral we have

$$
\begin{aligned}
&\int \left(\mathbf{H}_{y_*} * \mathcal{N}_{0,\frac{1}{k_0^2}}\right)(\phi_*)\ \exp\left(-\frac{1}{2s_*^2}(\phi_* - \boldsymbol{\sigma}(\mathbf{x}_*)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi})^2\right) d\phi_* \\
=&\iint \mathbf{H}_{y_*}(\xi_*)\mathcal{N}_{0,\frac{1}{k_0^2}}(\phi_* - x_B)d\xi_*\ \exp\left(-\frac{1}{2s_*^2}(\phi_* - \boldsymbol{\sigma}(\mathbf{x}_*)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi})^2\right) d\phi_* \\
=&\sqrt{2\pi s_*^2}\int \mathbf{H}_{y_*}(\xi_*)\underbrace{\int \mathcal{N}_{0,\frac{1}{k_0^2}}(x_B - \phi_*)\mathcal{N}_{\boldsymbol{\sigma}(\mathbf{x}_*)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi},s_*^2}(\phi_*)d\phi_*}_{\mathcal{N}_{\boldsymbol{\sigma}(\mathbf{x}_*)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi},s_*^2+\frac{1}{k_0^2}}(\xi_*)}\, d\xi_*
\end{aligned}
$$

which leads to

$$
\begin{aligned}
p(y_*|\mathbf{X},\mathbf{y},\mathbf{x}_*) \approx\,& c_3 \iint \mathbf{H}_{\mathbf{y}}(\boldsymbol{\xi})\mathcal{N}_{0,\frac{1}{k_0^2}\mathbf{I}}(\boldsymbol{\phi}-\boldsymbol{\xi})d\boldsymbol{\xi} \\
&\times \exp\left(-\frac{1}{2}\boldsymbol{\phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}\right)\int \mathbf{H}_{y_*}(\xi_*)\mathcal{N}_{\boldsymbol{\sigma}(\mathbf{x}_*)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi},s_*^2+\frac{1}{k_0^2}}(\xi_*)d\xi_*\, d\boldsymbol{\phi} \\
=\,& c_3 \iint \mathbf{H}_{\mathbf{y}}(\boldsymbol{\xi})\mathbf{H}_{y_*}(\xi_*) \\
&\times \int \mathcal{N}_{0,\frac{1}{k_0^2}\mathbf{I}}(\boldsymbol{\phi}-\boldsymbol{\xi})\exp\left(-\frac{1}{2}\boldsymbol{\phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}\right)\mathcal{N}_{0,s_*^2+\frac{1}{k_0^2}}(\boldsymbol{\sigma}(\mathbf{x}_*)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}-\xi_*)d\boldsymbol{\phi}\, d\boldsymbol{\xi}\, d\xi_*
\end{aligned}
$$

Defining $\mathbf{s} := \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}(\mathbf{x}_*)$ and $v := k_0^2/(s_*^2k_0^2+1)$, we obtain (up to a constant multiplier) for the integrand of the inner integral

$$
\begin{aligned}
&\exp\left(-\frac{1}{2}\left(\boldsymbol{\phi}-\boldsymbol{\xi}\right)^T k_0^2\left(\boldsymbol{\phi}-\boldsymbol{\xi}\right) - \frac{1}{2}\boldsymbol{\phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi} - \frac{1}{2}\left(\mathbf{s}^T\boldsymbol{\phi}-\xi_*\right)v\left(\mathbf{s}^T\boldsymbol{\phi}-\xi_*\right)\right) \\
=&\exp\left(-\frac{1}{2}\boldsymbol{\phi}^T\underbrace{\left(k_0^2\mathbf{I}+\boldsymbol{\Sigma}^{-1}+v\mathbf{s}\mathbf{s}^T\right)}_{:=\mathbf{R}}\boldsymbol{\phi} + \boldsymbol{\phi}^T\underbrace{\left(k_0^2\boldsymbol{\xi}+v\mathbf{s}\xi_*\right)}_{:=\mathbf{m}} - \frac{1}{2}\boldsymbol{\xi}^T k_0^2\boldsymbol{\xi} - \frac{1}{2}v\xi_*^2\right) \\
=&\exp\left(-\frac{1}{2}(\boldsymbol{\phi}-\mathbf{R}^{-1}\mathbf{m})^T\mathbf{R}(\boldsymbol{\phi}-\mathbf{R}^{-1}\mathbf{m})\right)\exp\left(\frac{1}{2}\mathbf{m}^T\mathbf{R}^{-1}\mathbf{m} - \frac{1}{2}\boldsymbol{\xi}^T k_0^2\boldsymbol{\xi} - \frac{1}{2}v\xi_*^2\right)
\end{aligned}
$$

The second factor is independent of $\boldsymbol{\phi}$ and the first factor is a Gaussian kernel function which integrates to a constant with respect to $\boldsymbol{\phi}$. Combining this constant with c_3 we obtain

$$
p(y_*|\mathbf{X},\mathbf{y},\mathbf{x}_*) \approx c_4 \iint \mathbf{H}_{\mathbf{y}}(\boldsymbol{\xi})\mathbf{H}_{y_*}(\xi_*)\exp\left(\frac{1}{2}\mathbf{m}^T\mathbf{R}^{-1}\mathbf{m} - \frac{1}{2}\boldsymbol{\xi}^T k_0^2\boldsymbol{\xi} - \frac{1}{2}v\xi_*^2\right)d\boldsymbol{\xi}\, d\xi_*
$$

When resubstituting $\mathbf{m}$ and reordering, the exponent becomes

$$
-\frac{1}{2}\boldsymbol{\xi}^T(k_0^2\mathbf{I} - k_0^4\mathbf{R}^{-1})\boldsymbol{\xi} + \boldsymbol{\xi}^T k_0^2\mathbf{R}^{-1}v\mathbf{s}\xi_* - \frac{1}{2}\xi_*\left(v - v^2\mathbf{s}^T\mathbf{R}^{-1}\mathbf{s}\right)\xi_*
$$

with $\mathbf{I}$ the identity matrix. This finally yields our principal result

$$
p(y_*|\mathbf{X},\mathbf{y},\mathbf{x}_*) \approx c_4 \int \mathbf{H}_{\mathbf{y}}(\boldsymbol{\xi})\mathbf{H}_{y_*}(\xi_*)\exp\left(-\frac{1}{2}\tilde{\boldsymbol{\xi}}^T\boldsymbol{\Lambda}\tilde{\boldsymbol{\xi}}\right)d\tilde{\boldsymbol{\xi}} \tag{26}
$$

where

$$
\tilde{\boldsymbol{\xi}} := (\boldsymbol{\xi}, \xi_*) \quad \text{and} \quad \boldsymbol{\Lambda} = \begin{pmatrix} k_0^2\mathbf{I} - k_0^4\mathbf{R}^{-1} & -k_0^2 v\mathbf{R}^{-1}\mathbf{s} \\ -k_0^2 v\mathbf{s}^T\mathbf{R}^{-1} & v - v^2\mathbf{s}^T\mathbf{R}^{-1}\mathbf{s} \end{pmatrix}
$$

The expression says that, to make a prediction under the doubly stochastic model, it suffices to compute the mass of an $(n+1)$-dimensional Gaussian distribution (centered at the origin and with precision matrix $\mathbf{\Lambda}$) in a given orthant. This is illustrated in Fig. 4. The covariance structure of the distribution is mainly given by the covariance matrix $\mathbf{\Sigma}$ and the vector $\sigma(x_*)$, i.e. by the relative position of the training points and the test point x_* in feature space. Moreover, the covariance structure also depends on the parameter k_0 which trades off prior and observed evidence. The orthant that is integrated over is picked by the observed training set labels (and setting $y_* = 0$ or $y_* = 1$). The normalizing constant c_4 can be determined by calculating not only the mass in the relevant but also in the adjacent orthant $\{(\boldsymbol{\xi}, \xi_*) \in \mathbb{R}^{n+1} : \mathbf{H}_{\mathbf{y}}(\boldsymbol{\xi})\mathbf{H}_{1-y_*}(\xi_*) = 1\}$ and then using the sum constraint $p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) + p(y_* = 0|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = 1$. In the left panel of Fig. 4, $\sigma(x_*)$ is relatively large and $\mathbf{y} = 0$. Hence, the posterior class prediction for class 0 is relatively large; here, $p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = 0.128$. In the right panel, $\sigma(x_*) = 0$. Consequently, the label of the training point does not influence the posterior prediction at x_* and hence, $p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = 0.5$.

For the actual computation of the integral of the Gaussian density, one can evaluate the multivariate error function at the origin after having adequately mirrored the normal distribution. The multivariate error function is e.g. implemented in R and Matlab based on methods of Genz and Bretz (2009).

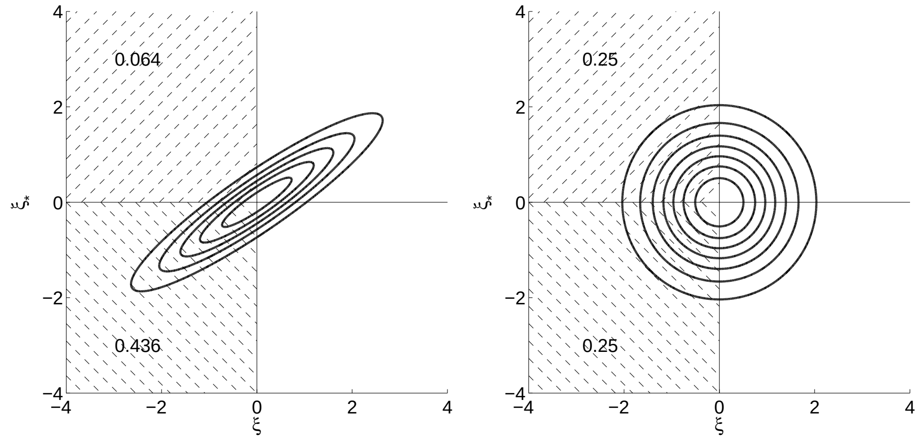

Figure 4: Computation of the posterior class probability $p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$ with the doubly stochastic Gaussian quadrature according to Eq. (26). Each panel shows the contour lines of the probability density function of an $(n+1)$-dimensional Gaussian distribution with 0 mean, where $\xi_* \in \mathbb{R}$ and $\boldsymbol{\xi} \in \mathbb{R}^n$ (obviously, $n = 1$ here). The covariance structure of the distribution mainly reflects the relative positions of the test and training points in feature space. The ratio $p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)/p(y_* = 0|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$ equals the ratio of integrals of the Gaussian density over two adjacent orthants, which are determined by the labels $\mathbf{y}$ of the training points. Here, the regions that are integrated over correspond to $\mathbf{y} = 0$, and $p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)/p(y_* = 0|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = 0.064/0.436$ and $p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)/p(y_* = 0|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = 0.25/0.25$ in the left and the right panel, respectively. Additionally using the sum constraint $p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) + p(y_* = 0|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = 1$ yields $p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = 0.128$ and $p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = 0.5$, respectively.

Remark. For both methods, the AMP and the DSGQ, there is a close relationship between the sill parameter (see e.g. Chilès and Delfiner, 1999, chap. 2.2), which affects the assumed covariance structure and therefore the computation of $\mathbf{\Sigma}$, and the parameter c_0.[1] As already mentioned in Remark 3, they together influence the variance of the prior in Eq. (12) and therefore govern the tradeoff between prior and evidence for the final prediction. The smaller c_0 and the smaller the sill, the higher the weight of the prior. This is particularly evident in Eq. (20).

5 Comparison of the presented algorithms

5.1 Data

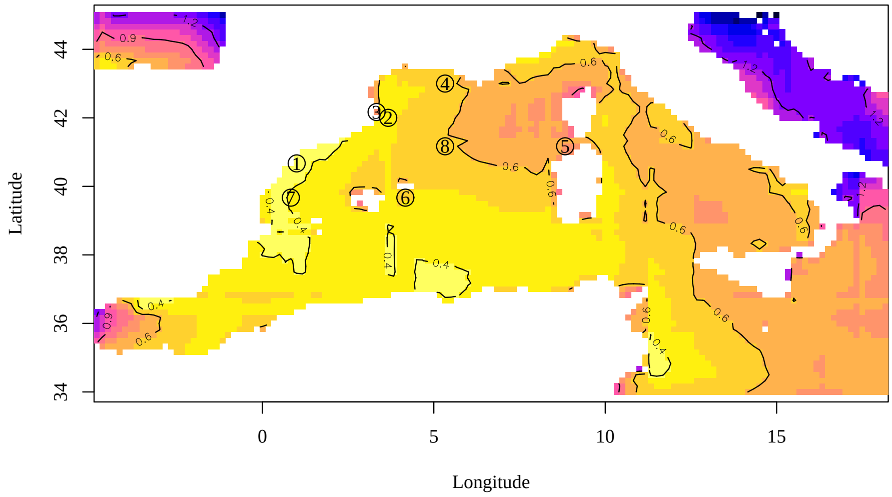

Figure 5: The Western Mediterranean with indication of the 8 explanatory features used to classify the forecast images. The contour map shows the variance along each possible feature, i.e. the variance of the logarithm of the wave heights at each pixel. Pixels are 16 × 16 km^2 approximately.

The several methods summarized or presented in this contribution will be illustrated and compared using a typical diagnostic problem: *given a static "image" of a system, can we decide whether it corresponds to a particular (dynamic) regime?* In this particular case, we want to use a map of significant wave height, provided by a numerical forecasting model of the Western Mediterranean Sea, to decide whether that is a *Llevant* storm (a storm with dominating winds from the East) or not.

We have available a set of $n = 114$ such images of past forecasts, for which we now know the dynamic situation. We manually select beforehand a small subset of $m = 8$ "informative" pixels. Subsampling of pixels is performed to avoid the "curse of dimensionality". Otherwise there would be $n = 114$ points in a space with several thousand dimensions m of which many correspond to uninformative locations in the East. As the empirical distribution of the individual wave heights is extremely skewed to the right, they are preprocessed by computing a logarithm. Then, we build a data set of feature vectors $\mathbf{x}_i \in \mathbb{R}^8, i = 1, \ldots, n$ (the logarithm of the wave heights at the selected pixel positions) and labels y_i (1 corresponds to a Llevant storm, 0 to "no Llevant storm") and apply the classification techniques to this set. Fig. 5 shows the variance

[1] Recall that the corresponding parameter k_0 of the DSGQ simply is proportional to c_0.

of logarithms of significant wave height for the whole forecasting area, using a larger set of 970 non-classified images. We do not consider all these images (but only 114) for the comparison to ensure a high degree of stochastic independence between the images, i.e., we selected the images in such a way that they are at least one week apart from each other. This figure also shows the locations of the 8 pixels chosen in this case as classification features. Note that the chosen features have moderate, fairly similar variances. Though this is not a necessary condition, it allows us to consider an isotropic variogram on $\mathbb{R}^8$ for the latent Gaussian process.

5.2 Experimental results

We compare the four methods – classical indicator kriging (CIK), simplicial indicator kriging (SIK), maximum density of the Aitchison posterior distribution (AMP), and the doubly stochastic Gaussian quadrature (DSGQ) – based on the data presented in the previous section. Throughout this section, we use a Matérn covariance function (see e.g. Rasmussen and Williams, 2006, chap. 4.2) for all methods and all experiments. Its one-dimensional correlogram is given by

$$\rho(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{l} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu} r}{l} \right), \quad \nu, l > 0,$$

where K_ν is the modified Bessel function of the second kind (Abramowitz and Stegun, 1965, chap. 9.6), ν is called a smoothness parameter and l a range parameter. Then, for a given nugget $s_0 > 0$ and a sill $s > s_0$, the covariance function is $h(r) = (s - s_0)\rho(r) + s_0 \mathbf{1}_0(r)$, where $\mathbf{1}_0(\cdot)$ is the indicator function at 0. Hence, $\mathbf{\Sigma}_{ij} = h(\|\mathbf{x}_i - \mathbf{x}_j\|)$ and $\sigma_i = h(\|\mathbf{x}_i - \mathbf{x}_*\|)$.

In the first experiment, we simply evaluate the classification performance of the 8-dimensional data using 5-fold cross-validation (CV): the data is divided into 5 folds; then, 4 of these are used for training to predict the posterior class probabilities for the samples in the remaining fold (test fold). This is repeated 5 times such that each sample is once in the test fold.

For both CIK and SIK the parameters are determined by standard variogram methods (Chilès and Delfiner, 1999) yielding a smoothness of $\nu = 10$, a range of $l = 0.4$, a sill of $s = 0.17$, and no nugget effect ($s_0 = 0$). For AMP and DSGQ, the function values of the underlying process are not observable, because the class labels are modeled as realizations of Bernoulli experiments. Hence, standard variogram methods are not applicable and we use *nested* CV for parameter estimation. In order to predict posterior probabilities for a test fold in the outer CV, only the data in the respective training folds are used for parameter tuning. This is performed in an inner CV loop. Hence, for different test folds of the outer CV, different parameters may be used. Note that, in contrast to a simple (non-nested) CV, this does not yield overoptimistic estimates for classifier performance as the parameters for predicting probabilities for a test fold in the outer CV loop are tuned completely without using any information about this test fold (Varma and Simon, 2006). For computational reasons, we use the same values for ν, s and s_0 as in the other two methods and optimize l and k_0 only (c_0 in the case of the AMP).

The quality indicators of the methods are presented in Table 1. One one hand, the highest and second highest accuracy is achieved by the doubly stochastic methods DSGQ and AMP, respectively. However, note that the differences are not statistically significant. On the other hand, the running time of the DSGQ is much higher, by a factor of 1.4 with respect to the maximum posterior extraction, and more than 500 the time needed for IK techniques.

Next, in order to get more insight into the differences of the methods, we perform an experiment using only two dimensions of the 8-dimensional data. By visual inspection, we select the second and the third feature as these seem to be the most informative features for classification. The 2-dimensional data is plotted in all panels of Fig. 6. We use all samples for training and

Method	Accuracy	Computation time
Classical indicator kriging	0.868 ± 0.062	0.68 s
Simplicial indicator kriging	0.868 ± 0.062	0.88 s
Aitchison maximum posterior	0.877 ± 0.060	346.99 s
Doubly stochastic Gaussian quadrature	0.895 ± 0.056	481.18 s

Table 1: Relative accuracy and computation time of the four different methods for the classification of the 8-dimensional data. Results are obtained by 5-fold cross validation over 114 samples, "±" indicates the boundaries of the 95% interval. As the parameter estimation is performed differently across the methods, it is not considered for the computation time.

predict posterior class probabilities on a two-dimensional grid. We obtain $\nu = 0.5$, $l = 0.4$, $s = 0.17$ and $s_0 = 0$ for CIK and SIK using variogram methods. For AMP and DSGQ, we again use the same values for ν, s and s_0 as in CIK and SIK and optimize the remaining parameters with cross-validation using only the training samples. This leads to $l = 0.99$ and $c_0 = 6.81$ for AMP and $l = 3$ and $k_0 = 4$ for DSGQ. The resulting contour plots for all methods are shown in Fig. 6.

It can be observed that all points of the training set are classified correctly with CIK and SIK, in particular the dissenting points that are located in between a cloud of points with a different label. Here, as the variogram estimate yields $s_0 = 0$, the estimate for the posterior class probabilities at those points are p or $1 - p$ for SIK and even 1 or 0 for CIK. This prevents the assignment of opposite classes in the neighborhood of observed labels and thus limits the generalization ability of CIK and SIK. In contrast, in the doubly stochastic models, the dissenting points are considered unlikely realizations of a Bernoulli experiment. This explains why the squashed realization of the Gaussian process is much smoother for the doubly stochastic methods AMP and DSGQ, as can be inferred from the contour lines. Hence, the AMP and specially the DSGQ methods are more robust with respect to these dissenting points, even under $s_0 = 0$.

6 Conclusions

We have presented two new methods for the estimation of the class probabilities in a classification setting, based on a doubly stochastic process formalism. Seen from a Bayesian perspective, the two methods are obtained as the maximum posterior probability (AMP method) and the predictive probability (DSGQ method) of a prior random field updated by a Bernoulli likelihood obtained from the training set. The posterior happens to be an Aitchison distribution, known in the field of compositional data analysis. The distinctive characteristic of both methods is that the underlying estimation is deterministic and analytical up to a final step of iterative maximization or integration.

The underlying doubly stochastic model is consistent with a classification framework. In contrast, (classical) indicator kriging (CIK) (Journel, 1983) is theoretically inconsistent, as it uses a non-transformed Gaussian random field (with range $-\infty$ to $+\infty$) to describe a probability (bounded between 0 and 1). SIK uses a logistic-transformed Gaussian RF as reference to avoid negative probabilities. However, both CIK and SIK are interpolators, and thus do not reflect a two-step stochastic process. In particular, this becomes apparent in the presence of conflicting observations (successes surrounded by failures, or vice versa): the posterior probabilities estimated by CIK or SIK can only be exactly 0 or 1 at locations where there are observations. These methods hence categorically rule out the possibility to observe the opposite label at those

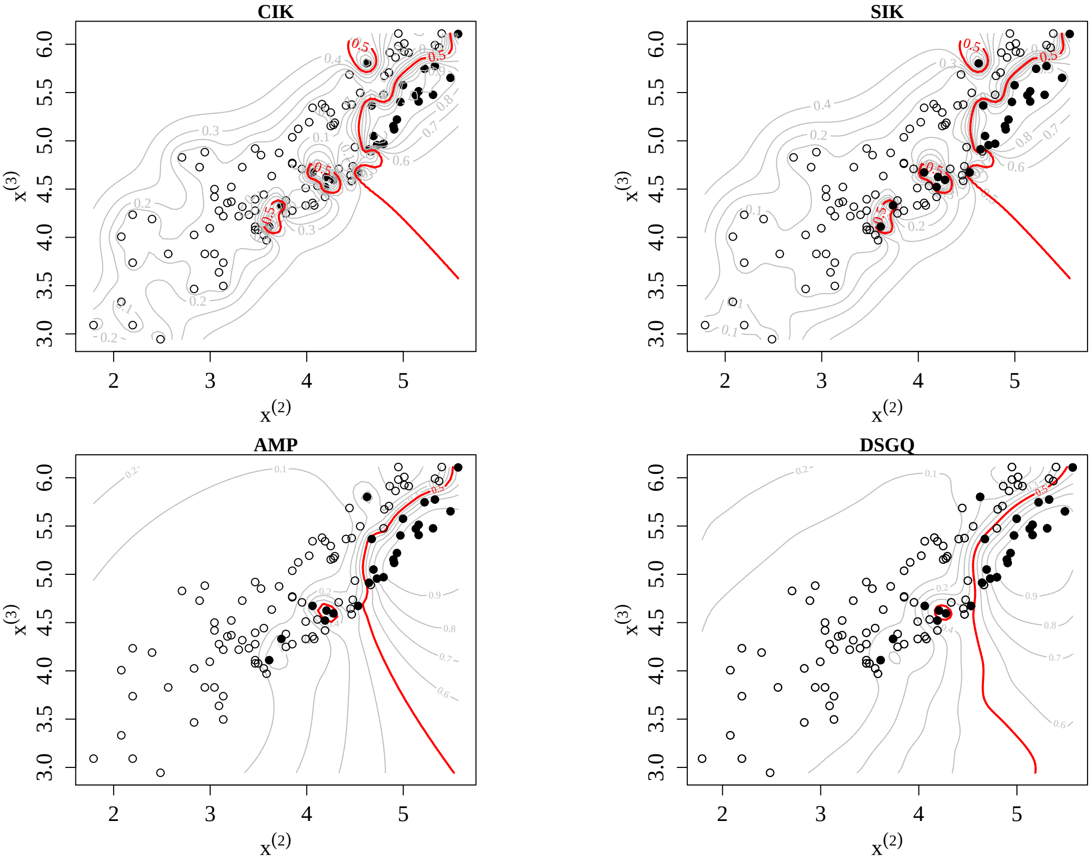

Figure 6: Contour plots of the posterior predictions for the four methods compared, based on a two-dimensional projection of the data on the $x^{(2)}$-$x^{(3)}$-plane, i.e. using these two features only. The decision boundary between the two classes, the level curve for $\{x_* : p(y_* = 0|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = 0.5\}$, is depicted with a thicker line. Samples of class 0 and 1 are represented by empty and filled circles, respectively. Note that the decision boundary (in contrast to the other contour lines) is equal for CIK and SIK. Both CIK and SIK make predictions that are compatible with each and every label from the training set which is prone to overfitting. In contrast, both AMP and DSGQ take dissenting points into account, but do not follow them unconditionally in their predictions.

locations. This is not realistic for typical prediction settings which are characterized by some class overlap. In contrast, both AMP and DSGQ can take observations from the neighborhood into account and produce more plausible predictions at the site of observations.

Although the accuracy of AMP and in particular that of DSGQ is higher than that of CIK and SIK for the 8-dimensional data used here, the difference is not significant. The fact that the doubly stochastic model is computationally more demanding than CIK and SIK without showing convincingly better performance is probably the reason why CIK, the classical approach in geostatistics for classification, still is very popular despite its inconsistency. Moreover, all parameters in the underlying statistical model of CIK can easily be interpreted in physical terms.

While the experiments show that the new computational scheme of the DSGQ works in principle, an alternative to the numerical integration is desirable because it may be too expensive or too inexact if n is large. For this, note that we only need to know the ratio of the probability and the complementary probability of obtaining label y_* at the point $\mathbf{x}_*$ to make a prediction:

$$\frac{c_4 \int_{\mathbb{R}^{n+1}} \mathbf{H}_{\mathbf{y}}(\boldsymbol{\xi}) \mathbf{H}_{y_*}(\xi_*) e^{-\frac{1}{2}\tilde{\boldsymbol{\xi}}^T \Lambda \tilde{\boldsymbol{\xi}}} d\tilde{\boldsymbol{\xi}}}{c_4 \int_{\mathbb{R}^{n+1}} \mathbf{H}_{\mathbf{y}}(\boldsymbol{\xi}) \mathbf{H}_{1-y_*}(\xi_*) e^{-\frac{1}{2}\tilde{\boldsymbol{\xi}}^T \Lambda \tilde{\boldsymbol{\xi}}} d\tilde{\boldsymbol{\xi}}} = \frac{\int_{\Omega_1} e^{-\frac{1}{2}(G\tilde{\boldsymbol{\xi}})^T (G\tilde{\boldsymbol{\xi}})} d\tilde{\boldsymbol{\xi}}}{\int_{\Omega_2} e^{-\frac{1}{2}(G\tilde{\boldsymbol{\xi}})^T (G\tilde{\boldsymbol{\xi}})} d\tilde{\boldsymbol{\xi}}} = \frac{\int_{G(\Omega_1)} e^{-\frac{1}{2}\tilde{\boldsymbol{\xi}}'^T \tilde{\boldsymbol{\xi}}'} d\tilde{\boldsymbol{\xi}}'}{\int_{G(\Omega_2)} e^{-\frac{1}{2}\tilde{\boldsymbol{\xi}}'^T \tilde{\boldsymbol{\xi}}'} d\tilde{\boldsymbol{\xi}}'} \tag{27}$$

where we have defined $\Omega_1 = \{\tilde{\boldsymbol{\xi}} : \mathbf{H}_{\mathbf{y}}(\boldsymbol{\xi})\mathbf{H}_{y_*}(\xi_*) = 1\}$, $\Omega_2 = \{\tilde{\boldsymbol{\xi}} : \mathbf{H}_{\mathbf{y}}(\boldsymbol{\xi})\mathbf{H}_{1-y_*}(\xi_*) = 1\}$ and have used the Cholesky decomposition $\Lambda = G^T G$ and the multidimensional substitution rule for integration. The regions $G(\Omega_i)$, over which we integrate, are convex cones with apices in the origin (because they are linear transformations of orthants) and the integrand $\exp(-\frac{1}{2}\tilde{\boldsymbol{\xi}}'^T \tilde{\boldsymbol{\xi}}')$ is a radially symmetric function. Hence, the value of the whole integral is proportional to the volume of the intersection of the cone with the unit sphere (called a spherical simplex). Thus, in order to evaluate the fraction (27), we need to compute the ratio of the volumes of the spherical simplices determined by $G(\Omega_1)$ and $G(\Omega_2)$ (Aomoto, 1977). Finding a tractable approximation to this ratio is an attractive avenue for future research.

Our final comment addresses the link between SIK and AMP. AMP can be seen as a two-step estimation process. First, one estimates the probabilities of success at the sampled locations maximizing the posterior Aitchison distribution. Second, one interpolates them to the estimation locations using SIK. Thus, SIK, AMP and DSGQ form a ladder of methods of increasing computational complexity paired with an increasingly better fit to the underlying two-step stochastic process hypothesis and reliability of results.

Acknowledgements

We gratefully acknowledge financial support by the Robert Bosch GmbH (J.R., F.A.H.), by the Deutsche Forschungsgemeinschaft (DFG) under grant no. HA4364/2-1 (F.A.H.) as well as by the Spanish Ministry of Science and Innovation through a “Juan de la Cierva” grant JCI-2008-1835 (R.T.), supported by the European Social Fund (ESF-FSE).

References

Abramowitz M, Stegun IA (1965) Handbook of mathematical functions. Dover, New York

Aitchison J (1982) The statistical analysis of compositional data (with discussion). J Royal Stat Soc, Ser B (Stat Methodol) 44(2):139–177

Aomoto K (1977) Analytic structure of the Schläfli function. Nagoya Math J 68:1–16

Billheimer D, Guttorp P, Fagan WF (2001) Statistical interpretation of species composition. J Am Stat Assoc 96:1205–1214

Bogaert P (2002) Spatial prediction of categorical variables: the Bayesian maximum entropy approach. Stoch Environ Res Risk Assess 16:425–448

Carle S, Fogg G (1996) Transition probability-based indicator geostatistics. Math Geol 28(4):453–476

Carr J, Mao N (1993) A general-form of probability kriging for estimation of the indicator and uniform transforms. Math Geol 25(4):425–438

Chilès JP, Delfiner P (1999) Geostatistics: Modeling Spatial Uncertainty. Wiley and Sons, New York

Christakos G (1990) A Bayesian/maximum entropy view to the spatial estimation problem. Math Geol 22(7):763–777

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics (with discussion). J Royal Stat Soc, Ser C (Appl Stat) 47(3):299–350

Genz A, Bretz F (2009) Computation of multivariate normal and t probabilities. Lecture Notes in Statistics, Vol. 195. Springer, Heidelberg

Gibbs MN (1997) Bayesian Gaussian processes for classification and regression. Dissertation, University of Cambridge

Gibbs MN, MacKay DJC (2000) Variational Gaussian process classifiers. IEEE Trans Neural Netw 11(6):1458–1464

Hsu Y, Tung Y, Kuo, J (2010) Evaluation of dam overtopping probability induced by flood and wind. Stoch Environ Res Risk Assess 25(1):35–49

Journel AG (1983) Nonparametric estimation of spatial distributions. Math Geol 15(3):445–468

Journel AG, Posa D (1990). Characteristic behavior and order relations for indicator variograms. Math Geol 22(8): 1011–1025

Kazianka H, Pilz J (2010) Copula-based geostatistical modeling of continuous and discrete data including covariates. Stoch Environ Res Risk Assess 24(5):661–673

Kuss M, Rasmussen CE (2005) Assessing approximate inference for binary Gaussian process classification. J of Mach Learn Res 6:1679–1704

Lim YB, Sacks J, Studdeen W, Welch W (2002) Design and analysis of computer experiments when the output is highly correlated over the input space. Can J Stat 30(1):109–126

Minka, TP (2001) A family of algorithms for approximate Bayesian inference. PhD thesis, Massachusetts Institute of Technology

Neal RM (1999) Regression and classification using Gaussian process priors. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM (ed) Bayesian Statistics 6, Oxford University Press, pp 475–501

Opper M, Winther O (2000) Gaussian processes for classification: mean field algorithms. Neural Comput 12(11):2655–2684

Pardo-Igúzquiza E, Dowd P (2005) Multiple indicator cokriging with application to optimal sampling for environmental monitoring. Comp & Geosci 31(1):1–13

Pawlowsky-Glahn V (2003) Statistical modelling on coordinates. In: Thió-Henestrosa S, Martín-Fernández JA (ed) Compositional data analysis workshop - CoDaWork'03, Universitat de Girona, http://dugi-doc.udg.edu/handle/10256/648

Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. Stoch Environ Res and Risk Assess (SERRA) 15:384–398

Petersen KB, Pedersen MS (2008) The Matrix Cookbook. http://matrixcookbook.com. Accessed: 11 February 2010

Rasmussen CE (1996) Evaluation of Gaussian processes and other methods for non-linear regression. Dissertation, Graduate Department of Computer Science, University of Toronto

Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. MIT Press, Cambridge

Rue H, Martino S, Chopin, N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J Roy Soc B 71:319–392

Suro-Perez V, Journel AG (1991) Indicator principal component kriging. Math Geol 23(5):759–788

Tolosana-Delgado R (2006) Geostatistics for constrained variables: positive data, compositions and probabilities. Application to environmental hazard monitoring. Dissertation, Universitat de Girona (Spain)

Tolosana-Delgado R, Pawlowsky-Glahn V, Egozcue JJ (2008) Indicator kriging without order relation violations. Math Geosci 40:327–347

Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7:91

Wiener N (1949) Extrapolation, interpolation and smoothing of stationary time series with engineering applications. Wiley and Sons, New York

Williams CKI, Barber D (1998) Bayesian classification with Gaussian processes. IEEE Trans on Pattern Anal and Mach Intell 20(12):1342–1351

Williams CKI, Rasmussen CE (1996) Gaussian processes for regression. In: Touretzky DS, Mozer MC, Hasselmo ME (ed) Advances in Neural Information Processing Systems 8, MIT Press, pp 514–520.

Yu H, Yang S, Yen H, Christakos G (2010) A spatio-temporal climate-based model of early dengue fever warning in southern Taiwan. Stoch Environ Res Risk Assess 25(4):485–494