Approximate Inference of the Bandwidth in Multivariate Kernel Density Estimation

Maurizio Filippone[a], Guido Sanguinetti[b]

[a] *University College London - Department of Statistical Science, 1-19 Torrington Place, London, WC1E 7HB.*

[b] *University of Edinburgh - School of Informatics, Informatics Forum 10, Crichton Street, Edinburgh, EH8 9AB.*

Abstract

Kernel density estimation is a popular and widely used non-parametric method for data-driven density estimation. Its appeal lies in its simplicity and ease of implementation, as well as its strong asymptotic results regarding its convergence to the true data distribution. However, a major difficulty is the setting of the bandwidth, particularly in high dimensions and with limited amount of data. An approximate Bayesian method is proposed, based on the Expectation-Propagation algorithm with a likelihood obtained from a leave-one-out cross validation approach. The proposed method yields an iterative procedure to approximate the posterior distribution of the inverse bandwidth. The approximate posterior can be used to estimate the model evidence for selecting the structure of the bandwidth and approach online learning. Extensive experimental validation shows that the proposed method is competitive in terms of performance with state-of-the-art plug-in methods.

Keywords: kernel density estimation, Bayesian inference, expectation propagation, multivariate analysis

1. Introduction

The estimation of a probability density function from a data set is an important statistical problem as it can provide useful insights on the properties of the system from which data are observed, and can form the basis

Email addresses: maurizio@stats.ucl.ac.uk (Maurizio Filippone), gsanguin@inf.ed.ac.uk (Guido Sanguinetti)

for supervised and unsupervised learning steps. Non-parametric approaches make no assumption about the form of the density to be estimated, and the aim is to obtain an empirical estimate from the data which can provably converge (in a suitable sense) to the true data generating density in the limit of infinite sample size. A popular and widely used technique for non-parametric density estimation is Kernel Density Estimation (KDE) (Silverman, 1986). In KDE a base (also known as kernel) function $K(\mathbf{x}|\lambda)$ parametrized by a shape parameter λ is selected, and the density is estimated as a superposition of kernel functions centered at the data points:

$$\hat{p}(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n} K(\mathbf{x} - \mathbf{x}_j|\lambda), \tag{1}$$

where $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is a set of observed data points in $\mathbb{R}^d$. The kernel function is a proper density function, thus guaranteeing that the estimate is an acceptable probability density. A common choice for the kernel function is the Gaussian kernel,

$$K(\mathbf{x}|\lambda) = \left(\frac{\lambda}{2\pi}\right)^{d/2} \exp\left(-\frac{\lambda}{2}\|\mathbf{x}\|^2\right).$$

In this case, the parameter λ is the precision of the Gaussian kernel and is the inverse of the square of the so called *bandwidth* σ, namely $\lambda = 1/\sigma^2$. It can be shown that the KDE estimate of a probability density function converges, in terms of integrated squared error, to the true density in the limit of infinite data, while simultaneously shrinking the bandwidth (Silverman, 1986).

In practical situations, when the amount of data is finite, the bandwidth is a parameter that needs to be properly tuned to guarantee a good performance (Jones, 1991; Hazelton and Turlach, 2007). Several approaches have been proposed to estimate it (Cao et al., 1994; Jones et al., 1996; Silverman, 1986). The quality of a bandwidth selector is usually evaluated on the basis of the Integrated Square Error (ISE)

$$\text{ISE} = \int_{\mathbb{R}^d} \left[p(\mathbf{x}) - \hat{p}(\mathbf{x})\right]^2 d\mathbf{x} \tag{2}$$

or its expectation (MISE). The direct optimization of any of these functions, or an asymptotic expansion of the MISE, requires some knowledge about the true density, that is usually unknown. Many bandwidth selectors are therefore designed to optimize them by plugging in some estimates of the missing characteristics of the true density (Loader, 1999; Sheather and Jones, 1991; Terrell, 1990; Duong and Hazelton, 2005; Chiu, 1991; Hall, 1982; Jones and Henderson, 2009). Some probabilistic approaches based on a maximum likelihood (ML) approach have been proposed, where the probability of the data itself under the KDE estimate (1) is used as the likelihood. To deal with the unbounded likelihood when σ approaches zero, cross-validated or penalized forms of the likelihood have been considered (Silverman, 1986; Zhang et al., 2006). In Zhang et al. (2006, 2009), the cross-validated likelihood is used along with a prior on the bandwidth to sample from its posterior distribution using Markov chain Monte Carlo (MCMC) methods. It is worth noting that this approach allows to deal with multivariate data in the same way it does with univariate data, as it is a likelihood-based approach. Other Bayesian approaches can be found in Brewer (2000), where an approach to Bayesian local smoothing for univariate data is considered; in particular, the bandwidth is locally adjusted to reflect the local density of the data. Also, Gangopadhyay and Cheung (2002) and de Lima and Atuncar (2011) provide closed form expressions for posterior densities over the bandwidth for univariate and multivariate data respectively, but the likelihood used in these approaches is not the probability of the observed data given the bandwidth. Note that in the particular case of censored data the inference of the bandwidth can be done in closed form (Kulasekera and Padgett, 2006). Also, recent advances in multivariate KDE can be found in Tran et al. (2006); Duong et al. (2008).

In this paper, we propose an approximate Bayesian approach to the estimation of the bandwidth for KDE, where the kernel function is the Gaussian density. We place a prior distribution on the kernel precision (inverse bandwidth), and obtain a joint likelihood by combining the prior with an approximate likelihood obtained from leave-one-out cross-validation (van der Laan et al., 2004; Zhang et al., 2006). This yields a posterior distribution over the inverse bandwidth up to a proportionality constant. We propose a method to approximate the posterior that directly interprets the joint likelihood as a *cavity distribution*, leading to a simple application of the Expectation-Propagation (EP) algorithm (Minka, 2001; Opper and Winther, 2000).

The motivation for a Bayesian approach is that it has many useful properties. In particular, obtaining a posterior distribution over the bandwidth allows to quantify the uncertainty on the estimate of the bandwidth, thus allowing to consider different scenarios when analyzing data. If KDE is employed as the first step of a supervised/unsupervised learning problem, such as, e.g., Filippone et al. (2010); Rinaldo and Wasserman (2010), then a distribution of decision functions can be obtained. Also, online learning can be approached quite naturally, as the posterior can be used as a prior when new data are observed. Finally, this provides a principled way to select the covariance structure for the bandwidth, namely full, diagonal, or isotropic, as the model evidence can be used to estimate Bayes factors (Kass and Raftery, 1995).

Although these are main reasons why one would consider a Bayesian approach, most of these aspects have been overlooked in previous studies on KDE (Zhang et al., 2006, 2009). One of the contributions of this paper is to study, through extensive experiments, how the Bayesian approach can be used to select the structure of the bandwidth in multivariate KDE and perform online learning. The main contribution is to present an approximate method to infer the bandwidth. In general, it can be argued that approximate methods could lead to very poor approximations of the posterior distribution, especially in cases of multimodalities or inadequate assumptions on the approximating distribution. MCMC methods would therefore seem to be more flexible. However, especially in multivariate cases, the application of MCMC methods can be rather challenging. Designing efficient MCMC methods for sampling multivariate variables, ensuring good independence between samples and with a reasonable computational cost requires some fine tuning. For example, in the Metropolis-Hastings method (Metropolis et al., 1953; Hastings, 1970) it is required to specify the parameters of a proposal distribution, or in Hamiltonian-based Monte Carlo methods (Neal, 1993) the mass matrix. Also, assessing the convergence of MCMC methods or deal with multimodal posterior densities usually involves the running of multiple chains, as discussed, e.g., in Gelman and Rubin (1992); Kass et al. (1998); Calderhead and Girolami (2009). Clearly, these considerations imply a computational overhead that has to be accounted for in the design of an inference method. In the inference of the bandwidth in KDE, we found that the proposed approximate method yields a posterior density that is very close to the true one. Given the speed of the deterministic approximation scheme compared to MCMC methods, we propose it as an efficient alternative to infer the bandwidth in multivariate KDE. The advantages of the deterministic approximation through EP are that convergence can be monitored during the optimization and that the model evidence is readily available at the end of the algorithm. MCMC estimates of the model evidence, instead, could be rather challenging as discussed e.g., in Friel and Pettitt (2008) and Skilling

(2006).

The paper is organized as follows. In Section 2 we introduce the ideas underlying the Bayesian approach to multivariate KDE with a full precision matrix, and in Section 3 we present the approximation of the posterior based on EP. In Section 4 we illustrate how to extend our method to online learning scenarios, while in Section 5 we give some details of implementation along with a discussion on the use of simpler structure for the precision matrix. Section 6 reports an extensive experimental comparison showing the performance of the proposed method, and Section 7 concludes the paper. In Appendix A, we detail the derivation of the special cases of diagonal and isotropic precision matrices.

2. Bayesian KDE

We consider multivariate Gaussian kernels

$$K(\mathbf{x}|\Lambda) = \frac{|\Lambda|^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\Lambda\mathbf{x}\right),$$

where Λ is the precision matrix (inverse bandwidth) capturing the correlations among dimensions. The estimated pdf in $\mathbf{x} \in \mathbb{R}^d$ is therefore

$$\hat{p}(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n} K(\mathbf{x} - \mathbf{x}_j|\Lambda) = \frac{1}{n}\sum_{j=1}^{n} \mathcal{N}\left(\mathbf{x}|\mathbf{x}_j, \Lambda\right), \tag{3}$$

where $\mathcal{N}\left(\mathbf{x}|\mathbf{x}_j, \Lambda\right)$ denotes a normal distribution with mean $\mathbf{x}_j$ and precision matrix Λ. Assuming i.i.d. data, one can define the log-likelihood:

$$\mathcal{L} = \log[p(X|\Lambda, \mathcal{M}_X)] = \sum_{i=1}^{n} \log\left[\sum_{j=1}^{n} \mathcal{N}\left(\mathbf{x}_i|\mathbf{x}_j, \Lambda\right)\right] + \text{const.},$$

where $\mathcal{M}_X$ denotes the fact that the model is based on placing kernels at the data points in X. It can be easily seen, however, that $\mathcal{L}$ is unbounded in Λ, achieving infinite value when the determinant of Λ goes to infinity.

In order to overcome this limitation, while retaining a ML approach, it has been proposed to maximize a cross-validated type of likelihood (Silverman, 1986)

$$\mathcal{L}_{\text{cv}} = \log[p(X|\Lambda, \mathcal{M}_X)] = \sum_{i=1}^{n} \log\left[\sum_{j\neq i, j=1}^{n} \mathcal{N}(\mathbf{x}_i|\mathbf{x}_j, \Lambda)\right] + \text{const.}. \tag{4}$$

Provided $\mathbf{x}_j \neq \mathbf{x}_i \, \forall i, j$, $\mathcal{L}_{\text{cv}}$ is a well behaved function of Λ and will be used in the remainder of this paper as the likelihood in the inference problem. Note that the definition in (4) does not define a proper joint probability distribution over X, as it is a mixture of degenerate distributions. In the case of ties instead, namely when $\mathbf{x}_j = \mathbf{x}_i$ for some $i \neq j$, it is possible to use ideas from Żychaluk and Patil (2008), where theoretical studies are reported on the asymptotics of estimates of the bandwidth based on the cross-validation method for minimizing the ISE (Hall, 1982).

To cast the problem in a Bayesian framework, we need to place a prior over the precision matrix Λ. To exploit conjugacy with the Gaussian kernel, we will use a Wishart prior

$$\begin{aligned} p\,(\Lambda) =& \mathcal{W}(\Lambda|\Lambda_0, \nu_0) \\ \propto& |\Lambda|^{\frac{\nu_0}{2} - \frac{d+1}{2}} \exp\left(-\frac{1}{2}\text{Tr}(\Lambda_0^{-1}\Lambda)\right). \end{aligned} \tag{5}$$

In absence of prior knowledge about the bandwidth to use, we can use an uninformative improper Wishart prior $p\,(\Lambda) = |\Lambda|^{-\frac{d+1}{2}}$ or a reasonably flat Wishart prior, for instance, by setting $\nu_0 = d + 1$ and $\Lambda_0 = \beta I$ with large β. The posterior is then

$$p(\Lambda|X, \mathcal{M}_X) \propto p(X|\Lambda, \mathcal{M}_X)p(\Lambda),$$

and the so called *model evidence* is

$$p(X|\mathcal{M}_X) = \int p(X|\Lambda, \mathcal{M}_X)p(\Lambda)d\Lambda.$$

The model evidence gives the probability of the observed data given the model assumption only, as the parameters are integrated out, and can therefore be used for model selection. Given the complex analytical form of the likelihood, exact inference of Λ is intractable. In the next section we will show an approximation method based on EP.

The approximate inference process leads to a deterministic approximation of the posterior distribution over Λ and therefore allows to estimate the model evidence. Model comparison can be done using the so called *Bayes factor* (Kass and Raftery, 1995) that is computed as the ratio between the evidence for the two models to compare. In Bayesian KDE, we can use this idea to compare different multivariate structures of the bandwidth, thus having a principled way of choosing whether simpler structures can be used in place of more complicated ones.

3. Expectation Propagation Estimation of the Bandwidth

3.1. Expectation Propagation

We briefly review here the EP algorithm (Minka, 2001; Opper and Winther, 2000); for more details see e.g. Bishop (2007). Consider a data modeling problem, where the observed data is denoted by X and the set of model parameters is $\theta \in \Theta$. EP aims at finding an approximation $q(\theta)$ to the posterior $p(\theta|X)$ by minimizing of the Kullback-Leibler divergence between the posterior and the approximating distribution:

$$\mathrm{KL}[p(\theta|X)||q(\theta)] = \int_{\Theta} \log\left(\frac{p(\theta|X)}{q(\theta)}\right) p(\theta|X) d\theta.$$

It can be easily shown that this is equivalent to matching the moments between the two distributions.

Generally, the direct computation of the moments of the posterior is intractable, and an iterative procedure for matching the moments is necessary. In the case of i.i.d. data, the posterior distribution over the parameters factorizes as

$$p(\theta|X) = \frac{1}{Z}\prod_{i=0}^{n} f_i(\theta),$$

where $f_0(\theta)$ is the prior and $f_i(\theta)$ is the likelihood of the i-th observation. Let the approximating distribution $q(\theta)$ factorize as a product of tractable distributions:

$$q(\theta) = \frac{1}{\tilde{Z}}\prod_{i=0}^{n} \tilde{f}_i(\theta).$$

The goal is to optimize

$$\mathrm{KL}\left[\frac{1}{Z}\prod_{i=0}^{n} f_i(\theta)\,\middle\|\,\frac{1}{\tilde{Z}}\prod_{i=0}^{n} \tilde{f}_i(\theta)\right]$$

with respect to the approximating distributions $\tilde{f}_i$. In order to do that, we start from the so called *cavity distribution*, where we remove the j-th factor from $q(\theta)$

$$q^{\backslash j}(\theta) \propto \prod_{i \neq j, i=0}^{n} \tilde{f}_i(\theta),$$

and match the moments of the normalized version of $f_j(\theta)q^{\backslash j}(\theta)$ and $q_{\text{new}}(\theta)$. Once we have a revised version of $q_{\text{new}}(\theta)$, we can obtain the updated version of $\tilde{f}_j(\theta)$ as

$$\tilde{f}_j'(\theta) \propto \frac{q_{\text{new}}(\theta)}{q^{\backslash j}(\theta)}.$$

Starting from an initialization of the factors $\tilde{f}_i(\theta)$, we can update them in turn until convergence (Minka, 2001). In many occasions, EP has been used to approximate the posterior using a Gaussian distribution (Opper and Winther, 2000). In our problem, since the bandwidth is constrained to be positive definite, such an approximation would not be acceptable. Therefore, we will model the factors as Wishart distributions.

Matching the moments of Wishart distributions would require matching the normalization constant, the mean, and the expectation of $\log|\Lambda|$. The resulting equation for $\log|\Lambda|$, however, cannot be solved analytically and requires an iterative optimization method. For simplicity, we will therefore pursue an approximation of the moment of second order (variance) based on the squared Frobenius norm

$$\|\Lambda\|_{\text{F}}^2 = \text{Tr}(\Lambda\Lambda^{\text{T}}) = \sum_{ij}(\Lambda_{ij})^2.$$

3.2. EP approximation of the posterior of the bandwidth of the kernel

In this section, we present the application of the EP framework for the approximation of the posterior distribution over the precision. In order to keep the notation uncluttered, we define the unnormalized Wishart distribution as

$$\tilde{\mathcal{W}}(\Lambda|\Lambda_i, \nu_i) = |\Lambda|^{\frac{\nu_i}{2}-\frac{d+1}{2}} \exp\left[-\frac{1}{2}\text{Tr}(\Lambda_i^{-1}\Lambda)\right].$$

The normalized Wishart distribution, instead, is defined as

$$\mathcal{W}(\Lambda|\Lambda_i, \nu_i) = B_{\mathcal{W}}(\Lambda_i, \nu_i)\tilde{\mathcal{W}}(\Lambda|\Lambda_i, \nu_i),$$

with

$$B_{\mathcal{W}}(\Lambda_i, \nu_i)^{-1} = 2^{\frac{\nu_i d}{2}}|\Lambda_i|^{\frac{\nu_i}{2}}\pi^{\frac{d(d-1)}{4}}\prod_{i=1}^{d}\Gamma\left(\frac{\nu_i+1-i}{2}\right).$$

The true posterior is

$$p(\Lambda|X, \mathcal{M}_X) \propto \prod_{i=0}^{n} f_i(\Lambda),$$

where $f_0(\Lambda) = \mathcal{W}(\Lambda|\Lambda_0, \nu_0)$ denotes the prior, and the components $f_j(\Lambda)$ read:

$$f_j(\Lambda) = p(\mathbf{x}_j|\Lambda, \mathcal{M}_X) = \frac{1}{n-1} \sum_{r\neq j, r=1}^{n} \mathcal{N}(\mathbf{x}_j|\mathbf{x}_r, \Lambda).$$

We propose to approximate the posterior over Λ as

$$q(\Lambda) = \prod_{i=0}^{n} \tilde{f}_i(\Lambda).$$

We consider an approximating form in which the factors are proportional to an unnormalized Wishart distribution

$$\tilde{f}_i(\Lambda) = \gamma_i \tilde{\mathcal{W}}(\Lambda|\Lambda_i, \nu_i),$$

with $\tilde{f}_0(\Lambda) = f_0(\Lambda)$. With this choice, it can be easily verified that

$$q(\Lambda) = \left(\prod_{i=0}^{n} \gamma_i\right) \tilde{\mathcal{W}}\left(\Lambda \,\middle|\, (\sum_{i=0}^{n} \Lambda_i^{-1})^{-1}, \sum_{i=0}^{n} \nu_i - (d+1)n\right).$$

The cavity distribution, when we remove the j-th component from $q(\Lambda)$ is

$$\begin{aligned} q^{\setminus j}(\Lambda) &= \prod_{i\neq j, i=0}^{n} \tilde{f}_i(\Lambda) \\ &= \left(\prod_{i\neq j, i=0}^{n} \gamma_i\right) \tilde{\mathcal{W}}\left(\Lambda \,\middle|\, (\sum_{i\neq j, i=0}^{n} \Lambda_i^{-1})^{-1}, \sum_{i\neq j, i=0}^{n} \nu_i - (d+1)(n-1)\right). \end{aligned} \tag{6}$$

Let's define the following quantities:

$$(U_r)^{-1} = \sum_{i\neq j, i=0}^{n} \Lambda_i^{-1} + (\mathbf{x}_r - \mathbf{x}_j)(\mathbf{x}_r - \mathbf{x}_j)^{\mathrm{T}}, \tag{7}$$

$$\alpha = \sum_{i\neq j, i=0}^{n} \nu_i - (d+1)(n-1) + 1, \tag{8}$$

$$\psi = \left(\prod_{i\neq j, i=0}^{n} \gamma_i\right) \frac{1}{(n-1)(2\pi)^{d/2}}, \tag{9}$$

$$z_r = B_{\mathcal{W}}(U_r, \alpha)^{-1}. \tag{10}$$

With the above definitions, we see that

$$f_j(\Lambda) q^{\backslash j}(\Lambda) = \psi \sum_{r \neq j, r=1}^{n} z_r \mathcal{W}(\Lambda | U_r, \alpha). \tag{11}$$

Now we have to match the moments of

$$q(\Lambda | \Lambda_{\text{new}}, \nu_{\text{new}}) = \gamma_{\text{new}} B_{\mathcal{W}}(\Lambda_{\text{new}}, \nu_{\text{new}})^{-1} \mathcal{W}(\Lambda | \Lambda_{\text{new}}, \nu_{\text{new}})$$

with those of $f_j(\Lambda) q^{\backslash j}(\Lambda)$. The normalization constant of (11) is

$$E_0 = \int f_j(\Lambda) q^{\backslash j}(\Lambda) d\Lambda = \psi \sum_{r \neq j, r=1}^{n} z_r. \tag{12}$$

The mean and the variance of $f_j(\Lambda) q^{\backslash j}(\Lambda)$ can be computed noticing that the resulting density is a weighted sum of Wishart

$$f_j(\Lambda) q^{\backslash j}(\Lambda) \propto \sum_{r \neq j, r=1}^{n} w_r \tilde{\mathcal{W}}(\Lambda | U_r, \alpha),$$

where the weights are

$$w_r = \frac{z_r}{\sum_{s \neq r, s=1}^{n} z_s}.$$

Therefore, the expectation results in

$$E_1 = \mathrm{E}_{f_j(\Lambda) q^{\backslash j}(\Lambda)}[\Lambda] = \alpha \sum_{r \neq j, r=1}^{n} w_r U_r. \tag{13}$$

The variance of Λ can be easily computed, given that

$$\mathrm{E}_{f_j(\Lambda) q^{\backslash j}(\Lambda)}[\Lambda_{lk}^2] = \sum_{r \neq j, r=1}^{n} w_r \left\{ (\alpha + \alpha^2)(U_r)_{lk}^2 + \alpha (U_r)_{ll} (U_r)_{kk} \right\}.$$

Therefore,

$$E_2 = \mathrm{var}(\Lambda) = \sum_{r \neq j, r=1}^{n} w_r \left\{ (\alpha + \alpha^2) \|U_r\|_{\mathrm{F}}^2 + \alpha \left[\mathrm{Tr}(U_r) \right]^2 \right\} - \|E_1\|_{\mathrm{F}}^2. \tag{14}$$

Matching the moments results in:

$$\gamma_{\text{new}} B_{\mathcal{W}}(\Lambda_{\text{new}}, \nu_{\text{new}})^{-1} = E_0,$$

$$\nu_{\text{new}} \Lambda_{\text{new}} = E_1,$$

$$\nu_{\text{new}} \|\Lambda_{\text{new}}\|_{\text{F}}^2 + \nu_{\text{new}} \left[\text{Tr}(\Lambda_{\text{new}})\right]^2 = E_2,$$

that can be solved with respect to ν_{new}, Λ_{new}, and γ_{new}, obtaining:

$$\nu_{\text{new}} = \frac{\|E_1\|_{\text{F}}^2 + \left[\text{Tr}(E_1)\right]^2}{E_2}, \tag{15}$$

$$\Lambda_{\text{new}} = \frac{E_1}{\nu_{\text{new}}}, \tag{16}$$

$$\gamma_{\text{new}} = E_0 B_{\mathcal{W}}(\Lambda_{\text{new}}, \nu_{\text{new}}). \tag{17}$$

The updated version of $\tilde{f}_j(\Lambda)$ is the ratio between $q(\Lambda|\Lambda_{\text{new}}, \nu_{\text{new}})$ and $q^{\setminus j}(\Lambda)$, resulting in

$$\tilde{f}_j'(\Lambda) = \gamma_j' \tilde{\mathcal{W}}(\Lambda|\Lambda_j', \nu_j'),$$

where

$$\gamma_j' = \frac{\gamma_{\text{new}}}{\prod_{i\neq j, i=0}^{n} \gamma_i}, \tag{18}$$

$$\left(\Lambda_j'\right)^{-1} = \Lambda_{\text{new}}^{-1} - (\sum_{i\neq j, i=0}^{n} \Lambda_i^{-1})^{-1}, \tag{19}$$

$$\nu_j' = \nu_{\text{new}} - \sum_{i\neq j, i=0}^{n} \nu_i + n(d+1). \tag{20}$$

When the moment matching procedure satisfies a predefined convergence criterion, the resulting approximating posterior is

$$q(\Lambda) = \gamma\, \tilde{\mathcal{W}}(\Lambda|\Lambda, \nu),$$

that allows to obtain the approximate model evidence as

$$\text{evidence} = \gamma\, B_{\mathcal{W}}(\Lambda, \nu)^{-1}.$$

4. Online Bayesian KDE

Online learning arises quite naturally in Bayesian inference. In our case, let's assume that the posterior over the precision, after having observed a data set $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, has been approximated using EP. This means that we have an approximation to $p(\Lambda|X, \mathcal{M}_X)$ based on a model $\mathcal{M}_X$ where X is the set of centers of the kernels. When we observe new data, and we want to include them in the inference of the bandwidth, we need to extend the model to account for the new centers of the kernels. In particular, the final goal will be to approximate $p(\Lambda|X, Y, \mathcal{M}_{X\cup Y})$ starting from $p(\Lambda|X, \mathcal{M}_X)$.

Let's analyze the following expression of the posterior:

$$p(\Lambda|X, Y, \mathcal{M}_{X\cup Y}) \propto p(Y|\Lambda, \mathcal{M}_{X\cup Y}) p(X|\Lambda, \mathcal{M}_{X\cup Y}) p(\Lambda).$$

We immediately see that this expression is based on a different model with respect to

$$p(\Lambda|X, \mathcal{M}_X) \propto p(X|\Lambda, \mathcal{M}_X) p(\Lambda).$$

Therefore, the posterior $p(\Lambda|X, \mathcal{M}_X)$ cannot be used directly as a prior for the second stage of inference. This problem, however, can be easily solved in the following way. First we need to extend the model from $\mathcal{M}_X$ to $\mathcal{M}_{X\cup Y}$. This can be done by approximating again $p(\Lambda|\mathcal{M}_{X\cup Y})$ as a product of $n+1$ factors

$$p(X|\Lambda, \mathcal{M}_{X\cup Y}) p(\Lambda) \simeq \prod_{i=0}^{n} \tilde{g}_i(\Lambda),$$

where each factor is approximated by

$$\tilde{g}_j(\Lambda) = \frac{1}{n+m-1} \left((n-1)\tilde{f}_j(\Lambda) + \sum_{r=1}^{n} \mathcal{N}(\mathbf{x}_j|\mathbf{y}_r, \Lambda) \right),$$

and $\tilde{g}_0(\Lambda)$ is the prior. This approximation does not require operations among vectors in X, that were used to obtain the factors $\tilde{f}_j(\Lambda)$. In practice, the first approximation stage yields

$$(n-1)\tilde{f}_j(\Lambda) \simeq \sum_{r\neq j, r=1}^{n} \mathcal{N}(\mathbf{x}_j|\mathbf{x}_r, \Lambda).$$

At the end of this stage, we have $p(X|\Lambda, \mathcal{M}_{X\cup Y})p(\Lambda)$ as an unnormalized Wishart. Now we just need to include $p(Y|\Lambda, \mathcal{M}_{X\cup Y})$ to obtain the updated posterior. Given that

$$p(Y|\Lambda, \mathcal{M}_{X\cup Y}) = \frac{1}{n+m-1}\left(\sum_{r=1}^{n} \mathcal{N}(\mathbf{y}_j|\mathbf{x}_r, \Lambda) + \sum_{r\neq j, r=1}^{n} \mathcal{N}(\mathbf{y}_j|\mathbf{y}_r, \Lambda)\right),$$

again we do not need to perform operations among vectors in X. Now the approximation to $p(\Lambda|\mathcal{M}_{X\cup Y})$ will be a product of $n+m+1$ factors:

$$q(\Lambda|\mathcal{M}_{X\cup Y}) = \prod_{i=0}^{n+m} \tilde{f}_i(\Lambda) = \prod_{i=0}^{n} \tilde{g}_i(\Lambda) \prod_{i=1}^{m} \tilde{h}_i(\Lambda).$$

The first product is the Wishart obtained when we extended the model, so the final posterior is a product of $m+1$ factors. Again, we can employ an EP scheme to iteratively update the factors, obtaining an approximation to the posterior $p(\Lambda|X, Y, \mathcal{M}_{X\cup Y})$.

5. Simpler structures of precision matrix and implementation details

In this section we provide more details on how to implement the proposed method and we discuss how to reduce its complexity by considering simpler structures of precision matrix, namely diagonal and isotropic. The detailed derivation of these two special cases can be found in the appendix.

5.1. Implementation Notes

The presented method involves several matrix inversions. We can avoid them by working with the covariances Σ_i rather than the precisions Λ_i. In particular the computation of the matrices U_r can be now simplified by using the Sherman-Morrison formula:

$$(U_r)^{-1} = \sum_{i\neq j, i=0}^{n} \Sigma_i + (\mathbf{x}_r - \mathbf{x}_j)(\mathbf{x}_r - \mathbf{x}_j)^{\mathrm{T}}.$$

Defining

$$Q = \left(\sum_{i\neq j, i=0}^{n} \Sigma_i\right)^{-1}, \qquad \mathbf{a} = \mathbf{x}_r - \mathbf{x}_j,$$

we obtain

$$U_r = Q - \frac{Q\mathbf{a}\mathbf{a}^{\mathrm{T}}Q}{1 + \mathbf{a}^{\mathrm{T}}Q\mathbf{a}}. \tag{21}$$

We report the pseudo-code of the proposed algorithm in Tab. 2; note that we actually work with the logarithm of the normalization constants of the Wishart distributions and the evidence by using standard log-sum and log-diff operations in order to avoid numerical underflows.

5.2. Reduced forms for precision matrix and complexity analysis

The inference of the full precision matrix can be computationally intensive. The reason is that each main iteration involves a loop of n iterations (for the refinement of the n factors) each of which has time complexity in $O(nd^3)$. This complexity is dominated by the the computation of the normalization constants of the Wishart distributions with scale matrices U_r that involves the computation of $|U_r|$ that has time complexity in $O(d^3)$. We notice that the complexity does not change even if we precompute and store the vectors $\mathbf{a}$ in (21). In terms of spatial complexity, the dominant part is related to storing the matrices Λ_i and U_r that requires $O(nd^2)$.

As we have discussed in Section 4, the online version where m new data point have to be included in the inference of the precision, does not require operations among the n data points used in the first part of the inference stage. We recall that the online version comprises two steps. In the first step we need to extend the model to include the new m data points; one main iteration of this step, involves a loop over n factors and requires the computation of m determinants $|U_r|$. Therefore, a whole iteration of the first step has time complexity $O(nmd^3)$. One main iteration of the second step, contains a loop of m factors only (as the other n are merged in one unnormalized Wishart), and again will be dominated by the computation of the determinants of $|U_r|$, yielding a time complexity in $O(m^2d^3)$. For the space complexity, we can repeat the same considerations as in the offline case, accounting for the fact that we need to store the updated version of the n matrices Λ_i computed in the first stage, leading to $O((n+m)d^2)$.

Due to the cubic scaling of the time complexity with respect to d, it might be useful to consider simpler forms of precision matrices, such as diagonal or isotropic. In particular, their respective forms are $\Lambda = I\boldsymbol{\lambda}$ with $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$, and $\Lambda = \lambda I$. In these cases, we place a Gamma prior over the precisions for each component λ_k for the diagonal precision, or a Gamma prior over the precision λ.

In the diagonal case, the time complexity of one iteration of the main loop is linear in the number of dimensions d, and is quadratic in n, leading to $O(n^2d)$. The size of the largest objects is in $O(nd)$, as they are $n \times d$

Table 1: Time and Space complexity for one iteration (main loop) of the three versions of the approximate Bayesian KDE.

Covariance	Time complexity	Space complexity
Full	$O(n^2d^3)$	$O(nd^2)$
Full online	$O((n+m)md^3)$	$O((n+m)d^2)$
Diagonal	$O(n^2d)$	$O(nd)$
Diagonal online	$O((n+m)md)$	$O((n+m)d)$
Isotropic	$O(n^2d)$	$O(n)$
Isotropic online	$O((n+m)md)$	$O(n+m)$

matrices. With the same considerations as in the full case, the online version has time complexity in $O((n+m)md)$ and space complexity in $O((n+m)d)$.

In the isotropic case, in each iteration of the main loop we have to compute the distances between a data point and all the other n; this has time complexity in $O(nd)$. Therefore, one of the main iterations will have a time complexity in $O(n^2d)$. The space complexity will be in $O(n)$, as we will only need to allocate the vectors of size n. It is interesting to notice that precomputing the distances (time complexity in $O(n^2d)$, space complexity in $O(n^2)$) will reduce the computational complexity in each main iteration to $O(n^2)$. Time and space complexities for the online versions follow straightforwardly as $O((n+m)md)$ and $O(n+m)$ respectively.

The complexity of one iteration for the three versions of the algorithm are summarized in Tab. 1. After extensive trials and evaluations, we noticed that a number of iterations in the order of tenths or sometimes less are needed for the proposed method to converge.

5.3. Initialization and Convergence

Denoting the parameters of the initializing posterior over Λ with γ_{post}, Λ_{post}, and ν_{post}, in the case when we want to assign the same initial value to the Λ_i as well as to the ν_i, we get

$$\sum_{i=0}^{n} \Lambda_i^{-1} = (\Lambda_{\text{post}})^{-1} \Rightarrow \Lambda_i^{-1} = \frac{\Lambda_{\text{post}}^{-1}}{n} - \frac{\Lambda_0^{-1}}{n},$$

$$\sum_{i=0}^{n} \nu_i - (d+1)n = \nu_{\text{post}} \Rightarrow \nu_i = \frac{\nu_{\text{post}}}{n} - \frac{\nu_0}{n} + d + 1,$$

$$\sum_{i=0}^{n} \log(\gamma_i) = \log(\gamma_{\text{post}}) \Rightarrow \log(\gamma_i) = \frac{\log(\gamma_{\text{post}})}{n} - \frac{\log(\gamma_0)}{n}.$$

In the experiments, we noticed that the initialization is usually not critical, in the sense that many different strategies lead to the same approximation. The former equations can be straightforwardly extended in the cases of diagonal and isotropic matrices.

As far as the convergence is concerned, we can monitor the change in the parameters of the approximating distribution. When these parameters change less than a threshold, we stop the iterations. In our experiments, we monitored the number of degrees of freedom, leading to the following convergence criterion:

$$|\nu_{\text{new}} - \nu_{\text{old}}| < \varepsilon,$$

with $\varepsilon = 10^{-3}$.

We note here that Q in (21) is not positive semidefinite in general. In this case, the matrices U_r can have negative determinant and the whole procedure would fail to converge. This happens when few data are provided with respect to the dimensionality of the problem; in practice this means that the approximation using a unimodal Wishart is not suitable. We can fix this issue by increasing the diagonal elements of Q. In practice, we compute the smallest eigenvalue of Q, and if it is negative, we subtract it to the diagonal of Q. Given the form assumed by Q, we can view this operation as an increase in the prior covariance Σ_0. When this happens, usually the model evidence penalizes the choice of this model.

6. Experimental Validation

In this section we present an extensive validation of our method. First we show the quality of the approximation of the posterior over the precision, and how the model evidence can be used to select the structure of precision matrix in multivariate cases. We then report a study on sampling techniques, comparing their running time with the proposed method. We also study the quality of the approximation in online and offline learning scenarios by comparing their approximate posteriors against the true one in a univariate problem. Finally, we show the performance in terms of ISE when we select the inverse of the expectation of the precision as the bandwidth, comparing it with state of the art methods for selecting the bandwidth, to demonstrate that the proposed method is competitive with the advantage of allowing

Table 2: Pseudo-code of the approximate Bayesian KDE algorithm.

```
1. set the maximum number of iterations maxit and the convergence parameter $\varepsilon$
2. set the parameters $\nu_0$ and $\Lambda_0$ of the prior
3. initialize the values of $\log(\gamma_i)$, $\nu_i$, and $\Sigma_i$ as explained in Section 5
4. for $i$ in 1 : maxit
   (a) for $j$ in 1 : $n$
       i. compute $U_r$ for $r = 1, \ldots, n$ using (21)
       ii. compute $\alpha$, $\log(\psi)$, $\log(z_r)$ for $r = 1, \ldots, n$, using (8), (9), and (10)
       iii. compute $\log(E_0)$, $E_1$, $E_2$, using (12), (13), and (14)
       iv. compute $\log(\gamma_{\text{new}})$, $\nu_{\text{new}}$, $\Sigma_{\text{new}}$, using (15), (16), and (17)
       v. compute $\log(\gamma'_j)$, $\nu'_j$, $\Sigma'_j$, using (18), (19), and (20)
   (b) if convergence criterion is satisfied then exit the loop
```

online learning and model selection. All the algorithms have been run on a quad-core 2.66GHz Intel workstation with 8GB of RAM.

6.1. Quality of the posterior approximation

In this section, we report two illustrative examples to show the approximation of the posterior over the precision in two synthetic cases; the first is a univariate problem, and the second one is bivariate.

6.1.1. Illustrative examples - Univariate and bivariate cases

We considered a data set of 50 data points sampled from the univariate distribution in the left panel of Fig. 1. This problem has been studied in Marron and Wand (1992) where it is referred to the "asymmetric claw". We numerically computed the actual posterior distribution as the multiplication of the prior (we chose a rather flat prior with $a_0 = 1$, $b_0 = 0.01$) and the likelihood, normalized to integrate to one (black curve in the left panel of Fig. 2). In the same figure, we show the approximation obtained by EP in red. The approximation of the posterior is quite accurate in capturing its expectation, variance, and log evidence. In the left panel of Fig. 3 we show in red the pdf when we selected the expectation of the precision, and in black

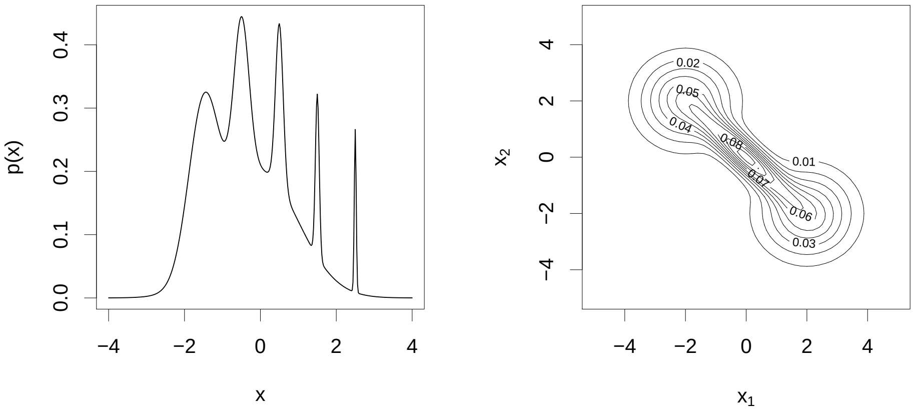

Figure 1: Left: the “asymmetric claw” density function - Right: a bivariate pdf.

dashed lines the 5th and 95th percentiles of the pdfs values when sampling the precision from the posterior.

We considered the synthetic distribution in the right panel of Fig. 1, that is a mixture of three Gaussians (Duong and Hazelton, 2005). We sampled 200 data points from it and we ran our method to approximate the precision. We chose a diagonal structure for the precision matrix as it is easy to visualize the resulting posterior, and we chose a rather flat prior with parameters $a_{0k} = 1$, $b_{0k} = 0.01$. Again, we numerically computed the actual posterior as the multiplication of the likelihood and the prior, normalized to integrate to one (black curve in the right panel of Fig. 2). In the same figure, we show the posterior approximation in red. As in the multivariate case, we see that the approximation captures the main characteristics of the posterior. In Fig. 3 we show in red the pdf when we selected the expectation of the approximate posterior over the precision in the two synthetic cases.

6.1.2. Model selection for choosing the structure of the precision matrix

We considered the bivariate pdfs in the first row of Fig. 4, so that it is easier to interpret the results. We ran the approximate version of Bayesian KDE using the three forms of precision matrix, we computed their respective model evidence, and we used them to compute the Bayes factors to compare the models. When models M_1 and M_2 were compared, we chose model M_1 when the ratio of its evidence over the evidence of M_2 was larger than

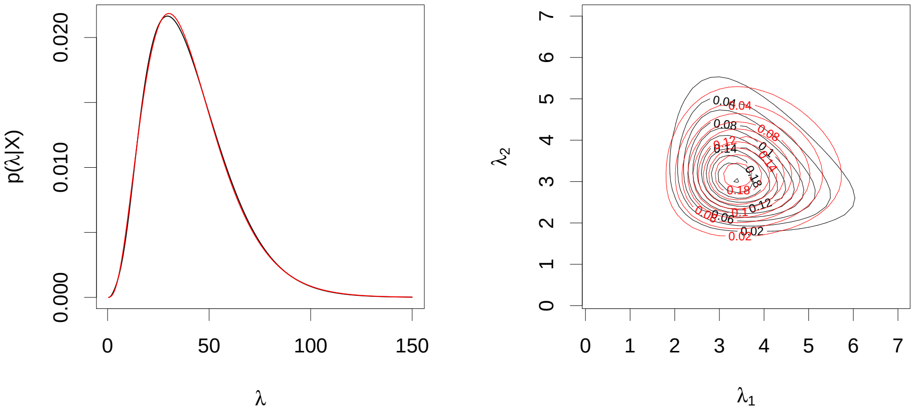

Figure 2: Approximation of the posterior distribution over the precision. Left: univariate case - Right: bivariate case

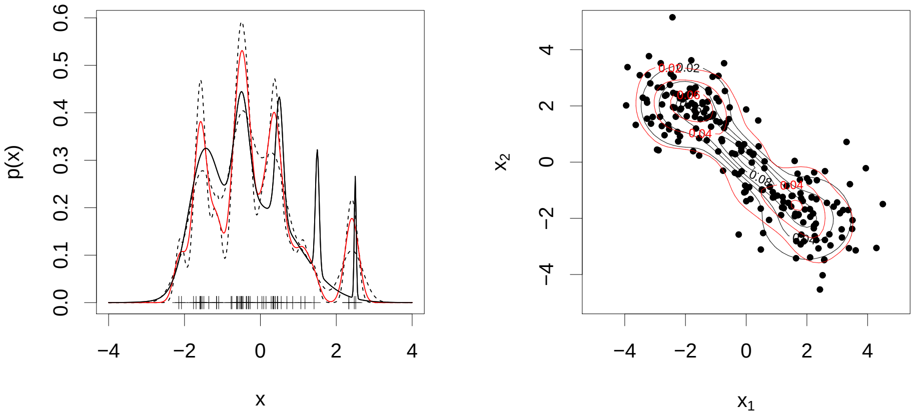

Figure 3: KDE densities when we select the precision as the expectation of the approximated posterior. Left: univariate case, where the dashed lines represent the 5th and 95th percentiles of the values taken by the pdfs when sampling the precision from the posterior - Right: bivariate case

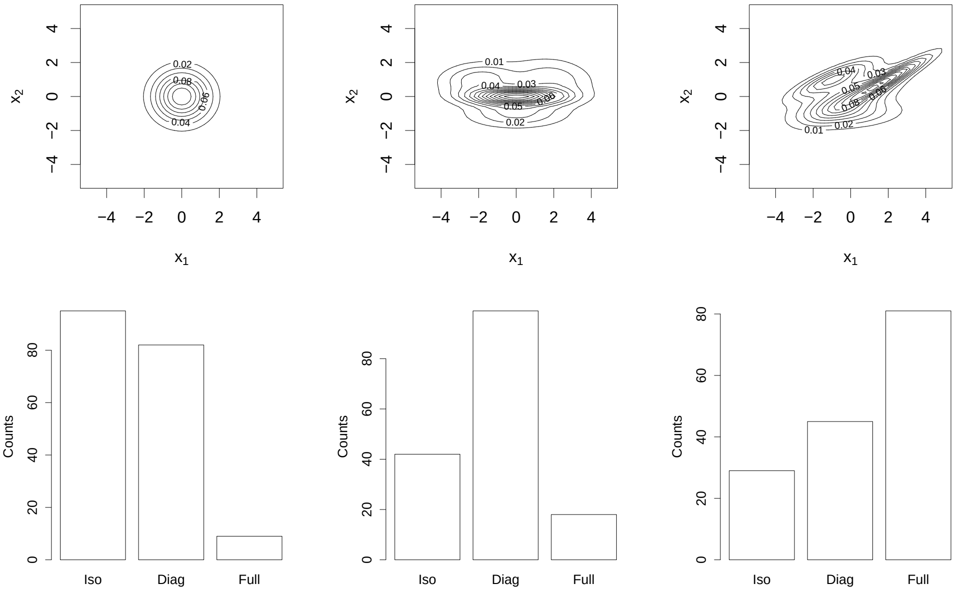

Figure 4: First row: Three bivariate densities - Second row: Counts of how many times the three different models were selected in the three cases above.

3, that is a common rule for preferring a model over another (Kass and Raftery, 1995). We report a study of model selection for the distributions shown in Fig. 4, where we sampled 100 times a set of $n = 100$ data points from them. We ran the three algorithms considering the same flat prior, that is $\nu_0 = d + 1$, $\Sigma_0 = 2I/10$ for the Wishart, and $a_0 = 1$, $b_0 = 0.1$ for the Gamma. We counted the model with the largest evidence as well as those whose evidence is within a factor of $1/3$ with respect to it. The results are reported in the second row of Fig. 4.

We can see that for the first pdf, on average, the isotropic precision is preferred as the problem is isotropic. In the second case, instead, the pdf is elongated along one of the axes, so a diagonal precision is selected more ofter than the isotropic and the full ones. Finally, the third case shows that a full covariance matrix is preferred by the Bayes factor when data have correlations between dimensions.

6.1.3. Comparison with sampling techniques

We compare our method with the one proposed in (Zhang et al., 2006), where Metropolis-Hastings sampling is used to sample from the posterior. In the case of a full precision matrix, the Authors consider the Cholesky decomposition

$$\Lambda = LL^{\mathrm{T}},$$

where L is lower triangular. Using this factorization, the prior used in their experiment for the nonzero elements of L was

$$p(l_{ij}) = \frac{1}{1 + \rho l_{ij}^2},$$

with ρ controlling the flatness of the prior. The likelihood is the cross-validated one in (4). As in Zhang et al. (2006), in all our experiments, we set $\rho = 1$, the number of burn-in samples to 5000, and the total number of samples to 25000. The parameters of the proposal were selected to keep the acceptance rate between 20% and 30%; this, of course, required few trials run that we have not accounted for in the comparison of running time, although they represent a cost in the whole inference process.

We studied the running time of EP and sampling methods in two cases. The first was the bivariate density in Fig. 1, and the second was a mixture of two Gaussians in 4 dimensions centered at $(0, 0, 0, 0)$ and $(1, 1, 1, 1)$ with identity covariance matrices. The ratio of running times for EP against sampling is reported in Tab. 3, where the algorithms have been run 10 times for each value of n. To ensure fair comparison, all the algorithms have been implemented in R (R Development Core Team, 2006). We can see how the approximation is considerably faster than sampling. As a check, we also compared the performance in terms of likelihood of unseen data, finding that they were comparable (results not shown). In the case of a full covariance, in high dimensions and with a limited amount of data, we noticed that sampling techniques were superior in terms of performance, given that the approximation of the posterior is not adequate. In these cases, a simpler covariance structure was favoured by Bayes factors. For the sake of brevity, we do not report comparisons in the case of simpler covariance structures; we just point out that the approximation is again considerably faster than sampling.

6.1.4. Online vs offline

We now show the quality of the approximation in an online scenario. We considered the univariate density in Fig. 1, from which we sampled 500 data

	$n = 200$	$n = 500$	$n = 1000$
$d = 2$	0.10	0.15	0.16
$d = 4$	0.20	0.19	0.31

Table 3: Ratio between running times of the approximation method vs sampling. As a frame of reference, sampling in the case of $d = 4$ and $n = 1000$ took a couple of hours, and for $d = 2$ and $n = 500$ about 6 minutes.

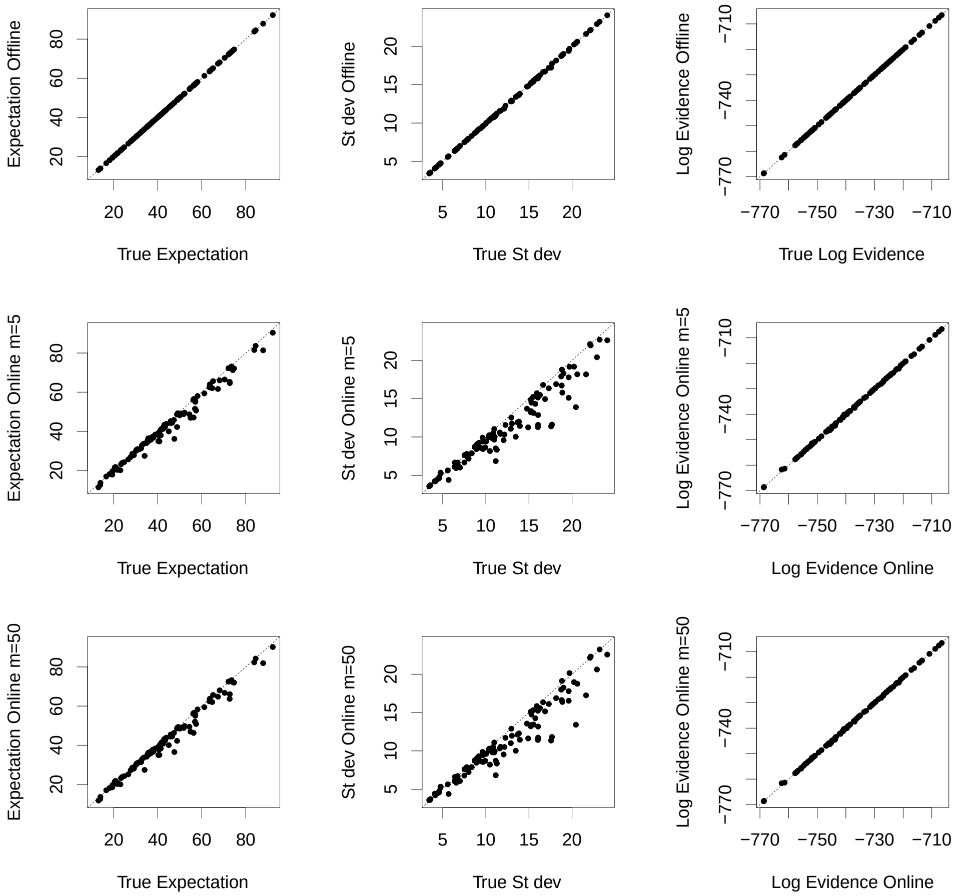

Figure 5: Expectation and standard deviation of the posterior over the precision and log evidence of the model - First row: offline vs true - Second row: online (m = 5) vs true - Third row: online (m = 50) vs true.

points that we used to run the offline version of our method. In order to compare the inference results with online learning, we ran the offline version on half data only and then we ran two online versions adding $m = 5$ and $m = 50$ data points at a time respectively. We compared the mean, variance, and log evidence of the true posterior against those of the approximations obtained by the offline and online versions (Fig. 5). We can see that the offline version is extremely accurate in capturing the moments of the posterior. The online versions do not differ very much from each other, and show an underestimation of the standard deviation; this is reasonable, given that the online version builds the solution upon an approximate form of the posterior obtained over the first batch of data. The evidence is still captured extremely well, while the expectation of the posterior is slightly underestimated on average.

6.2. Performance comparison

6.2.1. Synthetic data sets

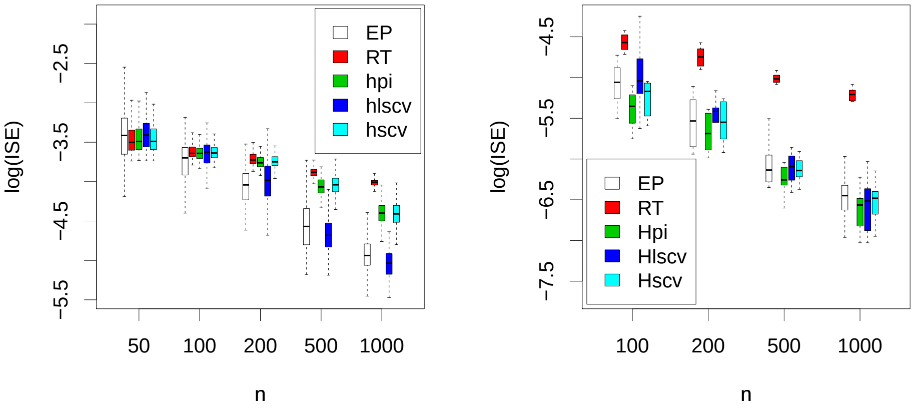

Figure 6: log(ISE) vs n over 100 repetitions for the univariate and bivariate data sets

In order to evaluate the performance of the proposed method, we carried out a study of the ISE in (2) with respect to the size of the data sampled from the true distribution. In particular, we sampled n data points from the true distribution and we estimated/inferred the bandwidth using several methods. Then, we computed the ISE score corresponding to the different estimates of the density functions. We repeated this procedure for different values of n and 100 times for each value of n.

We compared EP with the following methods for bandwidth estimation: Silverman (1986) rule of thumb (`RT`), which is $1.06\min(\mathrm{sd}(x), \mathrm{IQR}/1.34)n^{-1/5}$ with IQR denoting the interquartile range; the plug-in (`hpi`) method of Wand and Jones (1994); least-squares cross-validation (`hlscv`) of Bowman (1984); and smoothed cross-validation (`hscv`) of Jones et al. (1991). In the multivariate case, `RT` corresponds to selecting a bandwidth for each covariate to be $\sigma_i \left[\frac{4}{(d+2)n}\right]^{1/(d+4)}$, where σ_i is the estimated standard deviation of the i-th covariate. The multivariate versions of `hpi`, `hlscv`, and `hscv` will be denoted as `Hpi`, `Hlscv`, and `Hscv`; Duong and Hazelton (2005) presented descriptions of these methods and implementations can be found in the `R` package `ks`.

We considered the synthetic distributions in Fig. 1. The results are shown in the plots in Fig. 6. We can see that the performance of the proposed method in terms of ISE is comparable to the best among the comparing methods. Note also that none of the considered bandwidth selectors is superior to the others in general, while the proposed method achieves an ISE that is often comparable to the best among the methods considered in the comparison.

6.2.2. Real data sets

We applied the proposed methods to three real data set(s): Old Faithful, Unicef, and Earthquake. In cases of ties, we simply added a uniform random contribution of the order of the least significant digits to resolve them.

Old Faithful data set. This is a one dimensional data set containing 107 eruption lengths of Old Faithful geyser in Yellowstone National Park.

Unicef data set. The Unicef data set has two covariates: the number of deaths of children under 5 years of age per 1000 live births, and the average life expectancy (in years) at birth for 73 countries having gross national income less than \$1000 per annum per capita.

Earthquake data set. This data set contains 510 measurements of three covariates of an earthquake beneath the Mt. St. Helens volcano. The three covariates are respectively: longitude (in degrees), latitude, and depth (in km). We centered these data and scaled each covariate to have unit variance.

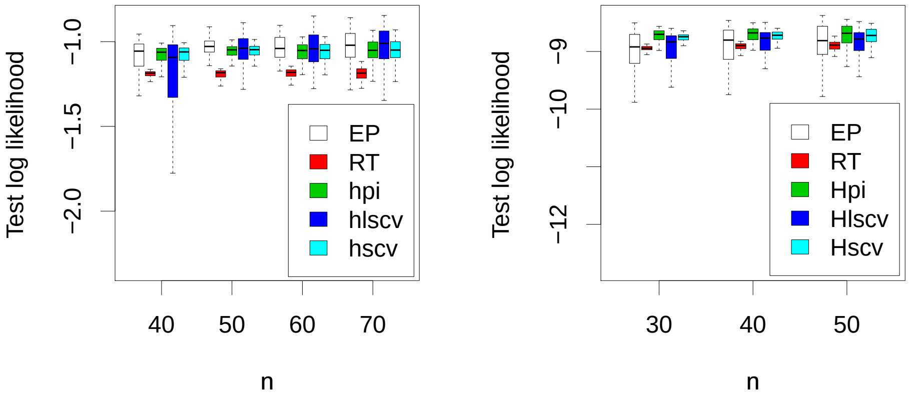

Figure 7: Average test log-likelihood on real data sets - Left: Old Faithful - Right: Unicef

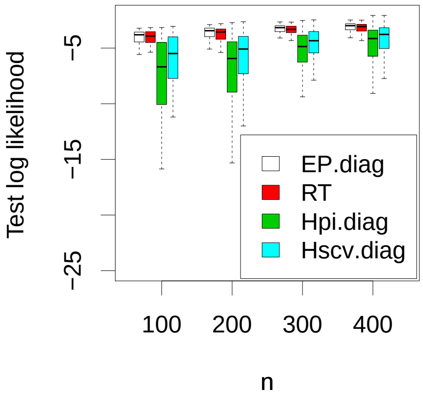

Figure 8: Average test log-likelihood on the Earthquake data sets

In order to evaluate the performance of the proposed method in real scenarios, we report a comparative study on its generalization capabilities. In particular, we sampled a subset of size n from the data set and we used it to estimate the density function. We used such an estimate to compute the average log-likelihood of the remaining data; this is a reasonable measure of the generalization capabilities of the procedure, as it provides an estimate of the goodness of the fit to unseen data. We repeated this procedure for different values of n and 100 times for each value of n. The results are shown in Figs. 7 and 8. We selected a full precision in the Unicef data and a diagonal precision for the Earthquake data. In fact, we could have used the estimated Bayes factors to select the most likely covariance structure at each iteration; since the comparing methods do not provide this feature, we preferred to choose one particular structure only and report the corresponding results. In the case of the earthquake data, `Hlscv` was not converging to any solution in some cases, and we do not report it in the plot.

The proposed method performs quite well in terms of generalization. Especially in the earthquake data set, the method is significantly better than the plug-in ones. Interestingly, the rule of thumb works quite well too. In the other two data sets, the performance is comparable to the other methods.

In general, considering all the experimentations presented in the last two sections, we see that the performance is at least as good as that of the comparing methods. We remark, however, that the proposed method is not based on the optimization of any sort of performance measure, but simply on the Bayesian principle.

7. Conclusions

We presented an approximate Bayesian approach to the selection of the bandwidth in KDE. The method is based on a cross-validated likelihood, and computes an approximation to the intractable posterior over the bandwidth. The approximation is carried out by the EP algorithm, which is based on matching the moments with appropriate approximating distributions. In our case we selected the Wishart distribution for the approximations, given its suitability for modeling distributions over positive definite matrices. We demonstrate that the proposed method is competitive in terms of accuracy with existing approaches and we extensively studied the results of the inference process. In particular, we showed how it is possible to choose the structure of covariance to use, and how to infer the bandwidth by updating the posterior distribution inferred from a former batch of data. Our approach is closely related to the sampling-based approach of Zhang et al. (2006), but exploits the typical speed of deterministic approximate methods, while still retaining an excellent approximation to the posterior distribution.

On a more general note, it is interesting to notice the similarity between KDE and non-parametric Bayesian approaches to density estimation. In particular, Dirichlet Process Mixture Models (DPMMs) (Ferguson, 1973) have received considerable attention in the machine learning community as they present a flexible method for clustering and density estimation. In DPMMs, the marginal data density (i.e. having marginalized the latent cluster membership variables) is also represented as a mixture of a (potentially infinite) number of Gaussians. However, unlike KDE, the number of Gaussian components inferred from a finite sample is in general not equal to the number of data points; also, we are not aware of L_2 convergence properties for DPMMs. In terms of inference, although an EP approach has been proposed for DPMMs (Minka and Ghahramani, 2003), block Gibbs sampling remains the most widely adopted choice within this context.

Appendix A. Special Cases for Expectation Propagation

Appendix A.1. Diagonal Precision Matrix

The unnormalized Gamma distribution is defined as

$$\tilde{\mathcal{G}}(\lambda|a,b) = \lambda^{a-1}\exp(-b\lambda),$$

and the Gamma distribution

$$\mathcal{G}(\lambda|a,b) = B_{\mathcal{G}}(a,b)\tilde{\mathcal{G}}(\lambda|a,b),$$

where

$$B_{\mathcal{G}}(a,b) = \frac{b^a}{\Gamma(a)}.$$

We consider a diagonal form for the precision Λ, where $\operatorname{diag}(\Lambda) = \boldsymbol{\lambda}$. The factors for the approximation of the posterior factorize between the covariates, and are defined as

$$\tilde{f}_i(\boldsymbol{\lambda}) = \gamma_i \prod_{k=1}^{d} \tilde{\mathcal{G}}(\lambda_k|a_{ik}, b_{ik}),$$

which gives

$$q(\boldsymbol{\lambda}) = \left(\prod_{i=0}^{n}\gamma_i\right)\prod_{k=1}^{d}\tilde{\mathcal{G}}\left(\lambda_k \,\middle|\, \sum_{i=0}^{n} a_{ik} - n, \sum_{i=0}^{n} b_{ik}\right).$$

The true posterior is

$$p(\boldsymbol{\lambda}|X) \propto \prod_{i=0}^{n} f_i(\boldsymbol{\lambda}),$$

where $f_0(\boldsymbol{\lambda}) = \prod_{k=1}^{d} \mathcal{G}(\lambda_k | a_{0k}, b_{0k})$ denotes the prior, and the components $f_j(\boldsymbol{\lambda})$ are

$$f_j(\boldsymbol{\lambda}) = p(\mathbf{x}_j | \boldsymbol{\lambda}, \mathcal{M}_X) = \frac{1}{n-1} \sum_{r \neq j, r=1}^{n} \mathcal{N}(\mathbf{x}_j | \mathbf{x}_r, \Lambda),$$

with Λ diagonal and $\text{diag}(\Lambda) = \boldsymbol{\lambda}$.

When we remove the j-th component from $q(\boldsymbol{\lambda})$, we obtain

$$q^{\setminus j}(\boldsymbol{\lambda}) = \prod_{i \neq j, i=0}^{n} \tilde{f}_i(\boldsymbol{\lambda}) = \left(\prod_{i \neq j, i=0}^{n} \gamma_i \right) \prod_{k=1}^{d} \tilde{\mathcal{G}} \left(\lambda_k \middle| \sum_{i \neq j, i=0}^{n} a_{ik} - n + 1, \sum_{i \neq j, i=0}^{n} b_{ik} \right).$$

Let's define:

$$\alpha_k = \sum_{i \neq j, i=0}^{n} a_{ik} - n + 1 + \frac{1}{2},$$

$$\beta_{rk} = \sum_{i \neq j, i=0}^{n} b_{ik} + \frac{1}{2} (\mathbf{x}_{rk} - \mathbf{x}_{jk})^2,$$

$$\psi = \left(\prod_{i \neq j, i=0}^{n} \gamma_i \right) \frac{1}{(n-1)(2\pi)^{d/2}},$$

$$z_r = \prod_{k=1}^{d} B_{\mathcal{G}}(\alpha_k, \beta_{rk})^{-1}.$$

With the definitions above, we see that

$$f_j(\boldsymbol{\lambda}) q^{\setminus j}(\boldsymbol{\lambda}) = \psi \sum_{r \neq j, r=1}^{n} z_r \prod_{k=1}^{d} \mathcal{G}(\lambda_k | \alpha_k, \beta_{kr})$$

Now we have to match the moments of

$$q(\Lambda | a_{\text{new}}, b_{\text{new}}) = \gamma_{\text{new}} \prod_{k=1}^{d} B_{\mathcal{G}}((a_k)_{\text{new}}, (b_k)_{\text{new}})^{-1} \mathcal{G}(\lambda_k | (a_k)_{\text{new}}, (b_k)_{\text{new}})$$

with those of $f_j(\Lambda) q^{\setminus j}(\boldsymbol{\lambda})$. The normalization constant is:

$$E_0 = \int f_j(\boldsymbol{\lambda}) q^{\setminus j}(\boldsymbol{\lambda}) d\boldsymbol{\lambda} = \psi \sum_{r \neq j, r=1}^{n} z_r$$

For the mean and the variance we notice that the resulting density is a weighted sum of products of Gamma across the covariates

$$f_j(\boldsymbol{\lambda})q^{\setminus j}(\boldsymbol{\lambda}) \propto \sum_{r\neq j, r=1}^{n} w_r \prod_{k=1}^{d} \mathcal{G}(\lambda_k|\alpha_k, \beta_{kr}),$$

where the weights are

$$w_r = \frac{z_r}{\sum_{s\neq r,s=1}^{n} z_s}.$$

Therefore, the expected value and the variance are

$$E_{1k} = \mathrm{E}_{f_j(\boldsymbol{\lambda})q^{\setminus j}(\boldsymbol{\lambda})}[\lambda_k] = \sum_{r\neq j, r=1}^{n} w_r \frac{\alpha_k}{\beta_{rk}},$$

$$E_{2k} = \mathrm{E}_{f_j(\boldsymbol{\lambda})q^{\setminus j}(\boldsymbol{\lambda})}[\lambda_k^2] - E_{1k}^2 = \sum_{r\neq j, r=1}^{n} w_r \left[\frac{\alpha_k}{(\beta_{rk})^2} + \left(\frac{\alpha_k}{\beta_{rk}}\right)^2\right] - E_{1k}^2.$$

For the updated version $q_{\text{new}}(\boldsymbol{\lambda})$ we have

$$E_0 = \int q_{\text{new}}(\boldsymbol{\lambda})d\boldsymbol{\lambda} = \gamma_{\text{new}} \prod_{k=1}^{d} B_{\mathcal{G}}((a_k)_{\text{new}}, (b_k)_{\text{new}})^{-1},$$

$$E_{1k} = \mathrm{E}_{q_{\text{new}}(\boldsymbol{\lambda})}[\lambda_k] = \frac{(a_k)_{\text{new}}}{(b_k)_{\text{new}}},$$

$$E_{2k} = \mathrm{var}_{q_{\text{new}}(\boldsymbol{\lambda})}[\lambda_k] = \frac{(a_k)_{\text{new}}}{((b_k)_{\text{new}})^2}.$$

This allows to compute γ_{new}, $(a_k)_{\text{new}}$, and $(b_k)_{\text{new}}$

$$(a_k)_{\text{new}} = \frac{E_{1k}^2}{v_k}, \quad (b_k)_{\text{new}} = \frac{E_{1k}}{v_k}, \quad \gamma_{\text{new}} = E_0 \prod_{k=1}^{d} B_{\mathcal{G}}((a_k)_{\text{new}}, (b_k)_{\text{new}}).$$

The updated version of $\tilde{f}_j(\Lambda)$ is the ratio between $q_{\text{new}}(\boldsymbol{\lambda})$ and $q^{\setminus j}(\boldsymbol{\lambda})$, resulting in

$$\tilde{f}_j'(\boldsymbol{\lambda}) = \gamma_j' \prod_{k=1}^{d} \tilde{\mathcal{G}}(\boldsymbol{\lambda}|a_{jk}', b_{jk}'),$$

with

$$\gamma'_{jk} = \frac{\gamma_{\text{new}}}{\prod_{i\neq j,i=0}^{n} \gamma_i},$$

$$a'_{jk} = (a_k)^{\text{new}} + n - \sum_{i\neq j,i=0}^{n} a_{ik},$$

$$b'_{jk} = (b_k)^{\text{new}} - \sum_{i\neq j,i=0}^{n} b_{ik}.$$

After the convergence criterion is satisfied, the resulting approximating posterior is

$$q(\boldsymbol{\lambda}) = \gamma \prod_{k=1}^{d} \tilde{\mathcal{G}}(\lambda_k | a_k, b_k),$$

that allows to obtain the approximate model evidence as

$$\text{evidence} = \gamma \prod_{k=1}^{d} B_{\mathcal{G}}(a_k, b_k)^{-1}.$$

Appendix A.2. Isotropic Precision Matrix

The proposed approximation of the posterior over the precision λ, in the case of a precision matrix $\Lambda = \lambda I$ is

$$q(\lambda) = \prod_{i=0}^{n} \tilde{f}_i(\lambda),$$

where the factors in the approximation are defined as

$$\tilde{f}_i(\lambda) = \gamma_i \tilde{\mathcal{G}}(\lambda | a_i, b_i)$$

and therefore

$$q(\lambda) = \left(\prod_{i=0}^{n} \gamma_i\right) \tilde{\mathcal{G}}\left(\lambda \,\middle|\, \sum_{i=0}^{n} a_i - n, \sum_{i=0}^{n} b_i\right).$$

The true posterior is

$$p(\lambda|X) \propto \prod_{i=0}^{n} f_i(\lambda),$$

where $f_0(\lambda) = \mathcal{G}(\lambda|a_0, b_0)$ denotes the prior, and the components $f_j(\lambda)$ are

$$f_j(\lambda) = p(\mathbf{x}_j|\lambda, \mathcal{M}_X) = \frac{1}{n-1} \sum_{r \neq j, r=1}^{n} \mathcal{N}(\mathbf{x}_j|\mathbf{x}_r, \lambda I).$$

When we remove the j-th component from $q(\lambda)$, we obtain

$$q^{\setminus j}(\lambda) = \prod_{i \neq j, i=0}^{n} \tilde{f}_i(\lambda) = \left(\prod_{i \neq j, i=0}^{n} \gamma_i \right) \tilde{\mathcal{G}} \left(\lambda \middle| \sum_{i \neq j, i=0}^{n} a_i - n + 1, \sum_{i \neq j, i=0}^{n} b_i \right).$$

Let's define the following quantities:

$$\alpha = \sum_{i \neq j, i=0}^{n} a_i - n + 1 + \frac{d}{2},$$

$$\beta_r = \sum_{i \neq j, i=0}^{n} b_i + \frac{1}{2} \|\mathbf{x}_r - \mathbf{x}_j\|^2,$$

$$\psi = \left(\prod_{i \neq j, i=0}^{n} \gamma_i \right) \frac{1}{(n-1)(2\pi)^{d/2}},$$

$$z_r = B_{\mathcal{G}}(\alpha, \beta_r)^{-1}.$$

With the definitions above, we see that

$$f_j(\lambda) q^{\setminus j}(\lambda) = \psi \sum_{r \neq j, r=1}^{n} z_r \mathcal{G}(\lambda|\alpha, \beta_r).$$

Now we have to match the moments of

$$q(\lambda|a_{\text{new}}, b_{\text{new}}) = \gamma_{\text{new}} B_{\mathcal{G}}(a_{\text{new}}, b_{\text{new}})^{-1} \mathcal{G}(\lambda|a_{\text{new}}, b_{\text{new}})$$

with those of $f_j(\lambda) q^{\setminus j}(\lambda)$. The normalization constant of $f_j(\lambda) q^{\setminus j}(\lambda)$ is

$$E_0 = \int f_j(\lambda) q^{\setminus j}(\lambda) d\lambda = \psi \sum_{r \neq j, r=1}^{n} z_r$$

For the mean and the variance, we notice that the resulting density is a weighted sum of Gamma

$$f_j(\lambda)q^{\setminus j}(\lambda) \propto \sum_{r\neq j, r=1}^{n} w_r \mathcal{G}(\lambda|\alpha, \beta_r)$$

with weights

$$w_r = \frac{z_r}{\sum_{s\neq r, s=1}^{n} z_s}.$$

Therefore

$$E_1 = \mathrm{E}_{f_j(\lambda)q^{\setminus j}(\lambda)}[\lambda] = \sum_{r\neq j, r=1}^{n} w_r \frac{\alpha}{\beta_r},$$

$$E_2 = \mathrm{E}_{f_j(\lambda)q^{\setminus j}(\lambda)}[\lambda^2] - E_1^2 = \sum_{r\neq j, r=1}^{n} w_r \left[\frac{\alpha}{(\beta_r)^2} + \left(\frac{\alpha}{\beta_r}\right)^2\right] - E_1^2.$$

For the updated version $q_{\text{new}}(\lambda)$ we have

$$E_0 = \int q_{\text{new}}(\lambda)d\lambda = \gamma_{\text{new}} B_{\mathcal{G}}(a_{\text{new}}, b_{\text{new}})^{-1},$$

$$E_1 = \mathrm{E}_{q_{\text{new}}(\lambda)}[\lambda] = \frac{a_{\text{new}}}{b_{\text{new}}},$$

$$E_2 = \mathrm{var}_{q_{\text{new}}(\lambda)}[\lambda] = \frac{a_{\text{new}}}{(b_{\text{new}})^2}.$$

This allows to compute γ_{new}, a_{new}, and b_{new} as

$$a_{\text{new}} = \frac{E_1^2}{E_2}, \quad b_{\text{new}} = \frac{E_1}{E_2}, \quad \gamma_{\text{new}} = E_0 B_{\mathcal{G}}(a_{\text{new}}, b_{\text{new}}).$$

The updated version of $\tilde{f}_j(\lambda)$ is the ratio between $q_{\text{new}}(\lambda)$ and $q^{\setminus j}(\lambda)$, resulting in

$$\tilde{f}_j'(\lambda) = \gamma_j' \tilde{\mathcal{G}}(\lambda|a_j', b_j'),$$

where

$$\gamma_j' = \frac{\gamma_{\text{new}}}{\prod_{i\neq j, i=0}^{n} \gamma_i},$$

$$a_j' = a^{\text{new}} + n - \sum_{i\neq j, i=0}^{n} a_i,$$

$$b_j' = b^{\text{new}} - \sum_{i \neq j, i=0}^{n} b_i.$$

After the convergence criterion is satisfied, the resulting approximating posterior is

$$q(\lambda) = \gamma\, \tilde{\mathcal{G}}(\lambda | a, b),$$

that allows to obtain the approximate model evidence as

$$\text{evidence} = \gamma\, B_{\mathcal{G}}(a, b)^{-1}.$$

References

Bishop, C. M., 2007. Pattern Recognition and Machine Learning (Information Science and Statistics), 1st Edition. Springer.

Bowman, A. W., 1984. An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71 (2), 353–360.

Brewer, M. J., 2000. A Bayesian model for local smoothing in kernel density estimation. Statistics and Computing 10 (4), 299–309.

Calderhead, B., Girolami, M., 2009. Estimating Bayes factors via thermodynamic integration and population MCMC. Computational Statistics & Data Analysis 53 (12), 4028–4045.

Cao, R., Cuevas, A., Manteiga, W. G., 1994. A comparative study of several smoothing methods in density estimation. Computational Statistics & Data Analysis 17 (2), 153–176.

Chiu, S. T., 1991. Bandwidth selection for kernel density estimation. The Annals of Statistics 19 (4), 1883–1905.

de Lima, M. S., Atuncar, G. S., 2011. A Bayesian method to estimate the optimal bandwidth for multivariate kernel estimator. Journal of Nonparametric Statistics, 137–148.

Duong, T., Cowling, A., Koch, I., Wand, M. P., 2008. Feature significance for multivariate kernel density estimation. Computational Statistics & Data Analysis 52 (9), 4225–4242.

Duong, T., Hazelton, M. L., 2005. Cross-validation bandwidth matrices for multivariate kernel density estimation. Scandinavian Journal of Statistics Theory and Applications 32 (3), 485–506.

Ferguson, T. S., 1973. A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1 (2), 209–230.

Filippone, M., Masulli, F., Rovetta, S., 2010. Applying the possibilistic c-means algorithm in kernel-induced spaces. IEEE Transactions on Fuzzy Systems 18 (3), 572–584.

Friel, N., Pettitt, A. N., 2008. Marginal likelihood estimation via power posteriors. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (3), 589–607.

Gangopadhyay, A., Cheung, K., 2002. Bayesian approach to the choice of smoothing parameter in kernel density estimation. Journal of Nonparametric Statistics 14 (6), 655–664.

Gelman, A., Rubin, D. B., 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7 (4), 457–472.

Hall, P., 1982. Cross-validation in density estimation. Biometrika 69 (2), 383–390.

Hastings, W. K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57 (1), 97–109.

Hazelton, M. L., Turlach, B. A., 2007. Reweighted kernel density estimation. Computational Statistics & Data Analysis 51 (6), 3057–3069.

Jones, M. C., 1991. On correcting for variance inflation in kernel density estimation. Computational Statistics & Data Analysis 11 (1), 3–15.

Jones, M. C., Henderson, D. A., 2009. Maximum likelihood kernel density estimation: On the potential of convolution sieves. Computational Statistics & Data Analysis 53 (10), 3726–3733.

Jones, M. C., Marron, J. S., Park, B. U., 1991. A simple root n bandwidth selector. The Annals of Statistics 19 (4), 1919–1932.

Jones, M. C., Marron, J. S., Sheather, S. J., 1996. A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association 91 (433), 401–407.

Kass, R. E., Carlin, B. P., Gelman, A., Neal, R. M., 1998. Markov chain Monte Carlo in practice: A roundtable discussion. The American Statistician 52 (2), 93–100.

Kass, R. E., Raftery, A. E., 1995. Bayes factors. Journal of the American Statistical Association 90 (430), 773–795.

Kulasekera, K. B., Padgett, W. J., 2006. Bayes bandwidth selection in kernel density estimation with censored data. Journal of Nonparametric Statistics 18 (2), 129–143.

Loader, C. R., 1999. Bandwidth selection: classical or plug-in? The Annals of Statistics 27 (2), 415–438.

Marron, J. S., Wand, M. P., 1992. Exact mean integrated squared error. The Annals of Statistics 20 (2), 712–736.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E., 1953. Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21 (6), 1087–1092.

Minka, T. P., 2001. Expectation propagation for approximate Bayesian inference. In: Uncertainty in artificial intelligence: proceedings of the seventeenth conference (2001), August 2-5, 2001, University of Washington, Seattle, Washington. Morgan Kaufmann, p. 362.

Minka, T. P., Ghahramani, Z., 2003. Expectation propagation for infinite mixtures. Tech. rep., NIPS'03 Workshop on Nonparametric Bayesian Methods and Infinite Models, 13, Whistler, BC Canada.

Neal, R. M., 1993. Probabilistic inference using Markov chain Monte Carlo methods. Tech. Rep. CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

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

R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rinaldo, A., Wasserman, L., 2010. Generalized density clustering. Annals of Statistics 38 (5), 2678–2722.

Sheather, S. J., Jones, M. C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society. Series B (Methodological) 53 (3), 683–690.

Silverman, B. W., 1986. Density Estimation for Statistics and Data Analysis (Chapman & Hall/CRC Monographs on Statistics & Applied Probability), 1st Edition. Chapman and Hall/CRC.

Skilling, J., 2006. Nested sampling for general Bayesian computation. Bayesian Analysis 1 (4), 833–860.

Terrell, G. R., 1990. The maximal smoothing principle in density estimation. Journal of the American Statistical Association 85 (410), 470–477.

Tran, T. N., Wehrens, R., Buydens, L. M. C., 2006. KNN-kernel density-based clustering for high-dimensional multivariate data. Computational Statistics & Data Analysis 51 (2), 513–525.

van der Laan, M. J., Dudoit, S., Keles, S., 2004. Asymptotic optimality of likelihood-based cross-validation. Statistical Applications in Genetics and Molecular Biology 3 (1).

Wand, M. P., Jones, M. C., 1994. Multivariate plug-in bandwidth selection. Computational Statistics 9, 97–116.

Zhang, X., Brooks, R. D., King, M. L., 2009. A Bayesian approach to bandwidth selection for multivariate kernel regression with an application to state-price density estimation. Journal of Econometrics 153 (1), 21–32.

Zhang, X., King, M. L., Hyndman, R. J., 2006. A Bayesian approach to bandwidth selection for multivariate kernel density estimation. Computational Statistics and Data Analysis 50 (11), 3009–3031.

Żychaluk, K., Patil, P., 2008. A cross-validation method for data with ties in kernel density estimation. Annals of the Institute of Statistical Mathematics 60 (1), 21–44.