Pattern Anal Applic
DOI 10.1007/s10044-015-0496-9

SHORT PAPER

Approximate variational inference based on a finite sample of Gaussian latent variables

Nikolaos Gianniotis[1,2] · **Christoph Schnörr**[3] · **Christian Molkenthin**[1] · **Sanjay Singh Bora**[1]

Received: 29 April 2014 / Accepted: 7 June 2015

Abstract Variational methods are employed in situations where exact Bayesian inference becomes intractable due to the difficulty in performing certain integrals. Typically, variational methods postulate a tractable posterior and formulate a lower bound on the desired integral to be approximated, e.g. marginal likelihood. The lower bound is then optimised with respect to its free parameters, the so-called variational parameters. However, this is not always possible as for certain integrals it is very challenging (or tedious) to come up with a suitable lower bound. Here, we propose a simple scheme that overcomes some of the awkward cases where the usual variational treatment becomes difficult. The scheme relies on a rewriting of the lower bound on the model log-likelihood. We demonstrate the proposed scheme on a number of synthetic and real examples, as well as on a real geophysical model for which the standard variational approaches are inapplicable.

✉ Nikolaos Gianniotis
nikos.gianniotis@h-its.org

Christoph Schnörr
schnoerr@math.uni-heidelberg.de

Christian Molkenthin
molkenthin@geo.uni-potsdam.de

Sanjay Singh Bora
sanjay.singh@geo.uni-potsdam.de

[1] Institute of Earth and Environmental Science, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany

[2] Present Address: Heidelberg Institute for Theoretical Studies, Astroinformatics Group, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany

[3] Image and Pattern Analysis Group, University of Heidelberg, Speyerer-Str. 6, 69115 Heidelberg, Germany

Keywords Bayesian inference · Posterior estimation · Expectation maximisation

1 Introduction

Bayesian inference is becoming the standard mode of inference as computational resources increase, algorithms advance and scientists across fields become aware of the importance of uncertainty. However, exact Bayesian inference is hardly ever possible whenever the model likelihood function deviates from mathematically convenient forms (i.e. conjugacy). Deterministic approximations are constantly gaining ground on the ubiquitous and computationally intensive Monte Carlo sampling methods that are capable of producing high-quality approximations to otherwise intractable quantities such as posterior densities or marginal likelihoods.

Often, however, deterministic schemes are tailored to a particular model, or family of models, and hence previously derived methods might not be transferable to a new setting (e.g. in [4], a specialised algorithm for Bayesian inference in neural networks is considered). This introduces practical difficulties, mostly in fields beyond machine learning, whenever an implementation of Bayesian inference is required, particularly in early explorative stages where the model formulation is likely to keep changing. In this work, we introduce a scheme for approximate Bayesian inference that aims to be general enough so that it can accommodate a variety of models. The proposed scheme is conceptually simple and requires only that the gradient of the model log-likelihood with respect to its parameters be available.

Specifically, we consider the task of computing a global Gaussian approximation $q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to a given

Published online: 30 June 2015

intractable posterior distribution representing a probabilistic model by maximizing a standard variational lower bound [14]. This lower bound involves the expectation of the log-likelihood of the model distribution with respect to the approximating distribution $q(\boldsymbol{w})$. To enable the computation of these expectations either in closed form or through computationally tractable numerical approximations, the likelihoods are typically restricted to conditionally factorized forms. This step of the approach specifically depends on the model at hand.

By contrast, our conceptually simple method presented below is more generally applicable to various models in the same way. We demonstrate this by working out examples for a variety of models. In particular, we demonstrate empirically that our approach results in Gaussian approximations that are superior to the basic Laplace approximation [6], which is the typical objective of variational approximation schemes. A formal comparison to related state-of-the-art methods for computing improved Gaussian approximations, e.g. through the nested Laplace approximation [18] or by expectation propagation [9], is however beyond the scope of this paper.

Our paper is organized as follows. In Sect. 2, we introduce in general terms the proposed scheme. Starting from a standard formulation that typically variational methods use, we postulate a Gaussian posterior $q(\boldsymbol{w})$ and show how to form an approximation to the lower bound of the marginal log-likelihood. The obtained approximation allows us to optimise the variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of $q(\boldsymbol{w})$ by gradient optimisation. For the reader who wishes to refresh her/his memory or obtain a more detailed explanation of the equations presented in Sect. 2, we refer to [5, 21].

In Sect. 3, we demonstrate the proposed scheme on a number of applications and compare it against the exact inference, Laplace and variational approximations. Specifically, in Sect. 3.1, we show a visual comparison of approximating flexible bivariate densities using the Laplace approximation and the proposed scheme. In Sect. 3.2 ,we apply our approach to the problem of Bayesian linear regression which actually does admit an analytical and exact solution. This is useful as it allows us to empirically verify the correctness of our scheme against the posterior obtained by exact inference. Subsequently, in Sects. 3.3 and 3.4, we compare the proposed scheme with approaches that take into account the functional forms of classification problems. We show that despite its general formulation, the proposed scheme performs up to par in this setting without exploiting any such problem-specific knowledge. In Sect. 3.5, we show how a change in the model likelihood of probabilistic principal component analysis [20], which renders inference problematic, can easily be accommodated by the proposed scheme. This demonstrates the versatility of our approach in handling such cases in a direct and simple manner. Finally, in Sect. 3.6, we show how the proposed scheme can be applied beyond the usual statistical models, namely on a real geophysical model [7]. We believe that the proposed method raises a range of interesting questions and directions of research; we briefly discuss them in Sect. 4.

2 Proposed scheme for approximate variational inference

Assume an observed dataset of inputs $\boldsymbol{X} = (x_1, \ldots, x_N)^{\mathrm{T}}$ and outputs $\boldsymbol{Y} = (y_1, \ldots, y_N)^{\mathrm{T}}$ modelled by a model f parametrised by $\boldsymbol{w} \in \mathbb{R}^M$. For observed outputs corrupted by Gaussian noise of precision β, the following likelihood[1] arises:

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n | f(x_n; \boldsymbol{w}), \beta^{-1}) \,, \\ = \mathcal{N}(\boldsymbol{Y} | f(\boldsymbol{X}; \boldsymbol{w}), \beta^{-1} \boldsymbol{I}_N), \tag{1}$$

where $f(\boldsymbol{X}; \boldsymbol{w}) = (f(x_1; \boldsymbol{w}), \ldots, f(x_N; \boldsymbol{w}))$ is the vector of model outputs calculated on all inputs x_n. Furthermore, assume a Gaussian prior on the parameters $\boldsymbol{w}$:

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}_M). \tag{2}$$

Our wish is to approximate the true posterior of the parameters $p(\boldsymbol{w}|\boldsymbol{Y}, \boldsymbol{X}, \alpha, \beta)$. We do not make any assumptions about the model having conjugate priors for the parameters $\boldsymbol{w}$. Model f may have a complex functional form that hinders exact Bayesian inference, or even the application of an approximate Bayesian scheme such as VBEM [5] with a factorised prior. However, we do have to make an assumption on the form of the posterior. We choose to postulate an approximate Gaussian posterior for the parameters $\boldsymbol{w}$:

$$q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{3}$$

Parameters $\boldsymbol{\mu} \in \mathbb{R}^M$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{M \times M}$ of the posterior are called variational parameters. For reasons that will become obvious shortly, we choose to parametrise the covariance matrix as $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$ with $\boldsymbol{L} \in \mathbb{R}^{M \times M}$. The postulated posterior now reads:

$$q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}). \tag{4}$$

Hence, the actual variational parameters are $\boldsymbol{\mu}$ and $\boldsymbol{L}$.

[1] Besides the likelihood based on the Gaussian density, others can also be accommodated as shown in Sects. 3.3–3.5. The choice of the Gaussian is made for ease of exposition.

2.1 Approximate lower bound

The first step in introducing the proposed scheme is writing the marginal log-likelihood and lower-bounding it in the standard way using Jensens' inequality [5, Eq. (2.46)]:

$$\begin{aligned}\log p(\boldsymbol{Y}|\boldsymbol{X},\alpha,\beta) &= \log \int p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w},\beta)p(\boldsymbol{w}|\alpha)\boldsymbol{dw}\\ &= \log \int \frac{q(\boldsymbol{w})}{q(\boldsymbol{w})} p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w},\beta)p(\boldsymbol{w}|\alpha)\boldsymbol{dw}\\ &\geq \int q(\boldsymbol{w}) \log \frac{p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w},\beta)p(\boldsymbol{w}|\alpha)}{q(\boldsymbol{w})}\boldsymbol{dw}\\ &= \int q(\boldsymbol{w}) \log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w},\beta)\boldsymbol{dw}\\ &\quad + \int q(\boldsymbol{w}) \log \frac{p(\boldsymbol{w}|\alpha)}{q(\boldsymbol{w})}\boldsymbol{dw}\\ &= \underbrace{\int q(\boldsymbol{w}) \log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w},\beta)\boldsymbol{dw}}_{(1)}\\ &\quad - \underbrace{\int q(\boldsymbol{w}) \log \frac{q(\boldsymbol{w})}{p(\boldsymbol{w}|\alpha)}\boldsymbol{dw}}_{(2)}\\ &\triangleq \mathcal{L}(\boldsymbol{\mu},\boldsymbol{L},\alpha,\beta).\end{aligned} \tag{5}$$

An alternative motivation of the lower bound is provided in [21, Eq. (15)]. Maximising the lower bound $\mathcal{L}$ in Eq. (5) with respect to the free variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{L}$ of $q(\boldsymbol{w})$ results in the best Gaussian approximation to the true posterior. Term (1), the integrated likelihood in Eq. (5), is a potentially intractable integral. We approximate term (1) using Monte Carlo sampling:

$$\frac{1}{S}\sum_{s=1}^{S} \log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w}_{(s)},\beta), \tag{6}$$

where we draw S samples $\boldsymbol{w}_{(s)}$ from the postulated posterior $q(\boldsymbol{w})$. Due to the sampling, however, the variational parameters no longer appear in the approximation Eq. (6) . Nevertheless, it is possible to re-introduce them by rewriting the sampled weights $\boldsymbol{w}_{(s)}$ as[2]:

$$\boldsymbol{w}_{(s)} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z}_{(s)}, \tag{7}$$

where variables $\boldsymbol{z}_{(s)}$ are sampled from the standard normal $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I}_M)$. We summarise all samples $\boldsymbol{z}_{(s)}$ by $Z = \{\boldsymbol{z}_{(1)}\ldots,\boldsymbol{z}_{(S)}\}$. Hence, we can rewrite Eq. (6) as:

$$\frac{1}{S}\sum_{s=1}^{S} \log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\mu}+\boldsymbol{L}\boldsymbol{z}_{(s)},\beta). \tag{8}$$

Hence, the variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{L}$ are now made explicit in this approximation. We expand the approximation of term (1) further:

$$\begin{aligned}&\frac{1}{S}\sum_{s=1}^{S} \log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\mu}+\boldsymbol{L}\boldsymbol{z}_{(s)},\beta)\\ &\quad = \frac{1}{S}\sum_{s=1}^{S} \log \mathcal{N}(\boldsymbol{Y}|f(\boldsymbol{X};\boldsymbol{\mu}+\boldsymbol{L}\boldsymbol{z}_{(s)}),\beta^{-1}\boldsymbol{I}_N)\\ &\quad = \frac{1}{S}\sum_{s=1}^{S} \frac{N}{2}\log(\beta) - \frac{\beta}{2}\|\boldsymbol{Y} - f(\boldsymbol{X};\boldsymbol{\mu}+\boldsymbol{L}\boldsymbol{z}_{(s)})\|^2\\ &\quad + \text{const.}\end{aligned} \tag{9}$$

Term (2) in Eq. (5) is simply the Kullback–Leibler divergence (KLD) between the two Gaussian densities $q(\boldsymbol{w})$ and $p(\boldsymbol{w}|\alpha)$, and can be calculated in closed form:

$$\frac{1}{2}\left(\mathrm{tr}(\alpha\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}) + \alpha\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu} - M - \ln|\alpha\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}|\right). \tag{10}$$

We can now put together the approximated term (1) in Eq. (9) and the KLD term (2) in Eq. (10), to formulate the following objective function[3] $\mathcal{L}_{(FS)}$. Discarding constants, the proposed approximate lower bound reads:

$$\begin{aligned}&\mathcal{L}_{(FS)}(\boldsymbol{\mu},\boldsymbol{L},\alpha,\beta,Z)\\ &\quad = \frac{1}{S}\sum_{s=1}^{S} \frac{N}{2}\log(\beta) - \frac{\beta}{2}\|\boldsymbol{Y} - f(\boldsymbol{X};\boldsymbol{\mu}+\boldsymbol{L}\boldsymbol{z}_{(s)})\|^2\\ &\quad - \frac{1}{2}\left(tr(\alpha\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}) + \alpha\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu} - \ln|\alpha\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}|\right).\end{aligned} \tag{11}$$

Objective $\mathcal{L}_{(FS)}$ is an approximation to the intractable lower bound $\mathcal{L}$ in Eq. (5). It consists of two parts, the approximation to the integrated likelihood [term (1)] and the exact KLD [term (2)]. The proposed lower bound $\mathcal{L}_{(FS)}$ becomes more accurate when the number S of samples $\boldsymbol{z}_{(s)}$ is large.

2.2 Optimisation of approximate lower bound

Gradients of $\mathcal{L}_{(FS)}$ can be calculated with respect to the variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{L}$ to find the best approximating posterior $q(\boldsymbol{w})$:

$$\begin{aligned}&\nabla_{\boldsymbol{\mu}}\mathcal{L}_{(FS)}(\boldsymbol{\mu},\boldsymbol{L},\alpha,\beta,Z)\\ &\quad = \frac{1}{S}\sum_{s=1}^{S} \beta(\nabla_{\boldsymbol{w}}f)^{\mathrm{T}}_{(s)}(\boldsymbol{Y} - f(\boldsymbol{X};\boldsymbol{\mu}+\boldsymbol{L}\boldsymbol{z}_{(s)})) - \alpha\boldsymbol{\mu},\end{aligned} \tag{12}$$

[2] This is where the parametrisation $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$ becomes useful.

[3] The subscript FS stands for finite sample.

$$\begin{aligned}\nabla_{\boldsymbol{L}}\mathcal{L}_{(FS)}&(\boldsymbol{\mu},\boldsymbol{L},\alpha,\beta,Z)\\ &=\frac{1}{S}\sum_{s=1}^{S}\beta(\nabla_{\boldsymbol{w}}f)^{\mathrm{T}}_{(s)}(\boldsymbol{Y}-f(\boldsymbol{X};\boldsymbol{\mu}+\boldsymbol{L}\boldsymbol{z}_{(s)}))\boldsymbol{z}^{\mathrm{T}}_{(s)}\\ &\quad-\alpha\boldsymbol{L}+\boldsymbol{L}^{+\mathrm{T}},\end{aligned}\tag{13}$$

where $\nabla_{\boldsymbol{w}}f$ denotes the Jacobian matrix of f and $\boldsymbol{L}^{+}$ is the pseudo-inverse of $\boldsymbol{L}$ due to the derivation of the log-determinant[4] in Eq. (11). Analogous equations for the case of exact variational inference can be found in [5, Eq. (2.64)]. Given the current posterior $q(\boldsymbol{w})$, hyperparameters α and β have analytical updates:

$$\alpha=\frac{M}{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}+\mathrm{tr}(\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}})},\tag{14}$$

$$\beta=\frac{SN}{\sum_{s=1}^{S}\|\boldsymbol{Y}-f(\boldsymbol{X};\boldsymbol{\mu}+\boldsymbol{L}\boldsymbol{z}_{(s)})\|^{2}}.\tag{15}$$

Again, analogous equations for the above hyperparameter updates can be found in [21, Eqs. (38), (39)].

The proposed scheme is summarised with the pseudocode in Algorithm 1. Convergence was established in the experiments by checking whether the difference between the objective function values $\mathcal{L}_{(FS)}$ between two successive iterations is less than 10^{-4}. Gradient optimisation of $\boldsymbol{\mu}$ and $\boldsymbol{L}$ was carried out using the scaled conjugate gradient algorithm [13]. The outcome of the above scheme is the approximation $\mathcal{L}_{(FS)}$ to the marginal log-likelihood $\log p(\boldsymbol{Y}|\boldsymbol{X},\alpha,\beta)$ (also called log-evidence). The scheme imparts us with the approximate Gaussian posterior $q(\boldsymbol{w})=\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu},\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}})$.

Algorithm 1 Proposed approximate variational inference

```
Initialisation:
  – Initialise variational parameters e.g. $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and
    $\boldsymbol{L} = c\boldsymbol{I}$ . % e.g. $c = 0.1$
    Initialise hyperparameters e.g. $\alpha = \beta = 0.1$.
  – Alternative to above, ML estimates may be useful as
    initial values.
  – Draw $S$ samples $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I})$ that remain fixed
    throughout the algorithm.
Alternating optimisation of $\mathcal{L}_{(FS)}$:
for iter = 1 : MaxIter do % e.g. $MaxIter = 1000$
  – Record $L_{prv} \leftarrow \mathcal{L}_{(FS)}$.
  – Optimise $\boldsymbol{\mu}$ for $J$ iterations using the gradient in Eq.
    (12). % e.g. $J = 10$
  – Optimise $\boldsymbol{L}$ for $J$ iterations using the gradient in Eq.
    (13).
  – Update $\alpha$ and $\beta$ using Eqs. (14) and (15) respectively.
  – Record $L_{new} \leftarrow \mathcal{L}_{(FS)}$.
  – Break if e.g. $L_{new} - L_{prv} < 10^{-4}$.
end for
Result:
  – Lower bound $\mathcal{L}_{(FS)}$ to marginal log-likelihood.
  – Gaussian Posterior $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^{T})$.
```

2.3 Monitoring generalisation performance

For large values of S, the proposed lower bound $\mathcal{L}_{(FS)}$ approximates the true bound $\mathcal{L}$ in Eq. (5) closely. Therefore, we expect that optimising $\mathcal{L}_{(FS)}$ will yield approximately the same optimal variational parameters $\boldsymbol{\mu},\boldsymbol{L}$ as the optimisation of the intractable true lower bound $\mathcal{L}$ would.

The proposed scheme exhibits some fluctuation as $\mathcal{L}_{(FS)}(\boldsymbol{\mu},\boldsymbol{L},\alpha,\beta,Z)$ is a function of the random set of samples Z. Hence, if the algorithm, as summarised in Algorithm 1, is run again, a new set $Z^{(\mathrm{new})}$ will be drawn and a different function $\mathcal{L}_{(FS)}(\boldsymbol{\mu},\boldsymbol{L},\alpha,\beta,Z^{(\mathrm{new}))}$ will be optimised. However, for large enough S the fluctuation due to Z will be innocuous and approximately the same variational parameters will be found for any drawn Z.[5]

However, if on the other hand, we choose a small S, then the variational parameters will overly depend on the small set of samples Z that happened to be drawn at the beginning of the algorithm. As a consequence, $\mathcal{L}_{(FS)}$ will approximate $\mathcal{L}$ poorly, and the resulting posterior $q(\boldsymbol{w})$ will also be a poor approximation to the true posterior. Hence, the variational parameters will be overfitted to the small set of samples Z that happened to be drawn.

Naturally, the question arises: how to choose a large enough S to avoid overfitting the variational parameters on Z. A practical answer to this question is the following: at the beginning of the algorithm, we draw a second independent set of samples $Z'=\{\boldsymbol{z}'_{(1)},\ldots,\boldsymbol{z}'_{(S')}\}$ where S' is preferably a number larger than S. At each (or every few) iteration(s), we monitor the quantity $\mathcal{L}_{(FS)}(\boldsymbol{\mu},\boldsymbol{L},\alpha,\beta,Z')$ on the independent[6] sample set Z'. If the variational parameters are not overfitting the drawn Z, then we should see that as the lower bound $\mathcal{L}_{(FS)}(\boldsymbol{\mu},\boldsymbol{L},\alpha,\beta,Z)$ increases, the quantity $\mathcal{L}_{(FS)}(\boldsymbol{\mu},\boldsymbol{L},\alpha,\beta,Z')$ should also display a tendency to increase. If on the other hand, the variational parameters overfit the drawn Z, then though $\mathcal{L}_{(FS)}(\boldsymbol{\mu},\boldsymbol{L},\alpha,\beta,Z)$ increases, we will notice that $\mathcal{L}_{(FS)}(\boldsymbol{\mu},\boldsymbol{L},\alpha,\beta,Z')$ actually deteriorates. This is a clear sign that a larger S is required.

The described procedure is reminiscent of monitoring the generalisation performance of a learning algorithm on a validation set during training. A significant difference, however, is that while validation sets are typically of limited size, here we can set S' arbitrarily large. For practical purposes, we found that $S'=5S$ was good enough to detect overfitting. An illustration of overfitting the variational parameters in provided in Sect. 3.2.

[4] See derivative rule 55 in [15].

[5] Discounting other sources of randomness like initialisation, etc.

[6] We stress that Z' is not used in training.

3 Applications

In this section, we apply the proposed approach on a variety of applications, namely regression, classification, denoising and geophysical modelling. In particular, the geophysical example shows how the method can be applied beyond standard statistical models.

3.1 Fitting bivariate posteriors

We test our proposed scheme on some artificially constructed posteriors by using a flexible parametric class of densities, according to [3], which reads:

$$f(w_1, w_2) = 2\mathcal{N}(w_1, w_2 | 0, \boldsymbol{I}_2)\Phi(h(w_1, w_2)), \tag{16}$$

where Φ is the cumulative distribution function of the standard normal distribution, and in general h is a real-value function such that $h(-w) = -h(w)$. Here, we take h to be the dot product of a row and column vector, $h(w_1, w_2) = (w_1, w_2, w_1 w_2^2, w_1^2 w_2, w_1^3, w_2^3)\boldsymbol{a}$, as in [3]. The goal is to find the best Gaussian approximation to instances of Eq. (16) for different column vectors $\boldsymbol{a}$. To that end, we tried to find the best Gaussian using the Laplace approximation and the proposed scheme. We used $S = 50$. The results are shown in Fig. 1. The Gaussian approximations are drawn as black dashed curves with their mean marked as a red cross. The goodness of each approximation has been measured as the Kullback–Leibler divergence and noted in the respective captions. We note that the proposed scheme fares better than the Laplace approximation, as the latter evidently focuses on the mode of the target density instead of where the volume of the density lies. The KLD in these examples was calculated numerically as there is no closed form between a Gaussian and a member of the densities in Eq. (16)

3.2 Bayesian linear regression

Bayesian linear regression constitutes a useful example for corroborating that the proposed scheme works correctly as we can compare the obtained posterior $q(\boldsymbol{w})$ with the posterior obtained by exact Bayesian inference [6].

We consider data targets y_n generated by the equation

$$y = 2\cos(x)\sin(x) - 0.1x^2, \tag{17}$$

with inputs x_n uniformly drawn in the range $[-6, 6]$. We add white Gaussian noise to the data targets with a standard deviation of $\sigma = 0.2$. We calculate a set of radial basis functions on the data inputs

$$\boldsymbol{\phi}_n = [\phi(x_n; r, c_1) \ldots \phi(x_n; r, c_{M-1}) 1]^{\mathrm{T}},$$

where $\phi(x_n; r, c_m) = \exp(-\frac{\|x_n - c_m\|^2}{2r^2})$. The last element 1 in $\boldsymbol{\phi}_n$ serves as a bias term. We set $r = 1$ and adopt the linear model $y = \boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{w}$, where $\boldsymbol{w} \in \mathbb{R}^M$. We complete the model by choosing the following densities:

- Likelihood: $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n | \boldsymbol{\phi}_n^{\mathrm{T}}\boldsymbol{w}, \beta^{-1})$.
- Prior: $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}_M)$.
- Postulated posterior: $q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}})$, where $\boldsymbol{\mu} \in \mathbb{R}^M$ and $\boldsymbol{L} \in \mathbb{R}^{M \times M}$.

We set the number of samples of variables $\boldsymbol{z}$ to $S = 100$. We inferred the Gaussian posterior of the weights $\boldsymbol{w}$ using both exact Bayesian inference [6] and the proposed scheme. In Fig. 2, we plot the mean predictions as obtained by the two schemes and note that they are very similar, especially in the areas where enough data items are present. Similarly, in Fig. 3, we plot the covariance matrices found by the two schemes and note their close similarity. Hence, we conclude that the proposed scheme stands in close agreement with the exact Bayesian solution.

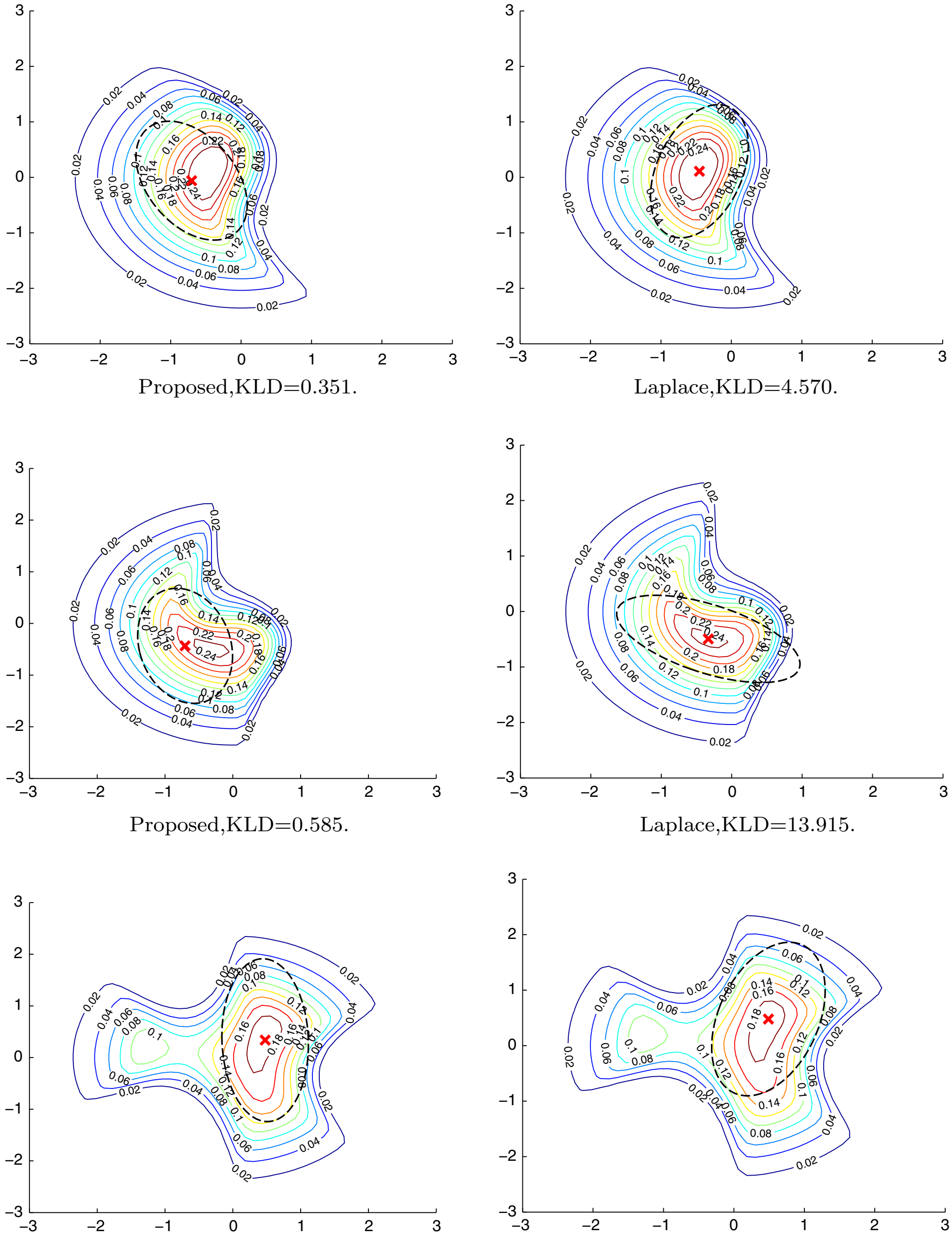
Finally, we demonstrate on the same dataset the effect of overfitting the variational parameters when S is set too low. In Fig. 4, we monitor the lower bounds $\mathcal{L}_{(FS)}(\boldsymbol{\mu}, \boldsymbol{L}, \alpha, \beta, Z)$ and $\mathcal{L}_{(FS)}(\boldsymbol{\mu}, \boldsymbol{L}, \alpha, \beta, Z')$; see Sect. 2.3. In the left, we run the algorithm for $S = 100$ and $S' = 500$: we see that as $\mathcal{L}_{(FS)}(\boldsymbol{\mu}, \boldsymbol{L}, \alpha, \beta, Z)$ increases with each iteration, so does $\mathcal{L}_{(FS)}(\boldsymbol{\mu}, \boldsymbol{L}, \alpha, \beta, Z')$. This means that the fitted variational parameters $\boldsymbol{\mu}, \boldsymbol{L}$ generalise well. On the right hand side, we run the algorithm for $S = 10$, but keep $S' = 500$. Here, we clearly see that while $\mathcal{L}_{(FS)}(\boldsymbol{\mu}, \boldsymbol{L}, \alpha, \beta, Z)$ increases, the lower bound $\mathcal{L}_{(FS)}(\boldsymbol{\mu}, \boldsymbol{L}, \alpha, \beta, Z')$ deteriorates. This a clear sign that a larger S is required and that the variational parameters are overfitted.

3.3 Bayesian logistic regression

In this section, we apply the proposed scheme to Bayesian logistic regression and compare with the variational approach presented in [11]. The data are input–label pairs $(\boldsymbol{x}, y)$ with $y \in \{0, 1\}$. Again, like in Sect. 3.2, we calculate basis functions $\boldsymbol{\phi}_n$ on the input data $\boldsymbol{x}_n$ and take $r = 0.5$. We set $S = 200$. We complete the model by choosing the following densities:

- Likelihood: $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}) = \prod_{n=1}^{N} \sigma(\boldsymbol{\phi}_n^{\mathrm{T}}\boldsymbol{w})^{y_n}$ $(1 - \sigma(\boldsymbol{\phi}_n^{\mathrm{T}}\boldsymbol{w}))^{1-y_n}$.
- Prior: $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}_M)$.
- Postulated posterior: $q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}})$.

Fig. 1 *Top* $\boldsymbol{a} = (-3, 1, -1, -1, -1, -1)^{\mathrm{T}}$, *middle* $\boldsymbol{a} = (0, -2, -4, -1, -3, 0)^{\mathrm{T}}$, *bottom* $\boldsymbol{a} = (1, 0, 2, 1, -1, 0)^{\mathrm{T}}$. The KLD values reveal (lower is better) that the proposed scheme fares better than the Laplace approximation



Proposed,KLD=0.351.

Laplace,KLD=4.570.

Proposed,KLD=0.585.

Laplace,KLD=13.915.

Proposed,KLD=1.103.

Laplace,KLD=1.384.

We evaluated both schemes on datasets preprocessed by Rätsch et al.[7] Each preprocessed dataset has been randomly partitioned into a 100 non-overlapping training and testing sets. Hence, the performance of both schemes was evaluated as the accuracy on the test set, that is the ratio of correctly classified test samples over all test samples. The predictive distribution for the proposed scheme was approximated using a Monte Carlo estimate. We drew 200 parameter samples from the fitted Gaussian posterior q and measured performance on the testing set as the average accuracy under each sample of parameters. The results reported in Table 1, mean squared error and standard deviation, show that both the proposed schemes and the variational bound in [11] perform virtually similarly. We note that, as opposed to [11] which exploits the functional form of logistic regression to design a bespoke lower bound, the proposed method does not take into account any such knowledge and still is capable of delivering comparable performance. Hence, we find the results in this section encouraging.

[7] http://www.raetschlab.org/Members/raetsch/benchmark.

3.4 Bayesian multiclass classification

In this section, we apply the proposed scheme on Bayesian multiclass classification. The data are input–label pairs $(\boldsymbol{x}, \boldsymbol{y})$. Vectors $\boldsymbol{y}$ are binary vectors encoding class labels using 1-of-K coding scheme, e.g. [0 1 0] encodes class label 2 in a three-class problem. A typical way of formulating multiclass classification is multiclass logistic regression (MLR); see [6, Chapter 4] for more details. MLR models the probability $p(C_k|\boldsymbol{\phi}_n)$ of the n-th data item belonging to class C_k via the softmax function $p(C_k|\boldsymbol{\phi}_n) = \frac{\exp(\boldsymbol{\phi}_n^{\mathrm{T}}\boldsymbol{w}_k)}{\sum_{\ell=1}^{K}\exp(\boldsymbol{\phi}_n^{\mathrm{T}}\boldsymbol{w}_\ell)}$. K denotes the total number of classes, and each class C_k is associated with a weight vector $\boldsymbol{w}_k$. Similarly to logistic regression, MLR does not allow direct Bayesian inference as the use of the softmax function renders integrals over the likelihood term intractable. Thus, Bayesian MLR is a good candidate problem for the proposed approach. We specify the following model:

- Likelihood: $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(C_k|\boldsymbol{\phi}_n)^{y_{nk}}$.
- Prior: $\prod_{k=1}^{K} p(\boldsymbol{w}_k) = \mathcal{N}(\boldsymbol{w}_k|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}_M)$.
- Postulated posterior: $q(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K) = \prod_{k=1}^{K} q(\boldsymbol{w}_k)$, with $q(\boldsymbol{w}_k) = \mathcal{N}(\boldsymbol{w}_k|\boldsymbol{\mu}_k, \boldsymbol{L}_k\boldsymbol{L}_k^{\mathrm{T}})$.

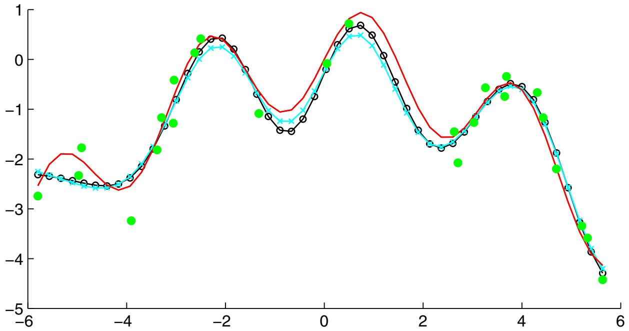
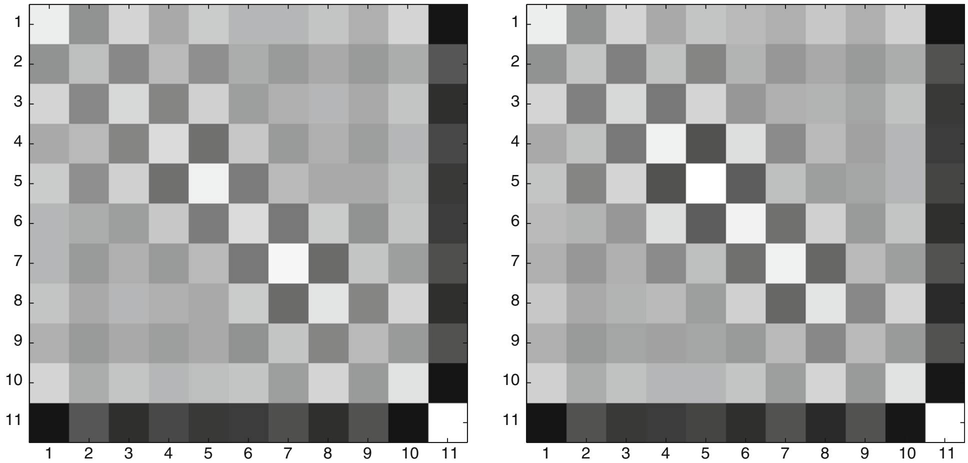


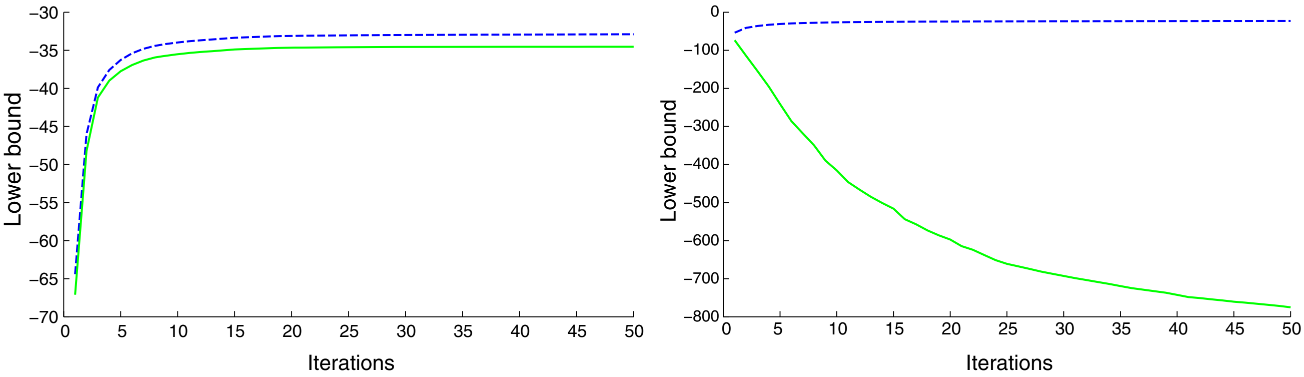
Fig. 2 Mean predictions. The *red line* is the true underlying function, and the *green points* are noisy realisations of it (data items). The *black* (*circles*) and *cyan* (*crosses*) *lines* are the mean predictions that correspond to exact and approximate inference, respectively. We see that the approximate scheme stands in close agreement to the exact solution



Exact inference.

Proposed.

Fig. 3 Posterior covariance matrices found by both schemes. We note their close similarity which indicates that the proposed scheme stands in close agreement to the exact solution



S=100, no overfitting.

S=10, overfitting.

Fig. 4 Monitoring generalisation performance. *Green solid line* is $\mathcal{L}_{(FS)}(\boldsymbol{\mu}, \boldsymbol{L}, \alpha, \beta, Z')$, and *blue dashed line* is $\mathcal{L}_{(FS)}(\boldsymbol{\mu}, \boldsymbol{L}, \alpha, \beta, Z)$. Overfitting of the variational parameters occurs when the number S of samples $\boldsymbol{z}$ is not large enough

Table 1 Classification performance for Bayesian logistic regression (higher is better)

Dataset	Bound in [11]	Proposed
Banana	0.8893 ± 0.0055	0.8893 ± 0.0054
Cancer	0.7119 ± 0.0456	0.7116 ± 0.0454
Heart	0.5528 ± 0.0445	0.5500 ± 0.0476
Solar	0.6449 ± 0.0172	0.6445 ± 0.0167

As a corroboration of the usefulness of our approximation to Bayesian MLR, we compare with the multiclass generalisation of the relevance vector machine (RVM) [19] presented in [16]. We use the multiclass UCI datasets suggested therein. Amongst the two generalisations of the RVM suggested in [16], we use the mRVM_2 version. We also follow the suggestion of [16] concerning the choice of kernels for the different datasets. We set $S = 200$. Table 2 summarises the results of our numerical simulations along with details of the datasets. While mRVM_2 designs a refined probabilistic model to make probabilistic multiclass classification possible, the proposed scheme does not take into account any kind of such knowledge and is still able to deliver competitive performance, in terms of predictive accuracy, as seen in Table 2. The good performance demonstrates both the usefulness and versatility of the proposed method.

3.5 Probabilistic image denoising

In this section, we further demonstrate how the proposed method can take in its stride a change in the model likelihood that complicates computations. The model considered here is the ubiquitous probabilistic principal component analysis (PPCA) introduced in [20]. PPCA assumes that the observed high-dimensional data $\boldsymbol{y} \in \mathbb{R}^d$ are manifestations of low-dimensional latent variables $\boldsymbol{x} \in \mathbb{R}^q$, under a linear mapping expressed by a matrix $\boldsymbol{W} \in \mathbb{R}^{d \times q}$ and an offset $\boldsymbol{\xi} \in \mathbb{R}^d$, corrupted by Gaussian noise $\boldsymbol{\epsilon}$:

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\xi} + \boldsymbol{\epsilon}. \tag{18}$$

PPCA formulates a computationally amenable linear-Gaussian model which allows integrating out the latent variables $\boldsymbol{x}$ and obtaining the marginal log-likelihood. Estimating $\boldsymbol{W}$ and $\boldsymbol{\mu}$ follows by maximising the marginal log-likelihood [20]. Various works extend PPCA by replacing the noise model with other choices, e.g. [2] uses the Student's t distribution, to deal with different types of noise. A recent interesting suggestion is the choice of the Cauchy density as the noise model [22], albeit in a non-probabilistic formulation. The Cauchy density with location x_0 and scale $\gamma > 0$ parameters reads:

$$\left(\pi\gamma\left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]\right)^{-1}. \tag{19}$$

Choosing the Cauchy density as the noise model leads to a version of PPCA where the marginal log-likelihood is no longer tractable and so the latent variables $\boldsymbol{x}$ cannot be integrated out. This is simply because the prior on $\boldsymbol{x}$ is not conjugate to the Cauchy likelihood. However, the proposed method can be used to approximate this intractable marginal log-likelihood. Formally, we specify the following Cauchy-PPCA model:

- Likelihood: $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\xi}, \gamma)$
$$= \prod_{n=1}^{N}\left(\pi\gamma\left[1 + \left(\frac{\boldsymbol{y}_n - \boldsymbol{W}\boldsymbol{x}_n - \boldsymbol{\xi}}{\gamma}\right)^2\right]\right)^{-1}.$$
- Prior: $p(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X}|\boldsymbol{0}, \boldsymbol{I}_N)$.
- Postulated posterior: $q(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}})$.

Parameters $\boldsymbol{W}$, $\boldsymbol{\xi}$ and γ are obtained by gradient optimisation of the proposed lower bound $\mathcal{L}_{(FS)}$.

We applied the original PPCA [20] and the proposed Cauchy-PPCA on a task concerning the denoising of images that have undergone pixel corruption. The aim here is to show the ease with which the proposed method can accommodate a change in the specification (i.e. change from normal to Cauchy likelihood) and deliver a well-performing model.

The data $\boldsymbol{Y}$ are 2,414 face images of 38 individuals obtained from the Extended Yale B Database [10]. There are 64 images per individual under 9 poses and 64

Table 2 Classification performance for Bayesian multiclass classification (higher is better)

Dataset	C	Suggested Kernel	mRVM_2	Proposed
Ecoli	8	Gaussian	0.855 ± 0.047	0.852 ± 0.060
Glass	6	Polynomial	0.581 ± 0.143	0.667 ± 0.112
Iris	3	Gaussian	0.913 ± 0.071	0.947 ± 0.053
Wine	3	Linear	0.959 ± 0.062	0.976 ± 0.030
Soybean (small)	4	Linear	1.000 ± 0.000	1.000 ± 0.000
Vehicle	4	Polynomial	0.481 ± 0.054	0.539 ± 0.103
Balance	3	Polynomial	0.931 ± 0.044	0.947 ± 0.029
Crabs	4	Linear	0.915 ± 0.071	0.950 ± 0.033

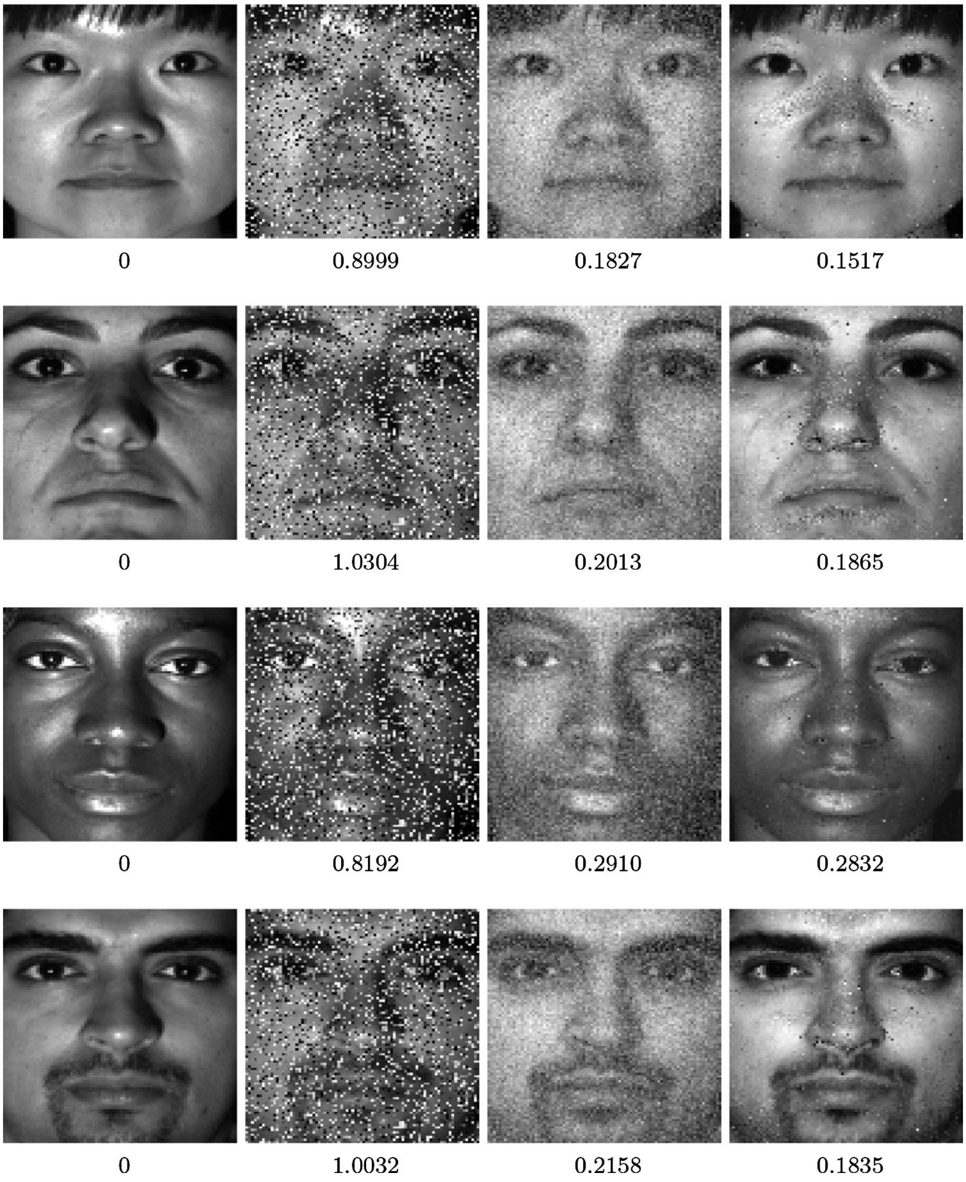


Fig. 5 *Columns* from *left* to *right*: original image, corrupted image, reconstruction by PPCA and reconstruction by Cauchy-PPCA. *Columns two* to *four*: below each image we quote its distance to the original image in the first column (lower is better), i.e. the quality of reconstruction

illumination conditions. The images are grayscale images whose pixels have values between 0 and 255. We rescale the images to 96 × 84 pixels. Hence, $d = 96 \times 84 = 8064$ and we set $q = 2$, i.e. both PCA schemes project the images to a latent space of dimension equal to 2. We corrupt 33.33% of the pixels in each image by drawing a new value uniformly in the range $[0, \ldots, 255]$. For each individual, we use half of the corrupted images as the training set and the other half as the test set.

Figure 5 presents results obtained on test images from four individuals. The figure shows from left to right the original and corrupted test image followed by the two reconstructions obtained by PPCA and Cauchy-PPCA, respectively. To quantify the quality of reconstruction, we use the following measure between the original and reconstructed images: $\|\boldsymbol{y}_{\mathrm{orig}} - \boldsymbol{y}_{\mathrm{rec}}\|^2 / \|\boldsymbol{y}_{\mathrm{orig}}\|^2$. This measure is quoted below each image. The results in Fig. 5 evidently show that Cauchy-PPCA achieves better denoising levels than PPCA. In actual fact, in our numerical experiments we found that Cauchy-PPCA outperformed PPCA on all 38 individuals.

The present numerical experiment demonstrates the versatility of the proposed method in how it can easily extend PPCA to incorporate a Cauchy likelihood. This is achieved without exploiting any particular knowledge pertaining to the probabilistic specification of the model.

3.6 Bayesian inference for the stochastic model by Boore

In this section, we apply the proposed scheme on a geophysical model called the stochastic model. The stochastic model, due to Boore [7], is used to predict ground motion at a given site of interest caused by an earthquake. Ground motion is simply the shaking of the earth and is a fundamental quantity in estimating the seismic hazard of structures. From a physical point of view, the stochastic model describes the radiated source spectrum and its amplitude changes in the frequency domain due to wave propagation from the source to the site of interest. The inputs to the stochastic model are the distance R of the site of interest to the seismic source, the magnitude M_w of the earthquake, and the frequency f of ground motion. The stochastic model, in its simple form, has a parameter associated with the seismic source known in the literature as stress parameter ($\Delta\sigma$), two parameters associated with the path attenuation called geometrical spreading (η) and quality factor (Q), and one more parameter associated with the site called near-surface attenuation (κ_0). All aforementioned parameters are bounded within a physical range. In the case of multiple seismic sources, each source is associated with its own distinct stress parameter. The scalar output of the model y is the mean Fourier amplitude of the ground motion. The type of ground motion we consider here is acceleration. We denote the stochastic model as a function $y = g(M_w, R, f; \boldsymbol{w})$, where $\boldsymbol{w} = [\Delta\sigma_1, \ldots, \Delta\sigma_E, \eta, Q, \kappa_0]$, in which E is the number of seismic sources. This situation is depicted in Fig. 6. We refer the interested reader to [7] for more details. Estimating the posterior uncertainty of the model parameters is important in seismic hazard analysis, as the propagation of uncertainty in the parameters can have an impact on the estimated hazard curve [17]. A discussion of how these posteriors can be utilised in later stages of seismic hazard analysis is beyond the scope of this work.

We specify the model by choosing the following densities:

- Likelihood: $p(\boldsymbol{Y}|\mathcal{D}, \boldsymbol{w}) = \prod_{n=1}^{N} \mathcal{N}(y_n | g(M_{wn}, R_n, f_n; \boldsymbol{w}), \sigma^2)$.
- Flat prior: $p(\boldsymbol{w}) \propto 1$.
- Postulated posterior: $q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}})$.

In contrast to the previous applications, here we choose a very flat prior. Ground motion data inputs are denoted by $\mathcal{D}$ and targets by $\boldsymbol{Y}$.

We performed experiments on a subset of the recently compiled RESORCE database [1] which is a pan-European database of strong motion recordings. In particular, we focused on data records originating from a station in the region of L'Aquilla for $E = 8$ seismic sources. Hence, the total number of free model parameters in $\boldsymbol{w}$ is 11. We experimented with varying numbers of data records, $N \in \{100, 200, 500, 1000\}$, to test the robustness of the Laplace and the proposed approximation in scenarios of limited data. Such situations arise in geophysical studies when data recordings are incomplete due to distortions in frequencies caused by instrumentation errors. The performance of Laplace and the proposed scheme was evaluated as the prediction error on test sets. Both schemes were run ten times, and each run involved a new random realisation of the training and testing set. Parameter S was set to 1000 for all experiments in this section. The predictive distribution for the proposed scheme was approximated using a Monte Carlo estimate. We drew 200 parameter samples from the Gaussian fitted posterior q and estimated performance on the testing set as the average of the mean-squared error under each parameter sample. The results are reported in Table 3.

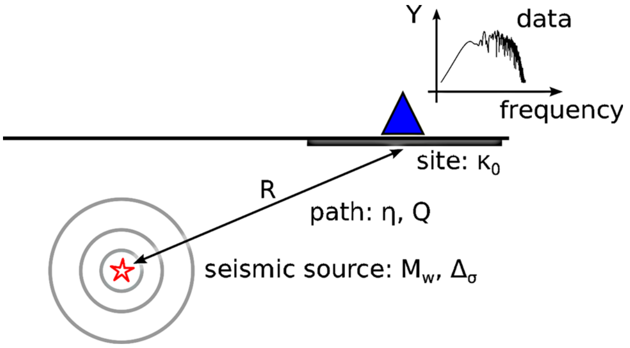


Fig. 6 Physical setting of seismic wave propagation from the source to the site of interest. Recorded at the site is the signal Fourier amplitude against frequency of ground motion

Table 3 Prediction error for ground motion problem (lower is better)

N	Laplace	Proposed
100	0.8550 ± 0.4008	0.6040 ± 0.0307
200	0.7411 ± 0.4925	0.5776 ± 0.0403
500	0.6926 ± 0.4790	0.5496 ± 0.0451
1000	0.5395 ± 0.0230	0.5323 ± 0.0273

The results show that the proposed approximation fares better than Laplace, although at $N = 1000$ the performances are virtually identical. For lower N, however, the Laplace approximation exhibits much higher variance than the proposed scheme.

4 Discussion and conclusion

We have presented a scheme for Bayesian variational inference that is applicable in cases where the likelihood function renders more standard approaches difficult. The scheme is conceptually simple, as it relies on a simple Monte Carlo average of the intractable part of the variational lower bound, see Eq. (5), and the re-introduction of the variational parameters resulting in the objective of Eq. (11). The scheme can thus be generally applied to other models where variational inference is difficult, requiring only the gradients of the log-likelihood function with respect to the parameters.

In the numerical experiments, we have shown that (a) the proposed scheme stands in close agreement with exact inference in Bayesian linear regression, (b) it performs on par in classification tasks against methods that design bespoke model formulations, (c) it fares better than the Laplace approximation in a number of cases, and (d) it is very versatile and can be applied to a variety of problems.

Future work will address the relationship of our approach to variational approaches [9, 18] that provide alternative ways to compute improved Gaussian approximations to intractable posteriors relative to the Laplace approximation. Another aspect concerns ways to cope with very large problems that would require a large number of samples S to obtain a sufficiently accurate approximation in Eq. (8). A natural choice would be to turn the scheme in Algorithm 1 into a recursive stochastic optimisation scheme [12] that employs small sample sets computed at each iteration, akin to stochastic gradient-based large-scale empirical risk minimisation [8]. These two approaches should not be confused, however. The latter employs subsampling of the data $(\boldsymbol{X}, \boldsymbol{Y})$ whereas our scheme generates samples based on current parameter estimates of the approximate posterior. Clearly, our scheme could incorporate sequential subsampling of the data as well. The problem of proving convergence of such an overall stochastic approximation approach in a suitable sense [12] seems to be open.

Acknowledgments The RESORCE database [1] was used in this work with the kind permission of the SIGMA project. [8] N. Gianniotis was partially funded by the BMBF project “Potsdam Research Cluster for Georisk Analysis, Environmental Change and Sustainability”. C. Molkenthin and S. S. Bora were funded by the graduate research school GeoSim of the Geo.X initiative.[9]

References

1. Akkar S, Sandikkaya M, Senyurt M, Azari Sisi A, Ay BO, Traversa P, Douglas J, Cotton F, Luzi L, Hernandez B, Godey S (2014) Reference database for seismic ground-motion in Europe (RESORCE). Bull Earthq Eng 12(1):311–339
2. Archambeau C, Delannay N, Verleysen M (2006) Robust probabilistic projections. In: Proceedings of the 23rd international conference on machine learning. ACM, pp 33–40
3. Azzalini A (2005) The skew-normal distribution and related multivariate families. Scand J Stat 32(2):159–188
4. Barber D, Bishop CM (1998) Ensemble learning in Bayesian neural networks. In: Generalization in neural networks and machine learning. Springer, pp 215–237
5. Beal MJ (2003) Variational algorithms for approximate bayesian inference. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London
6. Bishop CM (2006) Pattern recognition and machine learning. Springer, New York
7. Boore DM (2003) Simulation of ground motion using the stochastic method. Pure Appl Geophys 160(3–4):635–676
8. Bottou L (2012) Stochastic gradient tricks. In: Neural networks, tricks of the trade, reloaded, LNCS, vol 7700. Springer
9. Cseke B, Heskes T (2011) Approximate marginals in latent Gaussian models. J Mach Learn Res 12:417–454
10. Georghiades AS, Belhumeur PN, Kriegman D (2001) From few to many: illumination cone models for face recognition under variable lighting and pose. Pattern Anal Mach Intell IEEE Trans 23(6):643–660
11. Jaakkola T, Jordan M (2000) Bayesian parameter estimation via variational methods. Stat Comput 10(1):25–37
12. Kushner H, Yin G (2003) Stochastic approximation and recursive algorithms and applications, 2nd edn. Springer, New York
13. Møller MF (1993) A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw 6(4):525–533
14. Opper M, Archambeau C (2009) The variational Gaussian approximation revisited. Neural Comput 21:786–792
15. Petersen KB, Pedersen MS (2012) The matrix cookbook. http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3274
16. Psorakis I, Damoulas T, Girolami MA (2010) Multiclass relevance vector machines: sparsity and accuracy. IEEE Trans Neural Netw 21(10):1588–1598
17. Reiter L (1991) Earthquake hazard analysis: issues and insights. Colombia University Press, New York
18. Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc: Ser B 71(2):319–392
19. Tipping ME (2001) Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res 1:211–244
20. Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. J R Stat Socy: Ser B 61(3):611–622
21. Tzikas D, Likas C, Galatsanos N (2008) The variational approximation for Bayesian inference. Signal Process Mag IEEE 25(6):131–146
22. Xie P, Xing EP (2014) Cauchy principal component analysis. CoRR abs/1412.6506. http://arxiv.org/abs/1412.6506. Accessed 25 Mar 2015

[8] http://www.projet-sigma.com.

[9] http://www.geo-x.net.