Machine Learning (2021) 110:1105–1143
https://doi.org/10.1007/s10994-021-05971-2

Pseudo-marginal Bayesian inference for Gaussian process latent variable models

C. Gadd[2] · S. Wade[3,4] · A. A. Shah[1]

Received: 19 July 2019 / Revised: 23 December 2020 / Accepted: 5 March 2021 / Published online: 18 April 2021

Abstract
A Bayesian inference framework for supervised Gaussian process latent variable models is introduced. The framework overcomes the high correlations between latent variables and hyperparameters by collapsing the statistical model through approximate integration of the latent variables. Using an unbiased pseudo estimate for the marginal likelihood, the exact hyperparameter posterior can then be explored using collapsed Gibbs sampling and, conditional on these samples, the exact latent posterior can be explored through elliptical slice sampling. The framework is tested on both simulated and real examples. When compared with the standard approach based on variational inference, this approach leads to significant improvements in the predictive accuracy and quantification of uncertainty, as well as a deeper insight into the challenges of performing inference in this class of models.

Keywords Gaussian process · Latent variable model · Approximate inference · Variational · Collapsed Gibbs sampling

1 Introduction

Statistical Bayesian approaches to regression can be used to model nonlinear functions between inputs and outputs in a simple, flexible, nonparametric and probabilistic manner. In Gaussian process (GP) model approaches, the latent function is assumed to be a realisation of a Gaussian stochastic process. A GP is a family of random variables (with a common underlying probability space) ranging over an index set, such that any finite subset of the random variables has a joint Gaussian distribution with consistent parameters. A realisation of the GP is a deterministic function of the index variable. A GP is fully specified

Editor: Pradeep Ravikumar.

✉ A. A. Shah
akeelshah@cqu.edu.cn

[1] Key Laboratory of Low-grade Energy Utilization Technologies and Systems, Chongqing University, Chongqing 400044, China

[2] Warwick Centre for Predictive Modelling, University of Warwick, Coventry CV4 7AL, UK

[3] Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

[4] Present Address: School of Mathematics, University of Edinburgh, Edinburgh EH9 3FD, UK

by a mean function and a symmetric positive-definite covariance (kernel) function, which encapsulate any *a-priori* knowledge and/or assumptions in relation to the target function. The freedom to choose from a range of possible kernel functions introduces a degree of flexibility in the assumed complexity and smoothness of this underlying target function.

The Gaussian process latent variable model (GPLVM) introduced in Lawrence (2004) is a hierarchical model originally used to extend GPs to the (unsupervised) learning task of non-linear dimension reduction, in which the inputs are unobserved latent variables. The model places independent GP priors over the mappings from the latent space to each component in the observed output space. In Lawrence (2005), a Gaussian prior is placed over the latent variables, which are optimized according to their maximum a-posteriori (MAP) estimates (equivalent to the maximum likelihood estimates with L_2 regularisation). To capture uncertainty in the latent variables, Titsias and Lawrence (2010) developed a variational method for GPLVMs.

In this paper we consider the supervised version (sGPLVM) of the GPLVM, in which a GP prior is placed over latent variables *indexed by known and observable inputs*. This model was studied in a dynamical setting in Damianou et al. (2011), in which the only input is time, using a variational expectation maximization (VEM) approach to determine point estimates of the hyperparameters. Our contribution is a novel framework for robust, *fully Bayesian* inference for the sGPLVM. Towards this end, it is natural to explore the posterior distribution of the hyperparameters and latent variables with Markov Chain Monte Carlo (MCMC), but strong correlations between the hyperparameters and latent variables leads to low efficiency and poor mixing (Betancourt and Girolami 2015). A method that can break these correlations is required, which we obtain by using a *pseudo-marginal* scheme that approximately integrates out the latent variables. A major motivation for developing this method is the need to quantify hyperparameter uncertainty, and more accurately quantify latent uncertainty.

Indeed, variational methods make strong assumptions of independence, on the forms of distributions and, dependent on the choice of divergence, they necessarily under- or over-estimate the variance (Blei et al. 2017). When correlations between the latent variables (within variational factorisations) are large, and when these factors are over a non-trivial number of dimensions, this variational approximation becomes increasingly poor. Moreover, in the variational expectation-maximization (VEM) setting, crucial hyperparameters determining the noise level and smoothness of the latent functions are fixed to approximate maximum marginal likelihood (ML) estimates, obtained by optimizing the variational lower bound to the marginal likelihood. In this paper, the effects and performance of the variational approximation are studied in illustrative examples based on simulated and real data. These examples also demonstrate the benefits of a fully Bayesian inference on predictive performance when compared with VEM using the projected process approximation. Moreover, they shed light on the types of problems in which approximate ML estimates are poor.

The rest of this paper is structured as follows. Section 2 introduces the model. The state-of-the-art variational inference for the model is discussed in Sect. 3. A pseudo-marginal Monte Carlo scheme for fully Bayesian inference is then proposed in Sect. 4. In Sect. 4.2, an elliptical slice sampling scheme for the latent variables is described. This scheme is necessary to compute the predictions in Sect. 4.3. Section 5 demonstrates the advantages of the proposed inference framework over the variational method through a simulated and a real example. A discussion of the numerical computational cost is given in Sect. 6 and concluding remarks are provided in Sect. 7.

2 Statistical (supervised) model

Consider a set of N inputs, given by the rows of $\mathbf{X} := [\mathbf{x}_1 \dots \mathbf{x}_N]^T \in \mathbb{R}^{N \times k_x}$, and corresponding known outputs, given by the rows of $\mathbf{Y} := [\mathbf{y}_1 \dots \mathbf{y}_N]^T \in \mathbb{R}^{N \times k_y}$. Let, $y_{n,d}$, $d = 1, \dots, k_y$, denote the d-th coordinate (feature) of the output $\mathbf{y}_n$, $n = 1, \dots N$, and let y_d denote the d-th coordinate of the general output $\mathbf{y}$, corresponding to the general input $\mathbf{x}$. Further, let $\mathbf{y}_{:,d}$ denote the d-th column of $\mathbf{Y}$, i.e., the vector with components given by the d-th feature of each of the N samples. This compact matrix notation is used throughout. Consider also a set of unknown latent variable representations $\mathbf{Z} := [\mathbf{z}_1 \dots \mathbf{z}_N]^T \in \mathbb{R}^{N \times k_z}$ of the outputs $\mathbf{y}_n$, with $k_z \ll k_y$.

The assumed model is $y_d = f_d(\mathbf{z}) + \epsilon_d$, in which the noise terms ϵ_d are independent and (identically) normally distributed across d as $\epsilon_d \overset{iid}{\sim} \mathcal{N}(\epsilon_d|0, \beta^{-1})$. We could also write the model as $\mathbf{y} = \mathbf{f}(\mathbf{z}) + \boldsymbol{\epsilon}$, where $\mathbf{f}(\mathbf{z}) := (f_1, (\mathbf{z}), \dots, f_{k_y}(\mathbf{z}))^T$ is a latent (vector) function and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_{k_y})^T$. Independent GP priors (indexed by $\mathbf{z}$) are placed over the functions $f_d(\mathbf{z})$, namely $f_d(\mathbf{z}) \sim \mathcal{GP}(0, k_f(\mathbf{z}, \mathbf{z}'; \boldsymbol{\theta}))$, where $k_f(\mathbf{z}, \mathbf{z}'; \boldsymbol{\theta})$ is the common covariance/kernel function, with hyperparameters $\boldsymbol{\theta}$. The notation $\mathcal{GP}(\cdot, \cdot)$ in which the first argument is the mean function and the second is the covariance function is used throughout. The latent function values can be collected as the rows in a matrix $\mathbf{F} \in \mathbb{R}^{N \times k_y}$, with columns $\mathbf{f}_{:,d} = (f_{1,d}, \dots, f_{N,d})^T$, in which we use the notation $f_{n,d} = f_d(\mathbf{z}_n)$. By the independence assumption:

$$p(\mathbf{F}|\mathbf{Z}, \boldsymbol{\theta}) = \prod_{d=1}^{k_y} p(\mathbf{f}_{:,d}|\mathbf{Z}, \boldsymbol{\theta}),$$

and by the properties of GPs, we have $p(\mathbf{f}_{:,d}|\mathbf{Z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}_{:,d}|\mathbf{0}, \mathbf{K}_f)$, in which $\mathbf{K}_f$ is the N by N kernel (covariance) matrix with n, n'-th entry $k_f(\mathbf{z}_n, \mathbf{z}_{n'}; \boldsymbol{\theta})$. From $p(\mathbf{Y}, \mathbf{F}|\mathbf{Z}, \beta, \boldsymbol{\theta}) = p(\mathbf{Y}|\mathbf{F}, \beta)p(\mathbf{F}|\mathbf{Z}, \boldsymbol{\theta})$ we obtain the marginal likelihood:

$$\begin{aligned} p(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta}, \beta) &= \int p(\mathbf{Y}|\mathbf{F}, \beta)p(\mathbf{F}|\mathbf{Z}, \boldsymbol{\theta})d\mathbf{F} \\ &= \int \prod_{d=1}^{k_y} \prod_{n=1}^{N} p(y_{n,d}|f_{n,d}, \beta)p(\mathbf{f}_{:,d}|\mathbf{Z}, \boldsymbol{\theta})d\mathbf{F} \\ &= \prod_{d=1}^{k_y} \mathcal{N}(\mathbf{y}_{:,d}|\mathbf{0}, \mathbf{K}_f + \beta^{-1}\mathbf{I}_N), \end{aligned} \tag{1}$$

in which $p(\mathbf{y}_{:,d}|\mathbf{f}_{:,d}, \beta) = \mathcal{N}(\mathbf{y}_{:,d}|\mathbf{f}_{:,d}, \beta^{-1}\mathbf{I}_N)$ by virtue of the noise model. The model for the latent function coordinates $z_j(\mathbf{x})$ comes in the form of independent GP priors $z_j(\mathbf{x}) \sim \mathcal{GP}(0, k_z(\mathbf{x}, \mathbf{x}'; \boldsymbol{\sigma}))$, $j = 1, \dots, k_z$, with common covariance function $k_z(\mathbf{x}, \mathbf{x}'; \boldsymbol{\sigma})$ and kernel hyperparameters $\boldsymbol{\sigma}$. As a consequence:

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\sigma}) = \prod_{j=1}^{k_z} p(\mathbf{z}_{:,j}|\mathbf{X}, \boldsymbol{\sigma}) = \prod_{j=1}^{k_z} \mathcal{N}(\mathbf{z}_{:,j}|\mathbf{0}, \mathbf{K}_z), \tag{2}$$

where $\mathbf{K}_z$ is the N by N the kernel matrix, with n, n'-th entry equal to $k_z(\mathbf{x}_n, \mathbf{x}_{n'}; \boldsymbol{\sigma})$. The joint density for the observed data and latent variables is:

$$p(\mathbf{Y}, \mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \beta) = p(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta}, \beta)p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\sigma}).$$

The Bayesian model is completed by a prior on the precision β and both sets of kernel hyperparameters $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$. The prior is assumed to factorize as follows:

$$p(\beta, \boldsymbol{\theta}, \boldsymbol{\sigma}) = p(\beta) \prod_i p(\theta_i) \prod_{i\}} p(\sigma_{i\}}).$$

This model was studied in this supervised (dynamic) setting in Damianou et al. (2011). It can further be viewed as a deep GP model (Damianou and Lawrence 2013) with a single hidden layer.

3 Variational marginalization of the latent variables

In many Bayesian models, including GP-based models such as the GPLVM, posterior inference is sensitive to the choice of hyperparameters. There are generally two approaches to ameliorate this sensitivity (in the absence of strong prior knowledge): hierarchical Bayes, with a hyperprior assigned to account for uncertainty in the hyperparameters, and empirical Bayes, in which plug-in estimates of the hyperparameters are used. In empirical Bayes, these estimates are typically based on maximizing the marginal likelihood. For the GPLVM, however, computing the marginal likelihood requires integration with respect to the latent variables, which is analytically intractable since they appear nonlinearly in the kernel matrix. A major advance was made in Titsias and Lawrence (2010) by using a VEM approach, assuming a Gaussian variational posterior and utilising sparse GPs to obtain a closed form lower bound to the marginal likelihood. In an EM fashion, this lower bound can then be optimized with respect to the hyperparameters to obtain approximate type II maximum marginal likelihood estimates. This procedure can be generalised to the supervised case and is described fully in "Appendix A".

Considering the E-step of the VEM algorithm in isolation, the basic approach consists of using a proxy variational distribution over the latent variables in order to approximate the posterior distribution. The variational parameters of this distribution are chosen to minimise the Kullback-Leibler (KL) divergence between the proxy distribution and the posterior. When performing the expectation step in isolation (by optimizing the variational parameters conditional on the model hyperparameters), a latent variable posterior approximation is obtained with the following factorised Gaussian form:

$$p(\mathbf{Z}|\mathbf{Y}, \mathbf{X}, \sigma, \theta, \beta) \approx q(\mathbf{Z}) = \prod_{j=1}^{k_z} \mathcal{N}\big(\mathbf{z}_{:,j}|\boldsymbol{\mu}_j, \mathbf{S}_j\big), \tag{3}$$

where $\mathbf{S}_j$ is a diagonal $N \times N$ covariance matrix. Conditional dependence on the data and hyperparameters $(\boldsymbol{\sigma}, \boldsymbol{\theta}, \beta)$ enters through optimization of the variational parameters $\boldsymbol{\mu}_j \in \mathbb{R}^N$ and $\mathbf{S}_j \in \mathbb{R}^{N\times N}$ according to the evidence lower bound on the marginal likelihood, which is derived in "Appendix A". This evidence lower bound can then be used to perform the M-step to obtain new hyperparameter point-estimates, with the process repeating until convergence. However, this particular approach of inference by optimization comes with a number of caveats, as previously discussed.

4 Pseudo-marginal Monte Carlo for the GPLVM

This section introduces a novel framework for fully Bayesian inference of the supervised GPLVM. Model hyperparameters defining the covariance function have important implications for smoothness, complexity, and relevance of the inputs. It is common to obtain these parameters by optimizing an approximate ML (type II ML estimate), using gradient-based optimization. However, the likelihood as a function of these parameters is non-convex, and consequently practitioners often find that the optimization is highly dependent on initialisation, with no guarantee of a satisfactory local optimum (Bitzer and Williams 2010). This is particularly profound when the data set is small or has a low signal-to-noise ratio.

Moreover, it must be emphasised that the hyperparameter estimates in Sect. 3 do not optimize the marginal likelihood, but instead optimize a lower bound to the marginal likelihood. The consequences of this in simple examples was shown in Turner and Sahani (2011). Specifically, they found that it is not important for the lower bound to be as tight as possible to the marginal likelihood, but that it is equally tight everywhere. If this is not the case, the effect is to push estimates away from peaks in the likelihood and towards regions where the bound is tighter. Another interesting conclusion is that biases in the hyperparameter estimates increase considerably as the number of hyperparameters increases.

Additionally, while variational methods can substantially reduce computational time, this comes at the cost of strong assumptions and considerable bias. Assumptions are often made regarding the forms of distributions (e.g., they can be factorised), and/or highly simplified approximations of true posterior distributions are employed. Dependent upon the choice of divergence, variational methods also under- or over-estimate variance (Blei et al. 2017). This is particularly true when the posterior is highly correlated but the proxy distribution has a factorised form, or when the posterior is a mixture of Gaussians with well separated modes and the proxy is a single Gaussian. The reverse, however, may also be true. For example, when approximating a mixture of Gaussians with poorly separated modes with a single Gaussian (see Turner and Sahani 2011 for more details). In the case of the GPLVM, the quality of this approximation can easily be verified by sampling the true latent variable posterior (conditioned on the hyperparameters) with elliptical slice sampling. This is later discussed in more detail.

The above considerations motivate a fully Bayesian framework for inference with the sGPLVM. This Bayesian approach naturally regularises against overfitting by penalising unnecessary model complexity. Moreover, quantification of the hyperparameter uncertainty yields a more informative quantification of the uncertainty in the predictions (by integrating over the hyperparameters). The proposed framework overcomes the high correlations between latent variables and hyperparameters by using an unbiased pseudo estimate for the marginal likelihood that approximately integrates over the latent variables in a collapsed Gibbs sampler. This is used to construct a Markov Chain to explore the posterior of the hyperparameters; these samples can then be used alongside ESS to sample the latent variables. This overcomes issues associated with optimization of the hyperparameters, and avoids the strong assumptions of variational methods.

This framework is referred to as PM (Pseudo-Marginal) throughout this article. It is demonstrated on both simulated and real examples, which reveal improved accuracy and improved uncertainty quantification in predictions when compared with those obtained using the VEM approach. Another important contribution is to shed light upon situations in which the variational scheme works well and, conversely, when it does not, by considering simulated scenarios that are increasingly misspecified by the sGPLVM.

4.1 Collapsed pseudo-marginal Gibbs sampling

The natural choice to explore the posterior of the latent variables and hyperparameters is a Gibbs sampling algorithm, which alternates between sampling and fixing the latent variables and hyperparameters. In the GPLVM family of models the latent parameters and hyperparameters are strongly coupled, leading to sharp peaks in the posterior when latent variables are fixed. This results in poor MCMC mixing and slow convergence rates (Filippone and Girolami 2014). A method that can break these correlations is required.

Although analytical integration of the latent variables $\mathbf{Z}$ is intractable, since they appear nonlinearly in the kernel matrix $\mathbf{K_f}$, the correlation between the latent variables and hyperparameters can be broken by approximately integrating over the latent variables *via* a pseudo-marginal Monte Carlo scheme. The results of Andrieu and Roberts (2009) and Beaumont (2003) reveal that an unbiased estimate of the marginal likelihood can be used to sample from the correct hyperparameter posterior distribution.

Here importance sampling is used to obtain the unbiased approximation to the marginal likelihood based on the approximate distribution $q(\mathbf{Z}|\boldsymbol{\sigma}, \boldsymbol{\theta}, \beta) \approx p(\mathbf{Z}|\mathbf{Y}, \mathbf{X}, \boldsymbol{\sigma}, \boldsymbol{\theta}, \beta)$, which in this context is known as the *proposal, biased* or *sampling distribution*. Drawing Q importance samples, the unbiased estimate of the marginal is:

$$\tilde{p}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\sigma}, \boldsymbol{\theta}, \beta) \simeq \frac{1}{Q} \sum_{q=1}^{Q} \frac{p(\mathbf{Y}|\mathbf{Z}^{(q)}, \boldsymbol{\theta}, \beta) p(\mathbf{Z}^{(q)} \mid \mathbf{X}, \boldsymbol{\sigma})}{q(\mathbf{Z}^{(q)}|\boldsymbol{\sigma}, \boldsymbol{\theta}, \beta)}, \tag{4}$$

where $\mathbf{Z}^{(q)} \overset{iid}{\sim} q(\mathbf{Z}|\boldsymbol{\sigma}, \boldsymbol{\theta}, \beta)$, and $p(\mathbf{Y}|\mathbf{Z}^{(q)}, \boldsymbol{\theta}, \beta)$ and $p(\mathbf{Z}^{(q)}|\mathbf{X}, \boldsymbol{\sigma})$ are the GP models given by (1) and (2) respectively. For the proposal distribution $q(\mathbf{Z}|\boldsymbol{\sigma}, \boldsymbol{\theta}, \beta)$, the approximate variational posterior of Sect. 3 is utilised. In this setting, the hyperparameters $(\boldsymbol{\sigma}, \boldsymbol{\theta}, \beta)$ are fixed at the required sample and only the E-step of the variational scheme is performed, to optimize the lower bound with respect to the variational parameters. Importantly this avoids constraints on the tightness of the lower bound required to obtain good hyperparameter estimates. This pseudo-marginal can now be used to sample from the posterior of the hyperparameters in a Metropolis-Hastings (MH) algorithm.

To improve mixing, the set of hyperparameters $\boldsymbol{\xi} = (\boldsymbol{\sigma}, \boldsymbol{\theta}, \beta)$ are split into R disjoint subsets, $\boldsymbol{\xi}_r$, $r = 1, \ldots, R$. Each block of parameters can then be sampled from the its full conditional in a Metropolis-Hastings within the Gibbs algorithm. Specifically, the block $\boldsymbol{\xi}_r$ is updated with a random walk based on a transformation $\boldsymbol{\eta}_r = t_r(\boldsymbol{\xi}_r)$ to ensure full support on the real space of appropriate dimension, and a multivariate normal proposal distribution is used for the transformed parameter: $\pi(\boldsymbol{\eta}_r'|\boldsymbol{\eta}_r) \sim \mathcal{N}(\boldsymbol{\eta}_r, \boldsymbol{\Sigma}_r)$. This gives the proposal distribution $\pi(\boldsymbol{\xi}_r'|\boldsymbol{\xi}_r) = |\partial t_r / \partial \boldsymbol{\xi}_r(\boldsymbol{\xi}_r')| \pi(\boldsymbol{\eta}_r'|\boldsymbol{\eta}_r)$ in the original parameter space. The acceptance probability for a move from $\boldsymbol{\xi}_r$ to $\boldsymbol{\xi}_r'$ is therefore:

$$\tilde{\alpha}(\boldsymbol{\xi}_r, \boldsymbol{\xi}_r') = \min\left[1, \frac{\tilde{p}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\xi}') p(\boldsymbol{\xi}_r')}{\tilde{p}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\xi}) p(\boldsymbol{\xi}_r)} \frac{|\partial t_r / \partial \boldsymbol{\xi}_r(\boldsymbol{\xi}_r)|}{|\partial t_r / \partial \boldsymbol{\xi}_r(\boldsymbol{\xi}_r')|}\right], \tag{5}$$

where $\boldsymbol{\xi}'$ denotes the set of hyperparameters with the rth block updated to $\boldsymbol{\xi}_r'$.

In addition, we employ a variant of the adaptive random walk algorithm of Haario et al. (2001), in which the proposal covariance matrix $\boldsymbol{\Sigma}_r$ is adapted to approximate the target distribution's covariance matrix multiplied by a constant s_{d_r}. Following Haario et al. (2001), this constant is chosen to be $s_{d_r} = 2.38^2/d_r$ where d_r is the dimension of the block. The algorithm then begins with an initial proposal covariance matrix for each block, and

after g_0 iterations this matrix is updated by the sample covariance, with a small positive constant along the diagonal. The full procedure is outlined in Algorithm 1.

Note that when computing the acceptance probability of a move from $\boldsymbol{\xi}_r$ to $\boldsymbol{\xi}_r'$ in (5), the pseudo-marginal $\tilde{p}(\mathbf{Y}|\mathbf{X},\boldsymbol{\xi})$ must be recycled from the previous step to ensure convergence to the posterior (Andrieu and Roberts 2009). In general, this may result in the chain becoming stuck if the pseudo-marginal has high variance and overestimates the marginal likelihood. In order to confirm that this is not the case, multiple chains are run in parallel.

Algorithm 1 Pseudo-marginal adaptive MH in Gibbs.

Require: Number of states before adaptation: g_0. Burn in period: n_0.

for $g = 1, 2, \ldots$ **do**

Set $\boldsymbol{\xi}^{(g)} = \boldsymbol{\xi}^{(g-1)}$

for each $\boldsymbol{\xi}_r$, $r = 1, \ldots, R$ **do**

Sample $\boldsymbol{\eta}_r' = \boldsymbol{\eta}_r^{(g)} + \boldsymbol{\epsilon}_g$ where $\boldsymbol{\epsilon}_g \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_r^{(g-1)})$.

Find the unbiased approximation $\tilde{p}(\mathbf{Y}|\mathbf{X},\boldsymbol{\xi}')$ using importance sampling (4).

Set:

$$\boldsymbol{\xi}_r^{(g)} = \begin{cases} \boldsymbol{\xi}_r' & \text{with probability } \tilde{\alpha}(\boldsymbol{\xi}^{(g)}, \boldsymbol{\xi}') \\ \boldsymbol{\xi}_r^{(g-1)} & \text{with probability } 1 - \tilde{\alpha}(\boldsymbol{\xi}_r^{(g)}, \boldsymbol{\xi}'). \end{cases}$$

if $g > g_0$ **then**

$\boldsymbol{\Sigma}_r^{(g)} = \frac{s_{d_r}}{g-1} \left[\sum_{m=1}^{g} \boldsymbol{\eta}_r^{(m)} \boldsymbol{\eta}_r^{(m)^T} - g\bar{\boldsymbol{\eta}}_r \bar{\boldsymbol{\eta}}_r{}^T \right] + s_{d_r} \epsilon \mathbf{I}.$

end if

end for

return $\boldsymbol{\xi}^{(g)}$ for $g > n_0$.

end for

4.2 Uncollapsing with elliptical slice sampling

Given the hyperparameter posterior samples that are obtained as described in the previous section, samples of the latent variable posterior can be obtained using the ESS algorithm of Murray et al. (2010). These samples are required for the prediction procedure in Sect. 4.3. The target distribution for the sampler is the full conditional of the latent variables:

$$\begin{aligned} p(\mathbf{Z}|\mathbf{Y},\mathbf{X},\boldsymbol{\xi}) &\propto p(\mathbf{Y}|\mathbf{Z},\boldsymbol{\theta},\beta)p(\mathbf{Z}|\boldsymbol{\sigma},\mathbf{X}) \\ &\propto \prod_{d=1}^{k_y} \mathcal{N}\big(\mathbf{y}_{:,d}|\mathbf{0}, \mathbf{K}_\mathbf{f}(\mathbf{Z},\mathbf{Z};\boldsymbol{\theta}) + \beta^{-1}\mathbf{I}_N\big) \prod_{d=1}^{k_z} \mathcal{N}\big(\mathbf{z}_{:,d}|0, \mathbf{K}_\mathbf{z}(\mathbf{X},\mathbf{X};\boldsymbol{\sigma})\big), \end{aligned}$$

and the proposal distribution is given by:

$$\mathbf{Z}' = \mathbf{N}\sin\alpha + \mathbf{Z}\cos\alpha, \qquad \boldsymbol{\nu}_{:,d} \overset{iid}{\sim} \mathcal{N}\big(0, \mathbf{K}_\mathbf{z}\big), \qquad d = 1, \ldots, k_z.$$

This defines a full ellipse passing through the previous state $\mathbf{Z}$ and a prior sample $\mathbf{N} \in \mathbb{R}^{N \times k_z}$, with columns $\boldsymbol{\nu}_{:,d}$, as α varies. This proposal depends on a tuning parameter α which would be chosen a priori under a normal Metropolis-Hastings scheme. The algorithm of Murray et al. (2010) adaptively chooses this tuning parameter using slice sampling. The procedure for sampling $\mathbf{Z}$ using the elliptical slice sampler is given in Algorithm 2.

Algorithm 2 Elliptical slice sampler for the latent variables.

Require: current state $\mathbf{Z}$, and log-likelihood function.

Ensure: new state $\mathbf{Z}'$.

1: Sample: $\boldsymbol{\nu} \sim \prod_{d=1}^{k_z} \mathcal{N}(\boldsymbol{\nu}_{:,d}|0, \mathbf{K}_z)$.

2: Log-likelihood threshold:

$$u \sim \text{Uniform}[0,1], \quad \log h \leftarrow \log p(\mathbf{Y}|\mathbf{Z}; \boldsymbol{\theta}, \beta) + \log u.$$

3: Draw an initial proposal, define bracket on the ellipse:

$$\alpha \sim \text{Uniform}[0, 2\pi], \quad [\alpha_{\min}, \alpha_{\max}] \leftarrow [\alpha - 2\pi, \alpha].$$

4: **while** not returned **do**

5: Propose new latent variables:

$$\mathbf{Z}' \leftarrow \boldsymbol{\nu} \sin\alpha + \mathbf{Z} \cos\alpha.$$

6: **if:** $\log p(\mathbf{Y}|\mathbf{Z}', \boldsymbol{\theta}, \beta) > \log h$ (proposal lies in slice) **then:**

7: Accept: **return** $\mathbf{Z}'$.

8: **else:**

9: Shrink bracket and re-sample step size:

10: **if** $\alpha < 0$ **then:** $\alpha_{\min} \leftarrow \alpha$ **else:** $\alpha_{\max} \leftarrow \alpha$

11: $\alpha \sim \text{Uniform}[\alpha_{\min}, \alpha_{\max}]$.

12: **end while**

4.3 Predictions using MCMC

Predictions can now be made by marginalizing over the posterior samples, without the need for distributional assumptions or point estimates. The marginalized predictive density for a test point $\mathbf{x}_*$ is:

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{Y}, \mathbf{X}) = \int p(\mathbf{y}_*|\mathbf{z}_*, \mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}, \beta) p(\mathbf{z}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{Z}, \boldsymbol{\sigma}) p(\mathbf{Z}, \boldsymbol{\xi}|\mathbf{Y}, \mathbf{X}) d\mathbf{z}_* d\mathbf{Z} d\boldsymbol{\xi}. \quad (6)$$

The second term inside the integral of (6) is the predictive density of the latent variable $\mathbf{z}_*$ given the latent variables, hyperparameters and data, which is given by the noise-free GP predictive density:

$$p\big(\mathbf{z}_*|\mathbf{x}_*,\mathbf{X},\mathbf{Z},\boldsymbol{\sigma}\big)=\prod_{d=1}^{k_z}\mathcal{N}\big(z_{*d}|\mathbf{K}_{\mathbf{z}*}^T\mathbf{K}_{\mathbf{z}}^{-1}\mathbf{z}_{:,d},s_*\big),\tag{7}$$

with $s_*=k_z(\mathbf{x}_*,\mathbf{x}_*;\boldsymbol{\sigma})-\mathbf{K}_{\mathbf{z}*}^T\mathbf{K}_{\mathbf{z}}^{-1}\mathbf{K}_{\mathbf{z}*}$. Here $\mathbf{K}_{\mathbf{z}*}$ is the cross-covariance at the training inputs $\mathbf{X}$ and the test input $\mathbf{x}_*$. Similarly, the first term inside the integral of (6) is the predictive density of the test output $\mathbf{y}_*$ given $\mathbf{z}_*$, the latent variables, hyperparameters and data, which is given by the GP predictive density:

$$p\big(\mathbf{y}_*|\mathbf{z}_*,\mathbf{Z},\mathbf{Y},\boldsymbol{\theta},\beta\big)=\prod_{d=1}^{k_y}\mathcal{N}\big(y_{*d}|\mathbf{A}\mathbf{y}_{:,d},\mathbf{S}+\beta^{-1}\big),$$

where $\mathbf{A}=\mathbf{K}_{\mathbf{f}*}^T(\mathbf{K}_{\mathbf{f}}+\beta^{-1}\mathbf{I}_N)^{-1}$ and $\mathbf{S}=k_f(\mathbf{z}_*,\mathbf{z}_*;\boldsymbol{\theta})-\mathbf{K}_{\mathbf{f}*}^T(\mathbf{K}_{\mathbf{f}}+\beta^{-1}\mathbf{I}_N)^{-1}\mathbf{K}_{\mathbf{f}*}$. Here $\mathbf{K}_{\mathbf{f}*}$ corresponds to cross-covariance between latent function evaluations at $\mathbf{Z}$ and $\mathbf{z}_*$.

The MCMC samples can be used to obtain an approximation to the marginalized predictive density in (6), with asymptotic guarantees as the number of MCMC samples increases. However, the latent variable $\mathbf{z}_*$ cannot be marginalized analytically. Thus, given each sample of the chain $(\boldsymbol{\xi}^{(g)},\mathbf{Z}^{(g)})$, we sample the latent variable $\mathbf{z}_*^{(g)}$ based on its predictive distribution in (7). The predictive density estimate is:

$$p\big(\mathbf{y}_*|\mathbf{x}_*,\mathbf{Y},\mathbf{X}\big)\approx\frac{1}{G}\sum_{g=1}^{G}p\big(\mathbf{y}_*|\mathbf{z}_*^{(g)},\mathbf{Z}^{(g)},\mathbf{Y},\boldsymbol{\theta}^{(g)},\beta^{(g)}\big).\tag{8}$$

Similarly, the posterior mean function can be estimated from:

$$\mathbb{E}\big[\mathbf{y}_*|\mathbf{x}_*,\mathbf{Y},\mathbf{X}\big]\approx\frac{1}{G}\sum_{g=1}^{G}\mathbf{K}_{\mathbf{f}*}^{(g)\,T}(\mathbf{K}_{\mathbf{f}}^{(g)}+\beta^{(g)-1}\mathbf{I}_N)\mathbf{Y}.$$

5 Examples

In this section we present two examples that demonstrate both the improved predictive accuracy and uncertainty quantification of the pseudo-marginal inference framework in comparison to VEM. In Sect. 5.1, a simulated example is presented for three increasingly misspecified cases. In Sect. 5.2, we consider experimental data consisting of measurements of air quality over time.

For all models (in both examples) a squared exponential kernel is chosen to measure correlations in the input and latent spaces, with the addition of white noise for numerical stability:

$$\begin{aligned}k_z\big(x,x';\boldsymbol{\sigma}\big)&=\sigma_S\exp\Big(-\frac{1}{2}\sigma_1(x-x')^2\Big)+\epsilon\delta\big(x,x'\big),\\k_f(\mathbf{z},\mathbf{z}';\boldsymbol{\theta})&=\theta_S\exp\left(-\frac{1}{2}\sum_{d=1}^{k_z}\theta_d(z_d-z_d')^2\right),\end{aligned}\tag{9}$$

where ϵ is a small positive constant and $\delta(\cdot,\cdot)$ is the kronecker-delta function. To ensure identifiability, the magnitude σ_S is fixed to unity throughout. Additionally, in all models, a Gamma prior is placed over all hyperparameters (shown in Table 1) and a log

Table 1 The hyperprior distributions, where Ga is a Gamma distribution

Example	σ_1	σ_2	θ_1	θ_2	θ_S	β
Sinusoidal: case 1	Ga(2, 8)	–	Ga(1.25, 5)	Ga(1.25, 5)	Ga(1.5, 5)	Ga(3, 800)
Sinusoidal: case 2	Ga(1.5, 16)	–	Ga(2, 0.1)	Ga(2, 0.1)	Ga(2, 3)	Ga(3, 800)
Sinusoidal: case 3	Ga(1.5, 16)	–	Ga(2, 0.1)	Ga(2, 0.1)	Ga(2, 3)	Ga(3, 800)
Air quality	Ga(2, 8)	Ga(3, 3)	Ga(1.25, 5)	–	Ga(3, 2)	Ga(2, 8)

transformation in the random walk proposals is used. Each experiment was repeated for different prior parameterisations and it was found that predictive accuracy was not sensitive to the choice of prior. Four chains were run in parallel for 5000 iterations each, adapting after $g_0 = 200$ iterations, and discarding the first $n_0 = 1000$ as burn-in. Each chain was started at the approximate maximum marginal likelihood point-estimates with a small amount of noise added. The collapsed blocked Gibbs sampler uses two blocks, $\boldsymbol{\xi}_1 = \{\{\sigma_i\}_{i=1}^{k_x}, \{\theta_i\}_{i=1}^{k_z}\}$ and $\boldsymbol{\xi}_2 = \{\theta_S, \beta^{-1}\}$, using Algorithm 1. The parameters of the variational distribution are re-optimized at each Gibbs step, with hyperparameters fixed at the proposed values. optimization was performed until convergence, or until a maximum of 1000 scaled conjugate gradient iterations had been reached.

The variational distribution necessarily underestimates variance due to the choice of divergence, which can lead to a higher variance in the pseudo-marginal estimator. It is therefore necessary to qualify its use as a proposal distribution in the importance sampler. Following Doucet et al. (2015), in theory the variance of the log pseudo-marginal estimator should be less than 2. In both examples the summary statistics of the log pseudo-marginal (as a function of the number of importance samples and conditional on a single state of the Gibbs chain) are shown across 5000 approximations. That is, the accuracy of the estimator is demonstrated through summary statistics across the set:

$$\left\{\tilde{p}_r(\mathbf{Y}|\mathbf{X}, \boldsymbol{\xi}^{(i)})\right\}_{r=1}^{5000}, \tag{10}$$

in which each pseudo-marginal estimator is based on $Q = 1000$ importance samples. In order to show convergence the running values of these estimators are plotted for an increasing number of samples $q = 1, \ldots, Q$.

5.1 Simulated: sinusoidal data

In this section, a comparison between the variational and pseudo-marginal inference approaches is presented using a data set obtained from known trigonometric functions with artificially added noise. By simulating data in this way, each approach can be accurately compared to the truth. The data set is obtained by evaluating the data generating function:

$$f_{n,d}(\mathbf{x}_n) = \begin{cases} \zeta_d \cos(F_d \mathbf{x}_n) & \text{if } d = 1, 2, 3 \\ \zeta_d \sin(F_d \mathbf{x}_n) & \text{if } d = 4, 5, 6 \end{cases}, \tag{11}$$

Table 2 The data-generating specifications. Here, Unif$(\cdot,\cdot)$ is the uniform distribution

	F_d distribution	F_1	F_2	F_3	F_4	F_5	F_6
Well-specified case 1	Constant	1	1	1	1	1	1
Poorly-specified case 2	Unif(0.8, 1, 2)	1.03	0.92	1.11	0.99	0.87	1.02
Poorly-specified case 3	Unif(0.7, 1, 3)	1.04	0.88	1.15	0.99	0.80	1.03

at set of uniformly spaced inputs between 0 and 4π, where $\mathbf{x}_n \in \mathbb{R}$ denotes the n-th sample, and F_d is a periodicity factor. Amplitudes are uniformly sampled from $\zeta_d \sim \mathcal{U}(0,1)$ and kept consistent across examples. The noise corrupted responses are obtained from $y_{n,d} = f_{n,d} + \varepsilon_d$, where $\varepsilon_d \overset{iid}{\sim} \mathcal{N}\left(0, 0.05^2\right)$. The data sets are generated under multiple parameterisations, given in Table 2.

In each case, two latent dimensions $k_z = 2$ are considered with $N = 30$ samples. For comparison, the variational framework under two additional settings is also considered. In the first, the model is augmented with $k_z = 6$ latent dimensions (referred to as $\text{VEM}_{k_z=6}$) to demonstrate that two latent dimensions are sufficient for the first case,[1] and that making the model well-specified in the second two cases does not change the conclusions of the comparison. The second setting has $N = 60$ samples (this is referred to as $\text{VEM}_{N=60}$) to demonstrate that the advantages of PM are retained when optimization is performed with a larger sample.

In the first case, in which the periods are constant, it is expected that each inference approach will be able to make adequate predictions. In the two additional cases where F_d is sampled from increasing uniform intervals, it is expected that point estimates of the hyperparameters will give an inadequate predictive distribution in the poorly specified cases. These examples are also designed to demonstrate the ability of PM to capture multimodal, but connected, posteriors. The improved uncertainty quantification and accuracy of predictions using PM is then demonstrated.

Trace and autocorrelation plots for all cases are shown in Figs. 11 and 12 in "Appendix B", respectively, for each hyperparameter and across each chain. These plots demonstrate good mixing. If necessary, mixing can be further improved by splitting the hyperparameters into smaller blocks. The bivariate marginal posterior histograms, for different pairs of hyperparameters, are shown in Fig. 1, where the rows correspond to the three data generating cases. Specifically, the pairs include: the input lengthscale and model noise (σ_1, β^{-1}); latent lengthscales (θ_1, θ_2); and the signal variance and model noise (θ_S, β^{-1}). This shows multimodal posteriors for the poorly-specified cases, underlining the inappropriateness of point estimates and difficulties in optimization of the hyperparameters. When the maximum marginal likelihood value lies within the axis it is marked with a dot, showing the tendency for the point-estimates of the variational expectation-maximization approach to under-fit.

Bivariate marginal latent posterior distributions conditioned on the set of maximum marginal likelihood hyperparameters, $\theta^{(ML)}$, obtained from jointly optimizing over latent variables and hyperparameters, are shown for two pairs of samples in Fig. 2. This figure compare the quality of the variational approximation used in VEM to the true posterior used for predictions with the proposed PM inference scheme. Due to the high-dimensional nature of these spaces, only the bivariate contours corresponding to the splice of two

[1] Through automatic relevance determination the model should theoretically prune unnecessary dimensions.

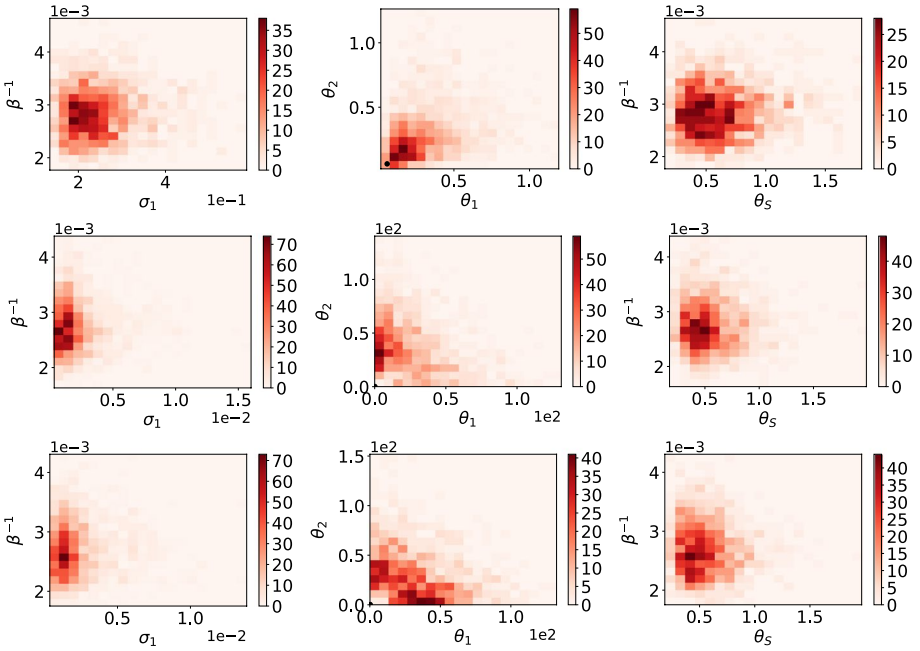


Fig. 1 The hyperparameter joint posterior distributions for different pairs. **Row 1**: Data generating case 1. **Row 2**: Data generating case 2. **Row 3**: Data generating case 3

training samples, can be visualised at a time. Even close to the ML, the approximation is poor and the correlation between hyperparameters and latent variables is high.

Similarly, the latent posterior distribution conditioned on the hyperparameters at randomly selected state 820 ($p\left(\mathbf{Z}|\mathbf{Y},\mathbf{X},\boldsymbol{\xi}^{(820)}\right)$) of the collapsed Gibbs sampler is shown in Fig. 3, alongside the variational approximation at this state in red. These plots demonstrates the extreme coupling between hyperparameters and latent variables, which makes iterative optimization extremely difficult. The joint optimization is also not trivial.

In Figs. 13 and 14 of "Appendix B", the full marginal latent posterior distributions for each sample are plotted, given $\boldsymbol{\xi}^{(820)}$ and $\boldsymbol{\xi}^{(ML)}$, respectively. These figures demonstrate a clear tendency for the variational approximation to underestimate the variance and approximate local modes due to the KL divergence, as expected. The true posterior of the latent variables is highly correlated and in many cases non-Gaussian, and the quality of variational approximation appears particularly poor with significant underestimation of the variance, especially as the model becomes increasingly misspecified. It is noted that while identifiability issues with latent variables may exaggerate the poor quality of the variational approximation in the second two cases, the issue is present in the first well-specified case.

The summary statistics of the log pseudo-marginal estimator (4) for increasing $Q \leq 1000$ (number of importance samples) conditioned on the hyperparameters of Gibbs state 820 of the respective chain are shown in Fig. 4. As can be observed from this figure, the variance is close to one for all cases. This was found to be consistent across each chain and is within the limits suggested by Doucet et al. (2015).

The accuracy of the predictive densities under the VEM and pseudo-marginal frameworks are compared in Table 3, in which the mean absolute error is reported. This is

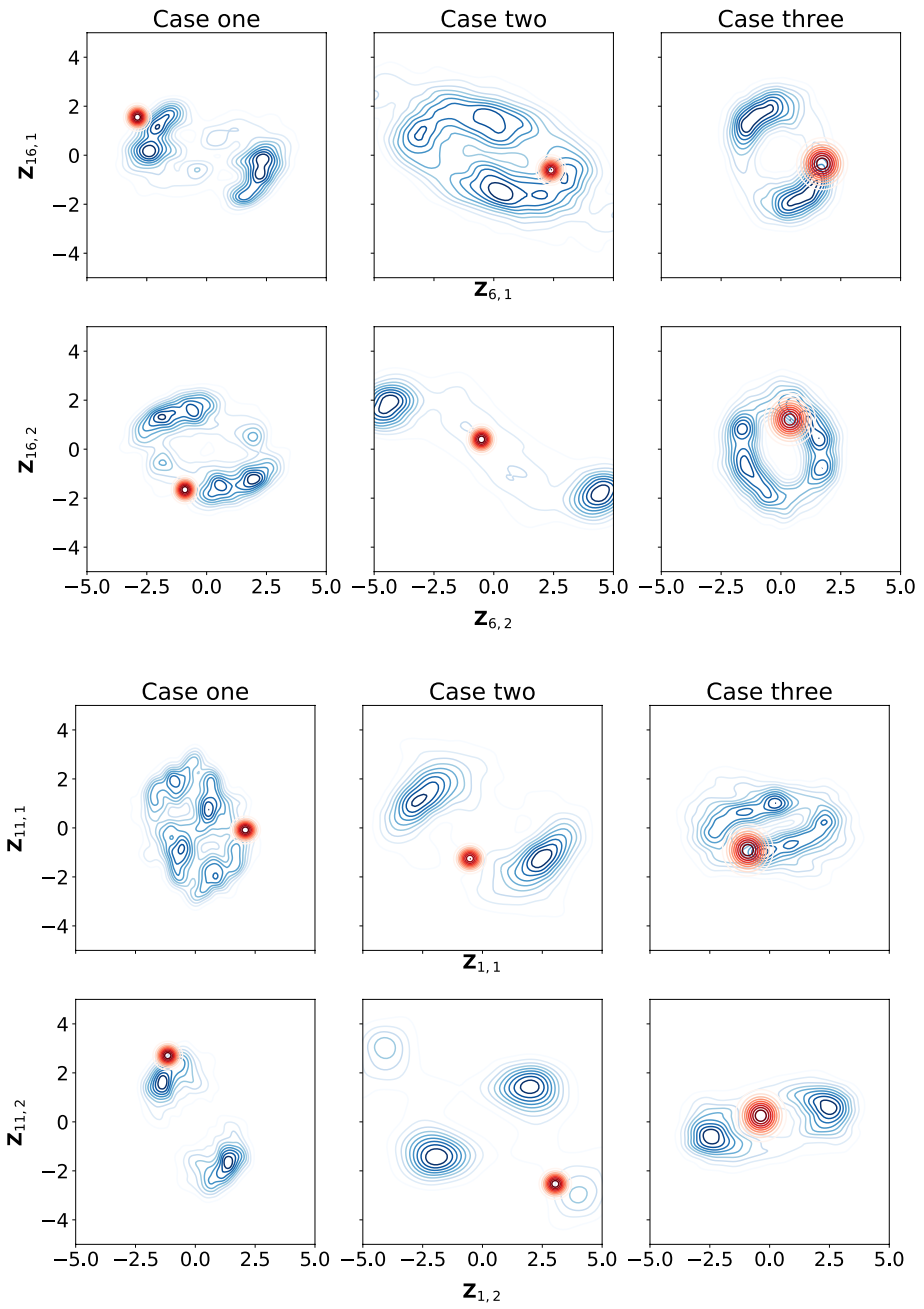


Fig. 2 Simulated example. Bivariate marginal latent posterior distributions for sample pairs (6, 16) and (1, 11), conditional on hyper-parameter posterior sample $\theta^{(ML)}$. The exact posterior (in blue) is obtained using kernel density estimation on 100, 000 elliptical slice samples, and the variational approximation (in red) is known analytically. The three columns correspond to the three data generating cases. The first row in each pair corresponds to the first latent dimension, while the second row corresponds to the second latent dimension

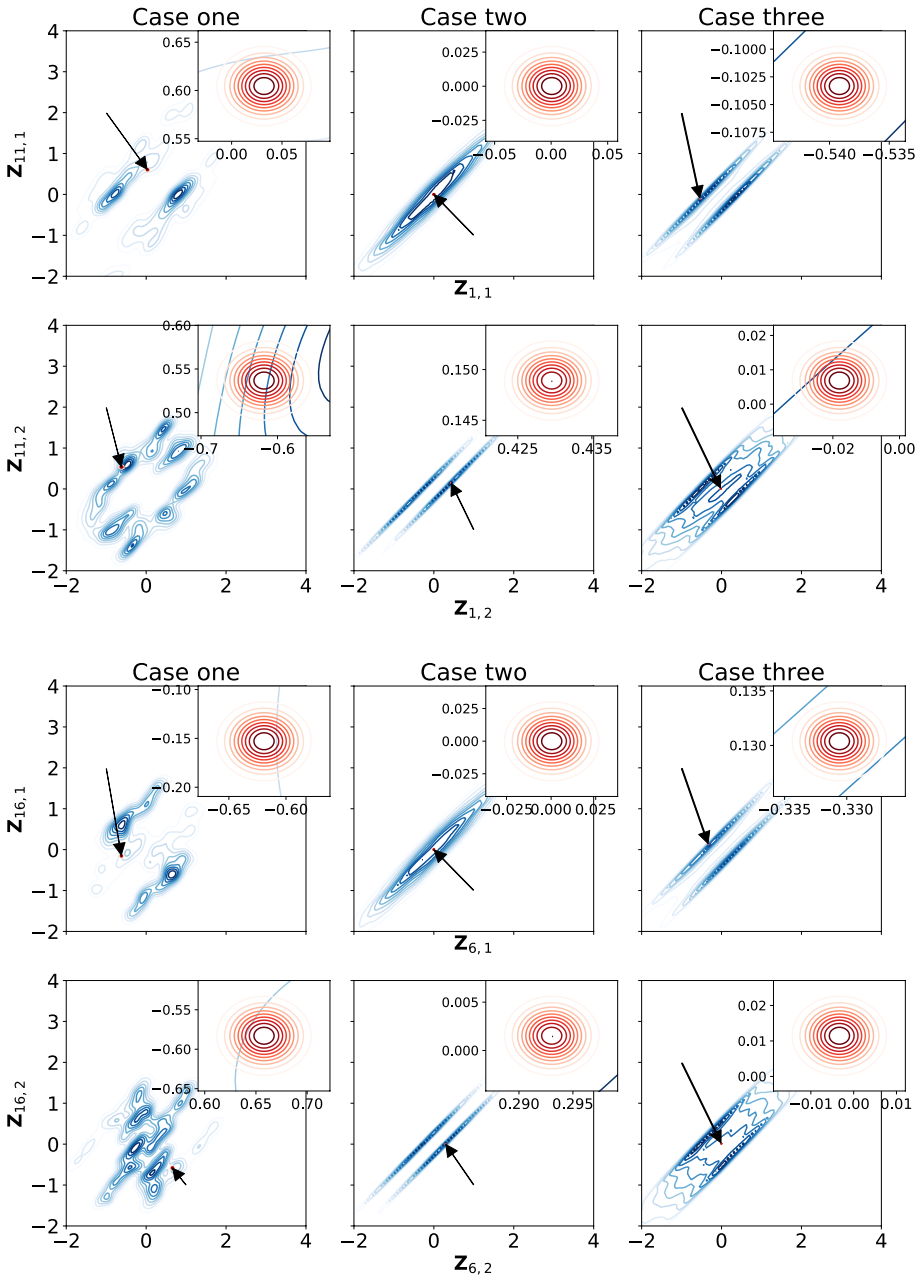


Fig. 3 PM inference scheme. Bivariate marginal latent posterior distributions for sample pairs (1, 11) (top two rows) and (6, 16) (bottom two rows), conditional on hyper-parameter posterior sample $\theta^{(820)}$. The exact posterior (in blue) is obtained using kernel density estimation on 100, 000 elliptical slice samples, and the variational approximation (in red) is known analytically. The three columns correspond to the three data generating cases. The first and third rows correspond to the first latent dimension, while the second and fourth rows correspond to the second latent dimension (Color figure online)

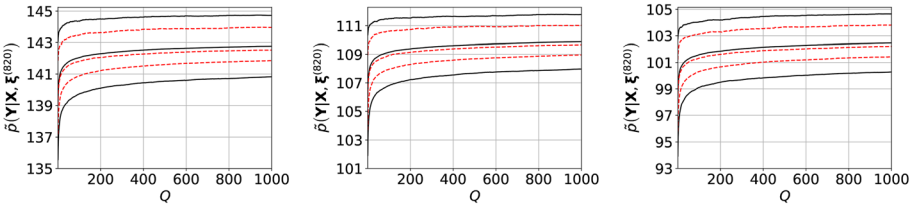


Fig. 4 Convergence plots for the log pseudo-marginal estimator, conditioned on the hyperparameters of Gibbs state 820 of the respective chain. Shown are summary statistics of the pseudo-marginal estimates. The mean±2 standard deviations is shown in black, while the median, 10th and 90th percentile are shown in red (Color figure online)

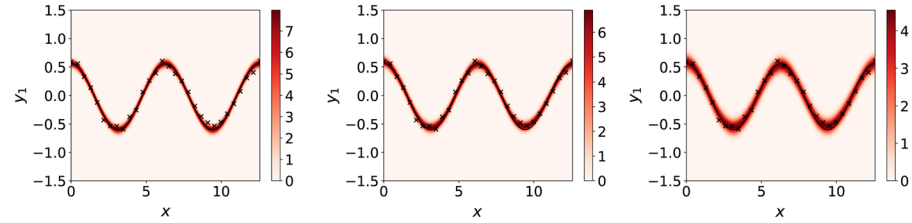


Fig. 5 Case 3 predictive densities for the first feature. The mean of the data generating function is given as a solid line, while scatter points depict the training data

Table 3 The mean absolute error between test samples of the true data generating distribution and the predictive distribution for each scheme and each case

		y_1	y_2	y_3	y_4	y_5	y_6
Case 1	PM	**0.057**	**0.052**	**0.056**	**0.056**	**0.055**	**0.056**
	VEM	0.070	0.069	0.082	0.067	0.087	0.069
	$\text{VEM}_{k_z=6}$	0.115	0.093	0.110	0.106	0.126	0.105
	$\text{VEM}_{N=60}$	0.071	0.068	0.081	0.069	0.079	0.074
Case 2	PM	**0.067**	**0.057**	**0.054**	**0.057**	**0.053**	**0.056**
	VEM	0.111	0.131	0.168	0.112	0.165	0.128
	$\text{VEM}_{k_z=6}$	0.107	0.097	0.108	0.087	0.110	0.098
	$\text{VEM}_{N=60}$	0.077	0.080	0.092	0.074	0.088	0.075
Case 3	PM	**0.066**	**0.057**	**0.054**	**0.058**	**0.053**	**0.056**
	VEM	0.181	0.204	0.173	0.138	0.164	0.178
	$\text{VEM}_{k_z=6}$	0.101	0.088	0.107	0.082	0.095	0.089
	$\text{VEM}_{N=60}$	0.119	0.119	0.124	0.101	0.117	0.107

Bold values are the lowest errors across all methods

defined between samples of the true data generating distribution and predictive distribution of each framework. For an output d, this is defined as:

$$\epsilon_d = \frac{1}{1000}\sum_{i=1}^{1000} |y^*_{i,d} - \tilde{y}^*_{i,d}|, \qquad y^*_{i,d} \sim \mathcal{N}\big(f_{i,d}(\mathbf{x}_i), 0.05^2\big), \tag{12}$$

in which $\mathbf{x}_i$ are linearly spaced between 0 and 4π, and $\tilde{y}^*_{i,d}$ are samples of the predictive distributions in (19) and (8) at $\mathbf{x}_i$, for the variational and pseudo-marginal frameworks respectively. The predictive densities for the first feature of case three are shown in Fig. 5, while all features of each case are shown in Figs. 15, 16, 17 of "Appendix B".

In addition, errors for the VEM with an increased latent dimension and with twice as many training samples are also reported. For the first case, increasing the latent dimension increases the error. Consequently, this appears to demonstrate that the VEM framework was unable to automatically prune unnecessary dimensions, despite the use of an automatic relevance determination kernel. Additionally, it shows that increasing the number of latent dimensions does not improve predictive performance in this example, perhaps due to the consequent optimizer search space dimension increase. For the later two cases, increasing the latent dimension allows for a well specified model. Despite this, the VEM approach on this model is still out performed by the PM approach on the misspecified model.

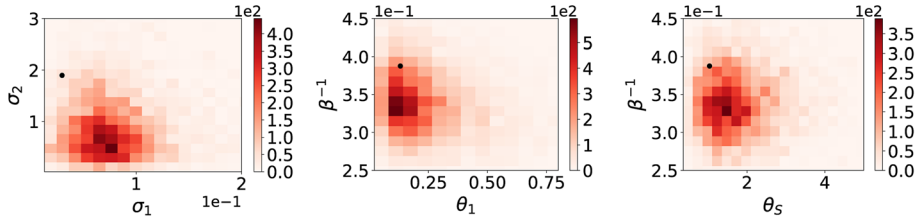
The variational approximation clearly leads to a model that overestimates the uncertainty by underfitting. We observe that PM gives a marked increase in accuracy across all features, particularly for the poorly specified examples. However, it must be noted that the VEM optimization is non-convex and may not reflect a global optimum despite convergence, particularly given that the optimum is very sensitive to initialisation, and, moreover, pertains to a lower bound on the marginal likelihood. This further illustrates the necessity for posterior sampling in many cases.

5.2 Experimental: New York air quality

Having compared the two methods on a simulated data set, a comparison is now presented using experimental data from the 'New York Air Quality Measurements' data set. This example will serve to demonstrate that the drawbacks of the variational scheme, particularly the underestimation of the latent variance and overestimation of the noise level, persist in real data. The air quality data was measured daily from May to September 1973 and is publicly available using the R datasets package Team R.C. contributors (2013). The two features include the log mean ozone in parts per billion (at Roosevelt Island) and the maximum daily temperature in degrees Fahrenheit (at La Guardia Airport). Two covariates are considered: the day since May 1*st* 1973, when the study began, and the month in which the sample was taken. Both variables are taken as integer values. The data set consists of 154 samples, of which the 116 with no missing values were used.

A single latent dimension is considered, with $k_z = 1$, and the kernels of the previous example are used (9). It is noted that the VEM model with $k_z = 2$ drastically over- or underfitted in the experiments for most initialisations. This is likely to be due to the decreased signal to noise ratio of this data set. As before, Gamma priors were assigned on the hyperparameters, also shown in Table 1, and a log transformation in the random walk proposal was used. Four chains were run in parallel for 5000 iterations, adapting after $g_0 = 200$ iterations, and the first 1000 were discarded (burn-in). Again, each chain was started at the approximate maximum marginal likelihood point estimates with addition of a small noise term.

The collapsed Gibbs sampler used two blocks, $\boldsymbol{\xi}_1 = \big(\sigma_1, \sigma_2, \theta_1\big)$ and $\boldsymbol{\xi}_2 = \big(\theta_S, \beta^{-1}\big)$, re-optimizing the variational distribution at each Gibbs step. optimization was performed until



(a) Case 1: σ_1 v. σ_2. **(b)** Case 1: θ_1 v. β^{-1}. **(c)** Case 1: θ_S v. β^{-1}.

Fig. 6 (Air quality example) The hyperparameter bivariate posterior histograms for different pairs. Approximate ML estimates are marked with a dot

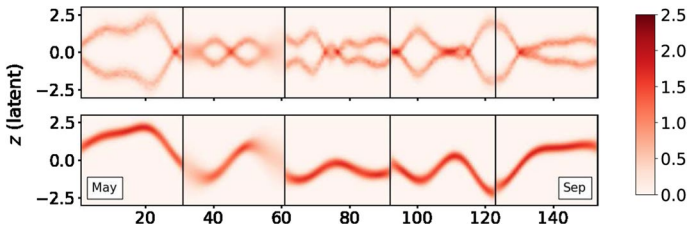


Fig. 7 (Air quality) The predictive latent density using the pseudo-marginal (PM) scheme and VEM scheme in each row, respectively

convergence, up to a maximum of 1000 scaled conjugate gradient iterations. Trace and autocorrelation plots demonstrate good mixing, as seen in Fig. 18 of "Appendix C". The bivariate marginal hyperposterior histograms, for different pairs of hyperparameters, are shown in Fig. 6. In this figure, the pairs include the input lengthscales (σ_1, σ_2); latent lengthscales and model noise (θ_1, β^{-1}); and the signal variance and model noise (θ_S, β^{-1}). When the maximum marginal likelihood value lies within the axis, it is again marked with a dot. In this example the point-estimates of the variational expectation-maximization approach slightly underfit, over-estimating the model noise and poorly estimating the lengthscales.

A comparison of the variational approximation to the posterior of the latent variables conditioned on hyperparameters is made (using ESS to obtain samples of the latent variables from the posterior) is first made. In Fig. 19 of "Appendix C", the true and variational posteriors (given $\boldsymbol{\xi}^{(820)}$) for pairs of latent variables corresponding to different samples are shown, while Fig. 20 shows the full marginal latent posterior distributions for each sample, again given $\boldsymbol{\xi}^{(820)}$. Although the true posterior is still clearly non-Gaussian, these figures reveal that the latent variables are less correlated and better approximated by a Gaussian distribution, and therefore the quality of variational approximation is better.

The predictive latent density (obtained using standard Gaussian process prediction) is shown for both frameworks in Fig. 7. These plots are composed of segments, where each segment is conditioned on the relevant month in which the 'days since study began' covariate belongs, and each segment is clearly separated by a vertical black line. Full plots are shown in Figs. 21 and 22, for PM and VEM, respectively. For example, the second segment (which encompasses the 50th day) is the predictive density of the latent variable as the 'day since study began' varies, but conditioned on the month of June. A reflection at $\mathbf{Z} = 0$ can clearly be seen in the PM predictive density as a consequence of symmetry in the latent kernel of the Gaussian process latent variable model; the variational approach

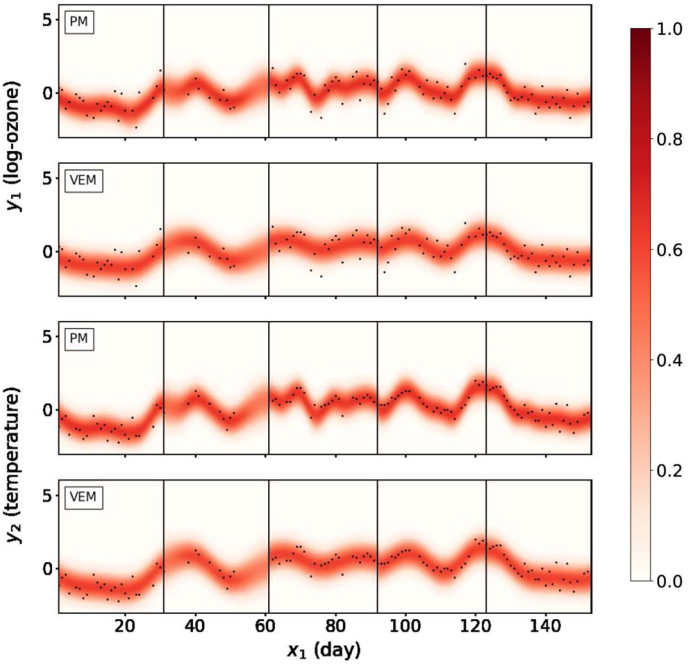
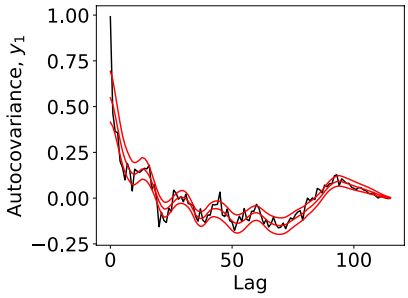


Fig. 8 (Air quality) The predictive output density using the pseudo-marginal (PM) scheme and VEM scheme in each row, respectively

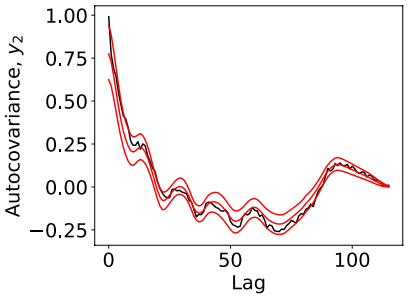
instead approximates this bimodal density with a single Gaussian. It can also be seen that at the start and end of June the predictive uncertainty increases. This is a consequence of missing data. Finally, the VEM predictive density lies within that obtained using PM, demonstrating that the manifold obtained using the approximate maximum marginal likelihood is contained within that of the sampled states.

In Fig. 8 the predictive output densities are shown for each feature for both approaches. Similarly, full plots are shown in Figs. 23, 24 for the PMMC approach, and Figs. 25, 26 for VEM. These are obtained as outlined in Sect. 4.3 and using Eq. 19, respectively. The VEM predictions are overly smooth functions, with longer correlations between time points when compared to the PM approach. In addition to this we can compare the uncertainty in these predictions. This comparison is best observed by inspecting the credible intervals, which are given by the width of the predictive density at a given input. We can see that the credible intervals of the VEM approximation are unnecessarily wide, while the PM approach results in a reduced credible interval. This is achieved without sacrificing empirical coverage, meaning that the predictions are both precise and accurate.

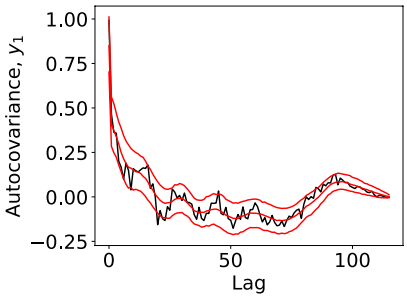
To further explore this over-smoothing, and the improved efficacy of the PMMC approach over VEM posterior predictive checks (PPC) were performed, in which the posterior predictive distribution of PMMC is used to sample replications of the training data set. Such checks then compare properties of the replicated data set with the training set, with the hope that



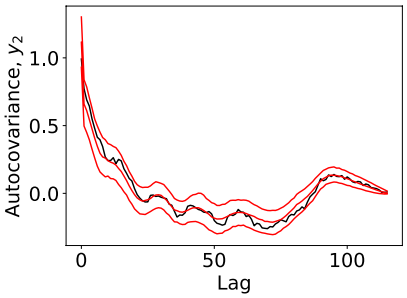
(a) PMMC autocovariance, feature 1.



(b) PMMC autocovariance, feature 2.

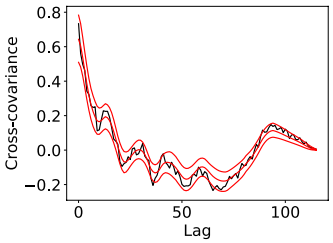


(c) VEM autocovariance, feature 1.

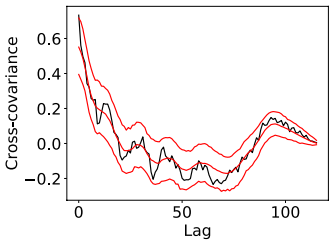


(d) VEM autocovariance, feature 2.

Fig. 9 Posterior predictive checks for New York Air Quality. Autocovariance of each feature. Replicate features are sampled under each model (red) and compared to the training data (black) (Color figure online)



(a) PMMC cross covariance PPC.



(b) VEM cross covariance PPC.

Fig. 10 Posterior predictive checks for New York Air Quality. Cross covariance between each feature. Replicate features are sampled under each model (red) and compared to the training data (black) (Color figure online)

the model accurately captures these statistical properties. In this case, they are performed as a form of model validation. Towards this end, y_{rep} feature replicates were sampled from the joint predictive posterior distribution of PMMC, given each of the training inputs. We considered two checks: (1) the autocovariance of each feature and (2) the cross covariance between

features. For the PMMC approach we sampled a set of replicates given thinned Gibbs samples and elliptical slice samples. For each replicated data set we then calculated each function and compared it with the truth. PPC results for (1) are given in Fig. 9, whilst PPC results for (2) are given in Fig. 10. We observe a clear difference between the two approaches, with VEM failing to accurately replicate either of the properties.

6 Numerical computation

The asymptotic convergence guarantees of the pseudo-marginal scheme come at the cost of an additional computational burden, requiring repeated variational approximations to the marginal likelihood at each Gibbs step. This cost can be reduced by using stochastic gradients, fewer optimizer iterations, more intelligent initialisation, or performing optimization with respect to certain variational parameters (e.g., the inducing locations) less frequently. However, for some examples these changes may also slow the convergence and mixing of the Markov chain.

Alternatively, the algorithm introduced in Sect. 4 may also be accelerated using the algorithm of Drovandi et al. (2018), in which a Gaussian process is used to approximate the marginal log likelihood. When the predictive variance is within a threshold, the Gaussian process can then be used to replace the variational approximation, avoiding an optimization procedure. Whilst this sacrifices the asymptotic convergence guarantees of the algorithm, the approach will still benefit by avoiding the strong distributional assumptions of the variational framework and performing full posterior inference for the hyperparameters.

Additionally, the MCMC scheme can be parallelised trivially, leading to a significant decrease in computational time. To scale to larger samples sizes, the proposed pseudo-marginal scheme can be combined with ideas from Hensman et al. (2015) and the approximate variational distribution used as a proposal in importance sampling can be replaced with the doubly stochastic variational scheme for deep Gaussian processes, recently proposed in Salimbeni and Deisenroth (2017). However, it is noted that this comes at the cost of approximations to the sGPLVM in the pseudo-marginal framework, in order to scale to larger data sets.

7 Discussion

In models with strong correlations between parameters, Gibbs sampling is known to perform poorly (Lawrence et al. 2009). Strong correlations between variables can result in inefficient mixing and slow convergence, and dependence in hierarchical models can lead to local behaviour of the tuning parameters, which cannot be adapted without breaking detailed balance.

Through the use of a pseudo-marginal scheme, the high correlations between latent variables and hyperparameters are broken. Simulated and experimental examples have demonstrated the significant improvements that can be obtained through the pseudo-marginal

inference scheme, particularly in the poorly-specified examples, in which point estimates of hyperparameters are inappropriate. In all experiments, the approximate ML value overestimates the model noise, due to underestimation of the latent variance; by removing the distribution assumptions on the posterior of the latent variables, the pseudo-marginal scheme is able to overcome this problem.

The underestimation of the posterior variance of the latent variables does not directly affect the pseudo-marginal algorithm, which has MCMC convergence guarantees. The closer the pseudo-marginal approximation is to the true marginal, the faster the chain converges, with fewer importance samples required, and therefore a reduced computational cost. Similarly, the predictions are unaffected since latent variables are sampled using ESS, after taking advantage of pseudo-marginalization to collapse the sampling algorithm.

Although not observed in this article, high variability in the pseudo-marginal estimates can induce 'stickiness' in the Markov chain, in which randomly estimating a larger pseudo-marginal leads to a state from which it can be improbable to transition. In this case, the variance of the pseudo-marginal estimates can be reduced using Pareto smoothed importance sampling (Vehtari et al. 2015), or annealed importance sampling (Filippone 2013). Alternatively, the pseudo-marginals can be re-estimated on each state transition, particularly in the initial burn in phase.

In recent years, deep learning has become a popular area of research. Many deep learning models, such as deep Gaussian processes, rely on variational approximations, both for scaling to large data sets and for analytic tractability. Notably a recent non-variational approach was developed (Havasi et al. 2018) but as with other approaches this method also relies on a point estimate of the hyperparameters. Although the methodology proposed here should readily extend to many such models, when the parameter space is of a higher dimension we would suggest the use of a pseudo Hamiltonian Monte Carlo scheme on the collapsed probability model for improved mixing (Lindsten and Doucet 2016).

Appendix A: Variational marginalization of latent variables

It is first noted that standard mean-field variational methodologies (as previously used in probabilistic principal component analysis and factor analysis models (Bishop 1999; Jordan et al. 1999) do not lead to an analytically tractable algorithm. Instead, the variational distribution is restricted to lie within a class. Specifically, consider a variational distribution $q(\mathbf{Z})$, which is taken to have the following factorised Gaussian form:

$$q(\mathbf{Z}) = \prod_{j=1}^{k_z} \mathcal{N}\left(\mathbf{z}_{:,j} | \boldsymbol{\mu}_j, \mathbf{S}_j\right), \tag{13}$$

where $\mathbf{S}_j$ is a diagonal $N \times N$ covariance matrix and conditional dependence on $\mathbf{X}$ and hyperparameters $(\boldsymbol{\sigma}, \boldsymbol{\theta}, \beta)$ enters through optimization of the variational parameters $\boldsymbol{\mu}_j \in \mathbb{R}^N$ and $\mathbf{S}_j \in \mathbb{R}^{N \times N}$. Using Jensen's inequality the evidence lower bound (ELBO) can be derived as:

$$\begin{aligned}\log p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta},\boldsymbol{\sigma},\beta) &= \log\left[\int p(\mathbf{Y}|\mathbf{Z},\boldsymbol{\theta},\beta)p(\mathbf{Z}|\mathbf{X},\boldsymbol{\sigma})d\mathbf{Z}\right]\\ &= \log\left[\int p(\mathbf{Y}|\mathbf{Z},\boldsymbol{\theta},\beta)\frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\sigma})}{q(\mathbf{Z})}q(\mathbf{Z})d\mathbf{Z}\right]\\ &\geq \int q(\mathbf{Z})\log p(\mathbf{Y}|\mathbf{Z},\boldsymbol{\theta},\beta)d\mathbf{Z} - \int q(\mathbf{Z})\log\frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\sigma})}d\mathbf{Z}\\ &:= \tilde{\mathcal{F}}(q(\mathbf{Z}),\boldsymbol{\theta},\beta) - \mathrm{KL}(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X},\boldsymbol{\sigma})).\end{aligned} \tag{14}$$

The second term is the negative Kullback-Leibler (KL) divergence between two Gaussian distributions and can, therefore, be evaluated with ease. Given that the data $\{\mathbf{y}_i\}_{i=1}^N$ is independent across features, the first term can be expanded as follows:

$$\tilde{\mathcal{F}}(q(\mathbf{Z}),\boldsymbol{\theta},\beta) = \sum_{d=1}^{k_y}\int q(\mathbf{Z})\log p\big(\mathbf{y}_{:,d}|\mathbf{Z},\boldsymbol{\theta},\beta\big)d\mathbf{Z} := \sum_{d=1}^{k_y}\tilde{\mathcal{F}}_d(q(\mathbf{Z}),\boldsymbol{\theta},\beta). \tag{15}$$

The term $\tilde{F}_d(q(\mathbf{Z}),\boldsymbol{\theta},\beta)$ is still analytically intractable since $\mathbf{Z}$ remains inside the kernel.

In order to formulate a tractable problem, the variational sparse GP approach of Titsias (2009) is applied. In this approach the probability model $p(\mathbf{Y},\mathbf{F}|\mathbf{Z},\beta,\boldsymbol{\theta})$ is augmented with M additional data points (*inducing points*), which are samples from the prior distribution placed over the latent function $\mathbf{f}(\cdot)$, evaluated at a set of M pseudo or inducing inputs (independent of the training inputs). The inducing points are collected as the rows in a matrix $\mathbf{U}$, with columns denoted $\mathbf{u}_{:,d} \in \mathbb{R}^M$, while the pseudo inputs form the rows of a matrix $\mathbf{Z}_\mathbf{u} \in \mathbb{R}^{M\times k_z}$. With the inducing variables $\mathbf{u}_{:,d}$, the augmented probability model is as follows:

$$p\big(\mathbf{y}_{:,d},\mathbf{f}_{:,d},\mathbf{u}_{:,d}|\mathbf{Z},\mathbf{Z}_\mathbf{u},\boldsymbol{\theta},\beta\big) = p\big(\mathbf{y}_{:,d}|\mathbf{f}_{:,d},\beta\big)p\big(\mathbf{f}_{:,d}|\mathbf{u}_{:,d},\mathbf{Z},\mathbf{Z}_\mathbf{u},\boldsymbol{\theta}\big)p\big(\mathbf{u}_{:,d}|\mathbf{Z}_\mathbf{u},\boldsymbol{\theta}\big),$$

The joint GP prior over the inducing and latent variables is factorized in the above equation, and the Gaussian prior over the latent variables conditioned on the inducing variables is given by:

$$p\big(\mathbf{f}_{:,d}|\mathbf{u}_{:,d},\mathbf{Z}_\mathbf{u},\mathbf{Z},\boldsymbol{\theta}\big) = \mathcal{N}\big(\mathbf{f}_{:,d}|\boldsymbol{\alpha}_d,\mathbf{K}_\mathbf{f} - \mathbf{K}_\mathbf{fu}\mathbf{K}_\mathbf{u}^{-1}\mathbf{K}_\mathbf{uf}\big), \tag{16}$$

in which $\mathbf{K}_\mathbf{u}$ is the covariance matrix corresponding to the inducing variables, $\mathbf{K}_\mathbf{fu} = \mathbf{K}_\mathbf{uf}^T$ is the cross-covariance between the inducing and the latent variables, and $\boldsymbol{\alpha}_d = \mathbf{K}_\mathbf{fu}\mathbf{K}_\mathbf{u}^{-1}\mathbf{u}_{:,d}$. The marginal Gaussian prior over the inducing variables is $p\big(\mathbf{u}_{:,d}|\mathbf{Z}_\mathbf{u},\boldsymbol{\theta}\big) = \mathcal{N}\big(\mathbf{u}_{:,d}|\mathbf{0},\mathbf{K}_\mathbf{u}\big)$. By marginalizing out $\mathbf{f}_{:,d}$ and $\mathbf{u}_{:,d}$, the likelihood $p\big(\mathbf{y}_{:,d}|\mathbf{Z},\boldsymbol{\theta},\beta\big)$ can then be recovered. This is true for any set of inducing points $\mathbf{Z}_\mathbf{u}$ and, consequently, they are considered variational parameters.

From this point onwards, the notation is simplified by suppressing the dependence on $\mathbf{Z}_\mathbf{u}$ in all expressions. Variational inference is now applied a second time to approximate the true posterior $p\big(\mathbf{f}_{:,d},\mathbf{u}_{:,d}|\mathbf{y}_{:,d},\mathbf{Z},\boldsymbol{\theta},\beta\big) = p\big(\mathbf{f}_{:,d}|\mathbf{u}_{:,d},\mathbf{y}_{:,d},\mathbf{Z},\boldsymbol{\theta},\beta\big)p\big(\mathbf{u}_{:,d}|\mathbf{y}_{:,d},\mathbf{Z},\boldsymbol{\theta},\beta\big)$, using the following sparse variational distribution:

$$q\big(\mathbf{f}_{:,d},\mathbf{u}_{:,d}\big) = p\big(\mathbf{f}_{:,d}|\mathbf{u}_{:,d},\mathbf{Z},\boldsymbol{\theta}\big)\phi\big(\mathbf{u}_{:,d}\big),$$

where $p(\mathbf{f}_{:,d}|\mathbf{u}_{:,d}, \mathbf{Z}, \boldsymbol{\theta})$ is the conditional GP prior given in (16) and $\phi(\mathbf{u}_{:,d})$ is the variational distribution over inducing variables. The lower bound of the log likelihood term in the integrand of $\tilde{\mathcal{F}}_d$ in (15) is given by:

$$\begin{aligned}
&\log p(\mathbf{y}_{:,d}|\mathbf{Z}, \boldsymbol{\theta}, \beta) \\
&\geq \int \phi(\mathbf{u}_{:,d}) \left[\int p(\mathbf{f}_{:,d}|\mathbf{u}_{:,d}, \mathbf{Z}, \boldsymbol{\theta}) \log \frac{p(\mathbf{y}_{:,d}|\mathbf{f}_{:,d}, \beta) p(\mathbf{u}_{:,d}|\boldsymbol{\theta})}{\phi(\mathbf{u}_{:,d})} d\mathbf{f}_{:,d} \right] d\mathbf{u}_{:,d} \\
&= \int \phi(\mathbf{u}_{:,d}) \left[\int p(\mathbf{f}_{:,d}|\mathbf{u}_{:,d}, \mathbf{Z}, \boldsymbol{\theta}) \log p(\mathbf{y}_{:,d}|\mathbf{f}_{:,d}, \beta) d\mathbf{f}_{:,d} + \log \frac{p(\mathbf{u}_{:,d}|\boldsymbol{\theta})}{\phi(\mathbf{u}_{:,d})} \right] d\mathbf{u}_{:,d} \\
&= \int \phi(\mathbf{u}_{:,d}) \log \frac{p(\mathbf{u}_{:,d}|\boldsymbol{\theta}) \mathcal{N}(\mathbf{y}_{:,d}|\boldsymbol{\alpha}_d, \beta^{-1} I_N)}{\phi(\mathbf{u}_{:,d})} d\mathbf{u}_{:,d} - \frac{\beta}{2} \mathrm{Tr}(\mathbf{K_f} - \mathbf{K_{fu}} \mathbf{K_u}^{-1} \mathbf{K_{uf}}).
\end{aligned} \tag{17}$$

since:

$$\begin{aligned}
\int p(\mathbf{f}_{:,d}|\mathbf{u}_{:,d}, \mathbf{Z}, \boldsymbol{\theta}) \log p(\mathbf{y}_{:,d}|\mathbf{f}_{:,d}, \beta) d\mathbf{f}_{:,d} &= \log \mathcal{N}(\mathbf{y}_{:,d}|\boldsymbol{\alpha}_d, \beta^{-1} I_N) \\
&\quad - \frac{\beta}{2} \mathrm{Tr}(\mathbf{K_f} - \mathbf{K_{fu}} \mathbf{K_u}^{-1} \mathbf{K_{uf}}).
\end{aligned}$$

In contrast to Titsias (2009), it is necessary to force independence of the distribution $\phi(\mathbf{u}_{:,d})$ from $\mathbf{Z}$. Combining the lower bound above with (15) gives:

$$\begin{aligned}
\tilde{\mathcal{F}}_d(q(\mathbf{Z}), \boldsymbol{\theta}, \beta) \geq \int q(\mathbf{Z}) \Big[\int \phi(\mathbf{u}_{:,d}) \log \frac{p(\mathbf{u}_{:,d}|\boldsymbol{\theta}) \mathcal{N}(\mathbf{y}_{:,d}|\boldsymbol{\alpha}_d, \beta^{-1} I_N)}{\phi(\mathbf{u}_{:,d})} d\mathbf{u}_{:,d} \\
- \frac{\beta}{2} \mathrm{Tr}(\mathbf{K_f}) + \frac{\beta}{2} \mathrm{Tr}(\mathbf{K_u}^{-1} \mathbf{K_{uf}} \mathbf{K_{fu}}) \Big] d\mathbf{Z},
\end{aligned}$$

using the standard properties of the trace of a matrix. Under the factorisation assumption, $\phi(\mathbf{u}_{:,d})$ does not depend on $\mathbf{Z}$ and so the integrations can be interchanged:

$$\begin{aligned}
\tilde{\mathcal{F}}_d(q(\mathbf{Z}), \boldsymbol{\theta}, \beta) \geq \int \phi(\mathbf{u}_{:,d}) \left[\left\langle \log \mathcal{N}(\mathbf{y}_{:,d}|\boldsymbol{\alpha}_d, \beta^{-1} I_N) \right\rangle_{q(\mathbf{Z})} + \log \frac{p(\mathbf{u}_{:,d}|\boldsymbol{\theta})}{\phi(\mathbf{u}_{:,d})} \right] d\mathbf{u}_{:,d} \\
- \frac{\beta}{2} \mathrm{Tr}(\langle \mathbf{K_f} \rangle_{q(\mathbf{Z})}) + \frac{\beta}{2} \mathrm{Tr}\Big(\mathbf{K_u}^{-1} \langle \mathbf{K_{uf}} \mathbf{K_{fu}} \rangle_{q(\mathbf{Z})} \Big),
\end{aligned}$$

where $\langle \cdot \rangle_{q(\mathbf{Z})}$ denotes an expectation under $q(\mathbf{Z})$. Now the lower bound under the distribution $\phi(\mathbf{u}_{:,d})$ can be maximized analytically. The optimal setting of this distribution is:

$$\phi(\mathbf{u}_{:,d}) \propto e^{\left\langle \log \mathcal{N}(\mathbf{y}_{:,d}|\boldsymbol{\alpha}_d, \beta^{-1} I_N) \right\rangle_{q(\mathbf{Z})}} p(\mathbf{u}_{:,d}|\boldsymbol{\theta}),$$

and the lower bound that incorporates such an optimal setting is obtained by inserting $\phi(\mathbf{u}_{:,d})$ into the lower bound expression:

$$\begin{aligned}\tilde{\mathcal{F}}_d(q(\mathbf{Z}),\boldsymbol{\theta},\beta) \geq \log\left(\int e^{\left\langle \log \mathcal{N}(\mathbf{y}_{:,d}|\boldsymbol{\alpha}_d,\beta^{-1}I_N)\right\rangle_{q(\mathbf{Z})}} p(\mathbf{u}_{:,d}|\boldsymbol{\theta}) d\mathbf{u}_{:,d}\right) \\ -\frac{\beta}{2}\mathrm{Tr}\left(\langle \mathbf{K_f}\rangle_{q(\mathbf{Z})}\right) + \frac{\beta}{2}\mathrm{Tr}\left(\mathbf{K_u}^{-1}\langle \mathbf{K_{uf}}\mathbf{K_{fu}}\rangle_{q(\mathbf{Z})}\right).\end{aligned} \tag{18}$$

For a number of kernels this can now be computed in closed form. optimization may be performed on the tractable variational lower bound according to (18), with respect to the variational parameters $(\{\boldsymbol{\mu}_j, \mathbf{S}_j\}_{j=1}^{k_z}, \mathbf{Z_u})$, in the expectation step, and model hyperparameters $(\boldsymbol{\theta}, \boldsymbol{\sigma}, \beta)$ to obtain approximate ML estimates in the maximization step.

Following Damianou (2015), in this paper the variational parameters $(\{\boldsymbol{\mu}_j, \mathbf{S}_j\}_{j=1}^{k_z}, \mathbf{Z_u})$ are treated as free parameters, and optimized directly with scaled conjugate gradients (jointly with model hyperparameters where relevant), using a reparameterisation. While this approach mitigates against local optima, it does not guarantee a globally optimal solution or aid against other problems associated with optimization of hyperparameters. This is implemented using Sheffield (2017).

The analytic computations can be found in Titsias and Lawrence (2010) and the gradient derivations can be found in Damianou (2015), alongside derivations for the predictive density:

$$p(\mathbf{Y}_*|\mathbf{Y}) \approx \int p(\mathbf{Y}_*|\mathbf{F}_*) q(\mathbf{F}_*|\mathbf{Z}_*) q(\mathbf{Z}_*) d\mathbf{Z}_* d\mathbf{F}_*, \tag{19}$$

where $q(\mathbf{Z}_*)$ is obtained using standard GP regression, and $q(\mathbf{F}_*|\mathbf{Z}_*)$ is expressed as a product of terms with the same form as the projected process approximation. This integral is a non-Gaussian multivariate density that cannot be computed. Consequently, the variational scheme instead computes the first and second moments which are available in closed-form. However, in order to sample and evaluate the predictive distribution this is assumed to be a multivariate Gaussian with corresponding first and second moments.

Appendix B: Additional results for example 1

See Figs. 11, 12, 13, 14, 15, 16, and 17.

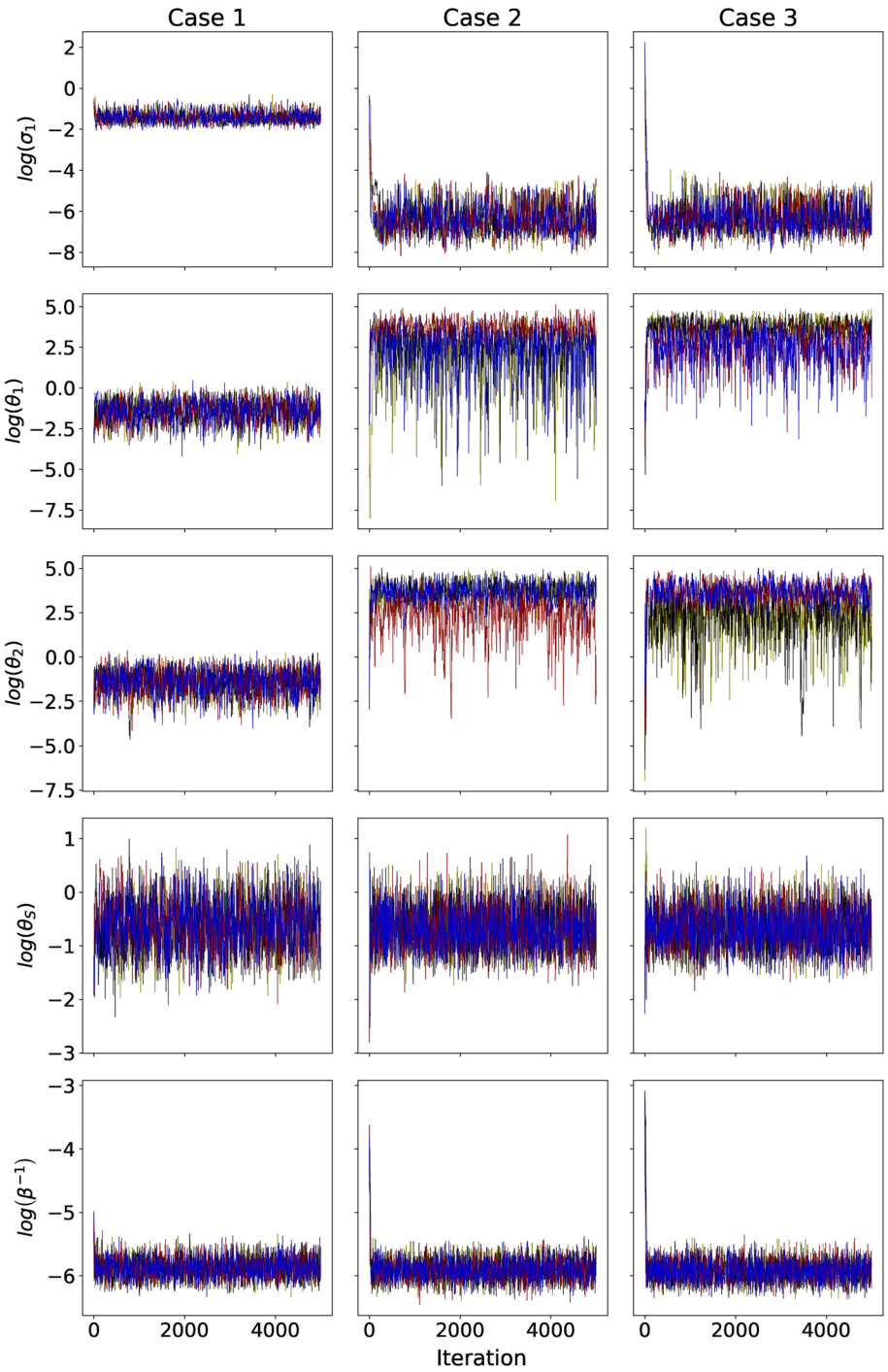


Fig. 11 Trace plots for the collapsed Gibbs hyperparameter posterior samples (with no thinning applied). The three columns correspond to the three data generating cases, while each row corresponds to a different hyperparameter

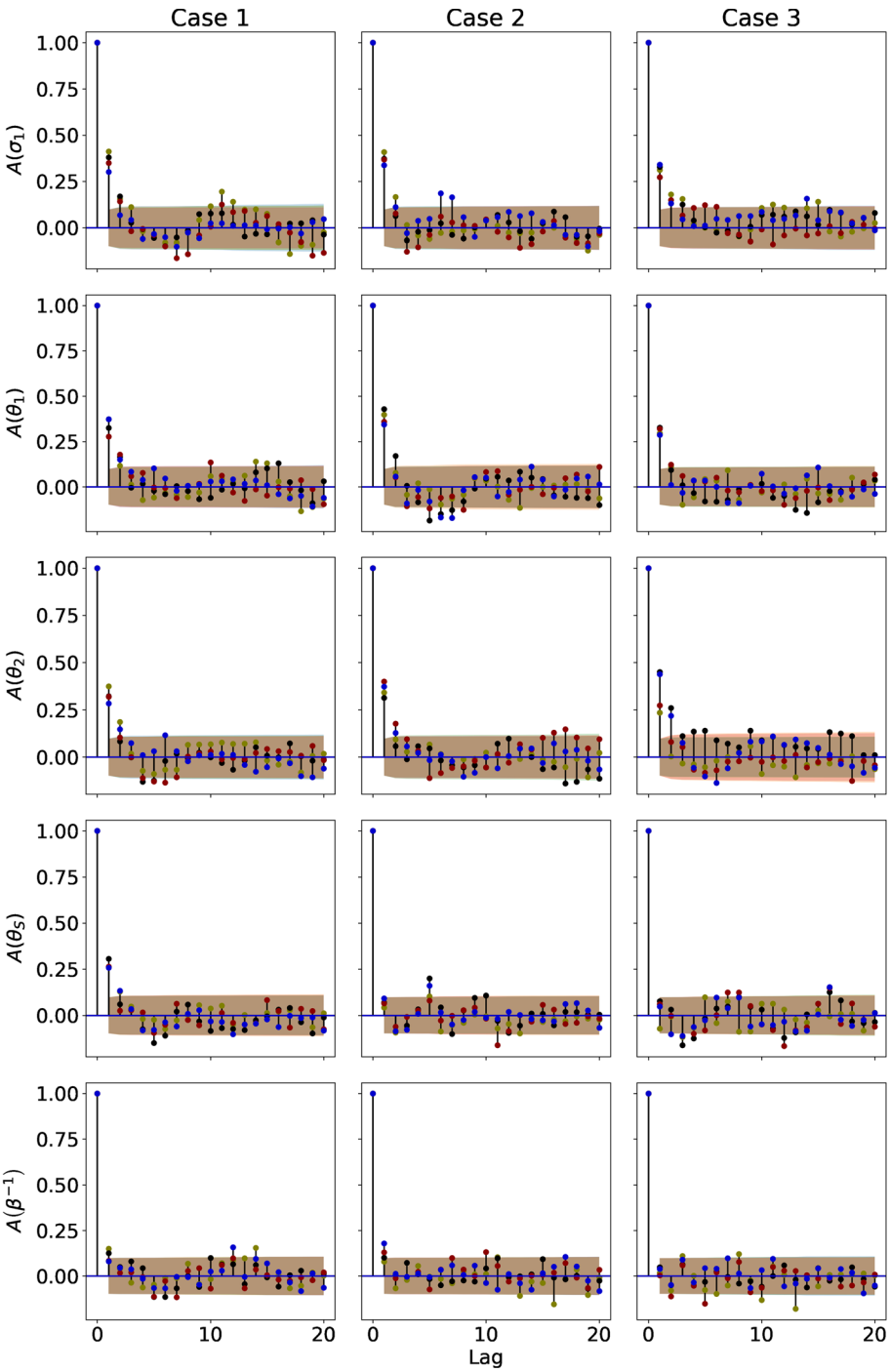


Fig. 12 Auto-correlation lag plots of the collapsed Gibbs hyperparameter posterior samples after a thinning factor 10 is applied. The three columns correspond to the three data generating cases, while each row corresponds to a different hyperparameter

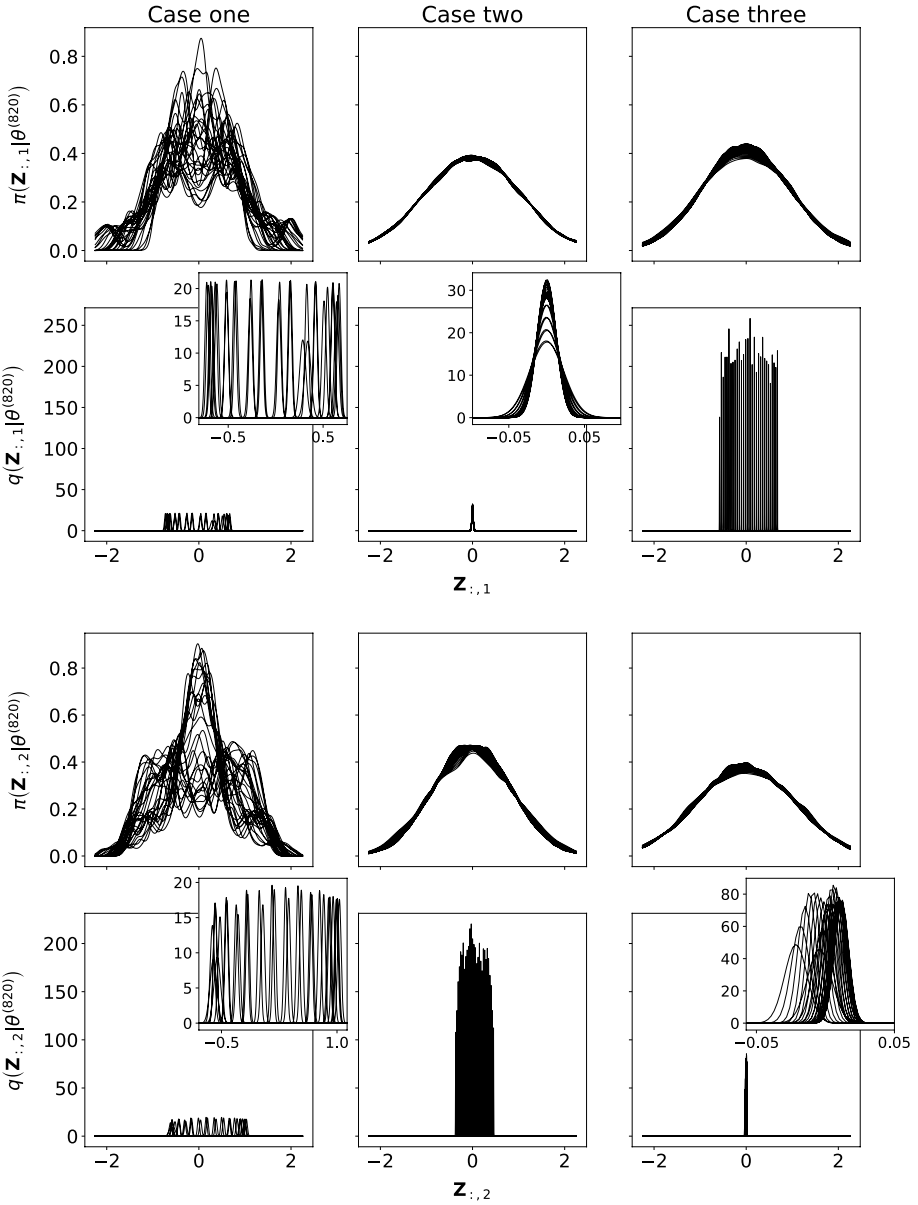


Fig. 13 PM inference scheme. Marginal latent posterior distributions for all samples, conditional on hyper-parameter posterior sample $\theta^{(820)}$. The exact posterior (first and third row) is obtained using Kernel Density Estimation on the ESS samples, and the variational approximation (second and fourth row) is known analytically. The three columns correspond to the three data generating cases. The first two rows corresponds to the first latent dimension, whilst the last two rows refer to the second latent dimension

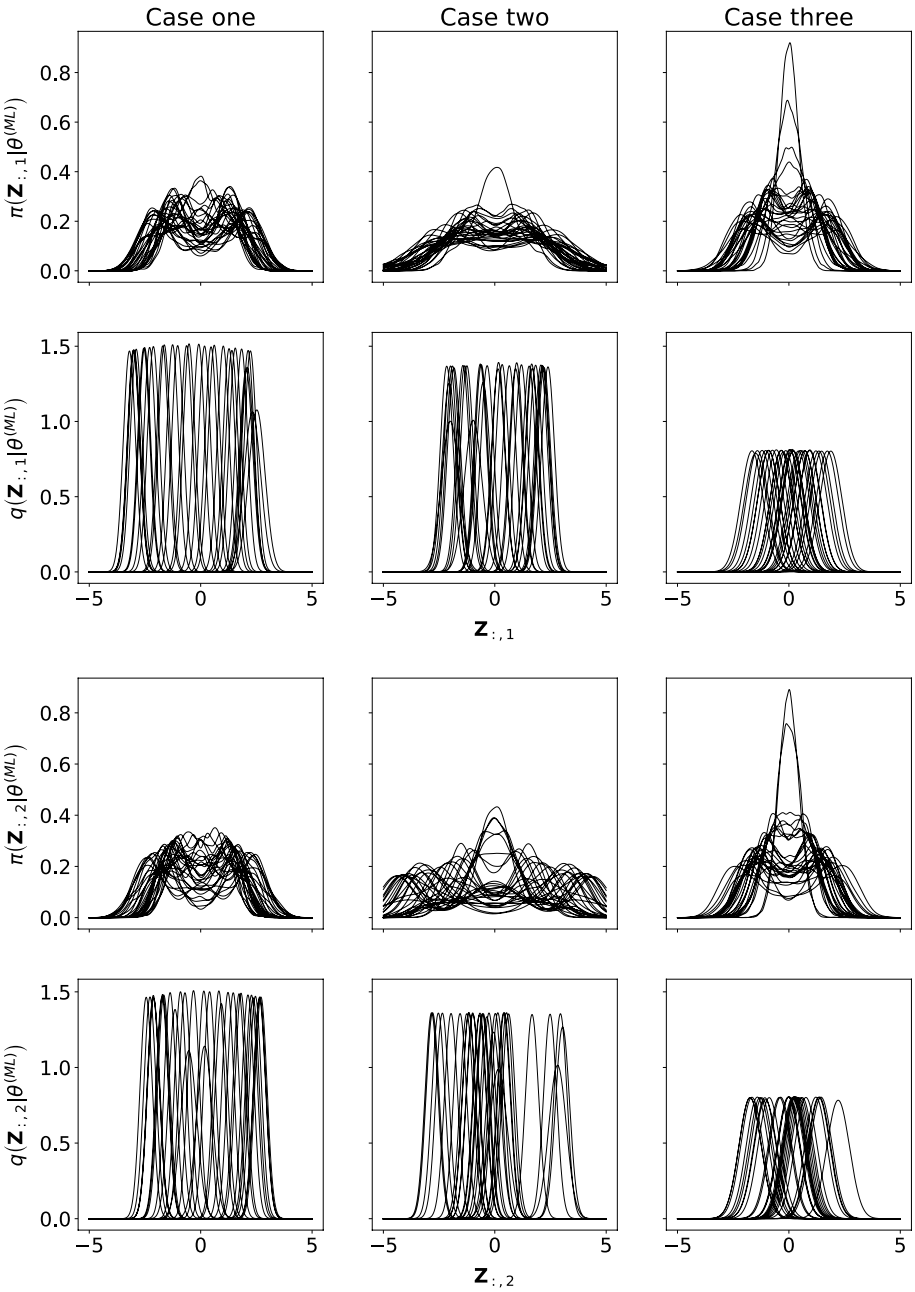


Fig. 14 VEM inference scheme. Marginal latent posterior distributions for all samples, conditional on the approximate maximum marginal likelihood hyper-parameters $\theta^{(ML)}$. The exact posterior (first and third row) is obtained using Kernel Density Estimation on the ESS samples, and the variational approximation (second and fourth row) is known analytically. The three columns correspond to the three data generating cases. The first two rows corresponds to the first latent dimension, whilst the last two rows refer to the second latent dimension

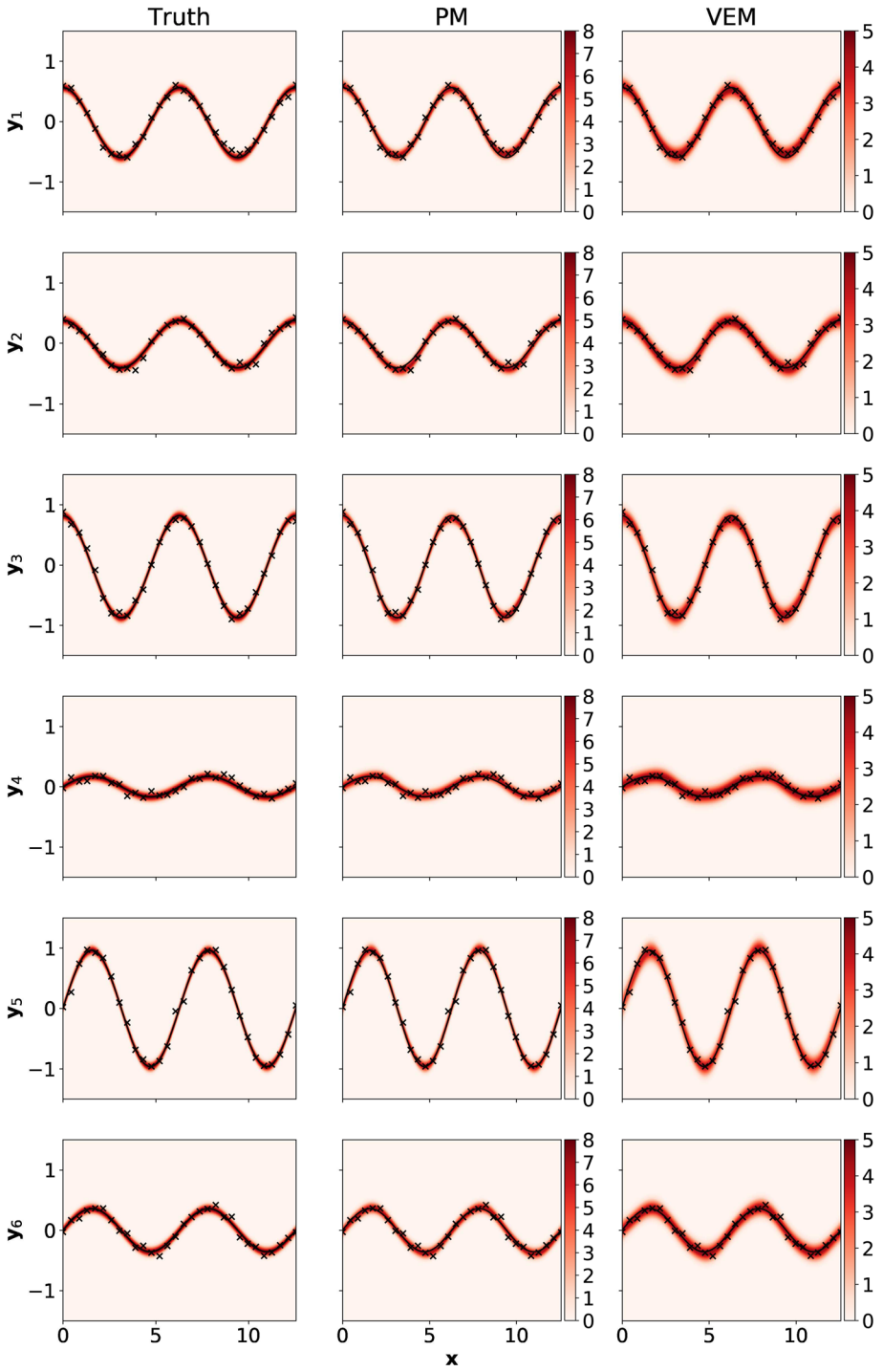


Fig. 15 Case 1 predictive densities. Each row corresponds to a different output dimension. The first column is the true density, the second is the PM approximation and the last is the VEM approximation

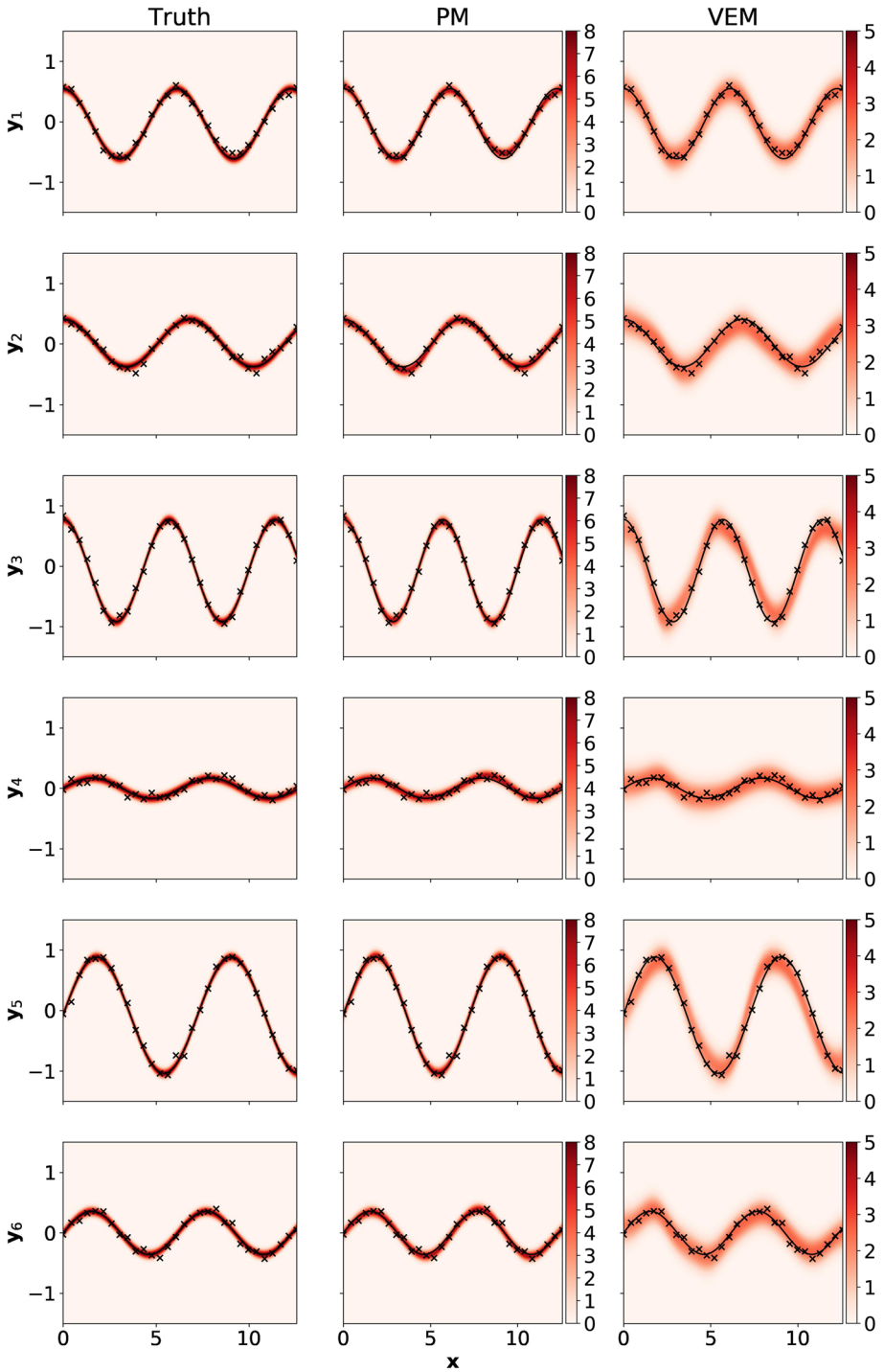


Fig. 16 Case 2 predictive densities. Each row corresponds to a different output dimension. The first column is the true density, the second is the PM approximation and the last is the VEM approximation

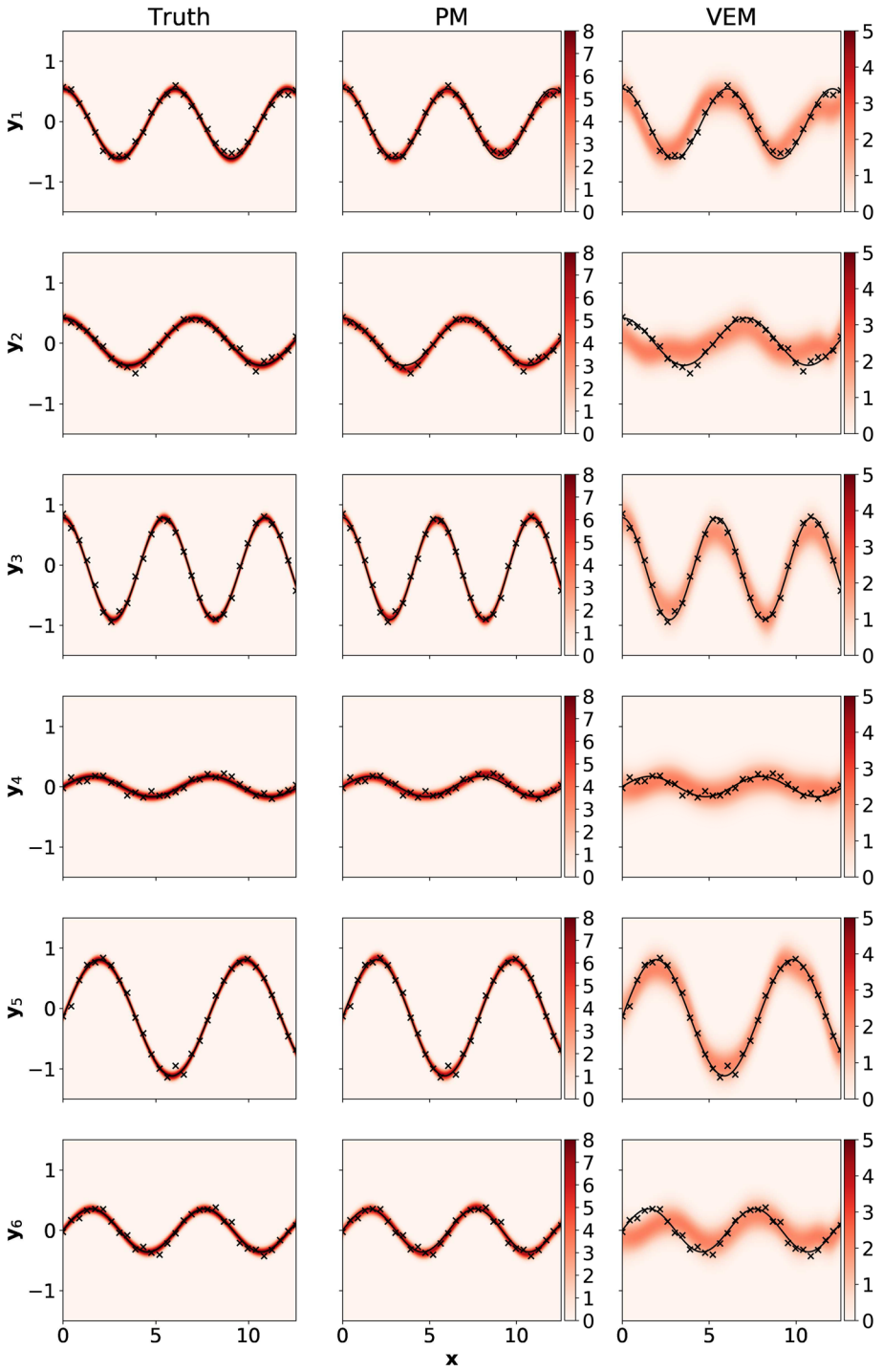


Fig. 17 Case 3 predictive densities. Each row corresponds to a different output dimension. The first column is the true density, the second is the PM approximation and the last is the VEM approximation

Appendix C: Additional results for example 2

See Figs. 18, 19, 20, 21, 22, 23, 24, 25, and 26.

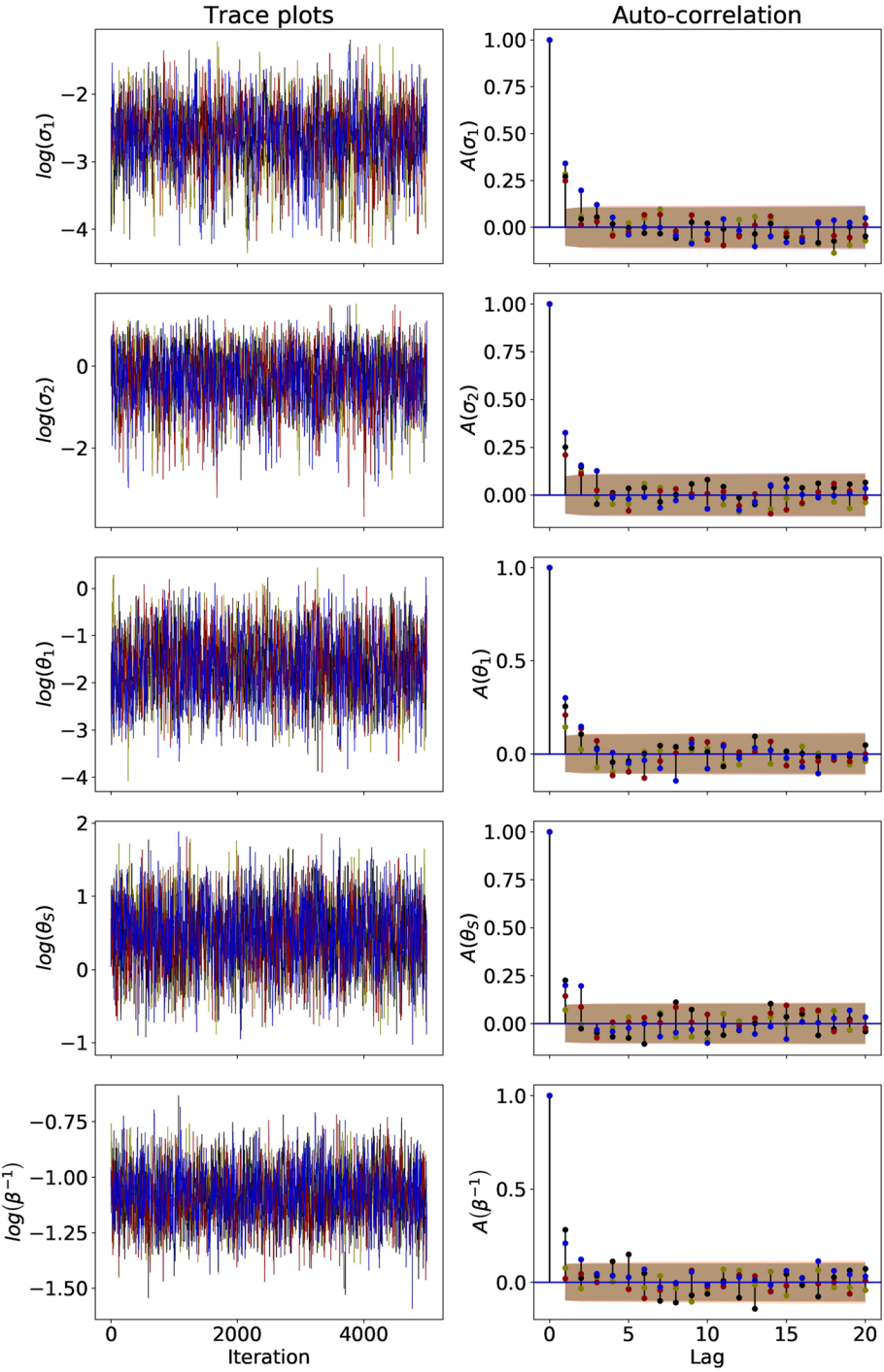


Fig. 18 (Air quality)Trace and auto-correlation lag plots of the collapsed Gibbs hyper-parameter posterior samples from each chain after a thinning factor 10 is applied. The two columns correspond to trace and autocorrelation plots, respectively, while each row corresponds to a different hyper-parameter

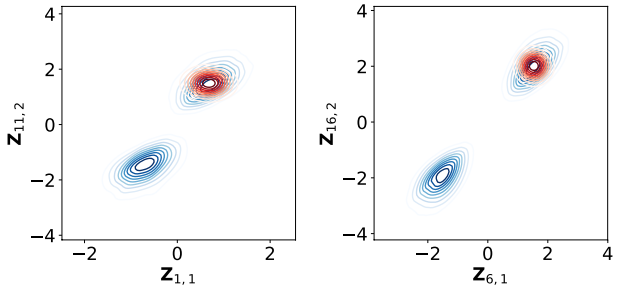


Fig. 19 (Air quality) PM inference scheme. Bivariate marginal latent posterior distributions for sample pairs (1, 11) (left) and (6, 16) (right), conditional on hyper-parameter posterior sample $\theta^{(820)}$. The exact posterior (in blue) is obtained using kernel density estimation on 100, 000 elliptical slice samples, and the variational approximation (in red) is known analytically (Color figure online)

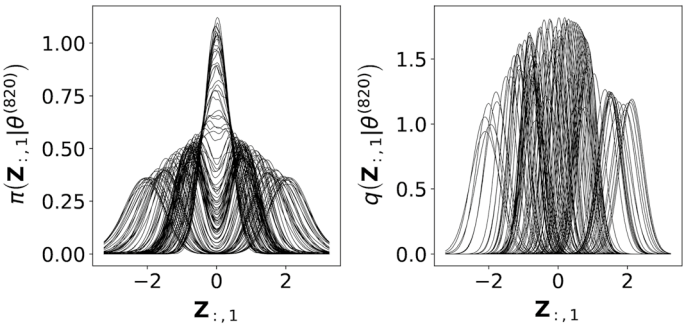


Fig. 20 (Air quality) PM inference scheme. Marginal latent posterior distributions for all samples, conditional on hyper-parameter posterior sample $\theta^{(820)}$. The exact posterior (left) is obtained using Kernel Density Estimation on the ESS samples, and the variational approximation (right) is known analytically

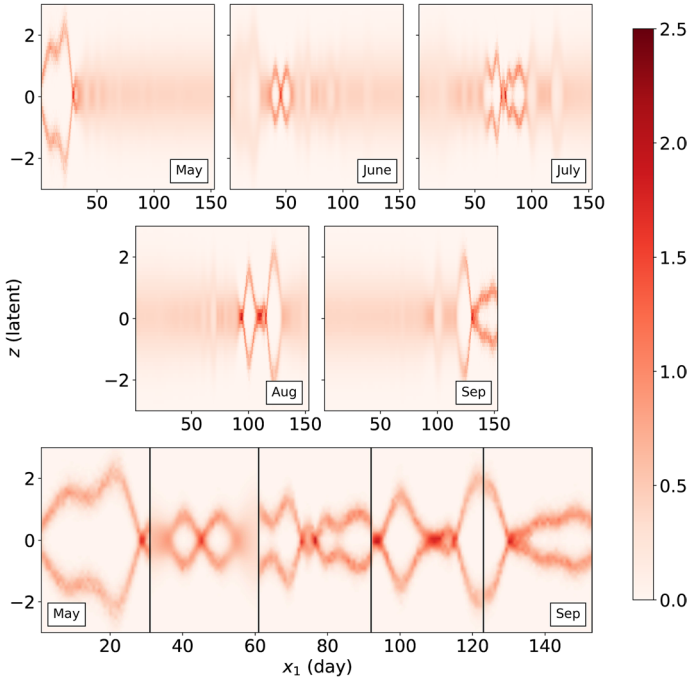


Fig. 21 (Air quality) The latent predictive densities for the pseudo-marginal (PM) scheme

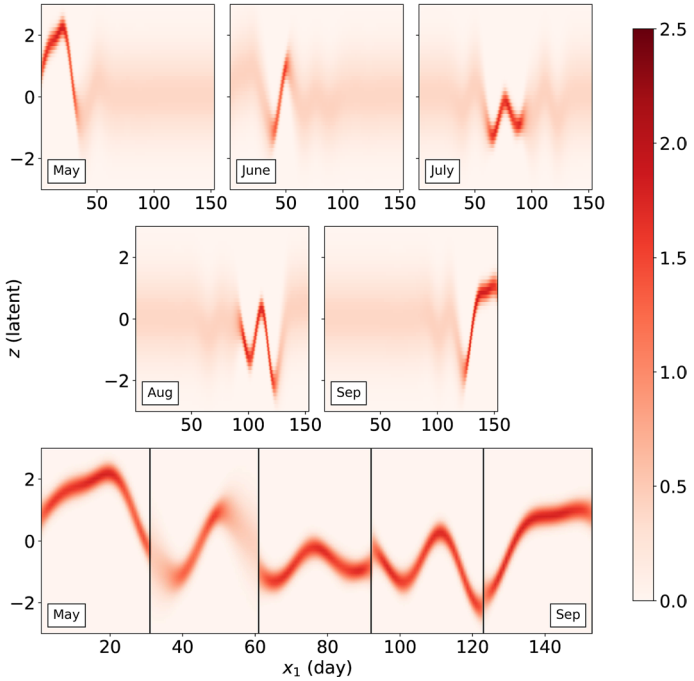


Fig. 22 (Air quality) The predictive densities for the variational expectation-maximization (VEM) scheme

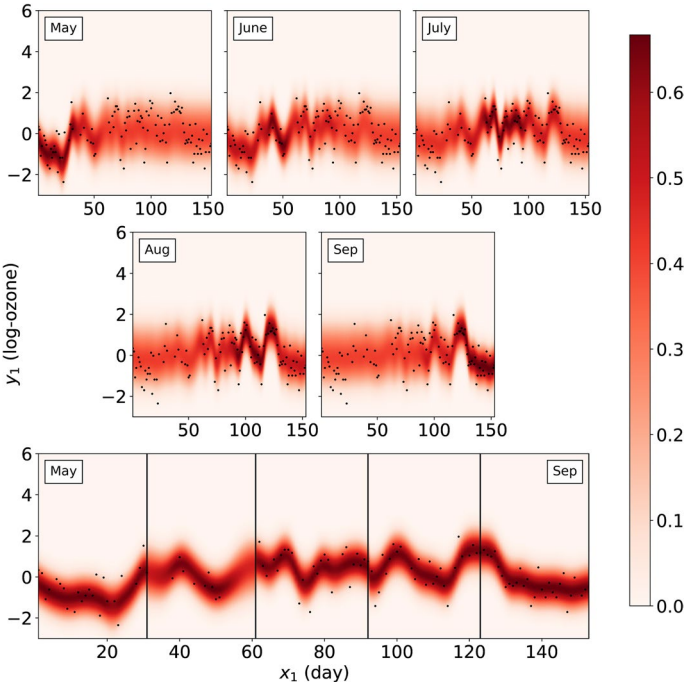


Fig. 23 (Air quality) The predictive density of the first feature using the pseudo-marginal (PM) scheme

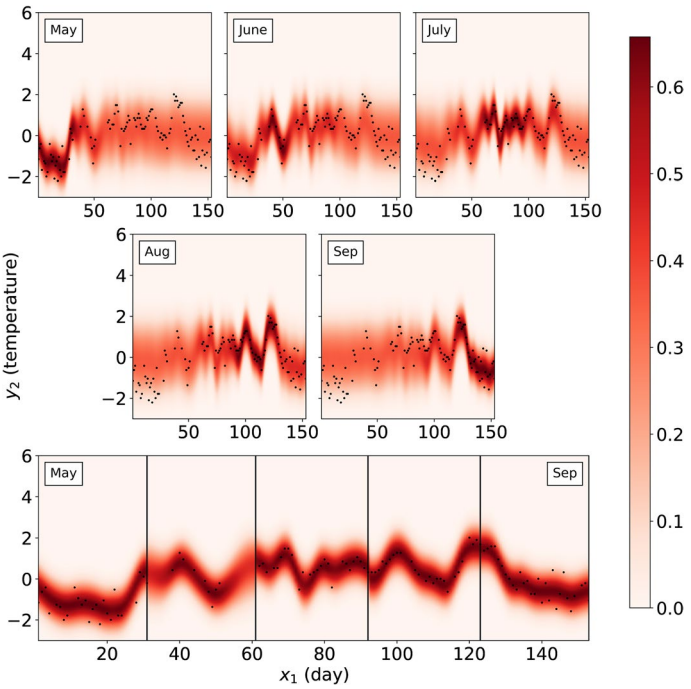


Fig. 24 (Air quality) The predictive density of the second feature using the pseudo-marginal (PM) scheme

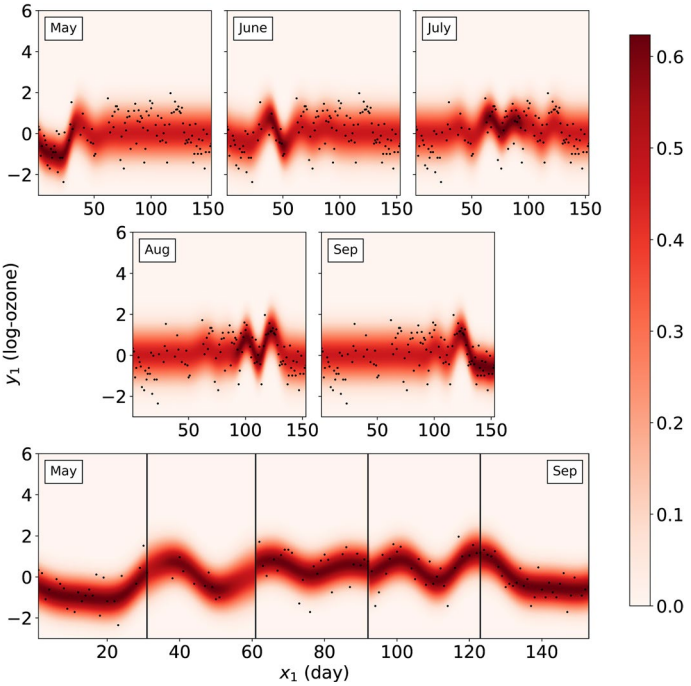


Fig. 25 The predictive density of the first feature using the variational expectation-maximization (VEM) scheme

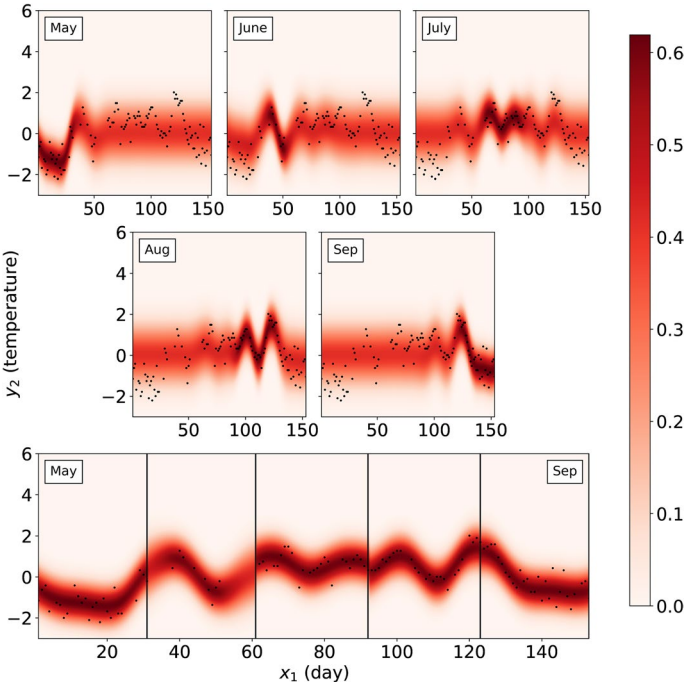


Fig. 26 The predictive density of the second feature using the variational expectation-maximization (VEM) scheme

Acknowledgements This work was partially supported by the National Key Research Development Program of China (Grant No. 2017YFB0701700).

References

Andrieu, C., & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics, 37,* 697–725.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics, 164*(3), 1139–1160.

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. *Current Trends in Bayesian Methodology with Applications, 79,* 30.

Bishop, C.M. (1999). Variational principal components.

Bitzer, S., Williams, C.K. (2010). Kick-starting GPLVM optimization via a connection to metric MDS. In: NIPS 2010 Workshop on Challenges of Data Visualization.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association, 112,* 859–877.

Damianou, A. (2015). Deep Gaussian processes and variational propagation of uncertainty. Ph.D. thesis, University of Sheffield.

Damianou, A., Lawrence, N. (2013). Deep Gaussian processes. In: Artificial Intelligence and Statistics, pp. 207–215.

Damianou, A., Titsias, M.K., Lawrence, N.D. (2011). Variational Gaussian process dynamical systems. In: Advances in Neural Information Processing Systems, pp. 2510–2518.

Doucet, A., Pitt, M.K., Deligiannidis, G., Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. In: Biometrika, pp. 295–313.

Drovandi, C. C., Moores, M. T., & Boys, R. J. (2018). Accelerating pseudo-marginal mcmc using gaussian processes. *Computational Statistics & Data Analysis, 118,* 1–17.

Filippone, M. (2013). Bayesian inference for gaussian process classifiers with annealing and exact-approximate mcmc. arXiv preprint arXiv:1311.7320.

Filippone, M., & Girolami, M. (2014). Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(11), 2214–2226.

Haario, H., Saksman, E., Tamminen, J. (2001). An adaptive Metropolis algorithm. Bernoulli pp. 223–242.

Havasi, A., Hernandez-Lobato, J. M, & Murillo-Fuentes, J. J. (2018). Inference in deep Gaussian processes using stochastic gradient Hamiltonian Monte Carlo. *Advances in Neural Information Processing Systems*, pp. 7506–7516.

Hensman, J., Matthews, A.G., Filippone, M., Ghahramani, Z. (2015). Mcmc for variationally sparse gaussian processes. In: Advances in Neural Information Processing Systems, pp. 1648–1656.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning, 37*(2), 183–233.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *The Journal of Machine Learning Research, 6,* 1783–1816.

Lawrence, N.D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. In: Advances in neural information processing systems, pp. 329–336.

Lawrence, N.D., Rattray, M., Titsias, M.K. (2009). Efficient sampling for Gaussian process inference using control variables. In: Advances in Neural Information Processing Systems, pp. 1681–1688.

Lindsten, F., Doucet, A. (2016). Pseudo-Marginal Hamiltonian Monte Carlo. arXiv preprint arXiv:1607.02516.

Murray, I., Prescott Adams, R., MacKay, D.J. (2010). Elliptical slice sampling.

Salimbeni, H., Deisenroth, M. (2017). Doubly stochastic variational inference for deep gaussian processes. In: Advances in Neural Information Processing Systems, pp. 4588–4599.

Sheffield, M.L. (2017). vargplvm. https://github.com/SheffieldML/vargplvm.

Team, R.C., contributors (2013). The R Datasets Package (2013). R package version 3.6.0

Titsias, M.K. (2009). Variational learning of inducing variables in sparse Gaussian processes. In: International Conference on Artificial Intelligence and Statistics, pp. 567–574.

Titsias, M.K., Lawrence, N.D. (2010). Bayesian gaussian process latent variable model. In: International Conference on Artificial Intelligence and Statistics, pp. 844–851.

Turner, R. E., & Sahani, M. (2011). Two problems with variational expectation maximization for time-series models. *Bayesian Time series models, 1*(3.1), 3–5.
Vehtari, A., Gelman, A., & Gabry, J. (2015). Pareto smoothed importance sampling. arXiv preprint arXiv:1507.02646.