Tronarp, Filip; Karvonen, Toni; Särkkä, Simo

Mixture representation of the matérn class with applications in state space approximations and Bayesian quadrature

Published in:
Proceedings of the 2018 IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2018

DOI:
10.1109/MLSP.2018.8516992

Published: 31/10/2018

Document Version
Peer reviewed version

Please cite the original version:
Tronarp, F., Karvonen, T., & Särkkä, S. (2018). Mixture representation of the matérn class with applications in state space approximations and Bayesian quadrature. In N. Pustelnik, Z-H. Tan, Z. Ma, & J. Larsen (Eds.), *Proceedings of the 2018 IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2018* (Vol. 2018-September). [8516992] (IEEE International Workshop on Machine Learning for Signal Processing). https://doi.org/10.1109/MLSP.2018.8516992

Mixture representation of the Matérn class with applications in state space approximations and Bayesian quadrature

Filip Tronarp, Toni Karvonen, and Simo Särkkä

This article has been accepted for publication in *Proceedings of 28th IEEE International Workshop on Machine Learning for Signal Processing*, to be held in Aalborg, Denmark, on September 17–20, 2018. When citing this article, use the following reference:

> F. Tronarp, T. Karvonen, and S. Särkkä, "Mixture representation of the Matérn class with applications in state space approximations and Bayesian quadrature," in *28th IEEE International Workshop on Machine Learning for Signal Processing*, 2018.

2018 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2018, AALBORG, DENMARK

MIXTURE REPRESENTATION OF THE MATÉRN CLASS WITH APPLICATIONS IN STATE SPACE APPROXIMATIONS AND BAYESIAN QUADRATURE

Filip Tronarp, Toni Karvonen, and Simo Särkkä

Aalto University
Department of Electrical Engineering and Automation
Rakentajanaukio 2C, Espoo, Finland

ABSTRACT

In this paper, the connection between the Matérn kernel and scale mixtures of squared exponential kernels is explored. It is shown that the Matérn kernel can be approximated by a finite scale mixture of squared exponential kernels through a quadrature approximation which in turn allows for (i) state space approximations of the Matérn kernel for arbitrary smoothness parameters using established state space approximations of the squared exponential kernel and (ii) exact calculation of the Bayesian quadrature weights for the approximate kernel under a Gaussian measure. The method is demonstrated in inference in a log-Gaussian Cox process as well as in approximating a Gaussian integral arising from a financial problem using Bayesian quadrature.

***Index Terms*—** Gaussian process regression, Matérn covariance, scale mixture representation, state space approximation, Bayesian quadrature

1. INTRODUCTION

Gaussian processes [1] are a prominent tool in signal processing, statistics, and machine learning. Selection of the covariance kernel k of the Gaussian process can have a drastic effect on the performance in any application, be it for example regression [1, 2] or Bayesian quadrature [3, 4]. The Matérn kernel is often seen as the ideal choice since it allows to specify smoothness of the process and contains many other kernels as a special cases [1, Sec. 4.2]. However, the Matérn model becomes intractable in the aforementioned applications because computational cost of Gaussian process regression is cubic in the number of data points (see Eq. (1)) and integrals of the form $\int_\Omega k(x, x')\pi(x)\,\mathrm{d}x$ (see Sec. 4.1) need to be evaluated efficiently in Bayesian quadrature.

The aim of this paper is to develop an approximation to the Matérn model that addresses the above challenges. To this end, we establish a connection between Matérn kernels and scale mixtures of squared exponential kernels. This enables the approximation of the Matérn kernel in terms of a finite mixture of squared exponential kernels. Consequently, (i) state space approximations—with their associated linear computational complexity—of the Matérn kernel become possible when additional approximations to the squared exponential kernel [5, 6, 7] are employed and (ii) Bayesian quadrature for the approximate kernel is made tractable since integrals of squared exponential kernels are computable in many settings of interest. More specifically, our contributions include

1. Demonstration of a link between squared exponential scale mixtures and the Matérn kernel.
2. Construction of a finite *squared exponential mixture approximation* (SEMA) to the Matérn kernel based on generalised Gauss–Laguerre quadrature [8]; see Eq. (8). A similar approximation for the rational quadratic kernel has been developed by Solin and Särkkä [9]. In contrast to our approximation, theirs is based on a well-known scale mixture representation of the rational quadratic kernel [1, Eq. (4.20)].
3. Use of established state space approximations of the squared exponential kernel [5, 6, 7] for constructing state space approximations to the Matérn kernel for arbitrary smoothness parameters, thus allowing for inference in linear time complexity [10].
4. Application of the developed mixture approximation to Bayesian quadrature.

The method is demonstrated in a log-Gaussian Cox process regression and pricing a zero coupon bond under the Vasicek model using Bayesian quadrature.

2. SCALE MIXTURE REPRESENTATION OF THE MATÉRN CLASS

This sections develops a squared exponential scale mixture representation and finite mixture approximations for Matérn kernels defined in Eq. (2).

This work was supported by Academy of Finland projects 266940 and 304087 as well as Aalto ELEC Doctoral School.

2.1. Gaussian processes

A zero mean isotropic Gaussian process f is parametrised by its positive-definite covariance kernel k: $\mathbb{C}[f(x), f(x')] = k(x, x') = k(\|x - x'\|)$. For a given data set $\mathcal{D} = \{(x_n, f(x_n)\}_{n=1}^N = (X, f_X)$, the posterior process is again Gaussian with mean and covariance functions

$$\begin{aligned} \mu_{\mathcal{D}}(x) &= k_X^{\mathsf{T}}(x) K_X^{-1} f_X, \\ k_{\mathcal{D}}(x, x') &= k(x, x') - k_X^{\mathsf{T}}(x) K_X^{-1} k_X(x'), \end{aligned} \tag{1}$$

where $[k_X(x)]_n = k(x, x_n)$ and $[K_X]_{nm} = k(x_n, x_m)$. A major obstacle in applying Gaussian processes is the cubic computational cost of solving the linear systems in Eq. (1).

2.2. Scale mixture for the Matérn class

Matérn kernels are defined as

$$k_{\mathrm{M}}(\tau; \rho, \nu) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\tau}{\rho} \right)^{\nu} \mathrm{K}_{\nu} \left(\frac{\sqrt{2\nu}\tau}{\rho} \right), \tag{2}$$

where $\tau = \|x - x'\|$ and $\|\cdot\|$ is the Euclidean norm on the input space, $\Gamma(\cdot)$ the gamma function, $\mathrm{K}_{\nu}(\cdot)$ the Bessel function of the second kind, and ν a smoothness parameter that can be used in encoding prior beliefs about smoothness of the process being modelled. Denote the squared exponential covariance kernel by

$$k_{\mathrm{SE}}(\tau; \ell) = \exp\left(-\frac{\tau^2}{2\ell^2} \right),$$

where $\ell > 0$ is the length-scale parameter. A squared exponential scale mixture model is then given by

$$\begin{aligned} s &\sim p_S(s), \\ f \mid s &\sim \mathcal{GP}\big(0, \sigma^2 k_{\mathrm{SE}}(\cdot\,; \sqrt{s})\big), \end{aligned}$$

where p_S is a density on $[0, \infty)$. The marginal covariance at distance τ can then be calculated as

$$\mathbb{C}[f(x), f(x')] = \sigma^2 \int_0^{\infty} k_{\mathrm{SE}}(\tau; \sqrt{s}\,) p_S(s)\, \mathrm{d}s. \tag{4}$$

Now, if p_S is selected as a reciprocal gamma density, $p_S(s) = \mathrm{G}^{-1}(s; \alpha, \alpha\ell_{\mathrm{RQ}}^2)$, then the rational quadratic is recovered [1, Sec. 4.2]:

$$k_{\mathrm{RQ}}(\tau; \alpha, \ell_{\mathrm{RQ}}^2) = \int_0^{\infty} k_{\mathrm{SE}}(\tau; \sqrt{s}) \mathrm{G}^{-1}(s; \alpha, \alpha\ell_{\mathrm{RQ}}^2)\, \mathrm{d}s.$$

As it turns out, the integral in Eq. (4) is tractable for also for other choices of p_S. In particular, the Matérn class can be retrieved by setting p_S to a gamma density.

Theorem 1. *Let the process f and the random variable $s > 0$ be governed by the probabilistic model*

$$\begin{aligned} s &\sim \mathrm{G}(\nu, \nu/\rho^2), \\ f \mid s &\sim \mathcal{GP}\big(0, \sigma^2 k_{SE}(\cdot\,; \sqrt{s}\,)\big). \end{aligned}$$

Then the covariance at distance $\tau = \|x - x'\|$ is

$$\mathbb{C}[f(x), f(x')] = \sigma^2 k_M(\tau; \rho, \nu).$$

Proof. The covariance is

$$\begin{aligned} &\mathbb{C}[f(x), f(x')] \\ &= \sigma^2 \int_0^{\infty} k_{\mathrm{SE}}(\tau; \sqrt{s}\,) \mathrm{G}(s; \nu, \nu/\rho^2)\, \mathrm{d}s \\ &= \sigma^2 \frac{(\nu/\rho^2)^{\nu}}{\Gamma(\nu)} \int_0^{\infty} s^{\nu-1} \exp\left(-\frac{\tau^2}{2} s^{-1} - \frac{2\nu}{2\rho^2} s \right) \mathrm{d}s. \end{aligned}$$

The integral is then evaluated as in [11]:

$$\begin{aligned} &\int_0^{\infty} s^{\nu-1} \exp\left(-\frac{\tau^2}{2} s^{-1} - \frac{2\nu}{2\rho^2} s \right) \mathrm{d}s \\ &\quad = \frac{2\mathrm{K}_{\nu}\big(\sqrt{2\nu\tau^2/\rho^2}\big)}{(2\nu/\rho^2\tau^2)^{\nu/2}} \\ &\quad = 2\left(\frac{\rho^2\tau^2}{2\nu} \right)^{\nu/2} \mathrm{K}_{\nu}\left(\sqrt{\frac{2\nu\tau^2}{\rho^2}} \right). \end{aligned} \tag{6}$$

Putting everything together gives

$$\begin{aligned} \mathbb{C}[f(x), f(x')] &= 2\sigma^2 \frac{2^{-\nu}(2\nu/\rho^2)^{\nu}}{\Gamma(\nu)} \left(\frac{\rho\tau}{\sqrt{2\nu}} \right)^{\nu} \mathrm{K}_{\nu}\left(\sqrt{\frac{2\nu\tau^2}{\rho^2}} \right) \\ &= \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\tau}{\rho} \right)^{\nu} \mathrm{K}_{\nu}\left(\frac{\sqrt{2\nu}\tau}{\rho} \right), \end{aligned}$$

which is precisely the Matérn kernel in Eq. (2). □

2.3. Squared exponential mixture approximation

While there is no problem in evaluating the Matérn kernel using standard libraries, the kernel tends to be under-utilised (the exception being when ν is a half-integer: $\nu = n + 1/2$ for $n \in \mathbb{N}$) in applications such as sparse time series regression and Bayesian quadrature due to non-Markovianity and intractability of the kernel mean. Therefore, for general ν an approximation scheme would be beneficial. Here we employ generalised Gauss–Laguerre quadrature as was done in [9]. Making a change of variables $z = \nu/\rho^2 s$ in Eq. (6) gives the Matérn kernel as

$$k_{\mathrm{M}}(\tau; \rho, \nu) = \frac{1}{\Gamma(\nu)} \int_0^{\infty} z^{\nu-1} \exp\left(-z - \frac{\tau^2 \nu}{2\rho^2 z} \right) \mathrm{d}z.$$

Now, $z^{\nu-1}\exp(-z)$ corresponds to the weight function in generalised Gauss–Laguerre quadrature [8]. Hence, a *squared exponential mixture approximation* (SEMA) of the Matérn kernel is given by

$$\sigma^2 k_{\mathrm{M}}(\tau; \rho, \nu) \approx \frac{\sigma^2}{\Gamma(\nu)} \sum_{j=1}^{J} w_j \exp\left(-\frac{\tau^2 \nu}{2\rho^2 z_j} \right),$$

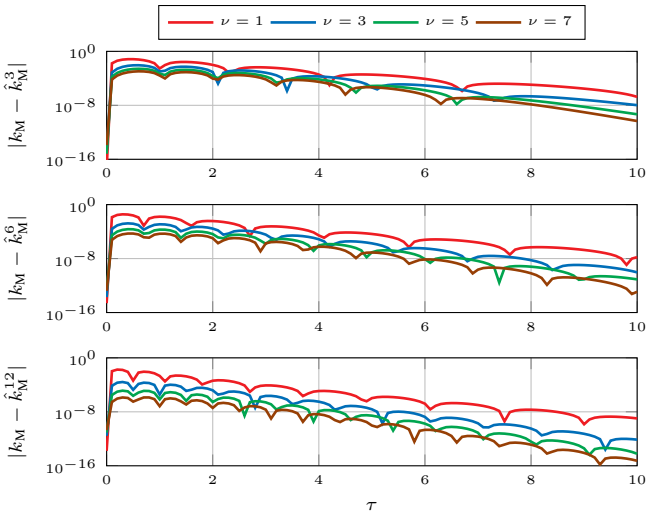

Fig. 1. Error between the exact Matérn kernel and the SEMA (8) for $\rho = \sigma = 1$ and $J = 3$ (top), $J = 6$ (middle), and $J = 12$ (bottom) Gauss–Laguerre nodes.

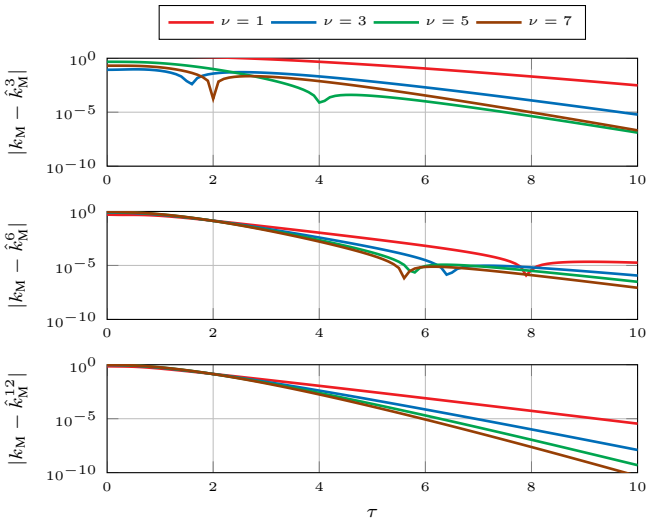

Fig. 2. The logarithm of the absolute value of the error between the exact Matérn kernel and the generalized Gauss–Laguerre approximation, using a state space dimension of $M = 6$ to approximate the squared exponential, for $\rho = \sigma = 1$ and $J = 3$ nodes (top), $J = 6$ nodes (middle), and $J = 12$ nodes (bottom).

where the nodes z_j are the roots of $\mathrm{L}_J^{\nu-1}$, the generalised Laguerre polynomial [12, Ch. 22] of degree J and index $\nu - 1$, and the weights w_j are

$$w_j = \frac{\Gamma(J+\nu) z_j}{\Gamma(J+1)(N+1)^2 [\mathrm{L}_{J+1}^{\nu-1}(z_j)]^2}.$$

Error of this approximation can be assessed using [13, Thm. 1]. Define

$$\hat{\sigma}_j^2 = \frac{\sigma^2 w_j}{\Gamma(\nu)} \quad \text{and} \quad \hat{\ell}_j^2 = \frac{z_j \rho^2}{\nu}. \tag{7}$$

Then the approximation takes the form of a sum of J independent Gaussian processes with squared exponential covariance kernels of varying length-scales:

$$\sigma^2 k_{\mathrm{M}}(\tau; \rho, \nu) \approx \sigma^2 \hat{k}_{\mathrm{M}}^J(\tau; \rho, \nu) = \sum_{j=1}^{J} \hat{\sigma}_j k_{\mathrm{SE}}(\tau; \hat{\ell}_j). \tag{8}$$

The accuracy of the approximation for various parameter selections is depicted in Fig. 1. As can be seen, the approximation quality improves as ν, τ, and J become large.

3. APPLICATION I: STATE SPACE APPROXIMATIONS

Our first application of the SEMA in Eq. (8) is to state space approximations of Gaussian processes using the Matérn kernel, enabling inference in linear time complexity [10].

3.1. Stationary Gaussian state space models

A large class of single input (here denoted by t) zero mean stationary Gaussian process can be represented as a stochastic differential equation as follows [14]:

$$\begin{aligned} \mathrm{d}u(t) &= Au(t)\,\mathrm{d}t + Q^{1/2}\,\mathrm{d}B(t), \\ f(t) &= Hu(t), \end{aligned}$$

where A is a matrix corresponding to an asymptotically stable dynamic system, $Q^{1/2}$ is the diffusion matrix, $B(t)$ is a vector of standard Brownian motions, and the initial distribution is

$$u(0) \sim \mathcal{N}(0, \Sigma(0)),$$

where the stationary covariance matrix $\Sigma(0)$ solves the Lyapunov equation $0 = A\Sigma(0) + \Sigma(0)A^{\mathsf{T}} + Q$. For a time grid $\{t_n\}_{n=1}^N$, the equivalent discrete-time system is

$$\begin{aligned} u(t_n) \mid u(t_{n-1}) &\sim \mathcal{N}\big(\Phi(\Delta_n) u(t_{n-1}), Q(\Delta_n)\big), \\ f(t_n) &= Hu(t_n), \end{aligned}$$

where $\Delta_n = t_n - t_{n-1}$ and

$$\begin{aligned} \Phi(\tau) &= \exp(A\tau), \\ Q(\tau) &= \int_0^{\Delta_n} \exp(As) Q \exp(A^{\mathsf{T}} s)\,\mathrm{d}s. \end{aligned}$$

Furthermore, the covariance kernel of $f(t)$ is given by

$$k(\tau) = \begin{cases} H\Sigma(0)\Phi^T(\tau)H^{\mathsf{T}}, & \tau \geq 0, \\ H\Phi(-\tau)\Sigma(0)H^{\mathsf{T}}, & \tau \leq 0. \end{cases} \tag{9}$$

3.2. State space approximation of the Matérn kernel for arbitrary smoothness parameters

While the Matérn kernels admit state space representations when the smoothness parameter is a half-integer [6, 10], this does not hold in general. Therefore, for the purposes of Gaussian process regression with linear time complexity, approximations need to be used. In essence this amounts to finding model matrices A, Q, and H such that the covariance kernel in Eq. (9) is a good approximation to the true kernel.

Finding suitable model matrices can be tricky in general. State space approximations for the squared exponential kernel go back at least to [5]. More recent approximations based on Taylor and Padé expansions appear in [6, 7]; see also [15]. This means that the sum of squared exponential kernels in Eq. (8), and thus the Matérn kernel for arbitrary smoothness parameter, can be approximated as a state space model using any of the aforementioned techniques.

To demonstrate the approximation quality, the Taylor series method [6, 7] with a state space dimension of $M = 6$ is used to approximate the squared exponential kernels. The resulting covariance kernel is compared to exact evaluations of the Matérn kernel for various parameter values in Fig. 2. It is again noted that the approximation quality improves when ν, τ, and J are increased. As expected, some accuracy is lost.

3.3. Numerical example

Here the preceding state space approximation of the Matérn kernel is evaluated on inference in a log-Gaussian Cox process. The data consists of 191 time stamps of coal mine explosion that killed ten or more people in Britain between the years 1851 and 1962 [16]. The time span is partitioned into 2^8 intervals and the number of events in each interval is modelled as an independent Poisson random variable conditioned on a latent intensity process that is modelled as a log-Gaussian process:

$$\begin{aligned} f(t) &\sim \mathcal{GP}(0, k_{\mathrm{M}}), \\ y(t_n) &\sim \mathrm{Po}(\lambda(t_n)), \quad \lambda(t) = \exp(f(t)). \end{aligned}$$

The smoothness parameter is fixed to $\nu = 1$, while the range $\rho = \exp(\theta_1)$ and standard deviation $\sigma = \exp(\theta_2)$ are estimated by the maximum marginal likelihood method for both exact evaluations of the Matérn covariance kernel and a state space approximation using $J = 6$ nodes in the generalised Gauss–Laguerre quadrature and the Taylor series method [6] with $M = 8$ state dimensions to approximate the squared exponential covariance kernels. A comparison between exact inference and the described state-space SEMA (SEMA-SSM) is shown in Fig. 3. While a disagreement between the estimates is visible, qualitative features are similar.

4. APPLICATION II: BAYESIAN QUADRATURE

This section presents an application of the SEMA to approximating kernel means needed in Bayesian quadrature.

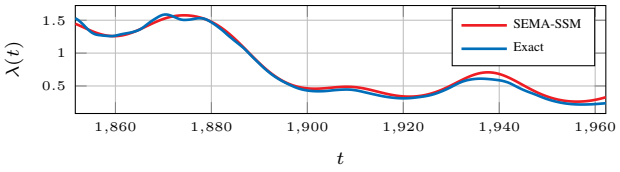

Fig. 3. The exponentiated posterior mean for SEMA-SSM (red) and exact covariance function evaluations (blue).

4.1. Bayesian quadrature

Bayesian quadrature [3, 4] is a probabilistic approach to computing a numerical approximation of the integral

$$\pi(f^\dagger) = \int_\Omega f^\dagger(x)\pi(x)\,\mathrm{d}x$$

of a deterministic function $f^\dagger \colon \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^d$, with respect to a probability density function π. The integrand is assigned a Gaussian process prior $f \sim \mathcal{GP}(0, k)$ that is conditioned on the noiseless data $\mathcal{D}$ of N evaluations of the integrand. Because integration is a linear operation, the integral $\pi(f)$ has a Gaussian posterior distribution: $\pi(f) \mid \mathcal{D} \sim \mathcal{N}(\mu_{\pi,\mathcal{D}}, \sigma^2_{\pi,\mathcal{D}})$. By the Gaussian process posterior equations (1), the mean and variance are

$$\begin{aligned} \mu_{\pi,\mathcal{D}} &= k_{\pi,X}^{\mathsf{T}} K_X^{-1} f_X^\dagger, \\ \sigma^2_{\pi,\mathcal{D}} &= \pi(k_\pi) - k_{\pi,X}^{\mathsf{T}} K_X^{-1} k_{\pi,X}, \end{aligned} \tag{10}$$

where $k_\pi(x) = \int_\Omega k(x, x')\pi(x)\,\mathrm{d}x$ is the *kernel mean function* and $[k_{\pi,X}]_n = k_\pi(x_n)$. The posterior mean $\mu_{\pi,\mathcal{D}}$ provides a point estimate for the unknown integral $\pi(f^\dagger)$ while $\sigma_{\pi,\mathcal{D}}$ can be used to assess uncertainty about this approximation [17, 18, 19]. In particular, the posterior mean takes the form of a quadrature rule:

$$\mu_{\pi,\mathcal{D}} = \sum_{n=1}^{N} w_n^* f^\dagger(x_n),$$

where the weights are

$$w^* = K^{-1} k_{\pi,X}. \tag{11}$$

4.2. Kernel mean approximation

One of the problems in Bayesian quadrature is the need to compute the kernel mean function. There are a number of pairs of π and k that result in closed-form expression for the kernel mean [4, Section 4.2]. Unfortunately, in the commonly occurring case of Gaussian densities there is no such expression for the Matérn kernel for arbitrary values of the smoothness parameter ν. The scale mixture approximation introduced in Sec. 2 provides a convenient and accurate approximation.

Let φ stand for the standard Gaussian density function:

$$\varphi(x) = (2\pi)^{-d/2} \exp\left(-\frac{\|x\|^2}{2}\right). \tag{12}$$

Consider the Matérn kernel $k(\tau) := \sigma^2 k_{\mathrm{M}}(\tau; \rho, \nu)$ that, as derived in Sec. 2.3, admits the approximation

$$k(\tau) \approx \sum_{j=1}^{J} \hat{\sigma}_j^2 k_{\mathrm{SE}}(\tau; \hat{\ell}_j)$$

as a sum of squared exponential kernels. As is well-known, these kernels can be integrated in closed form:

$$\hat{\sigma}_j^2 k_{\mathrm{SE},\varphi}(x; \hat{\ell}_j) = \hat{\sigma}_j^2 \left(\frac{\hat{\ell}_j^2}{1+\hat{\ell}_j^2}\right)^{d/2} \exp\left(-\frac{\|x\|^2}{2(1+\hat{\ell}_j^2)}\right),$$

$$\hat{\sigma}_j^2 \varphi\big(k_{\mathrm{SE},\varphi}(\cdot; \hat{\ell}_j)\big) = \hat{\sigma}_j^2 \left(\frac{\hat{\ell}_j^2}{2(1+\hat{\ell}_j^2)}\right)^{d/2}.$$

We can thus form the approximations

$$\sigma^2 k_{\varphi}(x; \rho, \nu) \approx \sum_{j=1}^{J} \hat{\sigma}_j^2 \left(\frac{\hat{\ell}_j^2}{1+\hat{\ell}_j^2}\right)^{d/2} \exp\left(-\frac{\|x\|^2}{2(1+\hat{\ell}_j^2)}\right),$$

$$\sigma^2 \varphi(k_{\varphi}(\cdot; \rho, \nu)) \approx \sum_{j=1}^{J} \hat{\sigma}_j^2 \left(\frac{\hat{\ell}_j^2}{2(1+\hat{\ell}_j^2)}\right)^{d/2},$$

where $\hat{\sigma}_j$ and $\hat{\ell}_j$ are defined in Eq. (7). What makes the above kernel mean approximation attractive is its positivity, following from positivity of the generalised Laguerre quadrature weights, for any $x \in \mathbb{R}$.

To obtain an approximate Bayesian quadrature rule for the Gaussian density φ, we merely replace the true Matérn kernel mean $k_{\mathrm{M},\varphi}(x; \rho, \nu, \sigma)$ and its integral $\varphi(k_{\mathrm{M},\varphi}(\cdot; \rho, \nu, \sigma))$ with the approximation above when computing the integral posterior mean and variance in Eq. (10). Next we assess accuracy of the resulting quadrature weight approximation.

4.3. Approximation accuracy

We experiment with accuracy of the Bayesian quadrature weight approximation developed in Sec. 4.2 in the case $d = 1$. We compute the Bayesian quadrature weights $w^* \in \mathbb{R}^N$ in Eq. (11) for $N = 12$ and $N = 24$, the standard Gaussian density function (12), and Matérn kernels with smoothness parameters $\nu = 3$ and $\nu = 7$ and $\rho = 1$. In the Gauss–Laguerre quadrature we use $J = 2, \ldots, 30$. For comparison, we also compute naive J-point Gauss–Hermite approximations to the intractable Matérn kernel mean. The reference value for the weights is computed to high precision using the MATLAB function `integral`. Approximation accuracy is measured in terms of the relative error

$$\sqrt{\sum_{i=1}^{N} \left(\frac{w_i^* - w_i^J}{w_i^*}\right)^2}, \tag{13}$$

where w_i^J are the approximate Bayesian quadrature weights obtained using J-point SEMA or Gauss–Hermite quadrature.

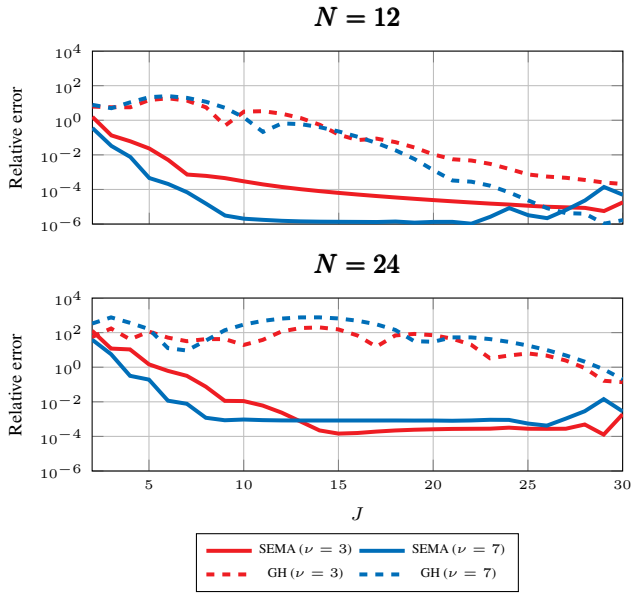

Fig. 4. Relative approximation errors (13) to N-point Bayesian quadrature weights with two different Matérn kernels using J-point SEMA and naive Gauss–Hermite (GH) integration of the kernel mean.

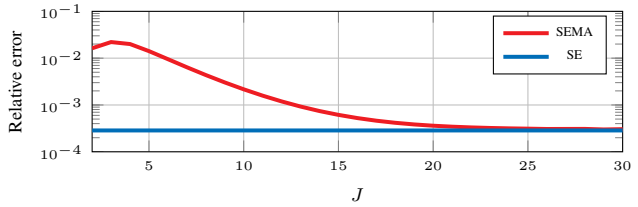

Fig. 5. Relative integration error for the 15-dimensional zero coupon bond integrand in Sec. 4.4. Plotted are errors by Bayesian quadrature with (i) a Matérn kernel whose kernel means are computed with J-point SEMA approximation and (ii) the squared exponential kernel.

The results are depicted in Fig. 4 where it is apparent that accuracy of the SEMA approximation is superior to direct numerical integration of the kernel mean using Gauss–Hermite quadrature.

4.4. Financial example

This section experiments with the weight approximation for a zero coupon bond example [20, Section 6.1]. The Gaussian integral that needs to be computed arises from a D-step uniform Euler–Maruyama discretisation of the Vasicek stochastic differential equation model and represents the price of a zero coupon bond with given maturity time. The integral is of dimension $D - 1$ and has an analytical solution, making this a convenient test problem. See the aforementioned reference for

more details.

We set $D = 16$ (i.e., $d = 15$) and draw 1,000 samples from the 15-dimensional standard normal distribution. Fig. 5 contains relative integration errors of Bayesian quadrature approximations that use (i) the Matérn kernel having parameters $\nu = 6$ and $\rho = d$ and with the kernel means approximated with J-point SEMA and (ii) the squared exponential kernel with $\ell = \rho$. It is seen that the SEMA-based approximation can be as accurate as one that has closed-form kernel means if J is sufficiently large.

5. CONCLUSIONS AND DISCUSSION

The connection between the Matérn covariance function and squared exponential scale mixtures was established and its approximation in terms of finite scale mixtures by means of generalised Gauss–Laguerre quadrature was presented. It was shown how this approximation can be used to achieve linear time complexity in regression by using state space approximations of the squared exponential kernel and in approximating the weights in Bayesian quadrature.

A topic for future investigation is to see if approximating the squared exponential kernel by matching the differentiability of the underlying Matérn kernel, which might be achieved by the Padé method [7], can reduce the loss of accuracy.

6. REFERENCES

[1] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[2] A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes, *Handbook of Spatial Statistics*, CRC press, 2010.

[3] A. O'Hagan, "Bayes–Hermite quadrature," *Journal of Statistical Planning and Inference*, vol. 29, no. 3, pp. 245–260, 1991.

[4] F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic, "Probabilistic integration: A role in statistical computation?," *Statistical Science*, 2018, Accepted for publication.

[5] R. L. Stratonovich, *Topics in the Theory of Random Noise*, Gordon and Breach, 1963.

[6] J. Hartikainen and S. Särkkä, "Kalman filtering and smoothing solutions to temporal Gaussian process regression models," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2010, pp. 379–384.

[7] S. Särkkä and R. Piché, "On convergence and accuracy of state-space approximations of squared exponential covariance functions," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2014.

[8] P. Rabinowitz and G. Weiss, "Tables of abscissas and weights for numerical evaluation of integrals of the form $\int_0^\infty e^{-x} x^n f(x) dx$," *Mathematical Tables and Other Aids to Computation*, vol. 13, no. 68, pp. 285–294, 1959.

[9] A. Solin and S. Särkkä, "Gaussian quadratures for state space approximation of scale mixtures of squared exponential covariance functions," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2014.

[10] S. Särkkä, A. Solin, and J. Hartikainen, "Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 51–61, 2013.

[11] O. Barndorff-Nielsen, J. Kent, and M. Sørensen, "Normal variance-mean mixtures and z distributions," *International Statistical Review*, vol. 50, pp. 145–159, 1982.

[12] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, vol. 55, Courier Corporation, 1964.

[13] G. Mastroianni and G. Monegato, "Error estimates for Gauss-Laguerre and Gauss-Hermite quadrature formulas," in *Approximation and Computation: A Festschrift in Honor of Walter Gautschi*, pp. 421–434. Springer, 1994.

[14] B. Øksendal, *Stochastic Differential Equations – An Introduction with Applications*, Springer, 2003.

[15] T. Karvonen and S. Särkkä, "Approximate state-space Gaussian processes via spectral transformation," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2016.

[16] D. J. Hand, F. Daly, K. McConway, D. Lunn, and E. Ostrowski, *A Handbook of Small Data Sets*, vol. 1, Chapman and Hall, 1993.

[17] F. M. Larkin, "Gaussian measure in Hilbert space and applications in numerical analysis," *The Rocky Mountain Journal of Mathematics*, vol. 2, no. 3, pp. 379–421, 1972.

[18] P. Diaconis, "Bayesian numerical analysis," in *Statistical Decision Theory and Related Topics IV*, vol. 1, pp. 163–175. Springer-Verlag New York, 1988.

[19] P. Hennig, M. A. Osborne, and M. Girolami, "Probabilistic numerics and uncertainty in computations," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 471, no. 2179, 2015.

[20] M. Holtz, *Sparse Grid Quadrature in High Dimensions with Applications in Finance and Insurance*, Springer, 2011.