

FAST AUTOMATIC BAYESIAN CUBATURE USING MATCHING KERNELS
AND DESIGN

BY

R. JAGADEESWARAN

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Applied Mathematics
in the Graduate College of the
Illinois Institute of Technology

Approved _____
Advisor

Chicago, Illinois
Received: date / Accepted: date

ACKNOWLEDGMENT

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENT	iii
LIST OF TABLES	vii
LIST OF FIGURES	viii
ABSTRACT	ix
CHAPTER	
1. INTRODUCTION	1
2. BAYESIAN CUBATURE	4
2.1 Bayesian posterior error	4
2.2 Parameter estimation	6
2.2.1 Empirical Bayes	6
Gradient descent to find optimal shape parameter	7
2.2.2 Full Bayes	8
2.2.3 Full Bayes with generic prior	10
2.2.4 Generalized Cross-Validation	14
2.3 The automatic Bayesian cubature algorithm	16
2.4 Example with the Matérn kernel	17
3. FAST AUTOMATIC BAYESIAN CUBATURE	21
3.0.1 Fast Transform Kernel	21
3.0.2 Empirical Bayes	23
3.0.3 Kernel derivative using fast transform	24
3.0.4 Full Bayes	25

3.0.5 Generalized Cross-Validation	26
4. INTEGRATION LATTICES AND SHIFT INVARIANT KERNELS	27
4.1 Extensible Integration Lattice Node Sets	27
4.2 Shift Invariant Kernels	28
4.2.1 Derivative of the kernel	29
4.2.2 Shape parameter search using steepest descent . . .	30
4.2.3 Eigenvectors	30
4.2.4 Iterative Computation of the Fast Transform	32
4.2.5 Overcoming Cancellation Error	34
5. SOBOLOV POINTS AND WALSH KERNELS	37
5.1 Higher order digital nets	37
5.1.1 Walsh functions	38
5.2 Higher order Walsh space	38
5.3 Walsh kernels	38
5.3.1 The worst-case error	39
5.3.2 Calculating the infinite sum	40
5.3.3 Closed form expressions	41
Consider the case $\alpha = 2.42$	
6. NUMERICAL IMPLEMENTATION	48
6.1 Periodizing Variable Transformations	48
7. NUMERICAL RESULTS AND OBSERVATIONS	50
7.1 Test Results and Observations	50
Multivariate Normal Probability.50 Keister's Example.51 Option Pricing.54	
7.2 Diagnostics for the Gaussian Process Assumption	59

7.2.1 Posterior	60
Non-informative prior60 TODO: How to interpret as t-distribution?61 Generic prior61 TODO: How to integrate further?62	
8. DISCUSSION AND FURTHER WORK	63
9. FUTURE WORK	64
9.1 Future work	64
Acknowledgments	64
APPENDIX	65
BIBLIOGRAPHY	65

LIST OF TABLES

Table	Page
-------	------

LIST OF FIGURES

Figure		Page
2.1	The $d = 3$ multivariate normal probability transformed to an integral of f_{Genz} with $d = 2$	18
2.2	Multivariate Normal probability estimated using Matern kernel in $d=2$ using empirical stopping criterion. Top: Guaranteed integration within error tolerance ε . Bottom: Computation time rapidly increases with increase of n	19
4.1	Example of a shifted integration lattice node set in $d = 2$	28
4.2	Fourier kernel	30
7.1	Guaranteed:	51
7.2	MVN guaranteed : FB	52
7.3	MVN guaranteed : GCV	53
7.4	Keister guaranteed:MLE	55
7.5	Keister guaranteed:FB	55
7.6	Keister guaranteed:GCV	56
7.7	Option pricing Guaranteed: MLE	57
7.8	OptPrice guaranteed : FB	58
7.9	OptPrice guaranteed : GCV	58

ABSTRACT

Automatic cubatures approximate multidimensional integrals to user-specified error tolerances. For high dimensional problems, it makes sense to fix the sampling density but determine the sample size, n , automatically. Bayesian cubature postulates that the integrand is an instance of a stochastic process. Here we assume a Gaussian process parameterized by a constant mean and a covariance function defined by a scale parameter and a function specifying how the integrand values at two different points in the domain are related. These parameters are estimated from integrand values or are given non-informative priors. This leads to credible interval for the integral. The sample size, n , is chosen to make the credible interval for the Bayesian posterior error no greater than the error tolerance.

However, the process just outlined typically requires vector-matrix operations with a computational cost of $O(n^3)$. Our innovation is to pair low discrepancy nodes with matching kernels that lower the computational cost to $O(n \log n)$. This approach is demonstrated using rank-1 lattice sequences and shift-invariant kernels. It is also implemented in the Guaranteed Automatic Integration Library (GAIL).

CHAPTER 1

INTRODUCTION

Cubature is the problem of inferring a numerical value for an integral, $\mu := \int_{\mathbb{R}^d} g(\mathbf{x}) \, d\mathbf{x}$, where μ has no closed form analytic expression. Typically, g is accessible through a black-box function routine. Cubature is a key component of many problems in scientific computing, finance, statistical modeling, and machine learning.

The integral may often be expressed as

$$\mu := \mu(f) := \mathbb{E}[f(\mathbf{X})] = \int_{[0,1]^d} f(\mathbf{x}) \, d\mathbf{x}, \quad (1.1)$$

where $f : [0, 1]^d \rightarrow \mathbb{R}$ is the integrand, and $\mathbf{X} \sim \mathcal{U}[0, 1]^d$. The process of transforming the original integral into the form of (1.1) is not addressed here. The cubature may be an affine function of integrand values:

$$\hat{\mu} := \hat{\mu}(f) := w_0 + \sum_{i=1}^n f(\mathbf{x}_i) w_i, \quad (1.2)$$

where the weights, w_0 , and $\mathbf{w} = (w_i)_{i=1}^n \in \mathbb{R}^n$, and the nodes, $\{\mathbf{x}_i\}_{i=1}^n \subset [0, 1]^d$, are chosen to make the error, $|\mu - \hat{\mu}|$, small. The integration domain $[0, 1]^d$ is convenient for the low discrepancy node sets that we use. The nodes are assumed to be deterministic.

We construct a reliable stopping criterion that determines the number of integrand values required, n , to ensure that the error is no greater than a user-defined error tolerance denoted ε , i.e.,

$$|\mu - \hat{\mu}| \leq \varepsilon. \quad (1.3)$$

Rather than relying on strong assumptions about the integrand, such as an upper bound on its variance or total variation, we construct a stopping criterion that is based on a credible interval arising from a Bayesian approach to the problem. We build upon

the work of Briol et al. [2], Diaconis [6], O’Hagan [23], Ritter [27], Rasmussen and Ghahramani [25], and others. Our algorithm is an example of *probabilistic numerics*.

The primary contribution of this article is to demonstrate how the choice of a family of covariance kernels that match the low discrepancy sampling nodes facilitates fast computation of the cubature and the data-driven stopping criterion. Our cubature requires n function values—at a cost of $\$(f)$ each—plus $\mathcal{O}(n \log(n))$ operations to check whether the error tolerance is satisfied. The total cost of our algorithm is then $\mathcal{O}(n[\$(f) + \log(n)])$. This is significantly fewer operations than the $\mathcal{O}(n^3)$ typically required for Bayesian cubature. If function evaluation is expensive, then $\$(f)$ might be similar in magnitude to $\log(n)$.

Hickernell [12] compares different approaches to cubature error analysis depending on whether the rule is deterministic or random and whether the integrand is assumed to be deterministic or random. Error analysis that assumes a deterministic integrand lying in a Banach space leads to an error bound that is typically impractical for deciding how large n must be to satisfy (1.3). The deterministic error bound includes a (semi-) norm of the integrand, which is often more complex to compute than the original integral.

Hickernell and Jiménez-Rugama [13, 17] have developed stopping criteria for cubature rules based on low discrepancy nodes by tracking the decay of the discrete Fourier coefficients of the integrand. The algorithm proposed here also relies on discrete Fourier coefficients, but in a different way. Although we only explore automatic Bayesian cubature for absolute error tolerances, the recent work by Hickernell, Jiménez-Rugama, and Li [14] suggest how one might accommodate more general error criteria, such as relative error tolerances.

Section 2 explains the Bayesian approach to estimate the posterior cubature

error and defines our automatic Bayesian cubature. Although much of this material is known, it is included for completeness. We end Section 2 by demonstrating why Bayesian cubature is typically computationally expensive. Section 3 introduces the concept of covariance kernels that match the nodes and expedite the computations required by our automatic Bayesian cubature. Section 4 implements this concept for shift invariant kernels and rank-1 lattice nodes. It also describes how to avoid cancellation error for kernels of product form. Numerical examples are provided in Section 6 to demonstrate our new algorithm. We conclude with a brief discussion.

CHAPTER 2

BAYESIAN CUBATURE

2.1 Bayesian posterior error

Suppose that the integrand, f , is drawn from a Gaussian process, i.e., $f \sim \mathcal{GP}(m, s^2 C_\theta)$. Specifically, f has real-valued constant mean m and covariance function $s^2 C_\theta$, where s is a non-negative scale factor, and $C_\theta : [0, 1]^d \times [0, 1]^d \rightarrow \mathbb{R}$ is a symmetric, positive-definite function and parameterized by θ :

$$C^T = C, \quad \mathbf{a}^T C \mathbf{a} > 0, \quad \text{where } C = (C_\theta(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n, \\ \text{for all } \mathbf{a} \neq 0, \quad n \in \mathbb{N}, \quad \mathbf{x}_1, \dots, \mathbf{x}_n \in [0, 1]^d. \quad (2.1)$$

The covariance function, C , and the Gram matrix, C depend implicitly on θ , but the notation may omit this for simplicity's sake.

For a Gaussian process, all vectors of linear functionals of f have a multivariate Gaussian distribution. Defining $\mathbf{f} := (f(\mathbf{x}_i))_{i=1}^n$ as the multivariate normal vector of function values, it follows that

$$\mathbf{f} \sim \mathcal{N}(m\mathbf{1}, s^2 C), \quad (2.2a)$$

$$\mu \sim \mathcal{N}(m, s^2 c_0), \quad (2.2b)$$

$$\text{where } c_0 = \int_{[0,1]^{2d}} C_\theta(\mathbf{x}, \mathbf{t}) \, d\mathbf{x} \, d\mathbf{t}, \quad (2.2c)$$

$$\text{cov}(\mathbf{f}, \mu) = \left(\int_{[0,1]^d} C(\mathbf{t}, \mathbf{x}_i) \, d\mathbf{t} \right)_{i=1}^n =: \mathbf{c}. \quad (2.2d)$$

We need the following lemma to derive the posterior error of our cubature.

Lemma 2.1.1 *[26, (A.6), (A.11–13)] If $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)^T \sim \mathcal{N}(\mathbf{m}, C)$, where \mathbf{Y}_1 and*

\mathbf{Y}_2 are random vectors of arbitrary length, and

$$\mathbf{m} = \begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{pmatrix} = \begin{pmatrix} \mathbb{E}(\mathbf{Y}_1) \\ \mathbb{E}(\mathbf{Y}_2) \end{pmatrix},$$

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{21}^T \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} = \begin{pmatrix} \text{var}(\mathbf{Y}_1) & \text{cov}(\mathbf{Y}_1, \mathbf{Y}_2) \\ \text{cov}(\mathbf{Y}_2, \mathbf{Y}_1) & \text{var}(\mathbf{Y}_2) \end{pmatrix}$$

then

$$\mathbf{Y}_1 | \mathbf{Y}_2 \sim \mathcal{N}(\mathbf{m}_1 + \mathbf{C}_{21}^T \mathbf{C}_{22}^{-1}(\mathbf{Y}_2 - \mathbf{m}_2), \quad \mathbf{C}_{11} - \mathbf{C}_{21}^T \mathbf{C}_{22}^{-1} \mathbf{C}_{21}).$$

Moreover, the inverse of the matrix \mathbf{C} may be partitioned as

$$\mathbf{C}^{-1} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{21}^T \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

$$\mathbf{A}_{11} = (\mathbf{C}_{11} - \mathbf{C}_{12} \mathbf{C}_{22}^{-1} \mathbf{C}_{21})^{-1}, \quad \mathbf{A}_{21} = -\mathbf{C}_{22}^{-1} \mathbf{C}_{21} \mathbf{A}_{11},$$

$$\mathbf{A}_{22} = \mathbf{C}_{22}^{-1} + \mathbf{C}_{22}^{-1} \mathbf{C}_{21} \mathbf{A}_{11} \mathbf{C}_{21}^T \mathbf{C}_{22}^{-1}.$$

It follows from Lemma 2.1.1 that the *conditional* distribution of the integral given observed function values, $\mathbf{f} = \mathbf{y}$ is also Gaussian:

$$\mu | (\mathbf{f} = \mathbf{y}) \sim \mathcal{N}(m(1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1}) + \mathbf{c}^T \mathbf{C}^{-1} \mathbf{y}, s^2(c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})). \quad (2.3)$$

The natural choice for the cubature is the posterior mean of the integral, namely,

$$\hat{\mu} | (\mathbf{f} = \mathbf{y}) = m(1 - \mathbf{1}^T \mathbf{C}^{-1} \mathbf{c}) + \mathbf{c}^T \mathbf{C}^{-1} \mathbf{y}, \quad (2.4)$$

which takes the form of (1.2). Under this definition, the cubature error has zero mean and a variance depending on the choice of nodes:

$$(\mu - \hat{\mu}) | (\mathbf{f} = \mathbf{y}) \sim \mathcal{N}(0, \quad s^2(c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})).$$

A credible interval for the integral is given given by

$$\mathbb{P}_f \left[|\mu - \hat{\mu}| \leq 2.58s \sqrt{c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}} \right] = 99\%. \quad (2.5)$$

Naturally, 2.58 and 99% can be replaced by other quantiles and credible levels.

2.2 Parameter estimation

The credible interval in (2.5) suggests how our automatic Bayesian cubature proceeds. Integrand data is accumulated until the width of the credible interval is no greater than the error tolerance. However, the credible interval still depends on the parameters m, s , and $\boldsymbol{\theta}$. These should not be assumed a priori but be inferred from the data. This section describes three approaches.

2.2.1 Empirical Bayes

One approach is to estimate the parameters is via maximum likelihood estimation (MLE). The log-likelihood function of the parameters given the function data \mathbf{y} is:

$$\begin{aligned} l(s, m, \boldsymbol{\theta} | \mathbf{y}) = & -\frac{1}{2} s^{-2} (\mathbf{y} - m\mathbf{1})^T \mathbf{C}^{-1} (\mathbf{y} - m\mathbf{1}) \\ & - \frac{1}{2} \log(\det \mathbf{C}) - \frac{n}{2} \log(s^2) + \text{constants}. \end{aligned}$$

Maximizing the log-likelihood first with respect to m , then with respect to s , and finally with respect to $\boldsymbol{\theta}$ yields

$$m_{\text{MLE}} = \frac{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{y}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}}, \quad (2.6)$$

$$\begin{aligned} s_{\text{MLE}}^2 = & \frac{1}{n} (\mathbf{y} - m_{\text{MLE}} \mathbf{1})^T \mathbf{C}^{-1} (\mathbf{y} - m_{\text{MLE}} \mathbf{1}) \\ = & \frac{1}{n} \mathbf{y}^T \left[\mathbf{C}^{-1} - \frac{\mathbf{C}^{-1} \mathbf{1} \mathbf{1}^T \mathbf{C}^{-1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \right] \mathbf{y}, \end{aligned} \quad (2.7)$$

$$\boldsymbol{\theta}_{\text{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ \log \left(\mathbf{y}^T \left[\mathbf{C}^{-1} - \frac{\mathbf{C}^{-1} \mathbf{1} \mathbf{1}^T \mathbf{C}^{-1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \right] \mathbf{y} \right) + \frac{1}{n} \log(\det(\mathbf{C})) \right\}. \quad (2.8)$$

The MLE estimate of $\boldsymbol{\theta}$ balances minimizing the covariance scale factor, s_{MLE}^2 , against minimizing $\det(\mathbf{C})$.

Under these estimates of the parameters, the cubature (2.4) and the credible interval (2.5) simplify to

$$\hat{\mu}_{\text{MLE}} = \left(\frac{(1 - \mathbf{1}^T \mathbf{C}^{-1} \mathbf{c}) \mathbf{1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} + \mathbf{c} \right)^T \mathbf{C}^{-1} \mathbf{y}, \quad (2.9)$$

$$\text{err}_{\text{MLE}}^2 := \frac{2.58^2}{n} \mathbf{y}^T \left[\mathbf{C}^{-1} - \frac{\mathbf{C}^{-1} \mathbf{1} \mathbf{1}^T \mathbf{C}^{-1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \right] \mathbf{y} (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}), \quad (2.10)$$

$$\mathbb{P}_f [|\mu - \hat{\mu}_{\text{MLE}}| \leq \text{err}_{\text{MLE}}] = 99\%. \quad (2.11)$$

Here c_0 , \mathbf{c} , and \mathbf{C} are assumed implicitly to be based on $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{MLE}}$.

Gradient descent to find optimal shape parameter

Searching for $\boldsymbol{\theta}_{\text{MLE}}$ as defined in (2.8) is not very efficient. Using the derivative of $l(s, m, \boldsymbol{\theta}|\mathbf{y})$, We can apply gradient descent to find the optimal value faster. Let us define the objective function for the same purpose by excluding the negative sign, which modifies the problem as minimization

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) := \frac{1}{n} \log(\det \mathbf{C}) + \log \left((\mathbf{y} - m_{\text{MLE}} \mathbf{1})^T \mathbf{C}^{-1} (\mathbf{y} - m_{\text{MLE}} \mathbf{1}) \right) + \text{constants}.$$

Taking derivative with respect to $\boldsymbol{\theta}_i$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}_i} \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) &= \frac{1}{n} \frac{\partial}{\partial \boldsymbol{\theta}_i} \log(\det \mathbf{C}) + \frac{\partial}{\partial \boldsymbol{\theta}_i} \log \left((\mathbf{y} - m_{\text{MLE}} \mathbf{1})^T \mathbf{C}^{-1} (\mathbf{y} - m_{\text{MLE}} \mathbf{1}) \right) \\ &= \frac{1}{n} \text{trace} \left(\mathbf{C}^{-1} \frac{\partial \mathbf{C}}{\partial \boldsymbol{\theta}_i} \right) - \frac{\left((\mathbf{y} - m_{\text{MLE}} \mathbf{1})^T \mathbf{C}^{-1} \right)^T \left(\frac{\partial \mathbf{C}}{\partial \boldsymbol{\theta}_i} \right) \left((\mathbf{y} - m_{\text{MLE}} \mathbf{1})^T \mathbf{C}^{-1} \right)}{(\mathbf{y} - m_{\text{MLE}} \mathbf{1})^T \mathbf{C}^{-1} (\mathbf{y} - m_{\text{MLE}} \mathbf{1})} \end{aligned}$$

where we used some of the results from [18]. This can be used further for steepest gradient descent as follows,

$$\boldsymbol{\theta}_i^{(j+1)} = \boldsymbol{\theta}_i^{(j)} - \nu \frac{\partial}{\partial \boldsymbol{\theta}_i} \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) \quad (2.12)$$

where ν is the stepping constant.

2.2.2 Full Bayes

Rather than use maximum likelihood to determine m and s one can treat them as hyper-parameters with a non-informative, conjugate prior, namely $\boldsymbol{\rho}_{m,s^2}(\xi, \lambda) \propto 1/\lambda$. Then the posterior density for the integral given the data using Bayes theorem is

$$\begin{aligned}
& \rho_\mu(z|\mathbf{f} = \mathbf{y}) \\
& \propto \int_0^\infty \int_{-\infty}^\infty \rho_\mu(z|\mathbf{f} = \mathbf{y}, m = \xi, s^2 = \lambda) \rho_f(\mathbf{y}|\xi, \lambda) \rho_{m,s^2}(\xi, \lambda) d\xi d\lambda \\
& \quad \text{by the properties of conditional probability} \\
& \propto \int_0^\infty \int_{-\infty}^\infty \rho_\mu(z|\mathbf{f} = \mathbf{y}, m = \xi, s^2 = \lambda) \rho_f(\mathbf{y}|\xi, \lambda) \rho_{m,s^2}(\xi, \lambda) d\xi d\lambda \\
& \quad \text{by Bayes' Theorem} \\
& \propto \int_0^\infty \frac{1}{\lambda^{(n+3)/2}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\lambda} \left\{ \frac{[z - \xi(1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1}) - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{y}]^2}{c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}} \right. \right. \\
& \quad \left. \left. + (\mathbf{y} - \xi \mathbf{1})^T \mathbf{C}^{-1} (\mathbf{y} - \xi \mathbf{1}) \right\} \right) d\xi d\lambda \\
& \quad \text{by (2.2), (2.3) and } \rho_{m,s^2}(\xi, \lambda) \propto 1/\lambda \\
& \propto \int_0^\infty \frac{1}{\lambda^{(n+3)/2}} \int_{-\infty}^\infty \exp\left(-\frac{\alpha \xi^2 - 2\beta \xi + \gamma}{2\lambda(c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})}\right) d\xi d\lambda,
\end{aligned}$$

where

$$\begin{aligned}
\alpha &= (1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1})^2 + \mathbf{1}^T \mathbf{C}^{-1} \mathbf{1} (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}), \\
\beta &= (1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1})(z - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{y}) + \mathbf{1}^T \mathbf{C}^{-1} \mathbf{y} (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}), \\
\gamma &= (z - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{y})^2 + \mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}).
\end{aligned}$$

In the derivation above and below, factors that are independent of ξ , λ , or z can be discarded since we only need to preserve the proportion. But, factors that depend on ξ , λ , or z must be kept. Completing the square $\alpha \xi^2 - 2\beta \xi + \gamma = \alpha(\xi - \beta/\alpha)^2 -$

$(\beta^2/\alpha) + \gamma$, allows us to evaluate the integrals with respect to ξ and λ :

$$\begin{aligned}
\rho_\mu(z|\mathbf{f} = \mathbf{y}) & \propto \int_0^\infty \frac{1}{\lambda^{(n+3)/2}} \exp\left(-\frac{\gamma - \beta^2/\alpha}{2\lambda(c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})}\right) \cdots \\
& \quad \cdots \int_{-\infty}^\infty \exp\left(-\frac{\alpha(\xi - \beta/\alpha)^2}{2\lambda(c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})}\right) d\xi d\lambda \\
& \propto \int_0^\infty \frac{1}{\lambda^{(n+2)/2}} \exp\left(-\frac{\gamma - \beta^2/\alpha}{2\lambda(c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})}\right) d\lambda \\
& \propto \left(\gamma - \frac{\beta^2}{\alpha}\right)^{-n/2} \propto (\alpha\gamma - \beta^2)^{-n/2}.
\end{aligned}$$

Finally, we simplify the key term:

$$\begin{aligned}
\alpha\gamma - \beta^2 &= \mathbf{1}^T \mathbf{C}^{-1} \mathbf{1} (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}) (z - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{y})^2 \\
& \quad - 2\mathbf{1}^T \mathbf{C}^{-1} \mathbf{y} (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}) (1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1}) (z - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{y}) \\
& \quad + (1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1})^2 \mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}) \\
& \quad + [\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1} \mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} - (\mathbf{1}^T \mathbf{C}^{-1} \mathbf{y})^2] (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})^2 \\
& \propto \mathbf{1}^T \mathbf{C}^{-1} \mathbf{1} \left(z - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{y} - \frac{(1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1}) \mathbf{1}^T \mathbf{C}^{-1} \mathbf{y}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \right)^2 \\
& \quad - \frac{[(1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1}) \mathbf{1}^T \mathbf{C}^{-1} \mathbf{y}]^2}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} + (1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1})^2 \mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} \\
& \quad (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}) [\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1} \mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} - (\mathbf{1}^T \mathbf{C}^{-1} \mathbf{y})^2] \\
& \propto \left(z - \left[\frac{(1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1}) \mathbf{1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} + \mathbf{c} \right]^T \mathbf{C}^{-1} \mathbf{y} \right)^2 \\
& \quad + \left[\frac{(1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1})^2}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} + (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}) \right] \times \mathbf{y}^T \left[\mathbf{C}^{-1} - \frac{\mathbf{C}^{-1} \mathbf{1} \mathbf{1}^T \mathbf{C}^{-1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \right] \mathbf{y} \\
& \propto (z - \hat{\mu}_{\text{full}})^2 + (n-1) \sigma_{\text{full}}^2 \\
& \propto \left(1 + \frac{1}{n-1} \frac{(z - \mu_{\text{full}})^2}{\hat{\sigma}_{\text{full}}^2} \right)
\end{aligned}$$

i.e.,

$$\alpha\gamma - \beta^2 \propto \left(1 + \frac{(z - \hat{\mu}_{\text{full}})^2}{(n-1) \hat{\sigma}_{\text{full}}^2} \right)$$

where $\hat{\mu}_{\text{full}} = \hat{\mu}_{\text{MLE}}$ and

$$\hat{\sigma}_{\text{full}}^2 := \frac{1}{n-1} \mathbf{y}^T \left[\mathbf{C}^{-1} - \frac{\mathbf{C}^{-1} \mathbf{1} \mathbf{1}^T \mathbf{C}^{-1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \right] \mathbf{y} \times \left[\frac{(1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1})^2}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} + (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}) \right].$$

This means that $\mu | (\mathbf{f} = \mathbf{y})$, properly centered and scaled, has a Student's t -distribution with $n - 1$ degrees of freedom. The estimated integral is the same as in the empirical Bayes case, $\hat{\mu}_{\text{full}} = \hat{\mu}_{\text{MLE}}$, but the confidence interval is wider:

$$\mathbb{P}_f [|\mu - \hat{\mu}_{\text{MLE}}| \leq \text{err}_{\text{full}}] = 99\%, \quad (2.13)$$

where

$$\text{err}_{\text{full}} := t_{n-1, 0.995} \hat{\sigma}_{\text{full}} > \text{err}_{\text{MLE}}. \quad (2.14)$$

Here $t_{n-1, 0.995}$ denotes the 99.5 percentile of a standard Student's t -distribution with $n - 1$ degrees of freedom. The stopping criterion for the full Bayes case is more conservative than that in the empirical Bayes case, (2.11).

Because the shape parameter, $\boldsymbol{\theta}$, enters the definition of the covariance kernel in a non-trivial way, the only way to treat it as a hyperparameter and assign a tractable prior would be for the prior to be discrete. We believe in practice that choosing such a prior involves more guesswork than using the empirical Bayes estimate of $\boldsymbol{\theta}$ in (2.8) or the cross-validation approach described next.

2.2.3 Full Bayes with generic prior

Rather than use non-informative, conjugate prior one can use generic prior, namely $\rho_{m, s^2}(\xi, \lambda) \propto g(\lambda)$, than can generalize to any generic function. Then the posterior density for the integral given the data using Bayes theorem is

$$\begin{aligned} & \rho_{\mu}(z | \mathbf{f} = \mathbf{y}) \\ & \propto \int_0^\infty \int_{-\infty}^\infty \rho_{\mu}(z | \mathbf{f} = \mathbf{y}, m = \xi, s^2 = \lambda) \rho_{\mathbf{f}}(\mathbf{y} | \xi, \lambda) \rho_{m, s^2}(\xi, \lambda) d\xi d\lambda \end{aligned}$$

by the properties of conditional probability

$$\propto \int_0^\infty \int_{-\infty}^\infty \rho_\mu(z|\mathbf{f} = \mathbf{y}, m = \xi, s^2 = \lambda) \rho_f(\mathbf{y}|\xi, \lambda) \rho_{m,s^2}(\xi, \lambda) d\xi d\lambda$$

by Bayes' Theorem

$$\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{(n+1)/2}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\lambda} \left\{ \frac{[z - \xi(1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1}) - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{y}]^2}{c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}} \right. \right. \\ \left. \left. + (\mathbf{y} - \xi \mathbf{1})^T \mathbf{C}^{-1} (\mathbf{y} - \xi \mathbf{1}) \right\} \right) d\xi d\lambda$$

by (2.2), (2.3) and $\rho_{m,s^2}(\xi, \lambda) \propto g(\lambda)$

$$\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{(n+1)/2}} \int_{-\infty}^\infty \exp\left(-\frac{\alpha \xi^2 - 2\beta \xi + \gamma}{2\lambda(c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})}\right) d\xi d\lambda,$$

where

$$\alpha = (1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1})^2 + \mathbf{1}^T \mathbf{C}^{-1} \mathbf{1} (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}),$$

$$\beta = (1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1})(z - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{y}) + \mathbf{1}^T \mathbf{C}^{-1} \mathbf{y} (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}),$$

$$\gamma = (z - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{y})^2 + \mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}).$$

In the derivation above and below, factors that are independent of ξ , λ , or z can be discarded since we only need to preserve the proportion. But, factors that depend on ξ , λ , or z must be kept. Completing the square $\alpha \xi^2 - 2\beta \xi + \gamma = \alpha(\xi - \beta/\alpha)^2 - (\beta^2/\alpha) + \gamma$, allows us to evaluate the integrals with respect to ξ and λ :

$$\begin{aligned} \rho_\mu(z|\mathbf{f} = \mathbf{y}) &\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{(n+1)/2}} \exp\left(-\frac{\gamma - \beta^2/\alpha}{2\lambda(c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})}\right) \cdots \\ &\quad \cdots \int_{-\infty}^\infty \exp\left(-\frac{\alpha(\xi - \beta/\alpha)^2}{2\lambda(c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})}\right) d\xi d\lambda \\ &\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{n/2}} \exp\left(-\frac{\gamma - \beta^2/\alpha}{2\lambda(c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})}\right) d\lambda. \end{aligned}$$

This can be interpreted as Laplace transform of $g(\lambda)$.

$$\rho_\mu(z|\mathbf{f} = \mathbf{y}) \propto \int_0^\infty \frac{g(\lambda)}{\lambda^{n/2}} \exp\left(-\frac{\gamma - \beta^2/\alpha}{2\lambda(c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})}\right) d\lambda$$

$$\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{n/2}} \exp\left(-\frac{1}{\lambda}\eta\right) d\lambda, \quad \text{where} \quad \eta = \frac{\gamma - \beta^2/\alpha}{2(c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})}$$

Let $\lambda = \frac{1}{w}$, $d\lambda = -w^{-2}dw$ then,

$$\begin{aligned} \rho_\mu(z|\mathbf{f} = \mathbf{y}) &\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{n/2}} \exp\left(-\frac{1}{\lambda}\eta\right) d\lambda \\ &= \int_0^\infty \frac{g(1/w)}{w^{-n/2}} \exp(-w\eta) (-w^{-2})dw \\ &= \int_\infty^0 -g(1/w)w^{\frac{n}{2}-2} \exp(-w\eta) dw \\ &= \int_0^\infty g(1/w)w^{\frac{n-4}{2}} \exp(-w\eta) dw \\ &= \mathcal{LT}\{g(1/\eta)\}^{(\frac{n-4}{2})'} \end{aligned}$$

where $\mathcal{LT}(\cdot)$ denotes the Laplace transform and $(\frac{n-4}{2})'$ indicates the derivative taken after the transform. Here we used frequency domain derivative property of the Laplace transform. Thus, $\rho_\mu(z|\mathbf{f} = \mathbf{y})$ is proportional to $(\frac{n-4}{2})$ th derivative of the Laplace transform of $g(1/\eta)$.

If $g(\lambda) = \frac{1}{\lambda}$, then

$$\begin{aligned} \rho_\mu(z|\mathbf{f} = \mathbf{y}) &= \int_0^\infty g(1/w)w^{\frac{n}{2}-2} \exp(-w\eta) dw \\ &= (\mathcal{LT}(g(1/t)))^{(\frac{n}{2}-2)'}|_{t=\eta} \\ &= (1/t^2)^{(\frac{n}{2}-2)'}|_{t=\eta} \\ &\propto \eta^{-n/2} = \left(\frac{\gamma - \beta^2/\alpha}{2(c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})}\right)^{-n/2} \\ &\propto \left(\gamma - \frac{\beta^2}{\alpha}\right)^{-n/2} \\ &\propto (\alpha\gamma - \beta^2)^{-n/2}. \end{aligned}$$

Laplace transform of $g(1/t)$ is $1/t^2$, taking $\frac{n}{2} - 2$ th derivative gives us

Finally, we simplify the key term η :

$$\alpha\gamma - \beta^2 = \mathbf{1}^T \mathbf{C}^{-1} \mathbf{1} (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}) (z - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{y})^2$$

$$\begin{aligned}
& - 2\mathbf{1}^T \mathbf{C}^{-1} \mathbf{y} (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}) (1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1}) (z - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{y}) \\
& + (1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1})^2 \mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}) \\
& + [\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1} \mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} - (\mathbf{1}^T \mathbf{C}^{-1} \mathbf{y})^2] (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c})^2 \\
& \propto \mathbf{1}^T \mathbf{C}^{-1} \mathbf{1} \left(z - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{y} - \frac{(1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1}) \mathbf{1}^T \mathbf{C}^{-1} \mathbf{y}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \right)^2 \\
& \quad - \frac{[(1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1}) \mathbf{1}^T \mathbf{C}^{-1} \mathbf{y}]^2}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} + (1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1})^2 \mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} \\
& \quad (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}) [\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1} \mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} - (\mathbf{1}^T \mathbf{C}^{-1} \mathbf{y})^2] \\
& \propto \left(z - \left[\frac{(1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1}) \mathbf{1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} + \mathbf{c} \right]^T \mathbf{C}^{-1} \mathbf{y} \right)^2 \\
& \quad + \left[\frac{(1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1})^2}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} + (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}) \right] \times \mathbf{y}^T \left[\mathbf{C}^{-1} - \frac{\mathbf{C}^{-1} \mathbf{1} \mathbf{1}^T \mathbf{C}^{-1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \right] \mathbf{y} \\
& \propto (z - \hat{\mu}_{\text{full}})^2 + (n - 1) \sigma_{\text{full}}^2 \\
& \propto \left(1 + \frac{1}{n - 1} \frac{(z - \mu_{\text{full}})^2}{\hat{\sigma}_{\text{full}}^2} \right)
\end{aligned}$$

i.e.,

$$\alpha\gamma - \beta^2 \propto \left(1 + \frac{(z - \hat{\mu}_{\text{full}})^2}{(n - 1) \hat{\sigma}_{\text{full}}^2} \right)$$

where $\hat{\mu}_{\text{full}} = \hat{\mu}_{\text{MLE}}$ and

$$\hat{\sigma}_{\text{full}}^2 := \frac{1}{n - 1} \mathbf{y}^T \left[\mathbf{C}^{-1} - \frac{\mathbf{C}^{-1} \mathbf{1} \mathbf{1}^T \mathbf{C}^{-1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \right] \mathbf{y} \times \left[\frac{(1 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1})^2}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} + (c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}) \right].$$

This means that $\mu |(\mathbf{f} = \mathbf{y})$, properly centered and scaled, has a Student's t -distribution with $n - 1$ degrees of freedom. The estimated integral is the same as in the empirical Bayes case, $\hat{\mu}_{\text{full}} = \hat{\mu}_{\text{MLE}}$, but the confidence interval is wider:

$$\mathbb{P}_f [|\mu - \hat{\mu}_{\text{MLE}}| \leq \text{err}_{\text{full}}] = 99\%, \quad (2.15)$$

where

$$\text{err}_{\text{full}} := t_{n_j - 1, 0.995} \hat{\sigma}_{\text{full}} > \text{err}_{\text{MLE}}. \quad (2.16)$$

2.2.4 Generalized Cross-Validation

A third parameter optimization technique is *leave-one-out cross-validation* (CV). Let $\tilde{y}_i = \mathbb{E}[f(\mathbf{x}_i) | \mathbf{f}_{-i} = \mathbf{y}_{-i}]$, where the subscript $-i$ denotes the vector excluding the i^{th} component. This is the conditional expectation of $f(\mathbf{x}_i)$ given all data but the function value at \mathbf{x}_i . The cross-validation criterion, which is to be minimized, is sum of squares of the difference between these conditional expectations and the observed values:

$$\text{CV} = \sum_{i=1}^n (y_i - \tilde{y}_i)^2. \quad (2.17)$$

Let $\mathbf{A} = \mathbf{C}^{-1}$, let $\boldsymbol{\zeta} = \mathbf{A}(\mathbf{y} - m\mathbf{1})$, and partition \mathbf{C} , \mathbf{A} , and $\boldsymbol{\zeta}$ as

$$\mathbf{C} = \begin{pmatrix} c_{ii} & \mathbf{C}_{-i,i}^T \\ \mathbf{C}_{-i,i} & \mathbf{C}_{-i,-i} \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} a_{ii} & \mathbf{A}_{-i,i}^T \\ \mathbf{A}_{-i,i} & \mathbf{A}_{-i,-i} \end{pmatrix}, \quad \boldsymbol{\zeta} = \begin{pmatrix} \zeta_i \\ \boldsymbol{\zeta}_{-i} \end{pmatrix},$$

where the subscript i denotes the i^{th} row or column, and the subscript $-i$ denotes all rows or columns except the i^{th} . Following this notation, Lemma 2.1.1 implies that

$$\begin{aligned} \tilde{y}_i &= m + \mathbf{C}_{-i,i}^T \mathbf{C}_{-i,-i}^{-1} (\mathbf{y}_{-i} - m\mathbf{1}) \\ \zeta_i &= a_{ii}(y_i - m) + \mathbf{A}_{-i,i}^T (\mathbf{y}_{-i} - m\mathbf{1}) \\ &= a_{ii}[(y_i - m) - \mathbf{C}_{-i,i}^T \mathbf{C}_{-i,-i}^{-1} (\mathbf{y}_{-i} - m\mathbf{1})] \\ &= a_{ii}(y_i - \tilde{y}_i). \end{aligned}$$

Thus, (2.17) may be re-written as

$$\text{CV} = \sum_{i=1}^n \left(\frac{\zeta_i}{a_{ii}} \right)^2, \quad \text{where} \quad \boldsymbol{\zeta} = \mathbf{C}^{-1}(\mathbf{y} - m\mathbf{1}).$$

The *generalized cross-validation* criterion (GCV) replaces the i^{th} diagonal element of \mathbf{A} in the denominator by the average diagonal element of \mathbf{A} [5, 10, 29]:

$$\text{GCV} = \frac{\sum_{i=1}^n \zeta_i^2}{\left(\frac{1}{n} \sum_{i=1}^n a_{ii} \right)^2}$$

$$= \frac{(\mathbf{y} - m\mathbf{1})^T \mathbf{C}^{-2} (\mathbf{y} - m\mathbf{1})}{\left(\frac{1}{n} \text{trace}(\mathbf{C}^{-1})\right)^2}.$$

The loss function GCV depends on m and $\boldsymbol{\theta}$, but not on s . Minimizing the GCV yields

$$m_{\text{GCV}} = \frac{\mathbf{1}^T \mathbf{C}^{-2} \mathbf{y}}{\mathbf{1}^T \mathbf{C}^{-2} \mathbf{1}},$$

$$\boldsymbol{\theta}_{\text{GCV}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \left\{ \log \left(\mathbf{y}^T \left[\mathbf{C}^{-2} - \frac{\mathbf{C}^{-2} \mathbf{1} \mathbf{1}^T \mathbf{C}^{-2}}{\mathbf{1}^T \mathbf{C}^{-2} \mathbf{1}} \right] \mathbf{y} \right) - 2 \log (\text{trace}(\mathbf{C}^{-1})) \right\}. \quad (2.18)$$

Plugging this value of m into (2.4) yields

$$\hat{\mu}_{\text{GCV}} = \left(\frac{(1 - \mathbf{1}^T \mathbf{C}^{-1} \mathbf{c}) \mathbf{C}^{-1} \mathbf{1}}{\mathbf{1}^T \mathbf{C}^{-2} \mathbf{1}} + \mathbf{c} \right)^T \mathbf{C}^{-1} \mathbf{y}. \quad (2.19)$$

An estimate for s may be obtained by noting that by Lemma 2.1.1,

$$\text{var}[f(\mathbf{x}_i) | \mathbf{f}_{-i} = \mathbf{y}_{-i}] = s^2 a_{ii}^{-1}.$$

Thus, we may estimate ‘ s ’ using an argument similar to that used in deriving the GCV and then substituting m_{GCV} for m :

$$\begin{aligned} s^2 &= \text{var}[f(\mathbf{x}_i) | \mathbf{f}_{-i} = \mathbf{y}_{-i}] a_{ii} \\ &\approx \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2 a_{ii} \\ &= \frac{1}{n} \sum_{i=1}^n \frac{\zeta_i^2}{a_{ii}} \\ &\approx \frac{\frac{1}{n} \sum_{i=1}^n \zeta_i^2}{\frac{1}{n} \sum_{i=1}^n a_{ii}} \\ &= \frac{(\mathbf{y} - m\mathbf{1})^T \mathbf{C}^{-2} (\mathbf{y} - m\mathbf{1})}{\text{trace}(\mathbf{C}^{-1})} = s_{\text{GCV}}^2, \end{aligned}$$

where

$$s_{\text{GCV}}^2 := \mathbf{y}^T \left[\mathbf{C}^{-2} - \frac{\mathbf{C}^{-2} \mathbf{1} \mathbf{1}^T \mathbf{C}^{-2}}{\mathbf{1}^T \mathbf{C}^{-2} \mathbf{1}} \right] \mathbf{y} [\text{trace}(\mathbf{C}^{-1})]^{-1}.$$

The confidence interval based on generalized cross-validation corresponds to (2.5) with the GCV estimates for m , s , and $\boldsymbol{\theta}$:

$$\text{err}_{\text{GCV}} = 2.58s_{\text{GCV}}\sqrt{c_0 - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}}, \quad (2.20)$$

$$\mathbb{P}_f [|\mu - \hat{\mu}_{\text{GCV}}| \leq \text{err}_{\text{GCV}}] = 99\%. \quad (2.21)$$

Looking back over the results of Sections 2.2.1–2.2.4, it is noted that if the original covariance function, C , is replaced by bC for some positive constant b , the cubature, $\hat{\mu}$, the estimates of $\boldsymbol{\theta}$, and the credible interval widths, err , all remain unchanged. The estimates of s^2 are multiplied by b^{-1} , as would be expected.

2.3 The automatic Bayesian cubature algorithm

The previous section presents three credible intervals, (2.11), (2.15), and (2.21), for the μ , the desired integral. Each credible interval is based on different assumptions about the hyperparameters m , s , and $\boldsymbol{\theta}$. We stress that one must estimate these hyperparameters or assume a prior distribution on them because the credible intervals are used as stopping criteria for our cubature rule. Since a credible intervals makes a statement about a typical function—not an outlier—one must try to ensure that the integrand is a typical draw from the assumed Gaussian process.

Our Bayesian cubature algorithm increases the sample size until the width of the credible interval is small enough. This is accomplished through successively doubling the sample size. The steps are detailed in Algorithm 1.

Algorithm 1 Automatic Bayesian Cubature

Require: a generator for the sequence $\mathbf{x}_1, \mathbf{x}_2, \dots$; a black-box function, f ; an absolute error tolerance, $\varepsilon > 0$; the positive initial sample size, n_0 ; the maximum sample size n_{\max}

```

1:  $n \leftarrow n_0, n' \leftarrow 0, \text{err} \leftarrow \infty$ 
2: while  $\text{err} > \varepsilon$  and  $n \leq n_{\max}$  do
3:   Generate  $\{\mathbf{x}_i\}_{i=n'+1}^n$  and sample  $\{f(\mathbf{x}_i)\}_{i=n'+1}^n$ 
4:   Compute  $\boldsymbol{\theta}$  by (2.8) or (2.18)
5:   Compute err according to (2.10), (2.16), or (2.20)
6:    $n' \leftarrow n, n \leftarrow 2n'$ 
7: end while
8: Sample size to compute  $\hat{\mu}$ ,  $n \leftarrow n'$ 
9: Compute  $\hat{\mu}$ , the approximate integral, according to (2.9) or (2.19)
10: return  $\hat{\mu}, n$  and err

```

2.4 Example with the Matérn kernel

To demonstrate automatic Bayesian cubature consider a Matérn covariance kernel:

$$C_\theta(\mathbf{x}, \mathbf{t}) = \prod_{k=1}^d \exp(-\theta |\mathbf{x}_k - \mathbf{t}_k|) (1 + \theta |\mathbf{x}_k - \mathbf{t}_k|).$$

Also, consider the integration problem of evaluating *multivariate normal probabilities*:

$$\mu = \int_{(\mathbf{a}, \mathbf{b})} \frac{\exp(-\frac{1}{2} \mathbf{t}^T \boldsymbol{\Sigma}^{-1} \mathbf{t})}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} d\mathbf{t}, \quad (2.22)$$

where (\mathbf{a}, \mathbf{b}) is a finite, semi-infinite or infinite box in \mathbb{R}^d . This integral does not have an analytic expression for general $\boldsymbol{\Sigma}$, so cubatures are required.

Genz [9] introduced a variable transformation to transform (2.22) into an integral on the unit cube. Let $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$ be the Cholesky decomposition where

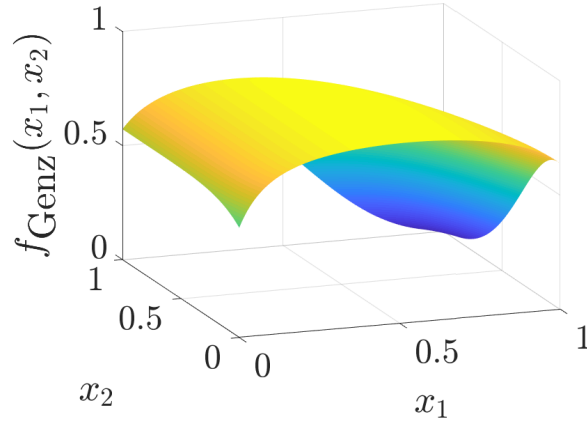


Figure 2.1: The $d = 3$ multivariate normal probability transformed to an integral of f_{Genz} with $d = 2$.

$\mathbf{L} = (l_{jk})_{j,k=1}^d$ is a lower triangular matrix. Iteratively define

$$\begin{aligned} \alpha_1 &= \Phi(a_1), & \beta_1 &= \Phi(b_1), \\ \alpha_j(x_1, \dots, x_{j-1}) &= \Phi \left(\frac{1}{l_{jj}} \left(a_j - \sum_{k=1}^{j-1} l_{jk} \Phi^{-1}(\alpha_k + x_k(\beta_k - \alpha_k)) \right) \right), & j &= 2, \dots, d, \\ \beta_j(x_1, \dots, x_{j-1}) &= \Phi \left(\frac{1}{l_{jj}} \left(b_j - \sum_{k=1}^{j-1} l_{jk} \Phi^{-1}(\alpha_k + x_k(\beta_k - \alpha_k)) \right) \right), & j &= 2, \dots, d, \\ f_{\text{Genz}}(\mathbf{x}) &= \prod_{j=1}^d [\beta_j(\mathbf{x}) - \alpha_j(\mathbf{x})]. \end{aligned} \quad (2.23)$$

where Φ is the cumulative standard normal distribution function. Then, $\mu = \int_{[0,1]^{d-1}} f_{\text{Genz}}(\mathbf{x}) \, d\mathbf{x}$.

We use the following parameter values in the simulation:

$$d = 3, \quad \mathbf{a} = \begin{pmatrix} -6 \\ -2 \\ -2 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 5 \\ 2 \\ 1 \end{pmatrix}, \quad \mathbf{L} = \begin{pmatrix} 4 & 1 & 1 \\ 0 & 1 & 0.5 \\ 0 & 0 & 0.25 \end{pmatrix}.$$

On our test computer, it took more than an hour to compute $\hat{\mu}_n$ with $n = 2^{14}$. As shown in Figure 2.2, computation time increases rapidly with n . Especially, max-

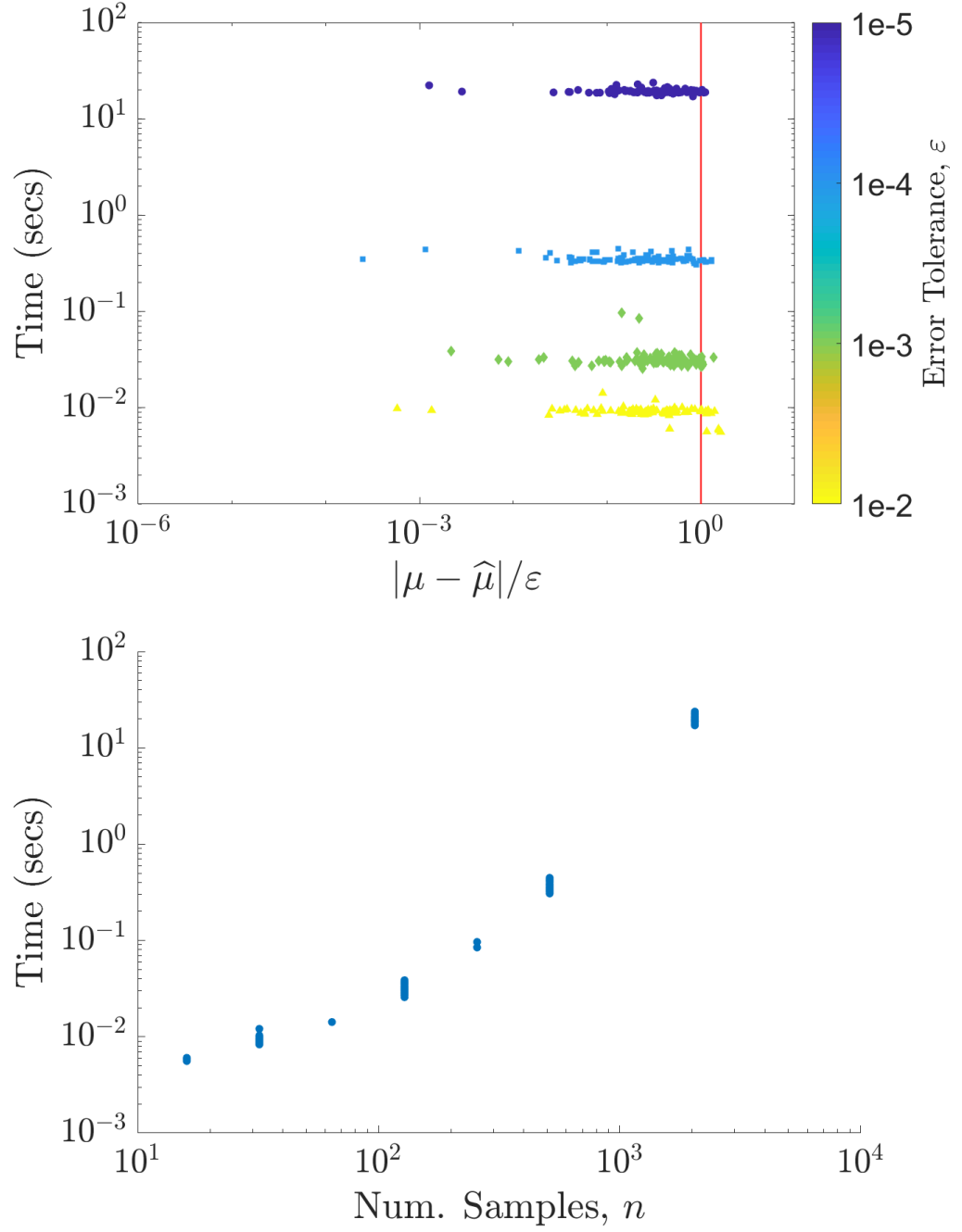


Figure 2.2: Multivariate Normal probability estimated using Matern kernel in $d=2$ using empirical stopping criterion. Top: Guaranteed integration within error tolerance ε . Bottom: Computation time rapidly increases with increase of n

imum likelihood estimation of θ , which needs the objective function, is the most time consuming of all. Because the objective function needs to be computed multiple times in every iteration to find its minimum. Not only the computational cost increases, also the C becomes highly ill-conditioned with increasing n . So, our algorithm in the current form is not straightaway usable for any practical applications.

CHAPTER 3

FAST AUTOMATIC BAYESIAN CUBATURE

The generic automatic Bayesian cubature algorithm described in the previous section requires $\mathcal{O}(n^3)$ operations to estimate $\boldsymbol{\theta}$, compute the credible interval width, and compute the cubature. This section explains how to speed up the calculations. A key is to choose kernels that match the design, $\{\mathbf{x}_i\}_{i=1}^n$, so that the vector-matrix operations required by Bayesian cubature can be accomplished using fast transforms at a cost of $\mathcal{O}(n \log(n))$.

3.0.1 Fast Transform Kernel

We make some assumptions about the relationship between the covariance kernel and the design, which will be shown to hold in Section 4 for rank-1 lattices and shift-invariant kernels. First we introduce the notation

$$\begin{aligned} \mathbf{C} &= \left(C_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j) \right)_{i,j=1}^n = (\mathbf{C}_1, \dots, \mathbf{C}_n) \\ &= \frac{1}{n} \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^H, \quad \mathbf{V}^H = n \mathbf{V}^{-1}, \\ \mathbf{V} &= (\mathbf{v}_1, \dots, \mathbf{v}_n)^T = (\mathbf{V}_1, \dots, \mathbf{V}_n) \\ \mathbf{C}^p &= \frac{1}{n} \mathbf{V} \boldsymbol{\Lambda}^p \mathbf{V}^H, \quad \forall p \in \mathbb{Z}, \end{aligned} \tag{3.1}$$

where \mathbf{V}^H is the Hermitian of \mathbf{V} . The columns of matrix \mathbf{V} are eigenvectors of \mathbf{C} , and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues of \mathbf{C} . For any $n \times n$ vector \mathbf{b} , define the notation $\tilde{\mathbf{b}} := \mathbf{V}^H \mathbf{b}$.

We make three assumptions that allow the fast computation:

$$\mathbf{V} \text{ may be identified analytically,} \tag{3.2a}$$

$$\mathbf{v}_1 = \mathbf{V}_1 = \mathbf{1}, \tag{3.2b}$$

$$\mathbf{V}^H \mathbf{b} \text{ requires only } \mathcal{O}(n \log(n)) \text{ operations } \forall \mathbf{b}. \quad (3.2c)$$

We call the transformation $\mathbf{b} \mapsto \mathbf{V}^H \mathbf{b}$ a *fast transform* and C_θ a *fast transform kernel*.

Under assumptions (3.2) the eigenvalues may be identified as the fast transform of the first column of \mathbf{C} :

$$\begin{aligned} \boldsymbol{\lambda} &= \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} = \boldsymbol{\Lambda} \mathbf{1} = \boldsymbol{\Lambda} \mathbf{v}_1^* = \underbrace{\left(\frac{1}{n} \mathbf{V}^H \mathbf{V} \right)}_{\mathbf{I}} \boldsymbol{\Lambda} \mathbf{v}_1^* \\ &= \mathbf{V}^H \left(\frac{1}{n} \mathbf{V} \boldsymbol{\Lambda} \mathbf{v}_1^* \right) = \mathbf{V}^H \mathbf{C}_1 = \tilde{\mathbf{C}}_1. \end{aligned} \quad (3.3)$$

Also note that the fast transform of $\mathbf{1}$ has a simple form

$$\tilde{\mathbf{1}} = \mathbf{V}^H \mathbf{1} = \mathbf{V}^H \mathbf{V}_1 = \begin{pmatrix} n \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Many of the terms that arise in the calculations in Algorithm 1 take the form $\mathbf{a}^T \mathbf{C}^p \mathbf{b}$ for real \mathbf{a} and \mathbf{b} and integer p . These can be calculated via the transforms $\tilde{\mathbf{a}} = \mathbf{V}^H \mathbf{a}$ and $\tilde{\mathbf{b}} = \mathbf{V}^H \mathbf{b}$ as

$$\mathbf{a}^T \mathbf{C}^p \mathbf{b} = \frac{1}{n} \mathbf{a}^T \mathbf{V} \boldsymbol{\Lambda}^p \mathbf{V}^H \mathbf{b} = \frac{1}{n} \tilde{\mathbf{a}}^H \boldsymbol{\Lambda}^p \tilde{\mathbf{b}} = \frac{1}{n} \sum_{i=1}^n \lambda_i^p \tilde{a}_i^* \tilde{b}_i,$$

In particular,

$$\begin{aligned} \mathbf{1}^T \mathbf{C}^{-p} \mathbf{1} &= \frac{n}{\lambda_1^p}, & \mathbf{1}^T \mathbf{C}^{-p} \mathbf{y} &= \frac{\tilde{y}_1}{\lambda_1^p}, \\ \mathbf{y}^T \mathbf{C}^{-p} \mathbf{y} &= \frac{1}{n} \sum_{i=1}^n \frac{|\tilde{y}_i|^2}{\lambda_i^p}, & \mathbf{c}^T \mathbf{C}^{-1} \mathbf{1} &= \frac{\tilde{c}_1}{\lambda_1}, \end{aligned}$$

$$\mathbf{c}^T \mathbf{C}^{-1} \mathbf{y} = \frac{1}{n} \sum_{i=1}^n \frac{\tilde{\mathbf{c}}_i^* \tilde{\mathbf{y}}_i}{\lambda_i}, \quad \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c} = \frac{1}{n} \sum_{i=1}^n \frac{|\tilde{\mathbf{c}}_i|^2}{\lambda_i},$$

where $\tilde{\mathbf{y}} = \mathbf{V}^H \mathbf{y}$ and $\tilde{\mathbf{c}} = \mathbf{V}^H \mathbf{c}$. For any real \mathbf{b} , with $\tilde{\mathbf{b}} = \mathbf{V}^H \mathbf{b}$, it follows that \tilde{b}_1 is real since the first row of \mathbf{V}^H is $\mathbf{1}$.

The covariance kernel used in practice also may satisfy an additional assumption:

$$\int_{[0,1]^d} C(\mathbf{t}, \mathbf{x}) d\mathbf{t} = 1 \quad \forall \mathbf{x} \in [0,1]^d, \quad (3.4)$$

which implies that $c_0 = 1$ and $\mathbf{c} = \mathbf{1}$. Under (3.4), the expressions above may be further simplified:

$$\mathbf{c}^T \mathbf{C}^{-1} \mathbf{1} = \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c} = \frac{n}{\lambda_1}.$$

3.0.2 Empirical Bayes

Under assumptions (3.2), the empirical Bayes parameters in (2.6), (2.7), (2.8) (2.9), and (2.10) can be expressed in terms of the fast transforms of the function data, the first column of the Gram matrix, and \mathbf{c} as follows:

$$\begin{aligned} m_{\text{MLE}} &= \frac{\tilde{y}_1}{n} = \frac{1}{n} \sum_{i=1}^n y_i, \\ s_{\text{MLE}}^2 &= \frac{1}{n^2} \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i}, \\ \boldsymbol{\theta}_{\text{MLE}} &= \underset{\boldsymbol{\theta}}{\text{argmin}} \left[\log \left(\sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i} \right) + \frac{1}{n} \sum_{i=1}^n \log(\lambda_i) \right], \\ \hat{\mu}_{\text{MLE}} &= \frac{\tilde{y}_1}{n} + \frac{1}{n} \sum_{i=2}^n \frac{\tilde{\mathbf{c}}_i^* \tilde{\mathbf{y}}_i}{\lambda_i}, \\ \text{err}_{\text{MLE}} &= \frac{2.58}{n} \sqrt{\sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i} \left(c_0 - \frac{1}{n} \sum_{i=1}^n \frac{|\tilde{\mathbf{c}}_i|^2}{\lambda_i} \right)}, \end{aligned} \quad (3.5)$$

Since all the quantities on the right hand sides can be obtained in $\mathcal{O}(n \log(n))$ operations by fast transforms, the left hand sides are all computable using the asymptotic

computational cost.

Under the further assumption (3.4) it follows that

$$\begin{aligned}\hat{\mu}_{\text{MLE}} &= \frac{\tilde{y}_1}{n} = \frac{1}{n} \sum_{i=1}^n y_i, \\ \text{err}_{\text{MLE}} &= \frac{2.58}{n} \sqrt{\sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i} \left(1 - \frac{n}{\lambda_1}\right)}.\end{aligned}\tag{3.6}$$

Thus, in this case $\hat{\mu}$ is simply the sample mean.

3.0.3 Kernel derivative using fast transform

If \mathbf{V} does not depend on $\boldsymbol{\theta}$ then one can fast compute the derivative of matrix \mathbf{C} .

Starting from the definition (3.1) and taking derivative w.r.t. θ_j ,

$$\frac{\partial \mathbf{C}}{\partial \theta_j} = \frac{1}{n} \mathbf{V} \frac{\partial \boldsymbol{\Lambda}}{\partial \theta_j} \mathbf{V}^H = \frac{1}{n} \mathbf{V} \bar{\boldsymbol{\Lambda}}_{(j)} \mathbf{V}^H$$

where $\bar{\boldsymbol{\Lambda}}_{(j)} = \text{diag}(\bar{\boldsymbol{\lambda}}_{(j)})$ and $\bar{\boldsymbol{\lambda}}_{(j)} = \left(\frac{\partial \lambda_i}{\partial \theta_j}\right)_{i=1}^n$. Also, note that $\bar{\boldsymbol{\lambda}}_{(j)} = \mathbf{V}^H \bar{\mathbf{C}}_{1(j)}$ computed using the fast transform, where $\bar{\mathbf{C}}_{1(j)}$ denotes the first row of the gram matrix after taking derivative, i.e.

$$\bar{\mathbf{C}}_{1(j)} = \left(\frac{\partial}{\partial \theta_i} C_{\boldsymbol{\theta}}(\mathbf{x}_1, \mathbf{x}_i)\right)_{i=1}^n.$$

First, let's define the objective function from (3.5),

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) &= \frac{1}{n} \log(\det \mathbf{C}) + \log((\mathbf{y} - m_{\text{MLE}} \mathbf{1})^T \mathbf{C}^{-1} (\mathbf{y} - m_{\text{MLE}} \mathbf{1})) \\ &= \mathcal{L}_{|\mathbf{C}|} + \mathcal{L}_{\mathbf{y}}\end{aligned}$$

Now, let's take the derivative

$$\frac{\partial}{\partial \theta_i} \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\partial}{\partial \theta_i} \mathcal{L}_{|\mathbf{C}|} + \frac{\partial}{\partial \theta_i} \mathcal{L}_{\mathbf{y}}$$

Let's simplify the individual terms

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} \mathcal{L}_{|\mathbf{C}|} &= \frac{\partial}{\partial \theta_i} \frac{1}{n} \log(\det \mathbf{C}) \\
&= \text{trace} \left(\mathbf{C}^{-1} \frac{\partial \mathbf{C}}{\partial \theta_i} \right) = \text{trace} \left(\frac{1}{n} \mathbf{V} \mathbf{\Lambda}^{-1} \mathbf{V}^H \frac{1}{n} \mathbf{V} \bar{\mathbf{\Lambda}} \mathbf{V}^H \right) \\
&= \text{trace} \left(\frac{1}{n} \mathbf{V} \mathbf{\Lambda}^{-1} \bar{\mathbf{\Lambda}} \mathbf{V}^H \right), \quad \text{where we used } \mathbf{V}^H \mathbf{V} = \mathbf{I} \\
&= \text{trace} \left(\frac{1}{n} \mathbf{V} \text{diag} \left(\frac{\bar{\lambda}_j}{\lambda_j} \right)_{j=1}^n \mathbf{V}^H \right) = \sum_{j=1}^n \frac{\bar{\lambda}_j}{\lambda_j}
\end{aligned}$$

Part of the $\mathcal{L}_{\mathbf{y}}$ was already simplified using the fast transform,

$$(\mathbf{y} - m_{\text{MLE}} \mathbf{1})^T \mathbf{C}^{-1} (\mathbf{y} - m_{\text{MLE}} \mathbf{1}) = \frac{1}{n} \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i}$$

Using the above result,

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} \mathcal{L}_{\mathbf{y}} &= \frac{\partial}{\partial \theta_j} \log \left(\frac{1}{n} \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i} \right) \\
&= \left(\frac{1}{n} \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i} \right)^{-1} \frac{\partial}{\partial \theta_j} \left(\frac{1}{n} \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i} \right) \\
&= \left(\frac{1}{n} \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i} \right)^{-1} \frac{1}{n} \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i^2} \left(-\frac{\partial \lambda_i}{\partial \theta_j} \right) \\
&= - \left(\sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i} \right)^{-1} \left(\sum_{i=2}^n |\tilde{y}_i|^2 \frac{\bar{\lambda}_{i(j)}}{\lambda_i^2} \right)
\end{aligned}$$

Finally, using the results

$$\frac{\partial}{\partial \theta_j} \mathcal{L}(\boldsymbol{\theta} | \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n \frac{\bar{\lambda}_{i(j)}}{\lambda_i} - \left(\sum_{i=2}^n \frac{|\tilde{\mathbf{y}}_i|^2 \bar{\lambda}_{i(j)}}{\lambda_i^2} \right) \left(\sum_{i=2}^n \frac{|\tilde{\mathbf{y}}_i|^2}{\lambda_i} \right)^{-1}$$

3.0.4 Full Bayes

For the full Bayes approach the cubature is the same as for empirical Bayes. We also defer to empirical Bayes to estimate the parameter $\boldsymbol{\theta}$. The width of the confidence interval is $\text{err}_{\text{full}} := t_{n_j-1, 0.995} \hat{\sigma}_{\text{full}}$, where $\hat{\sigma}_{\text{full}}^2$ can also be computed swiftly under

assumptions (3.2):

$$\hat{\sigma}_{\text{full}}^2 = \frac{1}{n(n-1)} \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i} \left[\frac{\lambda_1}{n} \left(1 - \frac{\tilde{c}_1}{\lambda_1} \right)^2 + \left(c_0 - \frac{1}{n} \sum_{i=1}^n \frac{|\tilde{c}_i|^2}{\lambda_i} \right) \right],$$

Under assumption (3.4) further simplification can be made:

$$\hat{\sigma}_{\text{full}}^2 = \frac{1}{n(n-1)} \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i} \left(\frac{\lambda_1}{n} - 1 \right),$$

It follows that

$$\text{err}_{\text{full}} = t_{n_j-1, 0.995} \sqrt{\frac{1}{n(n-1)} \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i} \left(\frac{\lambda_1}{n} - 1 \right)}. \quad (3.7)$$

3.0.5 Generalized Cross-Validation

GCV yields a different cubature, which nevertheless can also be computed quickly using the fast transform. Under assumptions (3.2):

$$\begin{aligned} m_{\text{GCV}} &= m_{\text{MLE}} = \frac{\tilde{y}_1}{n} = \frac{1}{n} \sum_{i=1}^n y_i, \\ s_{\text{GCV}}^2 &:= \frac{1}{n} \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i^2} \left[\sum_{i=1}^n \frac{1}{\lambda_i} \right]^{-1}, \\ \boldsymbol{\theta}_{\text{GCV}} &= \underset{\boldsymbol{\theta}}{\text{argmin}} \left[\log \left(\sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i^2} \right) - 2 \log \left(\sum_{i=1}^n \frac{1}{\lambda_i} \right) \right], \\ \hat{\mu}_{\text{GCV}} &= \hat{\mu}_{\text{MLE}} = \frac{\tilde{y}_1}{n} + \frac{1}{n} \sum_{i=2}^n \frac{\tilde{c}_i^* \tilde{y}_i}{\lambda_i}, \\ \text{err}_{\text{GCV}} &= \frac{2.58}{n} \left\{ \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i^2} \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{\lambda_i} \right]^{-1} \times \left(c_0 - \frac{1}{n} \sum_{i=1}^n \frac{|\tilde{c}_i|^2}{\lambda_i} \right) \right\}^{1/2}. \end{aligned} \quad (3.8)$$

Moreover, under further assumption (3.4) it follows that

$$\begin{aligned} \hat{\mu}_{\text{GCV}} &= \hat{\mu}_{\text{MLE}} = \hat{\mu}_{\text{full}} = \frac{\tilde{y}_1}{n} = \frac{1}{n} \sum_{i=1}^n y_i, \\ \text{err}_{\text{GCV}} &= \frac{2.58}{n} \left\{ \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i^2} \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{\lambda_i} \right]^{-1} \times \left(1 - \frac{n}{\lambda_1} \right) \right\}^{1/2}. \end{aligned}$$

In this case too, $\hat{\mu}$ is simply the sample mean.

CHAPTER 4

INTEGRATION LATTICES AND SHIFT INVARIANT KERNELS

The preceding sections lay out an automatic Bayesian cubature algorithm whose computational cost is only $\mathcal{O}(n \log(n))$ if n function values are used. However, this algorithm relies on covariance kernel functions, C and designs, $\{\mathbf{x}_i\}_{i=1}^n$ that satisfy assumptions (3.2). We will also satisfy assumption (3.4). To facilitate the fast transform, it is assumed in this section and the next that n is power of 2.

4.1 Extensible Integration Lattice Node Sets

The design or set of nodes used is defined by a shifted extensible integration lattice node sequence, which takes the form

$$\mathbf{x}_i = \mathbf{h}\phi(i-1) + \mathbf{\Delta} \pmod{\mathbf{1}}, \quad i \in \mathbb{N}.$$

Here, \mathbf{h} is a d -dimensional generating vector of positive integers, $\mathbf{\Delta}$ is some point in $[0, 1)^d$, often chosen at random, and $\{\phi(i)\}_{i=0}^n$ is the van der Corput sequence, defined by reflecting the binary digits of the integer about the decimal point, i.e.,

i	0	1	2	3	4	5	6	7	\dots
i	0_2	1_2	10_2	11_2	100_2	101_2	110_2	111_2	\dots
$\phi(i)$	2.0	2.1	2.01	2.11	2.001	2.101	2.011	2.111	\dots
$\phi(i)$	0	0.5	0.25	0.75	0.125	0.625	0.375	0.875	\dots

(4.1)

An example of 64 nodes is given in Figure 4.1. The even coverage of the unit cube is ensured by a well chosen generating vector. The choice of generating vector is typically done offline by computer search. See [8] and [15] for more on extensible integration lattices.

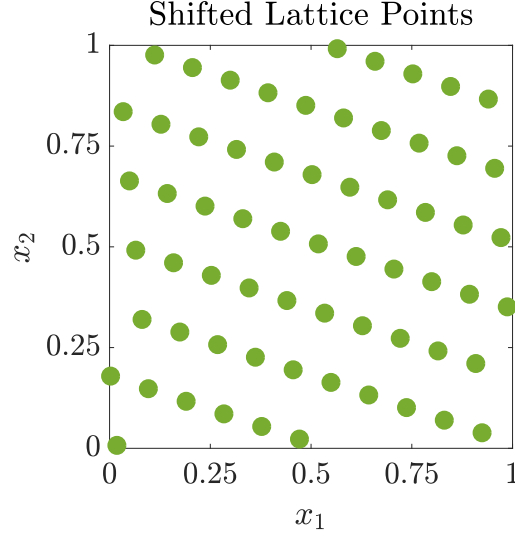


Figure 4.1: Example of a shifted integration lattice node set in $d = 2$

4.2 Shift Invariant Kernels

The covariance functions C that match integration lattice node sets have the form

$$C(\mathbf{t}, \mathbf{x}) = K(\mathbf{t} - \mathbf{x} \bmod \mathbf{1}). \quad (4.2)$$

This is called a *shift invariant kernel* because shifting both arguments of the covariance function by the same amount leaves the value unchanged. By a proper scaling of the kernel K it follows that assumption (3.4) is satisfied. Of course, K must also be of the form that ensures that C is symmetric and positive definite, as assumed in (2.1).

A family of shift invariant kernels is constructed via even degree Bernoulli polynomials:

$$C_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{t}) := \sum_{\mathbf{k} \in \mathbb{Z}^d} \alpha_{\mathbf{k}, \boldsymbol{\theta}} e^{2\pi\sqrt{-1}\mathbf{k}^T \mathbf{x}} e^{-2\pi\sqrt{-1}\mathbf{k}^T \mathbf{t}},$$

where d is number of dimensions and $\alpha_{\mathbf{k}}$ is a scalar. The Gram matrix formed by this kernel is Hermitian. The *shape parameter* $\boldsymbol{\theta}$ changes the kernel's shape, so that the function space spanned by the kernel closely resembles the space bearing

the integrand. This form of the kernel is very convenient to use in any analytical derivations, but not suitable for use with finite precision computers as this involves infinite sum. If the coefficients are chosen as

$$\alpha_{\mathbf{k}, \boldsymbol{\theta}} := \prod_{l=1}^d \frac{1}{\max(\frac{|k_l|}{\theta_l}, 1)_{\theta_l \leq 1}^r}, \quad \text{with } \alpha_{\mathbf{0}, \boldsymbol{\theta}} = 1,$$

then there exists a simpler closed form expression without infinite sum

$$C_{\boldsymbol{\theta}}(\mathbf{t}, \mathbf{x}) = \prod_{l=1}^d \left[1 - (-1)^r \gamma B_{2r}(|x_l - t_l|) \right], \quad \forall \mathbf{t}, \mathbf{x} \in [0, 1]^d,$$

$$\boldsymbol{\theta} = (r, \gamma), \quad r \in \mathbb{N}, \quad \gamma > 0.$$

Symmetric, periodic, positive definite kernels of this form appear in [8] and [11]. The Bernoulli polynomials $B_r(x)$ are described in [24, Chapter 24]

$$B_r(x) = \frac{-r!}{(2\pi\sqrt{-1})^r} \sum_{\substack{k \neq 0, \\ k=-\infty}}^{\infty} \frac{e^{2\pi\sqrt{-1}kx}}{k^r} \begin{cases} \text{for } r = 1, & 0 < x < 1 \\ \text{for } r = 2, 3, \dots & 0 \leq x \leq 1 \end{cases}$$

Larger ‘ r ’ implies a greater degree of smoothness of the kernel. Larger γ implies greater fluctuations of the output with respect to the input. Plots of $C(\cdot, 0.3)$ are given in Figure 4.2 for various r and γ values.

4.2.1 Derivative of the kernel

$$\begin{aligned} \frac{\partial}{\partial \gamma_i} C_{\boldsymbol{\theta}}(\mathbf{t}, \mathbf{x}) &= \frac{\partial}{\partial \gamma_i} \prod_{l=1}^d \left[1 - (-1)^r \gamma B_{2r}(|x_l - t_l|) \right] \\ &= \sum_{j=1}^d \prod_{l=1, l \neq j}^d \left[1 - (-1)^r \gamma B_{2r}(|x_l - t_l|) \right] \left(-(-1)^r B_{2r}(|x_j - t_j|) \right) \\ &= \prod_{l=1}^d \left[1 - (-1)^r \gamma B_{2r}(|x_l - t_l|) \right] \sum_{j=1}^d \frac{\left(-(-1)^r B_{2r}(|x_j - t_j|) \right)}{1 - (-1)^r \gamma B_{2r}(|x_j - t_j|)} \\ &= C_{\boldsymbol{\theta}}(\mathbf{t}, \mathbf{x}) \frac{1}{\gamma} \sum_{j=1}^d \frac{\left(1 - (-1)^r \gamma B_{2r}(|x_j - t_j| - 1) \right)}{1 - (-1)^r \gamma B_{2r}(|x_j - t_j|)} \end{aligned}$$

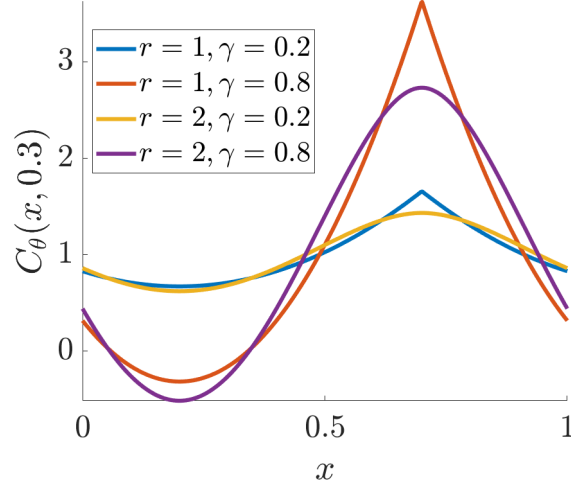


Figure 4.2: Shift invariant kernel in 1D shifted by 0.3 to show the discontinuity

$$\begin{aligned}
 &= C_{\theta}(\mathbf{t}, \mathbf{x}) \frac{1}{\gamma} \sum_{j=1}^d \left(1 - \frac{1}{1 - (-1)^r \gamma B_{2r}(|x_j - t_j|)} \right) \\
 &= \frac{d}{\gamma} \underbrace{\left(\prod_{l=1}^d \left[1 - (-1)^r \gamma B_{2r}(|x_l - t_l|) \right] \right)}_{C_{\theta}(\mathbf{t}, \mathbf{x})} \left(1 - \frac{1}{d} \sum_{j=1}^d \frac{1}{1 - (-1)^r \gamma B_{2r}(|x_j - t_j|)} \right)
 \end{aligned}$$

4.2.2 Shape parameter search using steepest descent

Now we can implement the steepest descent search introduced in (??)

$$\boldsymbol{\theta}_i^{(j+1)} = \boldsymbol{\theta}_i^{(j)} - \nu \frac{\partial}{\partial \boldsymbol{\theta}_i} \mathcal{L}(\boldsymbol{\theta} | \mathbf{y})$$

4.2.3 Eigenvectors

For general shift-invariance covariance functions the Gram matrix takes the form

$$\mathbf{C} = \left(C(\mathbf{x}_i, \mathbf{x}_j) \right)_{i,j=1}^n$$

$$= \left(K(\mathbf{h}(\phi(i-1) - \phi(j-1)) \bmod \mathbf{1}) \right)_{i,j=1}^n.$$

We now demonstrate that the eigenvector matrix for \mathbf{C} is

$$\mathbf{V} = \left(e^{2\pi n \sqrt{-1} \phi(i-1) \phi(j-1)} \right)_{i,j=1}^n. \quad (4.3)$$

Assumption (3.2b) follows automatically. Now, note that the k, j element of $\mathbf{V}^H \mathbf{V}$ is

$$\sum_{i=1}^n e^{2\pi n \sqrt{-1} \phi(i-1) [\phi(j-1) - \phi(k-1)]}.$$

Noting that the sequence $\{\phi(i-1)\}_{i=1}^n$ is a re-ordering of $0, \dots, 1 - 1/n$ for n a power of 2, this sum may be re-written by replacing $\phi(i-1)$ by $(i-1)/n$:

$$\sum_{i=1}^n e^{2\pi \sqrt{-1} (i-1) [\phi(j-1) - \phi(k-1)]}.$$

Since $\phi(j-1) - \phi(k-1)$ is some integer multiple of $1/n$, it follows that this sum is $n\delta_{j,k}$, where δ is the Kroneker delta function. This establishes that $\mathbf{V}^H = n\mathbf{V}^{-1}$ as in (3.1).

Next, let $\omega_{k,\ell}$ denote the k, ℓ element of $\mathbf{V}^H \mathbf{C} \mathbf{V}$, which is given by the double sum

$$\omega_{k,\ell} = \sum_{i,j=1}^n K(\mathbf{h}(\phi(i-1) - \phi(j-1)) \bmod \mathbf{1}) \times e^{-2\pi n \sqrt{-1} \phi(k-1) \phi(i-1)} e^{2\pi n \sqrt{-1} \phi(j-1) \phi(\ell-1)}$$

Noting that the sequence $\{\phi(i-1)\}_{i=1}^n$ is a re-ordering of $0, \dots, 1 - 1/n$ for n a power of 2, this sum may be re-written by replacing $\phi(i-1)$ by $(i-1)/n$ and $\phi(j-1)$ by $(j-1)/n$:

$$\omega_{k,\ell} = \sum_{i,j=1}^n K\left(\mathbf{h}\left(\frac{i-j}{n}\right) \bmod \mathbf{1}\right) \times e^{-2\pi \sqrt{-1} \phi(k-1)(i-1)} e^{2\pi \sqrt{-1} (j-1) \phi(\ell-1)}.$$

This sum also remains unchanged if i is replaced by $i+m$ and j is replaced by $j+m$ for any integer m :

$$\omega_{k,\ell} = \sum_{i,j=1}^n K\left(\mathbf{h}\left(\frac{i-j}{n}\right) \bmod \mathbf{1}\right) \times e^{-2\pi \sqrt{-1} \phi(k-1)(i+m-1)} e^{2\pi \sqrt{-1} (j+m-1) \phi(\ell-1)}$$

$$= \omega_{k,\ell} e^{2\pi\sqrt{-1}m(\phi(\ell-1)-\phi(k-1))}.$$

For this last equality to hold for all integers m , we must have $k = \ell$ or $\omega_{k,\ell} = 0$. Thus,

$$\begin{aligned} \omega_{k,\ell} &= \delta_{k,\ell} \sum_{i,j=1}^n K \left(\mathbf{h} \left(\frac{i-j}{n} \right) \bmod \mathbf{1} \right) \times e^{-2\pi\sqrt{-1}(i-j)\phi(k-1)} \\ &= n\delta_{k,\ell} \sum_{i=1}^n K \left(\left(\frac{i\mathbf{h}}{n} \right) \bmod \mathbf{1} \right) e^{-2\pi\sqrt{-1}i\phi(k-1)}. \end{aligned}$$

This establishes $\mathbf{V}^H \mathbf{C} \mathbf{V}$ as a diagonal matrix whose diagonal elements are n times the eigenvalues, i.e., $\lambda_k = \omega_{k,k}/n$. Furthermore, \mathbf{V} is the matrix of eigenvectors, which satisfies assumption (3.2a).

4.2.4 Iterative Computation of the Fast Transform

Assumption (3.2a) is that computing $\mathbf{V}^H \mathbf{b}$ requires only $\mathcal{O}(n \log(n))$ operations. Recall that we assume that n is a power of 2. This can be accomplished by an iterative algorithm. Let $\mathbf{V}^{(n)}$ denote the $n \times n$ matrix \mathbf{V} defined in (4.3). We show how to compute $\mathbf{V}^{(2n)H} \mathbf{b}$ quickly for all $\mathbf{b} \in \mathbb{R}^{2n}$ assuming that $\mathbf{V}^{(n)H} \mathbf{b}$ can be computed quickly for all $\mathbf{b} \in \mathbb{R}^n$.

From the definition of the van der Corput sequence in (4.1) it follows that

$$\phi(2i) = \phi(i)/2, \quad \phi(2i+1) = [\phi(i) + 1]/2, \quad i \in \mathbb{N}_0 \quad (4.4)$$

$$\phi(i+n) = \phi(i) + 1/(2n), \quad i = 0, \dots, n-1, \quad (4.5)$$

$$n\phi(i) \in \mathbb{N}_0, \quad i = 0, \dots, n-1, \quad (4.6)$$

still assuming that n is an integer power of two. Let $\tilde{\mathbf{b}} = \mathbf{V}^{(2n)H} \mathbf{b}$ for some arbitrary

$\mathbf{b} \in \mathbb{R}^{2n}$, and define

$$\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_{2n} \end{pmatrix}, \quad \mathbf{b}^{(1)} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}, \quad \mathbf{b}^{(2)} = \begin{pmatrix} b_{n+1} \\ \vdots \\ b_{2n} \end{pmatrix},$$

$$\tilde{\mathbf{b}} = \begin{pmatrix} \tilde{b}_1 \\ \vdots \\ \tilde{b}_{2n} \end{pmatrix}, \quad \tilde{\mathbf{b}}^{(1)} = \begin{pmatrix} b_1 \\ b_3 \\ \vdots \\ b_{2n-1} \end{pmatrix}, \quad \tilde{\mathbf{b}}^{(2)} = \begin{pmatrix} b_2 \\ b_4 \\ \vdots \\ b_{2n} \end{pmatrix}.$$

It follows from these definitions and the definition of \mathbf{V} in (4.3) that

$$\begin{aligned} \tilde{\mathbf{b}}^{(1)} &= \left(\sum_{j=1}^{2n} e^{-4\pi n \sqrt{-1} \phi(2i-2)\phi(j-1)} b_j \right)_{i=1}^n \\ &= \left(\sum_{j=1}^{2n} e^{-2\pi n \sqrt{-1} \phi(i-1)\phi(j-1)} b_j \right)_{i=1}^n \quad \text{by (4.4)} \\ &= \left(\sum_{j=1}^n e^{-2\pi n \sqrt{-1} \phi(i-1)\phi(j-1)} b_j \right)_{i=1}^n + \left(\sum_{j=1}^n e^{-2\pi n \sqrt{-1} \phi(i-1)\phi(n+j-1)} b_{n+j} \right)_{i=1}^n \\ &= \mathbf{V}^{(n)H} \mathbf{b}^{(1)} + \left(e^{-\pi \sqrt{-1} \phi(i-1)} s \sum_{j=1}^n e^{-2\pi n \sqrt{-1} \phi(i-1)\phi(j-1)} b_{n+j} \right)_{i=1}^n \quad \text{by (4.5)} \\ &= \mathbf{V}^{(n)H} \mathbf{b}^{(1)} + \left(e^{-\pi \sqrt{-1} \phi(i-1)} \right)_{i=1}^n \odot (\mathbf{V}^{(n)H} \mathbf{b}^{(2)}), \end{aligned}$$

where \odot denotes the Hadamard (term-by-term) product. By a similar argument,

$$\begin{aligned} \tilde{\mathbf{b}}^{(2)} &= \left(\sum_{j=1}^{2n} e^{-4\pi n \sqrt{-1} \phi(2i-1)\phi(j-1)} b_j \right)_{i=1}^n \\ &= \left(\sum_{j=1}^{2n} e^{-2\pi n \sqrt{-1} [\phi(i-1)+1]\phi(j-1)} b_j \right)_{i=1}^n \quad \text{by (4.4)} \\ &= \left(\sum_{j=1}^n e^{-2\pi n \sqrt{-1} [\phi(i-1)+1]\phi(j-1)} b_j \right)_{i=1}^n + \left(\sum_{j=1}^n e^{-2\pi n \sqrt{-1} [\phi(i-1)+1]\phi(n+j-1)} b_{n+j} \right)_{i=1}^n \end{aligned}$$

$$\begin{aligned}
&= \mathbf{V}^{(n)H} \mathbf{b}^{(1)} + \left(e^{-\pi\sqrt{-1}[\phi(i-1)+1]} \sum_{j=1}^n e^{-2\pi n\sqrt{-1}\phi(i-1)\phi(j-1)} b_{n+j} \right)_{i=1}^n \\
&\quad \text{by (4.5) and (4.6)} \\
&= \mathbf{V}^{(n)H} \mathbf{b}^{(1)} - \left(e^{-\pi\sqrt{-1}\phi(i-1)} \right)_{i=1}^n \odot (\mathbf{V}^{(n)H} \mathbf{b}^{(2)}).
\end{aligned}$$

The computational cost to compute $\mathbf{V}^{(2n)H} \mathbf{b}$ is then twice the cost of computing $\mathbf{V}^{(n)H} \mathbf{b}^{(1)}$ plus $2n$ multiplications plus $2n$ additions/subtractions. An inductive argument shows that $\mathbf{V}^{(n)H} \mathbf{b}$ requires only $\mathcal{O}(n \log(n))$ operations.

4.2.5 Overcoming Cancellation Error

For the kernels used in our computation it may happen that n/λ_1 is close to 1. Thus, the term $1 - n/\lambda_1$, which appears in the credible interval widths, err_{MLE} , err_{full} , and err_{GCv} , may suffer from cancellation error. We can avoid this cancellation error by modifying how we compute the Gram matrix and its eigenvalues.

Define a new function $\mathring{C} := C - 1$, and its associated Gram matrix $\mathring{\mathbf{C}} = \mathbf{C} - \mathbf{1}\mathbf{1}^T$. Note that \mathring{C} inherits the shift-invariant properties of C . Since $\mathbf{1}$ is the first eigenvector of \mathbf{C} it follows that the eigenvalues of $\mathring{\mathbf{C}}$ are $\mathring{\lambda}_1 = \lambda_1 - n, \lambda_2, \dots, \lambda_n$. Moreover,

$$1 - \frac{n}{\lambda_1} = \frac{\lambda_1 - n}{\lambda_1} = \frac{\mathring{\lambda}_1}{\mathring{\lambda}_1 + n},$$

where now the right hand side is free of cancellation error.

We show how to compute \mathring{C} without introducing round-off error. The covariance functions that we use are of product form, namely,

$$C(\mathbf{t}, \mathbf{x}) = \prod_{\ell=1}^d \left[1 + \mathring{C}_\ell(t_\ell, x_\ell) \right], \quad \mathring{C}_\ell : [0, 1]^2 \rightarrow \mathbb{R}.$$

Direct computation of $\mathring{C}(\mathbf{t}, \mathbf{x}) = C(\mathbf{t}, \mathbf{x}) - 1$ introduces cancellation error if the \mathring{C}_ℓ

are small. So, we employ the iteration

$$\begin{aligned}\mathring{C}^{(1)} &= \mathring{C}_1(t_1, x_1), \\ \mathring{C}^{(\ell)} &= \mathring{C}^{(\ell-1)}[1 + \mathring{C}_\ell(t_\ell, x_\ell)] + \mathring{C}_\ell(t_\ell, x_\ell), \quad \ell = 2, \dots, d, \\ \mathring{C}(\mathbf{t}, \mathbf{x}) &= \mathring{C}^{(d)}.\end{aligned}$$

In this way, the Gram matrix $\mathring{\mathbf{C}}$, whose i, j -element is $\mathring{C}(\mathbf{x}_i, \mathbf{x}_j)$ can be constructed with minimal round-off error.

Computing the eigenvalues of $\mathring{\mathbf{C}}$ via the procedure given in (3.3) yields $\mathring{\lambda}_1 = \lambda_1 - n, \lambda_2, \dots, \lambda_n$. The estimates of $\boldsymbol{\theta}$ are computed in terms of the eigenvalues of $\mathring{\mathbf{C}}$, so (3.5) and (3.8) become

$$\boldsymbol{\theta}_{\text{MLE}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \left[\log \left(\sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i} \right) + \frac{1}{n} \sum_{i=1}^n \log(\lambda_i) \right], \quad (4.7a)$$

$$\boldsymbol{\theta}_{\text{GCV}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \left[\log \left(\sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i^2} \right) - 2 \log \left(\sum_{i=1}^n \frac{1}{\lambda_i} \right) \right], \quad (4.7b)$$

where $\lambda_1 = n + \mathring{\lambda}_1$. The widths of the credible intervals in (3.6), (3.7), and (??) become

$$\text{err}_{\text{MLE}} = \frac{2.58}{n} \sqrt{\frac{\mathring{\lambda}_1}{\lambda_1} \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i}}, \quad (4.8a)$$

$$\text{err}_{\text{full}} = \frac{t_{n_j-1, 0.995}}{n} \sqrt{\frac{\mathring{\lambda}_1}{n-1} \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i}}, \quad (4.8b)$$

$$\text{err}_{\text{GCV}} = \frac{2.58}{n} \sqrt{\frac{\mathring{\lambda}_1}{\lambda_1} \sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\lambda_i^2} \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{\lambda_i} \right]^{-1}}. \quad (4.8c)$$

Since $\mathring{\lambda}_1 = \lambda_1 - n$ and $\lambda_1 \sim n$ it follows $\mathring{\lambda}_1/\lambda_1 \approx \mathring{\lambda}_1/(n-1)$ and is small for large n . Moreover, the credible intervals via empirical Bayes and full Bayes are similar, since $t_{n_j-1, 0.995}$ is approximately 2.58.

We summarize the results of this section and the previous one as follows:

Proposition 1 *Any periodic, symmetric, positive definite, shift-invariant covariance kernel of the form (4.2) scaled to satisfy (3.4), when matched with rank-1 lattice data-sites, must satisfy assumptions (3.2). The fast Fourier transform (FFT) can be used to expedite the estimates of $\boldsymbol{\theta}$ in (4.7) and the credible interval widths (4.8) in $\mathcal{O}(n \log(n))$ operations. The cubature, $\hat{\mu}$, is just the sample mean.*

We have implemented the fast adaptive Bayesian cubature algorithm in MATLAB as part of the Guaranteed Adaptive Integration Library (GAIL) [3] as `cubBayesLattice_g`. This algorithm uses the kernel defined in (??) with $r = 1, 2$ and the periodizing variable transforms in Section 6.1. The rank-1 lattice node generator is taken from [21] (`exod2_base2_m20`).

CHAPTER 5

SOBOL POINTS AND WALSH KERNELS

Another approach to formulate fast transform using matching kernel and point sets is presented in this section. Interlaced digital nets and smoother kernels are paired. Higher order digital nets with matching smoother kernels are expected to provide higher order error convergence $\mathcal{O}(N^{-\alpha+\epsilon})$ for $\alpha \geq 1$ and $\epsilon > 0$.

5.1 Higher order digital nets

Nets were developed to provide deterministic sample points for quasi-Monte Carlo rules [20]. The (t, m, s) -nets in base b introduced by Niederreiter are such point sets consist of b^m points in $[0, 1)^s$, whose quality is governed by t , in particular, lower values of t correspond to (t, m, s) -nets of higher quality [1]. Digital (t, m, s) -nets are a special case of (t, m, s) -nets, constructed using matrix-vector multiplications over finite fields.

Higher order digital nets are a modified (t, m, s) -nets, introduced in [7] can be used to numerically integrate smoother functions which are not necessarily periodic, but have square integrable mixed partial derivatives of order α , at a rate of $\mathcal{O}(N^{-\alpha})$ multiplied by a power of a $\log N$ factor using rules corresponding to the modified (t, m, s) -nets. The qualifier “higher order” emphasizes that quasi-Monte Carlo rules based on these point sets can achieve convergence rates faster than $\mathcal{O}(N^{-1})$.

Higher order digital nets are constructed using matrix-vector multiplications over finite fields. We use generating matrices from the magic point shop [22].

5.1.1 Walsh functions

Like Fourier transform naturally occurred with Lattice points (Section 4.2), Walsh transform happens for the nets. Define $\mathbb{N} := \{1, 2, \dots\}$ and $\mathbb{N}_0 := \{0, 1, 2, \dots\}$. The one-dimensional Walsh functions in base b are defined as

$$\text{wal}_{b,h}(x) := e^{2\pi i(x_1 h_0 + x_2 h_1 + \dots + x_n h_{n-1})/b} = e^{2\pi i[\vec{h}]_n^T [\vec{x}]_n / b}$$

for $x \in [0, 1)$ and $h \in \mathbb{N}_0$ and the unique base b expansions $x = \sum_{i \geq 1} x_i b^{-i} = (0.x_1 x_2 \dots)_b$, and $h = \sum_{i \geq 1} h_i b^{-i} = (0.h_1 h_2 \dots)_b$, $[\vec{h}] = (h_1, \dots, h_n)^T$, with n at least as large as the number of digits to represent x or h . Multivariate Walsh functions are defined as the product of the one-dimensional Walsh functions

$$\text{wal}_{b,h}(\mathbf{x}) := \prod_{j=1}^s \text{wal}_{b,h_j}(x_j)$$

5.2 Higher order Walsh space

5.3 Walsh kernels

Consider the kernels of the form

$$C(\mathbf{x}, \mathbf{t}) = 1 + \theta \omega_r(\mathbf{x} \ominus \mathbf{t})$$

where

$$\omega_r(x) = \sum_{h=1}^{\infty} \frac{\text{wal}_{b,h}(x)}{b^{qr \lfloor \log_b h \rfloor}}$$

$$\omega_1(\mathbf{x}) = \frac{1}{6} - 2^{\lfloor \log_2 \mathbf{x} \rfloor - 1},$$

$$\omega_2(\mathbf{x}) = s_1(\mathbf{x}) + \tilde{s}_2(\mathbf{x}),$$

$$\omega_3(\mathbf{x}) = s_1(\mathbf{x}) + s_2(\mathbf{x}) + \tilde{s}_3(\mathbf{x})$$

where

$$s_1(\mathbf{x}) = 1 - 2\mathbf{x},$$

$$s_2(\mathbf{x}) = 1/3 - 2(1 - \mathbf{x})\mathbf{x},$$

$$\tilde{s}_2(\mathbf{x}) = (1 - 5t_1)/2 - (a_1 - 2)\mathbf{x},$$

$$\tilde{s}_3(\mathbf{x}) = (1 - 43t_2)/18 + (5t_1 - 1)\mathbf{x} + (a_1 - 2)\mathbf{x}^2.$$

with, for $0 < \mathbf{x} < 1$,

$$a_1 = -\lfloor \log_2 \mathbf{x} \rfloor, \quad t_1 = 2^{-a_1}, \quad t_2 = 2^{-2a_1},$$

and $a_1 = 0, t_1 = 0$ and $t_2 = 0$ when $\mathbf{x} = 0$.

5.3.1 The worst-case error

Define the worst-case error of numerical integration using a cubature rule Q for functions in a Hilbert space \mathcal{F} by

$$e(Q, \mathcal{F}) := \sup_{f \in \mathcal{F}, \|f\|_{\mathcal{F}} \leq 1} |I(f) - Q(f)|$$

We now choose cubature rule Q to be a quasi-Monte Carlo using a higher order digital as its node set and denote it by Q^{net}

$$|I(f) - Q^{\text{net}}(f)| = \left| \sum_{\mathbf{0} \neq \mathbf{k} \in \mathcal{D}} \widehat{f}_{\mathbf{k}} \right| \leq \sum_{\mathbf{0} \neq \mathbf{k} \in \mathcal{D}} |\widehat{f}_{\mathbf{k}}|$$

Define $\mathcal{W}_{\alpha, d, \gamma}$, the function space consisting of all functions $f \in L_2([0, 1]^d)$ for which $\|f\|_{\mathcal{W}_{\alpha, d, \gamma}} < \infty$, where

$$\|f\|_{\mathcal{W}_{\alpha, d, \gamma}}^2 := \sum_{\mathbf{k} \in \mathbb{N}^d} \left(\frac{|\widehat{f}_{\mathbf{k}}|}{r_{\alpha}(\gamma, \mathbf{k})} \right)^2$$

where $\widehat{f}_{\mathbf{k}}$ are the walsh coefficients of f which satisfy

$$\left| \widehat{f}_{\mathbf{k}} \right|^2 \leq \|f\|_{\mathcal{W}_{\alpha,d,\gamma}} (\mathbf{r}_{\alpha}(\boldsymbol{\gamma}, \mathbf{k}))^2 \quad \forall \mathbf{k} \in \mathbb{N}_0^d$$

Given a fixed integer $\alpha > 1$ and a fixed sequence of positive weights $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \dots\}$, the decay function $\mathbf{r}_{\alpha}(\boldsymbol{\gamma}, \mathbf{k})$ is defined as

$$\mathbf{r}_{\alpha}^2(\boldsymbol{\gamma}, \mathbf{k}) = \prod_{j \in u} \gamma_j^2 r_{\alpha}^2(k_j), \quad r_{\alpha}(k) := 2^{-\sum_{i=1}^{\min(\#k, \alpha)} (a_i + 1)}$$

where for $\mathbf{k} = (k_1, \dots, k_d)$, $k = \sum_{i=1}^{\#k} k_{a_i} b^{a_i}$, we set $u = \{1 \leq j \leq d : k_j \neq 0\}$ and $\#k$ is the number of non-zero binary digits of integer k .

$$\zeta(\mathbf{x}) = \sum_{\mathbf{k} \in \mathbb{N}} \mathbf{r}_{\alpha}(\boldsymbol{\gamma}, \mathbf{k}) \text{wal}_{\mathbf{k}}(\mathbf{x})$$

5.3.2 Calculating the infinite sum

For practical usage with a computer, we can not have infinite sum. In this section we show how to calculate in finite terms. If x can be represented using n digits in base b , then $\omega_{\alpha}(x)$ can be computed in $\mathcal{O}(\alpha n)$ operations [16]. Specifically, we will compute explicit forms for $b = 2$ and $x \in [0, 1)$.]

$$\omega_{\alpha}(\mathbf{x}) := \sum_{\mathbf{k} \in \mathbb{N}} \mathbf{r}_{\alpha}^2(\mathbf{k}) \text{wal}_{\mathbf{k}}(\mathbf{x}) \tag{5.1}$$

$$= \sum_{\substack{\mathbf{k} \in \mathbb{N} \\ k = \sum_{i=1}^{\#k} k_{a_i} b^{a_i}}}^{\infty} \prod_{i=1}^{\min(\alpha, \#k)} b^{-2(a_i+1)} \exp(2\pi i k_{a_i} \xi_{a_i+1}/b) \underbrace{\prod_{i=\alpha+1}^{\#k} \exp(2\pi i k_{a_i} \xi_{a_i+1}/b)}_{\text{wal}_{k'}(x)} \tag{5.2}$$

We used the definition $k = \sum_{i=1}^{\#k} k_{a_i} b^{a_i}$ with $a_1 > \dots > a_{\#k}$, i.e., k_{a_1} is the most significant digit of k in base b . The second product in 5.1 can be seen as $\text{wal}_{k'}(x)$ where

$k' = k - \sum_{i=1}^{\min(\alpha, \#k)} k_{a_i} b^{a_i}$. The key technique used is to interpret the infinite sum in 5.1 as multiple sums over all possible digit expansions for all k 's with r digits for $r \geq 1$.

$$\sum_{\substack{k \in \mathbb{N} \\ k = \sum_{i=1}^{\#k} k_{a_i} b^{a_i}}} G(k, x) = \sum_{r=1}^{\infty} \underbrace{\sum_{a_1=r-1}^{\infty} \cdots \sum_{a_r=0}^{a_{r-1}-1}}_{r \text{ sums s.t. } \infty > a_1 \dots a_r \geq 0} \underbrace{\sum_{k_{a_1}=1}^{b-1} \cdots \sum_{k_{a_r}=1}^{b-1}}_{r \text{ independent sums}} G(k_{a_i} b^{a_i}, x)$$

where $G(k, x) = r_{\alpha}(k) \text{wal}_k(x)$. to simplify the notation and reveal the structure for further simplification, define the *triangular sum operator*

$$T_m^M(r)(g) := \sum_{a_1=m+r-1}^M \sum_{a_2=m+r-2}^{a_1-1} \cdots \sum_{a_r=m}^{a_{r-1}-1} g(a_1, \dots, a_r),$$

where $T_m^M(0)(g) := g$, i.e., no sums taken. Next, define the *concatenation* of two such operators as putting the sums next to each other:

$$\left(T_m^M(t) T_{m'}^{M'}(r-t) \right) (g) := \sum_{a_1=m+t-1}^M \cdots \sum_{a_t=m}^{a_{t-1}-1} \sum_{a_{t+1}=m'+(r-t)-1}^{M'} \cdots \sum_{a_r=m'}^{a_{r-1}-1} g(a_1, \dots, a_r),$$

5.3.3 Closed form expressions

$$\omega_{b,\alpha} := \sum_{\kappa=1}^{\infty} r_{b,\alpha} \text{wal}_{b,\kappa}(x)$$

for a given base b and number $x \in [0, 1)$. Given an integer κ , consider its base b expansion

$$\kappa = \sum_{i=0}^{\infty} \kappa_i b^i, \quad \kappa_i \in \mathbb{Z}_b := \{0, 1, \dots, b-1\}$$

We are interested in the non-zero digits of κ in base b . Denote the number of non-zero digits of κ by $\#\kappa$.

$$\kappa = \sum_{i=1}^{\#\kappa} \kappa_{a_i} b^{a_i}, \quad \kappa_{a_i} \in \mathbb{Z}_b \setminus \{0\}, \quad a_1 > a_2 \dots a_{\#\kappa} \geq 0$$

Now, define the function $r_{b,\alpha}(\kappa)$

$$r_{b,\alpha}(\kappa) := b^{-\sum_{i=1}^{\min(\#\kappa, \alpha)} (a_i+1)} = \begin{cases} 1, & \text{if } \kappa = 0, \\ b^{-((a_1+1)+\dots+(a_{\min(\#\kappa, \alpha)+1}))}, & \text{for } \kappa \text{ as given above} \end{cases}$$

Note that the function $r_{b,\alpha}(\kappa)$ only depends on the α most-significant base- b digit locations of κ , but in a such a way their digit values do not matter

Similarly, assume $x \in [0, 1)$ has a base- b expansion

$$x = \sum_{i=1}^{\infty} x_i b^{-i} = \sum_{i=1}^{\#x} x_{a_i} b^{-a_i}, \quad 1 \leq a_1 < a_2 < \dots < a_{\#x}, \quad x_i \in \mathbb{Z}_b, \quad x_{a_i} \in \mathbb{Z}_b \setminus \{0\}$$

where $-a_1$ is the most significant power of b . For non-zero x , we find $a_1 = -\lfloor \log_b(x) \rfloor$

Consider the case $\alpha = 2$.

$$\begin{aligned} \omega_{b,2} &= \sum_{l=0}^{\infty} \sum_{h_l=1}^{b-1} r_{b,2}(h_l b^l) \text{wal}_{b,h_l b^l}(x) \\ &\quad + \sum_{l=0}^{\infty} \sum_{h_l=1}^{b-1} \sum_{l'=0}^{l-1} \sum_{h_{l'}=1}^{b-1} \sum_{h'=0}^{b^{l'}-1} r_{b,2}(h_l b^l + h_{l'} b^{l'} + h') \text{wal}_{b,h_l b^l + h_{l'} b^{l'} + h'}(x) \\ &= s_1(x) + \tilde{s}_2(x) \end{aligned} \tag{5.3}$$

For $b = 2$ it gets much simpler

$$\begin{aligned} \omega_{b,2} &= \sum_{l=0}^{\infty} r_{b,2}(b^l) \text{wal}_{b,b^l}(x) + \sum_{l=0}^{\infty} \sum_{l'=0}^{l-1} \sum_{h'=0}^{b^{l'}-1} r_{b,2}(b^l + b^{l'} + h') \text{wal}_{b,b^l + b^{l'} + h'}(x) \\ &= \sum_{l=0}^{\infty} 2^{-(l+1)} \exp(2\pi i x_{l+1}/2) + \sum_{l=0}^{\infty} \sum_{l'=0}^{l-1} \sum_{h'=0}^{b^{l'}-1} r_{b,2}(b^l + b^{l'} + h') \text{wal}_{b,b^l + b^{l'} + h'}(x) \end{aligned}$$

Note that

$$\sum_{h_l=1}^{b-1} \text{wal}_{b,h_l b^l}(x) = \exp(2\pi i x_{l+1}/2) = (-1)^{x_{l+1}}$$

Let's start with s_1 ,

$$s_1(x) = \sum_{l=0}^{\infty} r_{b,2}(b^l) \text{wal}_{b,b^l}(x)$$

$$\begin{aligned}
&= \sum_{l=0}^{\infty} 2^{-(l+1)} \exp(2\pi\sqrt{-1}x_{l+1}/2) \\
&= \sum_{l=1}^{\infty} 2^{(-l)} \exp(\pi\sqrt{-1}x_l), \quad \text{where } x_l \in \{0, 1\} \\
&= \sum_{l=1}^{\infty} 2^{(-l)} (-1)^{x_l} \\
&= \sum_{l=1}^{\infty} 2^{-l} + \sum_{l=1}^{\infty} 2^{-l} ((-1)^{x_l} - 1) \\
&= 1 + \sum_{l=1}^{\infty} 2^{-l} ((-1)^{x_l} - 1) \\
&= 1 + \sum_{l=1}^{\infty} 2^{-l} (-2) [x_l = 1] \\
&= 1 - 2 \sum_{l=1}^{\infty} 2^{-l} [x_l = 1] \\
&= 1 - 2 \sum_{j=1}^{\#x} 2^{-a_j}, \quad \text{using } x = \sum_{j=1}^{\#x} 2^{-a_j} \\
&= 1 - 2x
\end{aligned}$$

Similarly, we can solve \tilde{s}_2

$$\tilde{s}_2(x) = \sum_{k=1; k=2^l+2^{l'}+k'}^{\infty} r_2(k) \text{wal}_k(x)$$

where $k = 2^l + 2^{l'} + k'$, $k' \in \{0, \dots, 2^{l'} - 1\}$

$$\begin{aligned}
\tilde{s}_2(x) &= \sum_{k=1; k=2^l+2^{l'}+k'}^{\infty} 2^{-(l+1)} (-1)^{x_{l+1}} 2^{-(l'+1)} (-1)^{x_{l'+1}} \text{wal}_{k'}(x) \\
&= \sum_{l=1}^{\infty} 2^{-(l+1)} (-1)^{x_{l+1}} \sum_{l'=0}^{l-1} 2^{-(l'+1)} (-1)^{x_{l'+1}} \sum_{k'=0}^{2^{l'}-1} \text{wal}_{k'}(x) \\
&= \sum_{l'=0}^{\infty} 2^{-(l'+1)} (-1)^{x_{l'+1}} \sum_{k'=0}^{2^{l'}-1} \text{wal}_{k'}(x) \sum_{l=l'+1}^{\infty} 2^{-(l+1)} (-1)^{x_{l+1}}
\end{aligned}$$

Using the character property or by the definition of wal function

$$\sum_{k=0}^{b^l-1} \text{wal}_k(x) = \sum_{k_0=0}^{b-1} \sum_{k_1=0}^{b-1} \cdots \sum_{k_{l-1}=0}^{b-1} \exp(2\pi i(x_1 k_0 + x_2 k_1 + \cdots + x_l k_{l-1})/b)$$

$$= \sum_{k_0=0}^{b-1} \exp(2\pi i(x_1 k_0)/b) \sum_{k_1=0}^{b-1} \exp(2\pi i(x_2 k_1)/b) \cdots \sum_{k_{l-1}=0}^{b-1} \exp(2\pi i(x_l k_{l-1})/b)$$

For $b = 2$

$$\begin{aligned} \sum_{k=0}^{b^l-1} \text{wal}_k(x) &= \sum_{k_0=0}^1 \exp(\pi i(x_1 k_0)) \sum_{k_1=0}^1 \exp(\pi i(x_2 k_1)) \cdots \sum_{k_{l-1}=0}^1 \exp(\pi i(x_l k_{l-1})) \\ &= \sum_{k_0=0}^1 (-1)^{x_1 k_0} \sum_{k_1=0}^1 (-1)^{x_2 k_1} \cdots \sum_{k_{l-1}=0}^1 (-1)^{x_l k_{l-1}} \\ &= 2^l \times [x < 2^{-l}] \end{aligned}$$

Continuing with \tilde{s}_2

$$\tilde{s}_2(x) = \sum_{l'=0}^{\infty} 2^{-(l'+1)} (-1)^{x_{l'+1}} 2^{l'} [x < 2^{-l'}] \sum_{l=l'+1}^{\infty} 2^{-(l+1)} (-1)^{x_{l+1}}$$

Now suppose $x = 0$, then

$$\begin{aligned} \tilde{s}_2(0) &= \sum_{l'=0}^{\infty} 2^{-(l'+1)} 2^{l'} \sum_{l=l'+1}^{\infty} 2^{-(l+1)} \\ &= \sum_{l'=0}^{\infty} 2^{-(l'+1)} 2^{l'} 2^{-l'} \sum_{l=1}^{\infty} 2^{-(l+1)} \\ &= \sum_{l'=0}^{\infty} 2^{-(l'+1)} 2^{-1} \sum_{l=1}^{\infty} 2^{-l} \\ &= \sum_{l'=0}^{\infty} 2^{-(l'+1)} 2^{-1} \\ &= 2^{-1} \sum_{l'=1}^{\infty} 2^{-l'} \\ &= 2^{-1} \end{aligned}$$

Now suppose $x = x_{a_1} b^{-a_1} \neq 0$

$$s_2(x) = \sum_{\substack{k=1 \\ k=2^{a_1}+2^{a_2}}}^{\infty} 2^{-(a_1+1)} (-1)^{\xi_{a_1+1}} 2^{-(a_2+1)} (-1)^{\xi_{a_2+1}}$$

$$\begin{aligned}
&= \sum_{a_1=1}^{\infty} 2^{-(a_1+1)}(-1)^{\xi_{a_1+1}} \sum_{a_2=0}^{a_1-1} 2^{-(a_2+1)}(-1)^{\xi_{a_2+1}} \\
&= \sum_{a_1=2}^{\infty} 2^{-(a_1)}(-1)^{\xi_{a_1}} \sum_{a_2=1}^{a_1-1} 2^{-(a_2)}(-1)^{\xi_{a_2}}, \quad \text{index shifted by 1} \\
&= \sum_{a_2=1}^{\infty} 2^{-(a_2)}(-1)^{\xi_{a_2}} \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)}(-1)^{\xi_{a_1}} \\
&= \left(\sum_{a_2=1}^{\infty} 2^{-(a_2)}(-1)^{\xi_{a_2}} [\xi_{a_2} = 0] + \sum_{a_2=1}^{\infty} 2^{-(a_2)}(-1)^{\xi_{a_2}} [\xi_{a_2} \neq 0] \right) \\
&\quad \left(\sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)}(-1)^{\xi_{a_1}} [\xi_{a_1} = 0] + \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)}(-1)^{\xi_{a_1}} [\xi_{a_1} \neq 0] \right) \\
&= \left(\sum_{a_2=1}^{\infty} 2^{-(a_2)}(-1)^{\xi_{a_2}} [\xi_{a_2} = 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)}(-1)^{\xi_{a_1}} [\xi_{a_1} = 0] \right) \\
&\quad + \left(\sum_{a_2=1}^{\infty} 2^{-(a_2)}(-1)^{\xi_{a_2}} [\xi_{a_2} = 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)}(-1)^{\xi_{a_1}} [\xi_{a_1} \neq 0] \right) \\
&\quad + \left(\sum_{a_2=1}^{\infty} 2^{-(a_2)}(-1)^{\xi_{a_2}} [\xi_{a_2} \neq 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)}(-1)^{\xi_{a_1}} [\xi_{a_1} = 0] \right) \\
&\quad + \left(\sum_{a_2=1}^{\infty} 2^{-(a_2)}(-1)^{\xi_{a_2}} [\xi_{a_2} \neq 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)}(-1)^{\xi_{a_1}} [\xi_{a_1} \neq 0] \right) \\
&= \left(\sum_{a_2=1}^{\infty} 2^{-(a_2)}(-1)^{\xi_{a_2}} [\xi_{a_2} = 0] \left(\sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)}(-1)^{\xi_{a_1}} - \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)}(-1)^{\xi_{a_1}} [\xi_{a_1} \neq 0] \right) \right) \\
&\quad + \left(\sum_{a_2=1}^{\infty} 2^{-(a_2)}(-1)^{\xi_{a_2}} [\xi_{a_2} = 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)}(-1)^{\xi_{a_1}} [\xi_{a_1} \neq 0] \right) \\
&\quad + \left(\sum_{a_2=1}^{\infty} 2^{-(a_2)}(-1)^{\xi_{a_2}} [\xi_{a_2} \neq 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)}(-1)^{\xi_{a_1}} [\xi_{a_1} = 0] \right) \\
&\quad + \left(\sum_{a_2=1}^{\infty} 2^{-(a_2)}(-1)^{\xi_{a_2}} [\xi_{a_2} \neq 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)}(-1)^{\xi_{a_1}} [\xi_{a_1} \neq 0] \right)
\end{aligned}$$

Alternative approach

$$\begin{aligned}
&= \sum_{a_2=1}^{\infty} 2^{-(a_2)}(-1)^{\xi_{a_2}} 2^{-a_2} \sum_{a_1=1}^{\infty} 2^{-(a_1)}(-1)^{\xi_{a_1}} \\
&= \sum_{a_2=1}^{\infty} 2^{-(2a_2)}(-1)^{\xi_{a_2}} (1 - 2x)
\end{aligned}$$

$$\begin{aligned}
&= (1 - 2x) \left(\sum_{a_2=1}^{\infty} 2^{-(2a_2)} (-1)^{\xi_{a_2}} [\xi_{a_2} = 0] + \sum_{a_2=1}^{\infty} 2^{-(2a_2)} (-1)^{\xi_{a_2}} [\xi_{a_2} \neq 0] \right) \\
&= (1 - 2x) \left(\sum_{a_2=1}^{\infty} 2^{-(2a_2)} + 2 \sum_{a_2=1}^{\infty} 2^{-(2a_2)} (-1)^{\xi_{a_2}} [\xi_{a_2} \neq 0] \right) \\
&= (1 - 2x) \left(\frac{1}{3} + 2 \sum_{j=1}^{\#x} 2^{-(2a_2)} \right)
\end{aligned}$$

Simplify the first part in s_2

$$\begin{aligned}
& \sum_{a_2=1}^{\infty} 2^{-(a_2)} (-1)^{\xi_{a_2}} [\xi_{a_2} = 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)} (-1)^{\xi_{a_1}} [\xi_{a_1} = 0] \\
&= \sum_{a_2=1}^{\infty} 2^{-(a_2)} [\xi_{a_2} = 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)} [\xi_{a_1} = 0] \\
&= \sum_{a_2=1}^{\infty} 2^{-(a_2)} [\xi_{a_2} = 0] \left(\sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)} - \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)} [\xi_{a_1} \neq 0] \right) \\
&= \sum_{a_2=1}^{\infty} 2^{-(a_2)} [\xi_{a_2} = 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)} - \sum_{a_2=1}^{\infty} 2^{-(a_2)} [\xi_{a_2} = 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)} [\xi_{a_1} \neq 0] \\
&= \sum_{a_2=1}^{\infty} (2^{-(a_2)} [\xi_{a_2} = 0]) 2^{-(a_2)} - \sum_{a_2=1}^{\infty} 2^{-(a_2)} [\xi_{a_2} = 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)} [\xi_{a_1} \neq 0] \\
&= \left(\sum_{a_2=1}^{\infty} 2^{-(a_2)} - \sum_{a_2=1}^{\infty} 2^{-(a_2)} [\xi_{a_2} \neq 0] \right) 2^{-(a_2)} \\
&\quad - \sum_{a_2=1}^{\infty} 2^{-(a_2)} [\xi_{a_2} = 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)} [\xi_{a_1} \neq 0] \\
&= \sum_{a_2=1}^{\infty} 2^{-(2a_2)} - \sum_{a_2=1}^{\infty} 2^{-(2a_2)} [\xi_{a_2} \neq 0] \\
&\quad - \sum_{a_2=1}^{\infty} 2^{-(a_2)} [\xi_{a_2} = 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)} [\xi_{a_1} \neq 0] \\
&= \frac{1}{3} - \sum_{a_2=1}^{\infty} 2^{-(2a_2)} [\xi_{a_2} \neq 0] - \sum_{a_2=1}^{\infty} 2^{-(a_2)} [\xi_{a_2} = 0] \sum_{a_1=a_2+1}^{\infty} 2^{-(a_1)} [\xi_{a_1} \neq 0]
\end{aligned}$$

CHAPTER 6

NUMERICAL IMPLEMENTATION

6.1 Periodizing Variable Transformations

The shift-invariant covariance kernels underlying our Bayesian cubature assume that the integrand has a degree of periodicity, with the smoothness assumed depending on the smoothness of the kernel. While integrands arising in practice may be smooth, they might not be periodic. Variable transformations can be used to ensure periodicity.

Suppose that the original integral has been expressed as

$$\mu := \int_{[0,1]^d} g(\mathbf{t}) \, d\mathbf{t}$$

where g has sufficient smoothness, but lacks periodicity. The baker's transform,

$$\Psi : \mathbf{x} \mapsto (\Psi(x_1), \dots, \Psi(x_d)),$$

$$\Psi(x) = 1 - 2|x - 1/2|, \quad (6.1)$$

allows us to write μ in the form of (1.1), where $f(\mathbf{x}) = g(\Psi(\mathbf{x}))$.

A family of variable transforms take the form

$$\Psi : \mathbf{x} \mapsto (\Psi(x_1), \dots, \Psi(x_d)), \quad \Psi : [0, 1] \mapsto [0, 1],$$

which allows us to write μ in the form of (1.1) with

$$f(\mathbf{x}) = g(\Psi(\mathbf{x})) \prod_{\ell=1}^d \Psi'(x_\ell).$$

If Ψ is sufficiently smooth, $\lim_{x \downarrow 0} x^{-r} \Psi'(x) = \lim_{x \uparrow 1} (x - 1)^{-r} \Psi'(x) = 0$ for $r \in \mathbb{N}_0$, and $g \in C^{(r, \dots, r)}[0, 1]^d$, then f has continuous, periodic derivatives up to order r in

each direction. Examples of this kind of transform include [28]:

$$\begin{aligned} \text{Sidi's } C^1 : \Psi(x) &= x - \frac{\sin(2\pi x)}{2\pi}, \\ \Psi'(x) &= 1 - \cos(2\pi x), \\ \text{Sidi's } C^2 : \Psi(x) &= \frac{8 - 9\cos(\pi x) + \cos(3\pi x)}{16}, \\ \Psi'(x) &= \frac{3\pi[3\sin(\pi x) - \sin(3\pi x)]}{16}. \end{aligned}$$

Periodizing variable transforms are used for the numerical examples below. In some cases they can speed the convergence of the Bayesian cubature.

CHAPTER 7

NUMERICAL RESULTS AND OBSERVATIONS

7.1 Test Results and Observations

Three integrals were evaluated using the GAIL algorithm `cubBayesLattice_g`: a multivariate normal probability, the Keister’s example, and an option pricing example. Four different error tolerances, ε , were set for each example, with the tolerances chosen depending on the difficulty of the problem. The nodes were the randomly shifted lattice rules supplied by GAIL. For each integral, each tolerance, and each of our stopping criteria—empirical Bayes, full Bayes, and generalized cross-validation—our algorithm is run 100 times. For each test, the execution time is plotted against $|\mu - \hat{\mu}|/\varepsilon$. We expect $|\mu - \hat{\mu}|/\varepsilon$ to be no greater than one, but hope that it is not too much smaller than one, which would indicate a stopping criterion that is too conservative.

Multivariate Normal Probability. This example was already introduced in Section 2.4, where we used the Matérn covariance kernel. Here we apply Sidi’s C^2 periodization to f_{Genz} (2.23), and chose $d = 3$ and $r = 2$. The simulation results for this example function are summarized in Figures 7.1, 7.2, and 7.3. In all cases the Bayesian cubature returns an approximation within the prescribed error tolerance. We used the same setting as before with generic slow Bayesian cubature in Section 2.4 for comparison. For error threshold $\varepsilon = 10^{-5}$ with empirical stopping criterion, our fast algorithm takes 0.001 seconds as shown in Figure 7.1 whereas the basic algorithm takes 30 seconds as shown in Figure 2.2. Amongst the three stopping criteria, GCV achieved the results faster than others. One can also observe from the figures, the

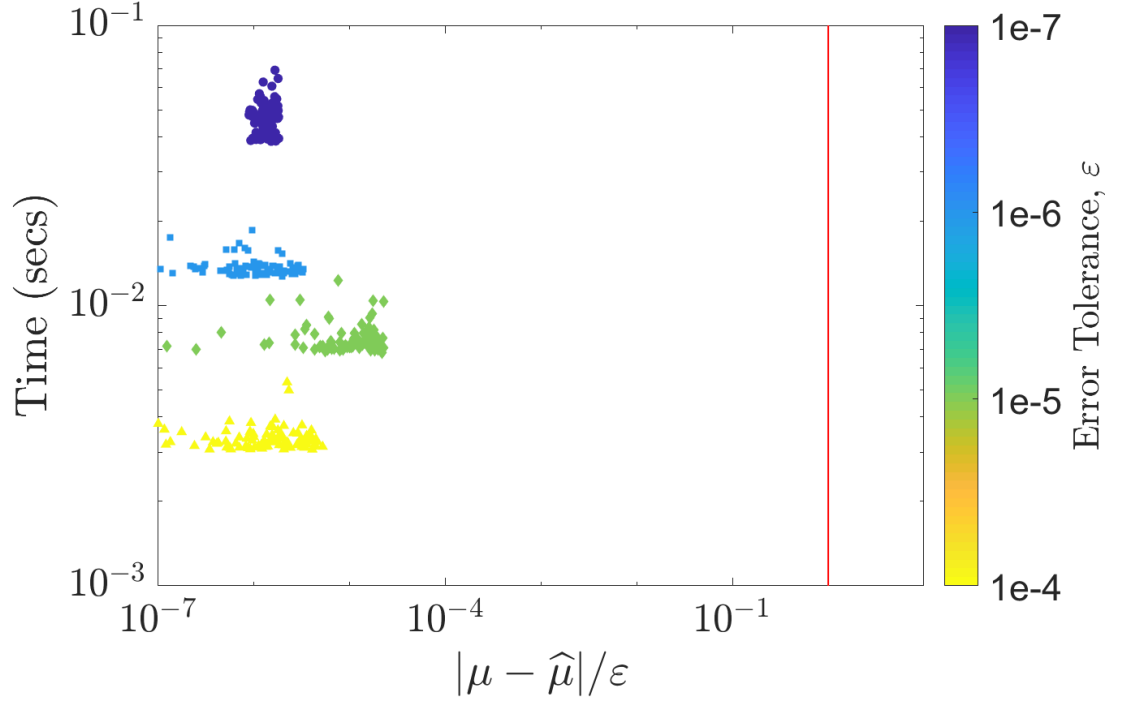


Figure 7.1: Multivariate normal probability example using the empirical Bayes stopping criterion.

credible intervals are wider, causing the true error much smaller than requested. This could be due to the periodization transformed f_{Genz} is smoother than the $r = 2$ kernel could approximate. Using a kernel of matching smoothness could produce right credible intervals.

Keister's Example. This multidimensional integral function comes from [19] and is inspired by a physics application:

$$\begin{aligned}\mu &= \int_{\mathbb{R}^d} \cos(\|\mathbf{t}\|) \exp(-\|\mathbf{t}\|^2) d\mathbf{t} \\ &= \int_{[0,1]^d} f_{\text{Keister}}(\mathbf{x}) d\mathbf{x},\end{aligned}$$

where

$$f_{\text{Keister}}(\mathbf{x}) = \pi^{d/2} \cos(\|\Phi^{-1}(\mathbf{x})/2\|),$$

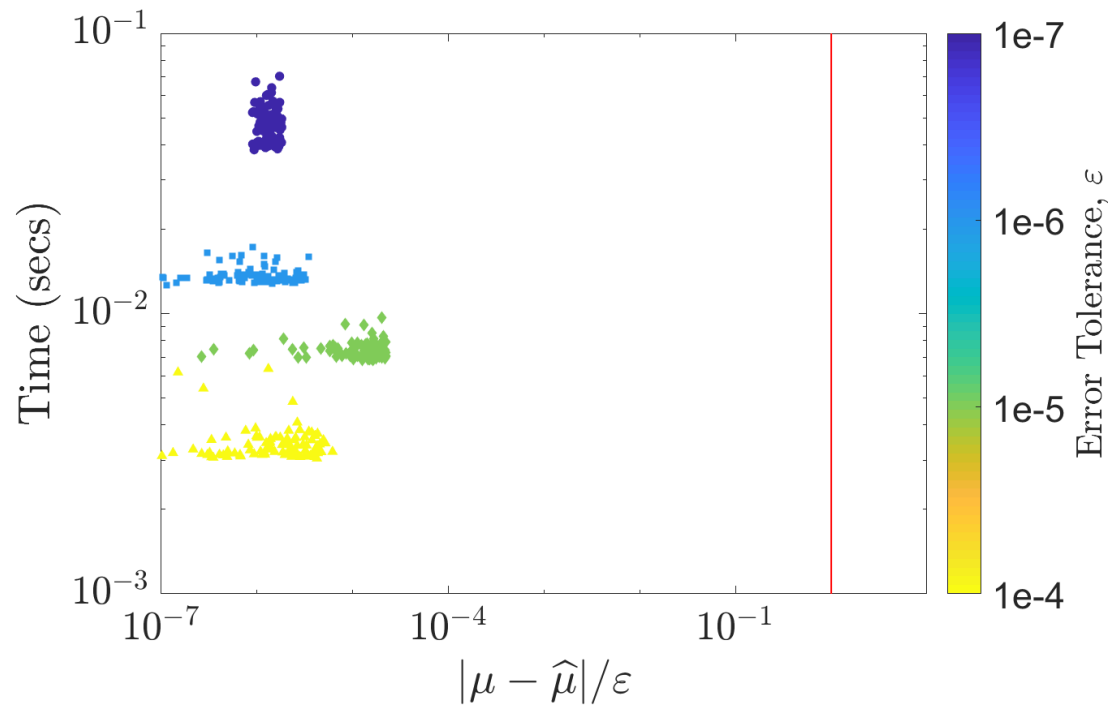


Figure 7.2: Multivariate normal probability example using the full Bayes stopping criterion.

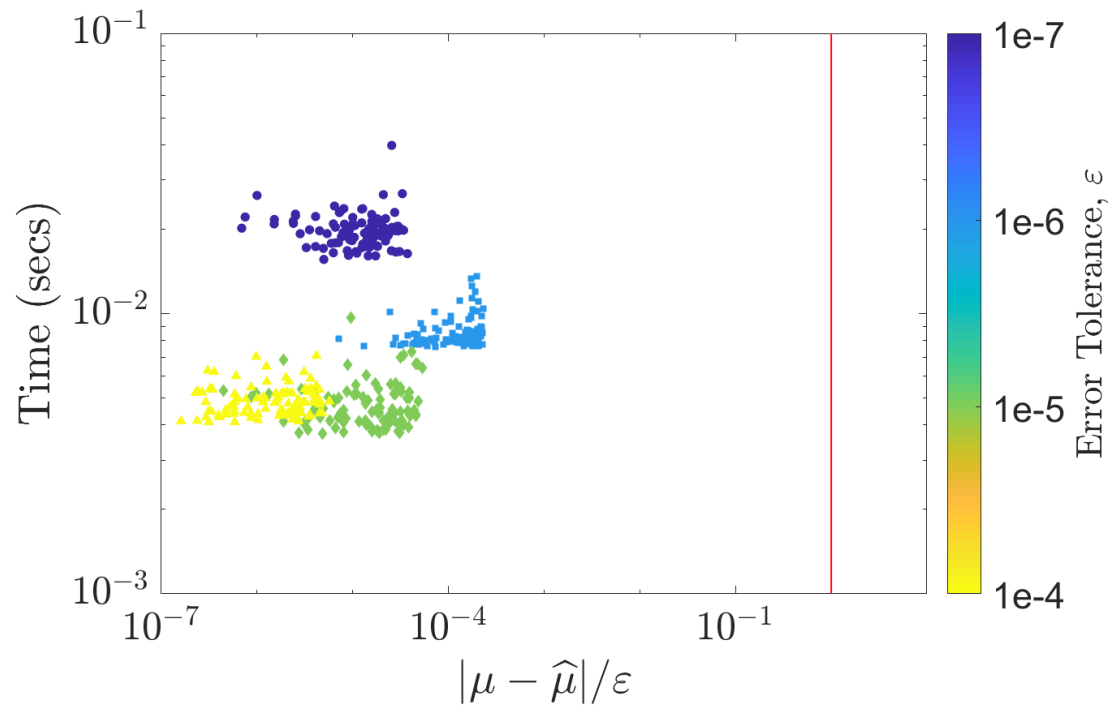


Figure 7.3: Multivariate normal probability example using the GCV stopping criterion.

and again Φ is the standard normal distribution. The true value of μ can be calculated iteratively in terms of a quadrature as follows:

$$\mu = \frac{2\pi^{d/2}I_c(d)}{\Gamma(d/2)}, \quad d = 1, 2, \dots$$

where Γ denotes the gamma function, and

$$\begin{aligned} I_c(1) &= \frac{\sqrt{\pi}}{2 \exp(1/4)}, \\ I_s(1) &= \int_{x=0}^{\infty} \exp(-\mathbf{x}^T \mathbf{x}) \sin(\mathbf{x}) d\mathbf{x} \\ &= 0.4244363835020225, \\ I_c(2) &= \frac{1 - I_s(1)}{2}, \quad I_s(2) = \frac{I_c(1)}{2} \\ I_c(j) &= \frac{(j-2)I_c(j-2) - I_s(j-1)}{2}, \quad j = 3, 4, \dots \\ I_s(j) &= \frac{(j-2)I_s(j-2) - I_c(j-1)}{2}, \quad j = 3, 4, \dots \end{aligned}$$

Figures 7.4, 7.5 and 7.6 summarize the numerical tests for this integral. We used the Sidi's C^1 periodization, dimension $d = 4$, and $r = 2$. As we can see the GCV stopping criterion achieved the results faster than the others similar to the multivariate normal case.

Option Pricing. The price of financial derivatives can often be modeled by high dimensional integrals. If the underlying asset is described in terms of a discretized geometric Brownian motion, then the fair price of the option is:

$$\mu = \int_{\mathbb{R}^d} \text{payoff}(\mathbf{z}) \frac{\exp(\frac{1}{2}\mathbf{z}^T \Sigma^{-1} \mathbf{z})}{\sqrt{(2\pi)^d \det(\Sigma)}} d\mathbf{z} = \int_{[0,1]^d} f(\mathbf{x}) d\mathbf{x},$$

where $\text{payoff}(\cdot)$ defines the discounted payoff of the option,

$$\Sigma = (T/d) (\min(j, k))_{j,k=1}^d = \mathbf{L}\mathbf{L}^T,$$

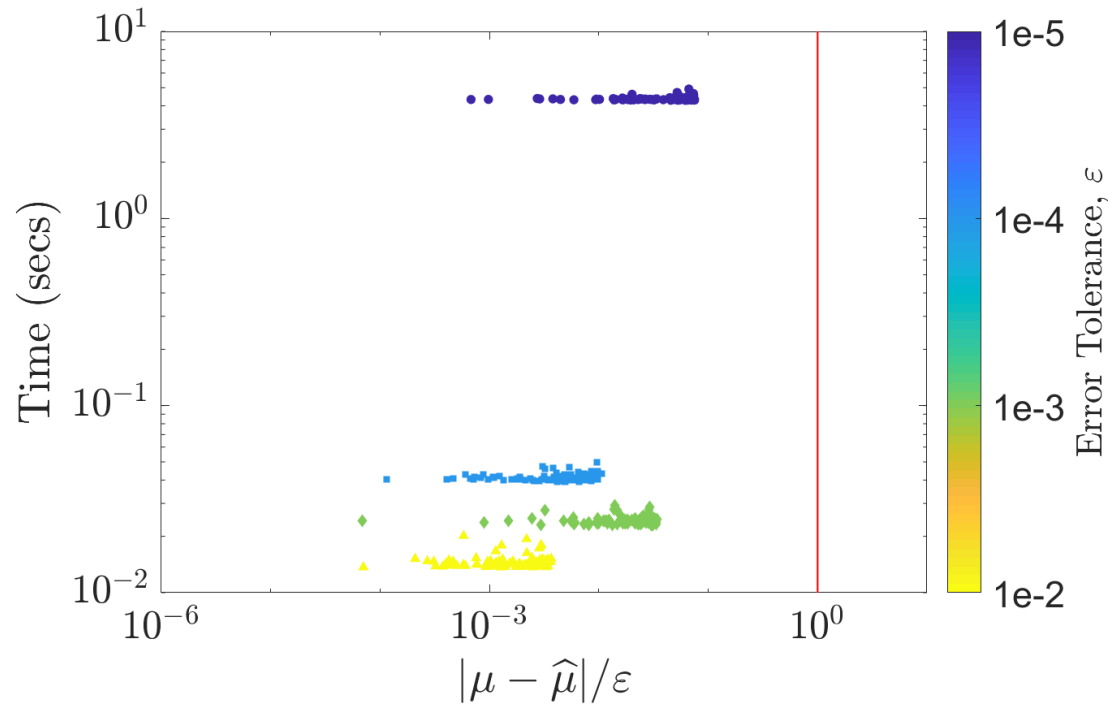


Figure 7.4: Keister example using the empirical Bayes stopping criterion.

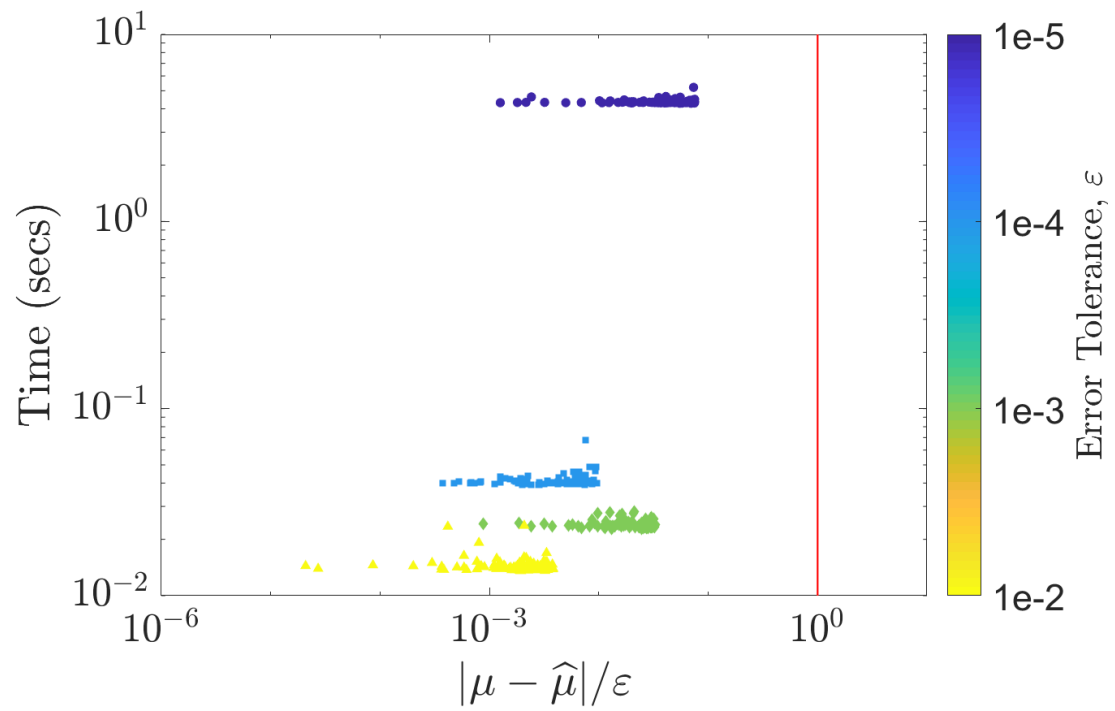


Figure 7.5: Keister example using the full Bayes stopping criterion.

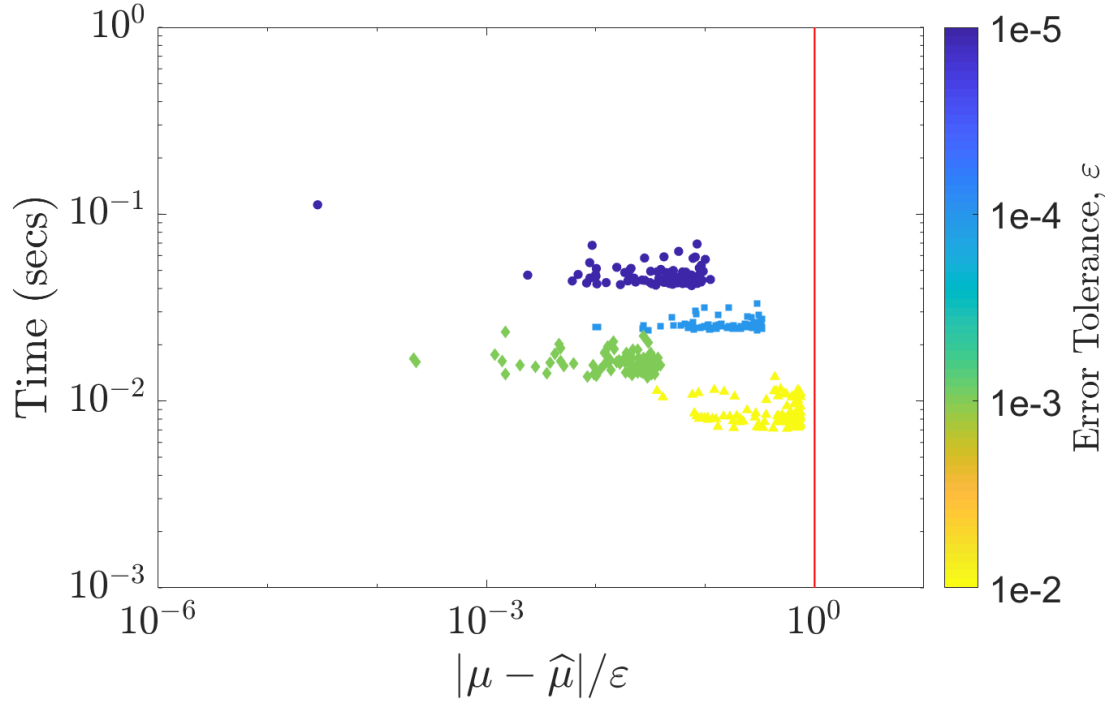


Figure 7.6: Keister example using the GCV stopping criterion.

$$f(\mathbf{x}) = \text{payoff} \left(\mathbf{L} \begin{pmatrix} \Phi^{-1}(x_1) \\ \vdots \\ \Phi^{-1}(x_d) \end{pmatrix} \right).$$

The Asian arithmetic mean call option has a payoff of the form

$$\text{payoff}(\mathbf{z}) = \max \left(\frac{1}{d} \sum_{j=1}^d S_j(\mathbf{z}) - K, 0 \right) e^{-rT},$$

$$S_j(\mathbf{z}) = S_0 \exp((r - \sigma^2/2)jT/d + \sigma\sqrt{T/d}z_j).$$

Here, T denotes the time to maturity of the option, d the number of time steps, S_0 the initial price of the stock, r the interest rate, σ the volatility, and K the strike price.

The Figures 7.7, 7.8 and 7.9 summarize the numerical results for this example using $T = 1/4$, $d = 13$, $S_0 = 100$, $r = 0.05$, $\sigma = 0.5$, $K = 100$. Moreover, \mathbf{L}

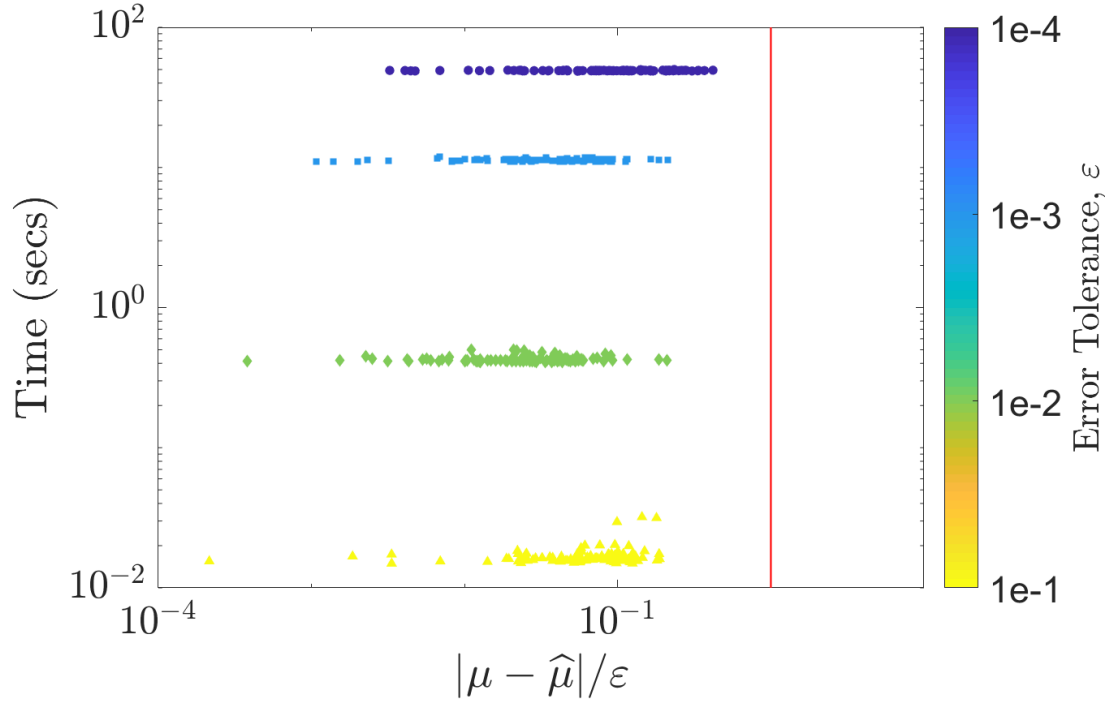


Figure 7.7: Option pricing using the empirical Bayes stopping criterion.

is chosen to be the matrix of eigenvectors of Σ times the square root of the diagonal matrix of eigenvalues of Σ . Because the integrand has a kink caused by the max function, it does not help to use a periodizing transform that is very smooth. We choose the baker's transform (6.1) and $r = 1$.

As shown in Figures 7.1 to 7.9, Bayesian cubature algorithm computes the integral within user specified threshold most of the times except on a few occasions. This is especially the case with option pricing example due to the complexity and high dimension of the integrand. Also notice the algorithm finished within 10seconds of time for Keister and multivariate Normal. For the option pricing it look closer to 70seconds. This is again due to the complexity of the integrand.

Another noticeable aspect from the plots is how much the error bounds differ from the true error. For option pricing example, the error bound is not as conservative as it is for the multivariate normal and Keister examples. A possible reason is that

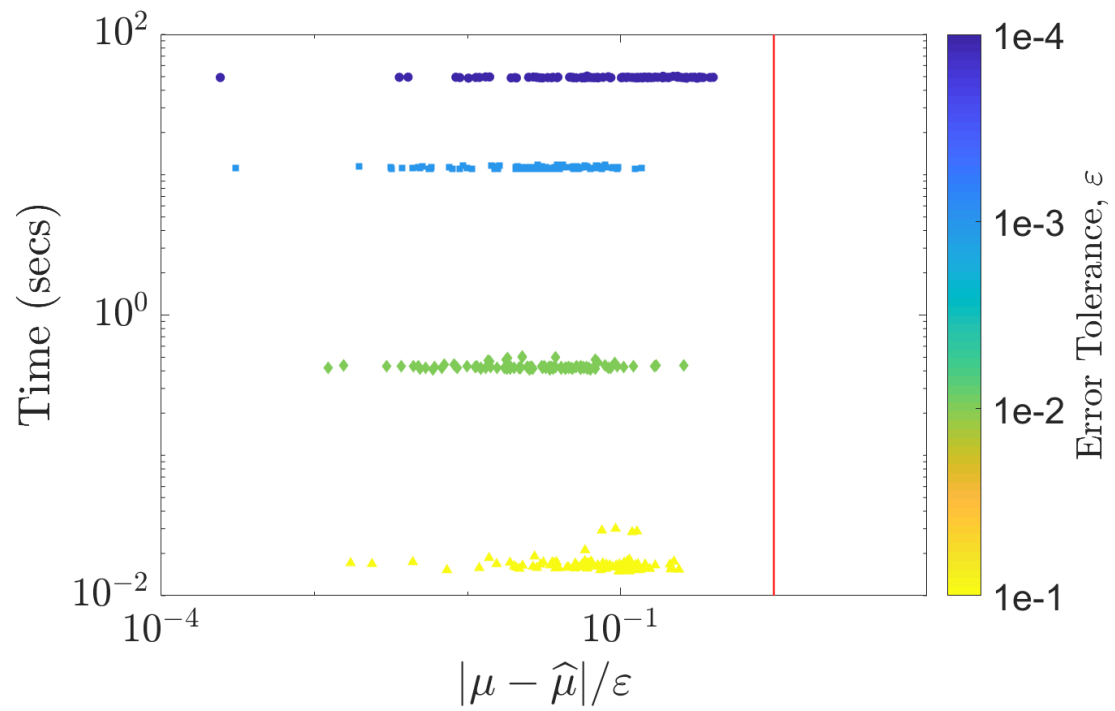


Figure 7.8: Option pricing using the full Bayes stopping criterion.

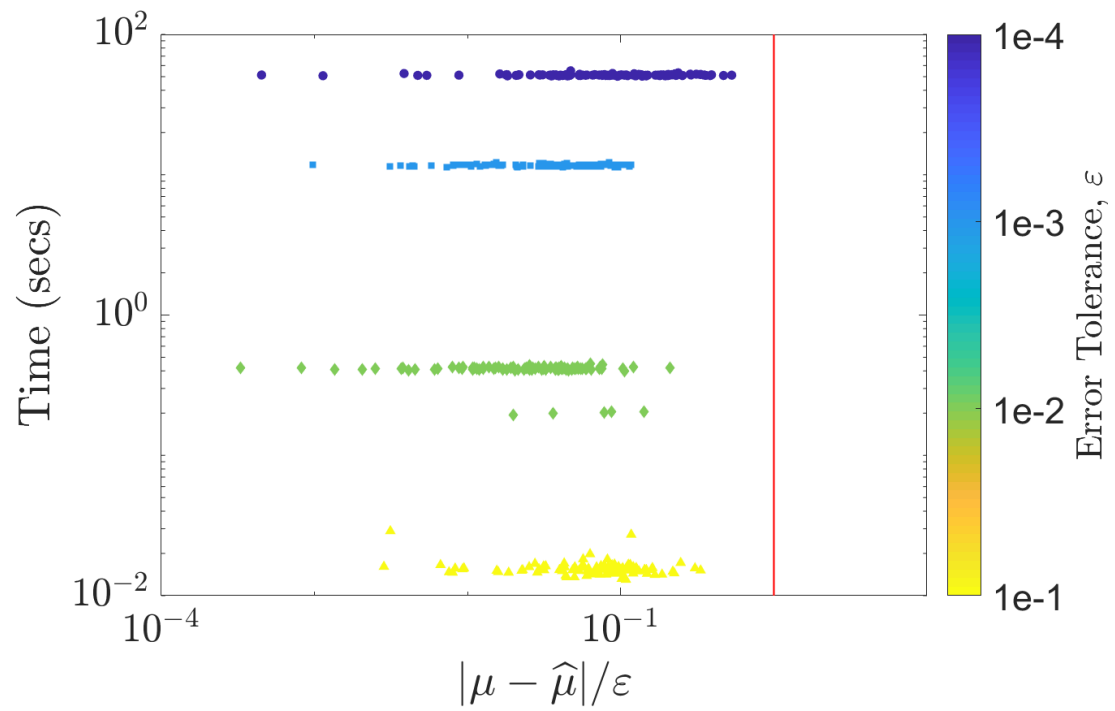


Figure 7.9: Option pricing using the GCV stopping criterion.

the latter integrands are significantly smoother than is assumed by our covariance kernel. This is a matter for further investigation.

7.2 Diagnostics for the Gaussian Process Assumption

The starting point for our Bayesian cubature is the assumption that the integrand arises from a Gaussian process. This means that the function data, \mathbf{f} , satisfy a multivariate Gaussian distribution, as in (2.2). The transformed data, $\mathbf{Z} = \frac{1}{n}\mathbf{V}\Lambda^{-\frac{1}{2}}\mathbf{V}^H(\mathbf{f} - m\mathbf{1})$, has zero mean and is also uncorrelated because

$$\begin{aligned}\mathbb{E}[\mathbf{Z}] &= \frac{1}{n}\mathbf{V}\Lambda^{-\frac{1}{2}}\mathbf{V}^H(\mathbb{E}[\mathbf{f}] - m\mathbf{1}) \\ &= 0\end{aligned}$$

$$\begin{aligned}\text{cov}(\mathbf{Z}) &= \frac{1}{n^2}\mathbb{E}\left[\mathbf{V}\Lambda^{-\frac{1}{2}}\mathbf{V}^H(\mathbf{f} - m\mathbf{1})(\mathbf{f} - m\mathbf{1})^T\mathbf{V}\Lambda^{-\frac{1}{2}}\mathbf{V}^H\right] \\ &= \frac{1}{n^2}\mathbf{V}\Lambda^{-\frac{1}{2}}\mathbf{V}^H\mathbb{E}\left[(\mathbf{f} - m\mathbf{1})(\mathbf{f} - m\mathbf{1})^T\right]\mathbf{V}\Lambda^{-\frac{1}{2}}\mathbf{V}^H \\ &= \frac{1}{n^2}\mathbf{V}\Lambda^{-\frac{1}{2}}\mathbf{V}^H\frac{1}{n}\mathbf{V}\Lambda\mathbf{V}^H\mathbf{V}\Lambda^{-\frac{1}{2}}\mathbf{V}^H \\ &= \frac{1}{n^3}\mathbf{V}\Lambda^{-\frac{1}{2}}(n)\Lambda(n)\Lambda^{-\frac{1}{2}}\mathbf{V}^H = \mathbf{I}\end{aligned}$$

Thus, the elements of \mathbf{Z} are IID standard Gaussian random variables. In practice, using the estimated m_{MLE} further simplifies,

$$\begin{aligned}\mathbf{Z} &= \frac{1}{n}\mathbf{V}\Lambda^{-\frac{1}{2}}\mathbf{V}^H(\mathbf{f} - m\mathbf{1}) \\ &= \frac{1}{n}\mathbf{V}\Lambda^{-\frac{1}{2}}(\mathbf{V}^H\mathbf{f} - \frac{\tilde{f}_1}{n}\mathbf{V}^H\mathbf{1}) \\ &= \frac{1}{n}\mathbf{V}\Lambda^{-\frac{1}{2}}\left(\mathbf{V}^H\mathbf{f} - \tilde{f}_1\begin{pmatrix} 1, 0, \dots, 0 \end{pmatrix}^T\right)\end{aligned}$$

Figure ?? and Figure ?? are normal probability plots of the Z_i using empirical Bayes estimates of m and $\boldsymbol{\theta}$. **more goes here.**

To verify this result, we build a toy function using the known Bernoulli polynomial series,

$$f(x) = \hat{f}_0 + \sqrt{\theta} \sum_{k \neq 0}^N \frac{f_{kR} \cos(2\pi kx) + f_{kI} \sin(2\pi kx)}{k}$$

where $f_0, f_{kR}, f_{kI} \sim \text{IIDN}(0, 1)$ and the series is truncated to length N for computability. It has zero mean and covariance

$$\begin{aligned} \text{cov}(f) &= \mathbb{E}[f(x)f(t)] \\ &= \text{cov}(f_0) + \theta \sum_{k=1}^N \left(\mathbb{E}(f_{kR}^2) \frac{\cos(2\pi kx) \cos(2\pi kt)}{k^2} + \mathbb{E}(f_{kI}^2) \frac{\sin(2\pi kx) \sin(2\pi kt)}{k^2} \right) \\ &= 1 + \theta \sum_{k=1}^N \frac{\cos(2\pi k(x-t))}{k^2} \end{aligned}$$

7.2.1 Posterior

Non-informative prior Assume the prior is of the form $\rho_{m,s^2}(\xi, \lambda) = \frac{1}{\lambda}$

$$\begin{aligned} \rho_{f(x)}(\mathbf{y}) &= \int_0^\infty \int_{-\infty}^\infty \rho_f(\mathbf{y}|m = \xi, s^2 = \lambda) \rho_{m,s^2}(\xi, \lambda) d\xi d\lambda \\ &= \int_0^\infty \int_{-\infty}^\infty \frac{1}{(\sqrt{2\pi})^n \sqrt{\det(\lambda \mathbf{C})}} \exp\left(-\frac{1}{2\lambda}(\mathbf{y} - \xi \mathbf{1})^T \mathbf{C}^{-1}(\mathbf{y} - \xi \mathbf{1})\right) \frac{1}{\lambda} d\xi d\lambda \\ &\propto \frac{1}{(\sqrt{2\pi})^n} \int_0^\infty \frac{1}{\lambda^{\frac{n}{2}+1}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\lambda}(\mathbf{y} - \xi \mathbf{1})^T \mathbf{C}^{-1}(\mathbf{y} - \xi \mathbf{1})\right) d\xi d\lambda \\ &\propto \int_0^\infty \frac{1}{\lambda^{\frac{n}{2}+1}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\lambda} \underbrace{(\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1})}_D \left[\xi^2 - 2\xi \underbrace{\frac{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{y}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}}}_A + \underbrace{\frac{\mathbf{y}^T \mathbf{C}^{-1} \mathbf{y}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}}}_B \right] \right) d\xi d\lambda \\ &\propto \int_0^\infty \frac{1}{\lambda^{\frac{n}{2}+1}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\lambda} D [\xi^2 - 2\xi A + A^2 - A^2 + B] \right) d\xi d\lambda \end{aligned}$$

$$\begin{aligned}
& \propto \int_0^\infty \frac{1}{\lambda^{\frac{n}{2}+1}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\lambda} D [\xi^2 - 2\xi A + A^2]\right) d\xi \exp\left(-\frac{1}{2\lambda} D[B - A^2]\right) d\lambda \\
& \propto \int_0^\infty \frac{1}{\lambda^{\frac{(n+1)}{2}}} \frac{1}{\sqrt{2\pi\lambda/D}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\lambda} D [\xi - A]^2\right) d\xi \exp\left(-\frac{1}{2\lambda} D[B - A^2]\right) d\lambda \\
& \propto \int_0^\infty \frac{1}{\lambda^{\frac{(n+1)}{2}}} \exp\left(-\frac{1}{2\lambda} D[B - A^2]\right) d\lambda \\
& \propto (D[B - A^2])^{-\frac{(n-1)}{2}} \\
& \propto \left(\mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} - \frac{(\mathbf{1}^T \mathbf{C}^{-1} \mathbf{y})^2}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}}\right)^{-\frac{(n-1)}{2}} \\
& = \left(\mathbf{y}^T \left[\mathbf{C}^{-1} - \frac{(\mathbf{C}^{-1} \mathbf{1} \mathbf{1}^T \mathbf{C}^{-1})}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}}\right] \mathbf{y}\right)^{-\frac{(n-1)}{2}} \\
& = \left((\mathbf{y} - m_{\text{MLE}})^T \mathbf{C}^{-1} (\mathbf{y} - m_{\text{MLE}})\right)^{-\frac{(n-1)}{2}}, \quad \text{where } m_{\text{MLE}} = \frac{(\mathbf{1}^T \mathbf{C}^{-1} \mathbf{y})}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}}
\end{aligned}$$

TODO: How to interpret as t-distribution?

Generic prior Assume the prior is of the form $\rho_{m,s^2}(\xi, \lambda) = g(\lambda)$

$$\begin{aligned}
\rho_{f(x)}(\mathbf{y}) &= \int_0^\infty \int_{-\infty}^\infty \rho_f(\mathbf{y} | m = \xi, s^2 = \lambda) \rho_{m,s^2}(\xi, \lambda) d\xi d\lambda \\
&= \int_0^\infty \int_{-\infty}^\infty \frac{1}{(\sqrt{2\pi})^n \sqrt{\det(\lambda \mathbf{C})}} \exp\left(-\frac{1}{2\lambda} (\mathbf{y} - \xi \mathbf{1})^T \mathbf{C}^{-1} (\mathbf{y} - \xi \mathbf{1})\right) g(\lambda) d\xi d\lambda \\
&\propto \frac{1}{(\sqrt{2\pi})^n} \int_0^\infty \frac{g(\lambda)}{\lambda^{\frac{n}{2}}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\lambda} (\mathbf{y} - \xi \mathbf{1})^T \mathbf{C}^{-1} (\mathbf{y} - \xi \mathbf{1})\right) d\xi d\lambda \\
&\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{\frac{n}{2}}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\lambda} \underbrace{(\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1})}_D \left[\xi^2 - 2\xi \underbrace{\frac{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{y}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}}}_A + \underbrace{\frac{\mathbf{y}^T \mathbf{C}^{-1} \mathbf{y}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}}}_B\right]\right) d\xi d\lambda \\
&\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{\frac{n}{2}}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\lambda} D [\xi^2 - 2\xi A + A^2 - A^2 + B]\right) d\xi d\lambda \\
&\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{\frac{n}{2}}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\lambda} D [\xi^2 - 2\xi A + A^2]\right) d\xi \exp\left(-\frac{1}{2\lambda} D[B - A^2]\right) d\lambda \\
&\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{\frac{(n-1)}{2}}} \frac{1}{\sqrt{2\pi\lambda/D}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\lambda} D [\xi - A]^2\right) d\xi \exp\left(-\frac{1}{2\lambda} D[B - A^2]\right) d\lambda
\end{aligned}$$

$$\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{\frac{(n-1)}{2}}} \exp\left(-\frac{1}{2\lambda}D[B - A^2]\right) d\lambda$$

This can be interpreted as Laplace transform of $g(\lambda)$. Let $\eta = \frac{1}{2}D[B - A^2]$ and $\lambda = \frac{1}{w}, d\lambda = -w^{-2}dw$

$$\begin{aligned} \rho_{f(x)}(\mathbf{y}) &\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{\frac{(n-1)}{2}}} \exp\left(-\frac{1}{2\lambda}D[B - A^2]\right) d\lambda \\ &= \int_0^\infty \frac{g(1/w)}{w^{-\frac{(n-1)}{2}}} \exp(-w\eta) (-w^{-2}) dw \\ &= \int_\infty^0 -g(1/w) w^{\frac{(n-1)}{2}-2} \exp(-w\eta) dw \\ &= \int_0^\infty g(1/w) w^{\frac{(n-5)}{2}} \exp(-w\eta) dw \\ &= \mathcal{LT}\{g(1/\eta)\}^{(\frac{n-5}{2})'} \end{aligned}$$

where $\mathcal{LT}(\cdot)$ denotes the Laplace transform and $(\frac{n-5}{2})'$ indicates the $(\frac{n-5}{2})$ th derivative taken after the transform. Here we used frequency domain derivative property of the Laplace transform. $\rho_{f(x)}(\mathbf{y})$ is proportional to $(\frac{n-5}{2})$ th derivative of the Laplace transform of $g(1/\eta)$.

$$\begin{aligned} D[B - A^2] &= \left(\mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} - \frac{(\mathbf{1}^T \mathbf{C}^{-1} \mathbf{y})^2}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \right) \\ &= \left((\mathbf{y} - m_{\text{MLE}})^T \mathbf{C}^{-1} (\mathbf{y} - m_{\text{MLE}}) \right), \quad \text{where} \quad m_{\text{MLE}} = \frac{(\mathbf{1}^T \mathbf{C}^{-1} \mathbf{y})}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \end{aligned}$$

TODO: How to integrate further?

CHAPTER 8

DISCUSSION AND FURTHER WORK

We have developed a fast, automatic Bayesian cubature that estimates the integral within a user defined error tolerance. The stopping criteria arise from assuming the integrand to be a Gaussian process. There are three versions: empirical Bayes, full Bayes, and generalized cross-validation. The computational cost of the automatic Bayesian cubature can be dramatically reduced if the covariance kernel matches the nodes. We have demonstrated one such match in practice: rank-1 lattice nodes and shift-invariant kernels. The matrix-vector multiplications can be accomplished using the fast Fourier Transform. Three integration problems illustrate the performance of our automatic Bayesian cubature.

Digital sequences and digital shift and/or scramble invariant kernels have the potential of being another match that satisfies the conditions in Section 3. The fast transform would correspond to a fast Walsh transform. For such kernels, periodicity is not assumed, however, special structure of both the sequences and the kernels are required to take advantage of integrand smoothness.

One should be able to adapt our Bayesian cubature to control variates, i.e., assuming

$$f = \mathcal{GP}(\beta_0 + \beta_1 g_1 + \cdots + \beta_p g_p, s^2 C),$$

for some choice of g_1, \dots, g_p whose integrals are known, and some parameters β_0, \dots, β_p in addition to the s and C . The efficacy of this approach has not yet been explored.

CHAPTER 9

FUTURE WORK

9.1 Future work

- Optimal kernel order selection for the integrand

Acknowledgments

This research was supported in part by the National Science Foundation grants DMS-1522687 and DMS-1638521 (SAMSI). The authors would like to thank the organizers of the SAMSI-Lloyds-Turing Workshop on Probabilistic Numerical Methods, where a preliminary version of this work was discussed. The authors also thank Chris Oates and Sou-Cheng Choi for valuable comments.

BIBLIOGRAPHY

- [1] J. F. Baldeaux. *Higher order nets and sequences*. PhD thesis, The School of Mathematics and Statistics at The University of New South Wales, June 2010.
- [2] F-X Briol, C. J. Oates, M. Griolami, M. A. Osborne, and D. Sejdinovic. Probabilistic integration: A role in statistical computation? *Statist. Sci.*, 2018+. to appear.
- [3] S.-C. T. Choi, Y. Ding, F. J. Hickernell, L. Jiang, Ll. A. Jiménez Rugama, D. Li, R. Jagadeeswaran, X. Tong, K. Zhang, Y. Zhang, and X. Zhou. GAIL: Guaranteed Automatic Integration Library (versions 1.0–2.2). MATLAB software, 2013–2017.
- [4] R. Cools and D. Nuyens, editors. *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, Belgium, April 2014*, volume 163 of *Springer Proceedings in Mathematics and Statistics*. Springer-Verlag, Berlin, 2016.
- [5] P. Craven and G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31:307–403, 1979.
- [6] P. Diaconis. Bayesian numerical analysis. In S. S. Gupta and J. O. Berger, editors, *Statistical Decision Theory and Related Topics IV, Papers from the 4th Purdue Symp., West Lafayette, Indiana 1986*, volume 1, pages 163–175. Springer-Verlag, New York, 1988.
- [7] J. Dick. Walsh spaces containing smooth functions an quasi-monte carlo rules of arbitrary high order. *SIAM J. Numer. Anal.*, 46(1519–1553), 2008.
- [8] J. Dick, F. Kuo, and I. H. Sloan. High dimensional integration — the Quasi-Monte Carlo way. *Acta Numer.*, 22:133–288, 2013.
- [9] A. Genz. Comparison of methods for the computation of multivariate normal probabilities. *Computing Science and Statistics*, 25:400–405, 1993.
- [10] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979.
- [11] F. J. Hickernell. Quadrature error bounds with applications to lattice rules. *SIAM J. Numer. Anal.*, 33:1995–2016, 1996. corrected printing of Sections 3-6 in *ibid.*, **34** (1997), 853–866.
- [12] F. J. Hickernell. The trio identity for quasi-Monte Carlo error analysis. In P. Glynn and A. Owen, editors, *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Stanford, USA, August 2016*, Springer Proceedings in Mathematics and Statistics, pages 13–37. Springer-Verlag, Berlin, 2018. arXiv:1702.01487.
- [13] F. J. Hickernell and Ll. A. Jiménez Rugama. Reliable adaptive cubature using digital sequences. In Cools and Nuyens [4], pages 367–383. arXiv:1410.8615 [math.NA].
- [14] F. J. Hickernell, Ll. A. Jiménez Rugama, and D. Li. Adaptive quasi-Monte Carlo methods for cubature. In J. Dick, F. Y. Kuo, and H. Woźniakowski, editors, *Contemporary Computational Mathematics — a celebration of the 80th birthday of Ian Sloan*, pages 597–619. Springer-Verlag, 2018.

- [15] F. J. Hickernell and H. Niederreiter. The existence of good extensible rank-1 lattices. *J. Complexity*, 19:286–300, 2003.
- [16] G. Leobacher, D. Nuyens, F. Pillichshammer, J. Baldeaux, J. Dick. Efficient calculation of the worst-case error and (fast) component-by-component construction of higher order polynomial lattice rules. *Numerical Algorithms*, 59:403–431, March 2012.
- [17] Ll. A. Jiménez Rugama and F. J. Hickernell. Adaptive multidimensional integration based on rank-1 lattices. In Cools and Nuyens [4], pages 407–422. arXiv:1411.1966.
- [18] H. Nickisch, D. Bindel, K. Dong, D. Eriksson and A. G. Wilson. Scalable log determinants for gaussian process kernel learning. *NIPS*, 2017. in press.
- [19] B. D. Keister. Multidimensional quadrature algorithms. *Computers in Physics*, 10:119–122, 1996.
- [20] H. Niederreiter. Constructions of (t, m, s) -nets and (t, s) -sequences. *Finite Fields Appl.*, 11:578–600, 2005.
- [21] D. Nuyens.
- [22] D. Nuyens.
- [23] A. O’Hagan. Bayes-Hermite quadrature. *J. Statist. Plann. Inference*, 29:245–260, 1991.
- [24] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, C. W. Clark, and A. B. O. Dalhousie. Digital library of mathematical functions, 2018.
- [25] C. E. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. In S. Thrun, L. K. Saul, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, pages 489–496. MIT Press, 2003.
- [26] C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts, 2006. (online version at <http://www.gaussianprocess.org/gpml/>).
- [27] K. Ritter. *Average-Case Analysis of Numerical Problems*, volume 1733 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000.
- [28] A. Sidi. Further extension of a class of periodizing variable transformations for numerical integration. *J. Comput. Appl. Math.*, 221:132–149, 2008.
- [29] G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1990.