Bachelor's Programme in Science and Technology

Central Composite Design Integration for Bayesian Inference

Georgy Ananov

Aalto University

BACHELOR'S THESIS

Aalto University
BACHELOR'S THESIS 2023

Central Composite Design Integration for Bayesian Inference

Georgy Ananov

Thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Technology.
Otaniemi, 28 April 2023

Supervisor: professor Maarit Korpi-Lagg
Advisor: professor Aki Vehtari

Aalto University
School of Science
Bachelor's Programme in Science and Technology

Author
Georgy Ananov

Title
Central Composite Design Integration for Bayesian Inference

School School of Science

Degree programme Bachelor's Programme in Science and Technology

Major Data Science **Code** SCI3095

Supervisor professor Maarit Korpi-Lagg

Advisor professor Aki Vehtari

Level Bachelor's thesis **Date** 28.04.2023 **Pages** 23+5 **Language** English

Abstract

As Bayesian statistics continues to gain popularity in applied fields, the development of computationally efficient numerical integration techniques is becoming increasingly important as a way to support accurate inference on complex models. In this thesis we investigate a deterministic numerical integration strategy employing the principles of central composite design as an alternative to popular stochastic methods, such as Markov chain Monte Carlo. We begin with a brief introduction to the Bayesian methodology and an outline of the role of numerical integration in Bayesian inference. We then introduce the central composite design (CCD) integration and explain how it addresses the shortcomings of existing methods, such as the loss of uncertainty associated with maximum a posteriori estimate and the high computational cost of Markov chain Monte Carlo. We then provide a detailed breakdown of the CCD integration algorithm, accompanied by a visual example based on a bioassay experiment. Subsequently, we demonstrate our implementation of the algorithm with a case study in motorcycle crash dynamics, showing how CCD integration can enable faster inference compared to Markov chain Monte Carlo. Our findings suggest that CCD integration has the potential to make Bayesian methodology more practical and accessible in scenarios where the computational cost of traditional stochastic methods is prohibitive.

Keywords Bayesian statistics, central composite design, numerical integration

urn https://aaltodoc.aalto.fi

Preface

I would like to thank professor Aki Vehtari for the wonderful opportunity to work on an exciting topic. I would also like to extend my gratitude to Nikolas Siccha, whose guidance and support were instrumental in the completion of this thesis.

Contents

Symbols and abbreviations

Symbols

$\mathcal{D}$	Observed data
$\mathcal{M}$	Model assumptions
$\mathcal{N}(\cdot, \cdot)$	Gaussian distribution with the specified location and scale
$p(\cdot)$	Probability density function
$p(\cdot, \cdot)$	Joint probability density function
$p(\cdot \vert \cdot)$	Conditional probability density function
θ	Random vector of model parameters
$\theta^{(i)}$	Individual model parameter
θ_i	A specific evaluation of model parameters
Δ_i	Integration weights for numerical integration
z	Standard Gaussian variable
l_p	Length scale parameter for the GP covariance function
σ_f^2	Magnitude parameter for the GP covariance function
σ_n	Normalized residual parameter

Abbreviations

CCD	central composite design
GP	Gaussian process
MAP	maximum a posteriori [estimate]
MCMC	Markov chain Monte Carlo

1. Introduction

The field of modern statistics widely recognizes two main approaches to generating inferences from data — the frequentist approach and the Bayesian approach. The frequentist school of thought relies on the assumption that the observed outcomes are generated by one 'true' model. It seeks to answer questions about that underlying model by examining the experimental data and uncovering a model that would fit those observations well. In contrast, the fundamental concept involved in Bayesian thinking is that the observed data can be explained by many different models, with some being more likely than others to serve as an explanation. The outcome of this modeling process can then be expressed as a range of model hypotheses, with each being assigned a relative probability value based on prior assumptions and the observed data. It then follows that the parameters of the model are considered to be a fixed unknown value under the frequentist approach, while the Bayesian approach treats the parameters as random variables.

In recent decades, the Bayesian methodology has gained in popularity in many applied fields, including psychology [1] and medicine [2]. The flexibility extended by the inclusion of prior assumptions into posterior inference makes the Bayesian approach an attractive choice in many data analysis tasks. However, in practice, the deployment of Bayesian methods poses some computational challenges. In particular, since model parameters are viewed as random variables, the marginalization over the parameters becomes an integral step of the Bayesian inference. The marginalization process often produces analytically intractable integrals, which call for the use of numerical integration methods.

A numerical integration technique that is often utilized for numerical integration in Bayesian inference is the Markov chain Monte Carlo (MCMC) method, which includes drawing a large number of points from the pos-

terior distribution to be used for numerical integration. While MCMC simulation circumvents the need to solve the integral analytically, the simulation can still be computationally unfeasible, particularly when the number of model parameters is large, or when the evaluation of the likelihood function is especially costly. In such cases, methods that rely on analytical approximations rather than full numerical integration may be appropriate [3]. A computationally straightforward approach to this issue would be to approximate the integral over the model parameters utilizing a single point estimate of said parameters. However, this strategy would exclude the uncertainty connected with model parameters from the inference and thus go against the central notion of the Bayesian methodology. We must then explore methods that preserve the uncertainty of the parameters, while still offering some analytical approximation approach to the parameter marginalization. One such method, which borrows techniques from the central composite design (CCD) methodology, is explored in this work.

Therefore, the goal of this thesis is to develop a software implementation of the CCD-based approximate integration of the model parameters in order to evaluate the viability of the technique by applying the approximation to an existing experimental study. Of particular interest is the impact on integration accuracy and computation times compared to both the single point estimate approach and the full MCMC integration approach.

The remainder of this thesis is divided into four sections. Section 2 provides the background information on Bayesian inference and central composite design methodology, as well as details the context in which CCD-based approximate integration may be used. Section 3 covers the implementation details of the approximate integration algorithm and presents an illustrative example. A concrete case study is presented in Section 4. Finally, Section 5 summarizes the work discussed so far and suggests possible avenues for future research.

2. Background

This section introduces the terms and concepts involved in the discussion of the CCD-based approximate integration algorithm. The discussion below offers a brief overview of the Bayesian inference process and a summary of the central composite design integration method.

2.1 Introduction to Bayesian inference

Oftentimes, scientists are faced with the task of obtaining insight into the nature of a random event based on observations of past outcomes and possibly some pre-existing knowledge or understanding related to the event. The process of drawing conclusions about the underlying probability distribution behind the event based on observational evidence is commonly referred to as *statistical inference*. The first step in such studies is typically the selection of the model type that best suits the examined event. The choice of model is usually informed by analytical reasoning. After the model type is selected, the specific parameters of the model are inferred from the data.

In Bayesian inference, the uncertainty in the model parameters is represented by a probability distribution over all possible parameter combinations. In other words, each potential hypothesis about the nature of the event is assigned a relative probability value, which can inform comparisons between the sensibility of different hypotheses based on the available data. The key advantages of the Bayesian methodology for statistical inference include the ability to reason about probabilities of different hypotheses for a wide range of events as well as the flexibility that comes with the incorporation of *a priori* beliefs into the posterior inference [4].

2.2 Bayes' Theorem

At the heart of Bayesian inference lies Bayes' theorem, also known as Bayes' rule. The theorem describes the primary mechanism by which prior information about the model parameters and the available observations are combined to form the posterior inference:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}, \tag{2.1}$$

where θ denotes model parameters, $\mathcal{D}$ represents the observed data, $p(\theta|\mathcal{D})$ describes the posterior distribution of model parameters, $p(\mathcal{D}|\theta)$ is the likelihood of observing the data $\mathcal{D}$ under the model with parameters θ, and, finally, $p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta$ represents the marginal probability of observing the data $\mathcal{D}$ and acts as a normalizing constant.

It is worth pointing out that all probability distributions described above are also conditioned on the model assumptions $\mathcal{M}$ selected as the first step in the inference process. Thus, a notation $p(\theta|\mathcal{D}, \mathcal{M})$ can be utilized to highlight the fact that the posterior distribution is inherently conditioned on the choice of model assumptions. However, since the model type typically remains unchanged throughout the inference process, $\mathcal{M}$ is often considered to be implied and is omitted from formulas and equations.

2.3 Marginalization over model parameters

The Bayesian methodology calls for factoring the uncertainty of model parameters into any inferences about the properties of the posterior distribution. Producing the inference then requires marginalization over the model parameters, i.e., integration of the inference function over the entirety of the parameter space. For example, formulating the posterior predictive distribution $p(y_*|\mathcal{D})$ of the next observation y_* requires marginalizing out the model parameters θ:

$$p(y_*|\mathcal{D}) = \int p(y_*, \theta|\mathcal{D})d\theta = \int p(y_*|\theta)p(\theta|\mathcal{D})d\theta, \tag{2.2}$$

where the probability density function $p(y_*|\theta)$ is defined by the model assumptions.

The integral that is produced by the marginalization process is often not analytically tractable, thus necessitating the application of approximate integration methods, which seek to estimate the outcome of integration with-

out requiring traditional full integration. This work focuses on discussing numerical integration techniques rather than analytical approaches to approximate integration.

2.4 Numerical integration methods

When an integration task $\int f(x)dx$ cannot be efficiently solved through analytically obtaining the antiderivative of the integrand $f(x)$, we can instead approximate the integral with a linear combination of the values of f taken at a finite number of locations [5].

$$\int f(x)dx \approx \Delta_1 f(x_1) + \Delta_2 f(x_2) + \cdots + \Delta_n f(x_n), \tag{2.3}$$

where $x_1, x_2, \ldots, x_n$ represent the locations or points selected for numerical integration, and $\Delta_1, \Delta_2, \ldots, \Delta_n$ are weights that correspond to those points.

The approximation can be utilized when conducting Bayesian inference. Continuing with the example of posterior predictive distribution (2.2), we can apply numerical integration to obtain an approximation with a finite sum:

$$p(y_*|\mathcal{D}) = \int p(y_*|\theta)p(\theta|\mathcal{D})d\theta \approx \sum_{i=1}^{n} p(y_*|\theta_i)p(\theta_i|\mathcal{D})\Delta_i, \tag{2.4}$$

where θ_i are points in the parameter space and Δ_i are the corresponding integration weights. The task of performing the approximation can then be reduced to choosing the appropriate integration points and computing the weights.

The accuracy and the computational complexity of the approximation greatly depend on the choice of integration points. On one hand, selecting an excessive number of points would demand the evaluation of the posterior $p(\theta_i|\mathcal{D})$ in many different locations, which can be computationally expensive. On the other hand, selecting too few points or neglecting to include points from regions of high density of the posterior can lead to increased approximation error. We thus seek such a set of integration points that covers the majority of the posterior mass, while also keeping the point count reasonable.

Single point estimate

An approximation for an integral over the model parameters θ can be obtained utilizing only a single point estimate in a technique referred to as empirical Bayes:

$$p(y_*|\mathcal{D}) = \int p(y_*|\theta)p(\theta|\mathcal{D})d\theta \approx p(y_*|\hat{\theta}), \tag{2.5}$$

where $\hat{\theta}$ is a maximum *a posteriori* (MAP) probability estimate for the location of the random vector of model parameters θ.

$$\hat{\theta} = \arg\max_{\theta} \ p(\theta|\mathcal{D}) = \arg\max_{\theta} \ \log p(\mathcal{D}|\theta) + \log p(\theta) \tag{2.6}$$

While the MAP estimation is capable of providing a meaningful approximate integration scheme with a minimal computational load, it fails to incorporate the uncertainty over the model parameters into the final inference. As such, the technique is at odds with the fundamental principles of the Bayesian approach. In cases where the uncertainty in the inference is dominated by the uncertainty in model parameters, the point estimate approximation can result in higher approximation error due to poor exploration of the parameter space and is thus not suitable.

Markov chain Monte-Carlo sampling

An approach widely used for approximate integration in Bayesian statistics is Markov chain Monte-Carlo (MCMC) integration, which includes stochastically drawing samples from the posterior distribution by forming a Markov chain of samples. The inference task can then be performed at each sampled location, similar to the single point estimate technique, and the results can be averaged to produce the final posterior inference.

While MCMC is a powerful and versatile family of methods, utilizing these techniques in practice is not always computationally feasible. In cases where the evaluation of the posterior probability density function is costly, such as when working with models involving Gaussian processes [6], the computational times can become prohibitively long.

The remainder of this thesis is dedicated to the exploration of an alternative numerical integration method, which aims to produce a reasonably accurate approximation of the posterior while limiting the number of costly posterior density evaluations.

Central composite design

The central composite design methodology offers a deterministic procedure for selecting points from the parameter space in a manner that allows for effective exploration of the posterior density. The proposed points can serve as effective candidates for use in numerical integration, since their selection is designed to capture the shape of the posterior distribution well.

A central composite design contains two sets of points, denoted as *design* points and *star* points. The star points lie on the principal axes of the posterior distribution, with two points selected on each axis. The number of star points then depends on the dimensionality of the posterior and is equal to twice the number of model parameters. A single point at the mode is also typically added to the star points. The design points are selected at locations in-between the principle axes and constitute a subset of a full factorial design, i.e., a fractional factorial design. The number of points in a full factorial design scales exponentially with the number of model parameters, but not all of them are needed to explore the shape of the posterior effectively. The exact numbers of design points corresponding to some of the possible dimensionalities of the model parameters are presented in Table 2.1.

Dimensions of θ	2	3	5	6	8	11	17
CCD design points	4	8	16	32	64	128	256
Full factorial design	4	8	32	64	256	2048	131072

Table 2.1. Number of design points in CCD fractional factorial design.

For a detailed overview of the design point selection algorithm, refer to [7].

3. Implementation

This section introduces the central composite design integration and provides an illustration with a visual example.

3.1 Breakdown of the algorithm steps

Below are the details of the CCD integration procedure, as implemented for this work based on publications by Rue et al. [3] and Vanhatalo et al. [6]:

(a) *Step 1:* Apply a log transformation to positive-restricted parameters.

In certain applications, some of the model parameters $\theta^{(i)}$ might be restricted to be positive by design. To improve the accuracy of the approximations described in further steps, it would be sensible to transform θ such that each component $\theta^{(i)}$ spans the entirety of $\mathbb{R}$. This effect can be achieved by applying a logarithmic transformation to those components of the model parameters that are bound to be positive:

$$\gamma_i = \begin{cases} \log \theta^{(i)} & \text{if } \theta^{(i)} \text{ is bound to be positive} \\ \theta^{(i)} & \text{otherwise.} \end{cases} \tag{3.1}$$

After the transformation, the newly acquired log-density function $p(\gamma|\mathcal{D})$ is often closer to a Gaussian distribution than the untransformed posterior [6]. Since the log-density function is approximated by a Gaussian during the subsequent steps, applying the transformation typically reduces the approximation error. When operating with non-linear reparametrizations, it is important to apply the Jacobian adjustment.

(b) *Step 2:* Locate the mode $\hat{\gamma}$ of the posterior log-density.

The mode $\hat{\gamma}$ of the distribution in question can be obtained by maximizing the log-posterior:

$$\log p(\gamma|\mathcal{D}) \propto \log p(\mathcal{D}|\theta) + \log p(\theta) \tag{3.2a}$$

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmin}}(\log p(\mathcal{D}|\theta) + \log p(\theta)) \tag{3.2b}$$

The modal parameter configuration is obtained using iterative techniques, such as the Newton method [3], which utilizes the gradient information of the distribution to locate its mode.

(c) *Step 3:* Approximate the posterior.

Once the mode is located, we approximate the posterior log-density with a Gaussian distribution $q(\gamma|\mathcal{D})$ in a technique known as Gaussian or Laplace approximation:

$$p(\gamma|\mathcal{D}) \sim q(\gamma|\mathcal{D}) = \mathcal{N}(\hat{\gamma}, H^{-1}), \tag{3.3}$$

where $H = -\nabla\nabla \log p(\gamma|\mathcal{D})$ is the negative Hessian matrix (i.e., a matrix consisting of the second derivatives of the log-density of the distribution) computed at the mode $\hat{\gamma}$ using finite differences. Here the inverse negative Hessian is serving as an approximation of the covariance matrix for the log-posterior around its mode.

(d) *Step 4:* Reparametrize the approximate posterior to a standard Gaussian distribution.

For ease of exploration, we can reparametrize the approximate posterior with the standard Gaussian variable z.

$$\gamma(z) = \hat{\gamma} + V\Lambda^{\frac{1}{2}}z, \tag{3.4}$$

where $H^{-1} = V\Lambda V^T$ is the eigenvalue decomposition of the inverse negative Hessian. A linear transformation that transforms a random vector into a new set of uncorrelated random variables, such as one described in Equation 3.4, is commonly referred to as a *whitening transformation*.

(e) *Step 5:* Utilize CCD to obtain the integration points from the parameter space.

We can now proceed to select the integration points. For each axis of the z-space, we select two axial points around the mode at a distance of $\pm f_0\sqrt{d}$ from the mode, where $f_0 > 1$ is an arbitrary scaling parameter set to 1.1 for this work and d is the number of model parameters. To the selected $2d$ star points, we add the modal point, as well as the design points informed by the CCD method.

(f) *Step 6:* Shift the points along principal directions.

In the previous step, we obtained a set of integration points that lie on an equidensity contour of the approximate posterior density $q(\gamma|\mathcal{D})$. While these points would be ideal for integration over the approximate posterior, they might not be as fitting for integrating over the original posterior $p(\gamma|\mathcal{D})$. To adjust for the difference between $q(\gamma|\mathcal{D})$ and $p(\gamma|\mathcal{D})$, we can employ a technique that involves computing a scaling coefficient for each of the principal directions of z [8].

In standard Gaussian density, when we step away from the mode by a distance of $\sqrt{2}$ standard deviations, the log-density reduces by by exactly 1 compared to the modal log-density. The scaling coefficients can then be computed in a such a way that this property approximately holds for each principal direction of the approximation. To determine the coefficient t_i for a particular direction, the original posterior log-density is probed at a point corresponding to a location in z-space that is positioned $\sqrt{2}$ away from the mode in the given direction. The coefficient is then calculated based on the difference between the modal log-density $\log p(\hat{\gamma}|\mathcal{D})$ and the probed log-density $\log p(\gamma_{probe}|\mathcal{D})$ as follows:

$$t_i = \sqrt{\frac{1}{\log p(\hat{\gamma}|\mathcal{D}) - \log p(\gamma_{probe}|\mathcal{D})}} \tag{3.5}$$

Once the coefficient is computed, all integration points that lie in the half-space corresponding to the given direction are shifted along that direction according to the value of coefficient t. If t_i is greater than 1, the points are pushed away from the mode, and if t_i is less than 1, points are pulled towards the mode. The process is repeated for the remaining $2d - 1$ principal directions of z.

The shifting technique lends the approximation greater flexibility without dramatically increasing computational complexity. The base

Gaussian approximation always yields points that are positioned symmetrically around the posterior mode, which increases the approximation error for skewed posterior distributions. With the added shifting step, we can better accommodate non-symmetrical posteriors.

(g) *Step 7:* Compute the integration weights.

Since all of the CCD points besides the mode are located on a d-dimensional sphere within z-space, the integration weights for those points are equal. After reparametrization, we assumed $q(\gamma|\mathcal{D})$ to be a standard d-variate Gaussian distribution; thus, $E[\gamma^T\gamma] = d$ and $\int q(\gamma)d\gamma = 1$. The formulas for the integration weights can then be expressed as follows:

$$\Delta = \left[(n_p - 1)(f_0^2 - 1)\left(1 + \exp(-\frac{df_0^2}{2})\right)\right]^{-1} \tag{3.6a}$$

$$\Delta_0 = 1, \tag{3.6b}$$

where Δ represents the integration weight for the design and axial points, while Δ_0 is the integration weight for the modal point. Once the integration points and weights are computed, the procedure is complete, and the results can be used in numeric integration.

3.2 Illustrative example — bioassay study

To obtain a better understanding of the mechanism of CCD integration, we can turn to a visual illustration of the algorithm's output when applied to a bivariate posterior distribution, as demonstrated on a *bioassay* toxicity study.

Experiment model

The primary goal of a typical bioassay experiment is to determine the potency of a particular compound when administered to a living organism or living cells at different doses. One way that the experimental data is commonly collected by conducting animal tests. Each of the k experimental observations can be characterized by three variables

$$(x_i, n_i, y_i); i = 1..k, \tag{3.7}$$

where x_i indicates the dosage level of the compound administered to a particular group of subjects and is typically measured on a logarithmic scale, n_i is the number of subjects in said group, and y_i is the number of subjects from the group that have responded to the treatment. In the context of toxicity testing, a response to treatment constitutes the manifestation of adverse effects in the experiment subject. The exact effects that are considered adverse differ from one study to another, and often include development of tumors or death.

Assuming that the treatment outcomes within each group are independent, it is reasonable to model the number of affected subjects y_i with a Binomial distribution:

$$y_i|p_i \sim Bin(n_i, p_i), \tag{3.8}$$

where p_i is the probability of developing adverse symptoms when administered the compound at dose x_i. A linear dose-response relationship between p_i and x_i can be modelled with the *logit* function:

$$\text{logit}(p_i) = \alpha + \beta x_i, \tag{3.9}$$

where $\text{logit}(p_i) = \log(p_i/(1-p_i))$. With the *logit* transformation applied and the linear dose-response function established, the model is now parameterized by variables α and β. The likelihood function can then be constructed based on the properties of the binomial distribution:

$$p(y_i|\alpha, \beta, n_i, x_i) \propto [\text{logit}^{-1}(\alpha + \beta x_i)]^{y_i} * (1 - [\text{logit}^{-1}(\alpha + \beta x_i)]^{n_i - y_i} \tag{3.10}$$

Finally, the joint posterior distribution of α and β can be expressed as

$$p(\alpha, \beta|x, y, n) \propto p(\alpha, \beta) \prod_{i=1}^{k} p(y_i|\alpha, \beta, n_i, x_i) \tag{3.11}$$

A metric commonly used to quantify the toxicity level of the compound is the LD50 parameter, which defines such a dose that the probability of developing adverse effects following treatment is equal to 0.5. The aim of the study is then to conduct inference about the LD50 dosage based on the observed data. With the transformations described above, the LD50 parameter can be described in terms of α and β as follows:

$$\begin{aligned} \text{LD50:} \quad E(\frac{y_i}{n_i}) &= \text{logit}^{-1}(\alpha + \beta x_i) = 0.5; \\ \alpha + \beta x_i &= \text{logit}(0.5) = 0 \\ x_i &= -\frac{\alpha}{\beta} \end{aligned} \tag{3.12}$$

For a more comprehensive overview of the bioassay experiment model, the motivations behind the model choice, and the derivations of the formulae, refer to [9].

Data

The data for the presented illustrative example is borrowed from a 1986 bioassay study documented in [10]. The experiment includes testing the toxicity of the compound at four dose levels, with five animals exposed to treatment and subsequently observed for symptoms at each level. The resulting observations are summarized in Table 3.1. For simplicity, a uniform prior probability distribution $p(\alpha, \beta) \propto 1$ is used.

Dose x_i, (log mg/ml)	Number of animals n_i	Number of deaths y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

Table 3.1. Bioassay experiment data.

Visualization of the posterior distribution

Before we begin visualizing the CCD integration algorithm steps, it would be beneficial to construct a plot of the joint posterior probability distribution of the two model parameters α and β. We can construct a visual representation of the posterior by evaluating it at points positioned in a rectangular grid covering most of the probability mass. The bounds for the region containing the grid were selected to be $[-5, 10]$ for α and $[-10, 40]$ for β, and the grid included a total of 10000 points (100 breaks in each dimension), as suggested in [9]. The resulting plot is presented in Figure 3.1.

In real-world scenarios where CCD integration would be applicable, it is rarely computationally practical to evaluate the posterior at such a large number of points. However, within the context of this illustrative

example, the computations are entirely feasible, and the resulting plot is useful in grounding the subsequent visualizations. We can also use the computed densities to conduct numerical integration, the results of which can later be used as reference to evaluate the performance of other integration methods.

Following the steps of CCD integration

The first step of the CCD integration process is to locate the mode of the posterior and to estimate the Hessian matrix of the posterior at that location. For the presented experiment, the mode was determined using the Nelder-Mead method [11], while the Hessian estimation was carried out utilizing the finite differences method. The two procedures required a total of 85 evaluations of the posterior log-density, with 16 of those evaluations being used for the Hessian estimation. The path traced by the mode locating algorithm is visualized on Figure 3.1.

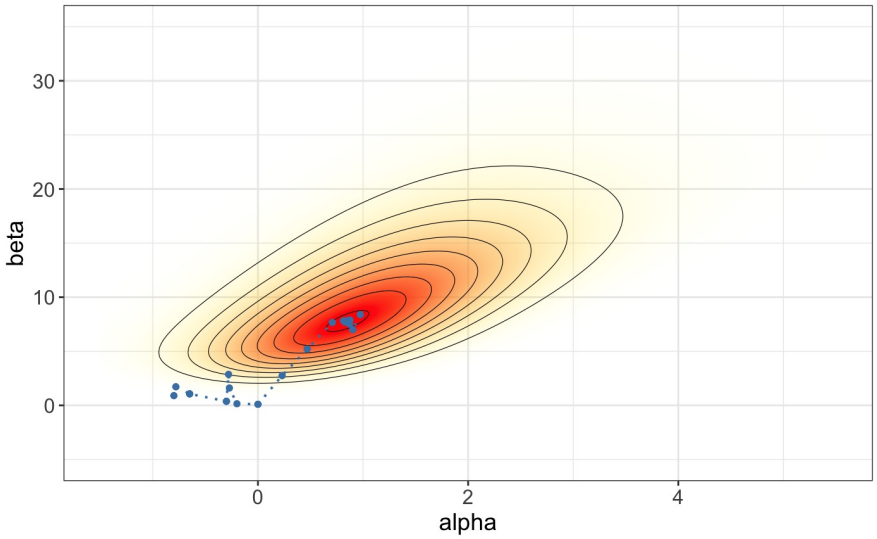
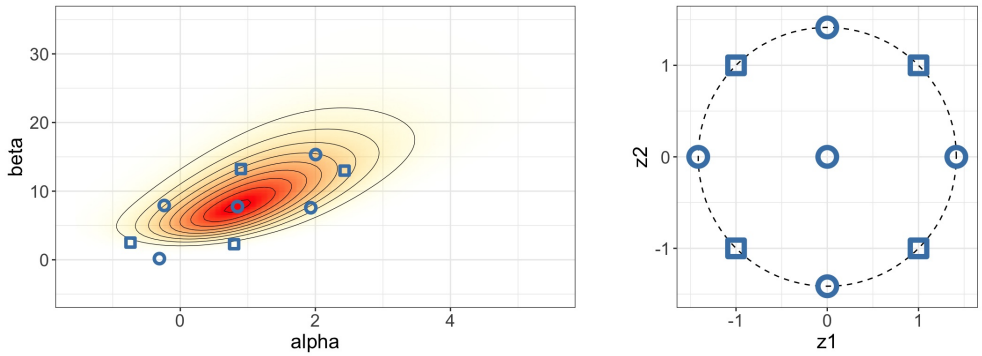


Figure 3.1. *Posterior probability density heatmap for the model parameters in the bioassay example. Contour lines at 0.9, 0.8,..., 0.1 of the modal density are added. The sequence of points at which the posterior was evaluated when locating the mode is plotted on top.*

With the estimates for the mode and the Hessian of the posterior obtained, we have collected all the necessary parameters to compute the whitening transformation, as described in Equation 3.4. The next step in the algorithm would then be to generate the star and design points in z-space. Since the experimental model is parameterized by two variables, five star points are generated (two points per principal axis plus the mode). Four additional design points are added with the central composite design method. Figure 3.2 depicts the resulting point configuration in both the

original parameter space and the z-space.



(a) *θ-space.* **(b)** *z-space.*

Figure 3.2. *Integration points generated by CCD before shifting. The star points are denoted with circular glyphs, while the design points are denoted with square glyphs.*

Next, we perform the shifting of the obtained points along each of the principal directions. This operation requires one additional evaluation of the log-density for computing each of the scaling coefficients, totalling 8 additional evaluations. The resulting nine points generated with CCD are depicted in Figure 3.3. We can note that after shifting, the points lie roughly on an equidensity circuit of the posterior.

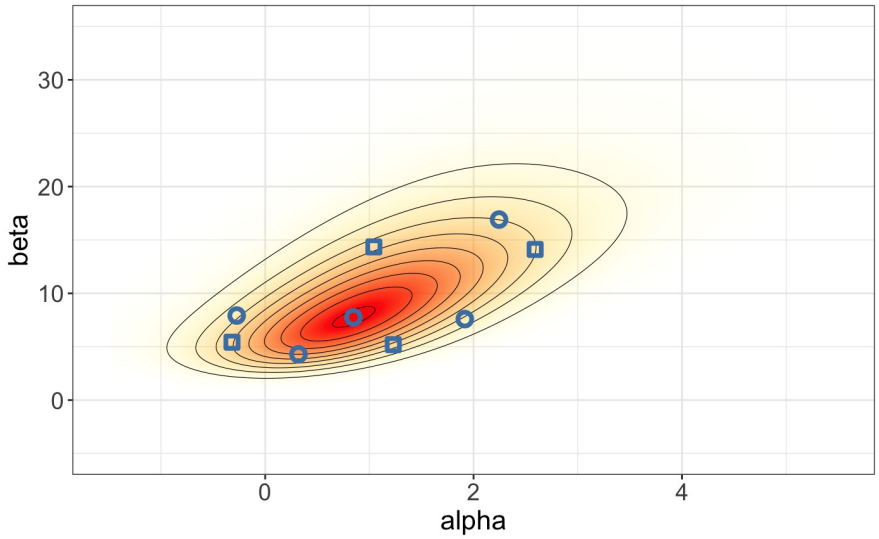


Figure 3.3. *Final integration points obtained with CCD.*

The final step of the CCD integration procedure is to compute the integration weights as well as the posterior densities at the generated points. The latter computation requires 8 additional evaluations of the posterior log-density. With the weights and densities obtained, we can now proceed to conducting inference about the parameter of interest — LD50. The expectation of LD50, for instance, can be estimated as follows:

$$E[\text{LD50}] \approx \sum_{i=1}^{9} -\frac{\alpha_i}{\beta_i} p(\alpha_i, \beta_i | x, n, y) \Delta_i, \tag{3.13}$$

where α_i and β_i are the coordinates of points generated with CCD.

Results and conclusions

The estimations for the expectation of LD50, along with the number of posterior log-density evaluations needed to compute the estimate, are presented in Table 3.2. For the sake of comparison, results for the MAP estimate, as well as the grid-based numerical integration, are also provided.

Integration strategy	Log-density evaluations	$E[\text{LD50}]$	$E[\alpha]$	$E[\beta]$
MAP estimate	69	-0.1093	0.8471	7.7495
CCD integration	98	-0.1079	1.0431	9.1145
Reference value	10000	-0.1068	1.3128	11.6132

Table 3.2. *Bioassay experiment results. Reference values obtained through grid-based numerical integration.*

The differences in the obtained estimates can hardly lead to conclusive judgement on the performance of the three presented integration methods due to the limited scale of the underlying experiment. This study was presented as a means of illustrating the mechanism by which CCD generates integration points rather than as a proper evaluation of the efficacy of the integration techniques.

Nevertheless, we can observe that the CCD integration strategy has provided an estimation that is noticeably closer to the reference than that produced by the MAP estimate, while still requiring only a modest number of log-density evaluations.

4. Case study — motorcycle accident simulation

This section details the outcome of utilizing the central composite design integration as part of an existing experimental study. The study presented in this work explores motion data collected from simulated motorcycle accidents created for helmet testing.

4.1 Experiment model and data

The goal of this experiment is to investigate the forces that are applied to the head of a motorcycle driver during the first few moments of a crash based on simulated measurements of head acceleration [12]. The data set contains 133 measurements of acceleration at moments of time between 0 and 60 milliseconds after the impact. The model for the acceleration function is selected to be normal with the prior on the mean being a Gaussian process (GP) function. For simplicity, the variance of the residual σ is assumed to be constant, with a normal prior. The entire model can then be described as below:

$$\begin{aligned} y &\sim \mathcal{N}(f(x), \sigma) \\ f &\sim GP(0, K) \\ \sigma &\sim \mathcal{N}^{+}(0, 1), \end{aligned} \tag{4.1}$$

where y describes the predicted head acceleration at a time x, and K is squared exponential kernel utilized by the GP. In this model, the covariance function that defines K is parameterized by the length scale parameter l_p and magnitude parameter σ_f^2. The overall model is then described by three parameters $\theta = \{l_p, \sigma_f^2, \sigma_n\}$, where $\sigma_n = \sigma / sd(\mathcal{D}_y)$ is the normalized residual parameter. Since each of the parameters is constrained to $\mathbb{R}^+$, a logarithmic transformation is applied. For a comprehensive introduction to

Gaussian processes, refer to [13]. The general structure of the experiment and exact model specification have been borrowed from [14]. The latter is provided as Stan code in the appendix A.1.

4.2 Method

For this experiment, three numerical integration strategies were used for conducting inference in the GP model: the MAP estimate, CCD integration, and MCMC sampling. Each of the methods was applied to the posterior fitted on the full data set, and the estimates of the mean values for the three model parameters were recorded. To obtain a reference value for each of the estimates, an extended round of MCMC sampling was conducted using a large number of draws. In this experiment, the number of draws for the extended sampling round was selected to be 400,000.

In addition, the experiment was repeated with a truncated data set to evaluate the impact of a reduction in the volume of observations on the performance of the integration strategies.

4.3 Results

Full data set experiment

Before we begin analyzing the quality of the predictions provided by the CCD technique, we can first examine the locations of the integration points proposed by the algorithm to verify adequate exploration of the parameter space. A visualization is provided in Figure 4.1. We can observe that the proposed points cover the posterior mass reasonably well and capture the details of the posterior shape.

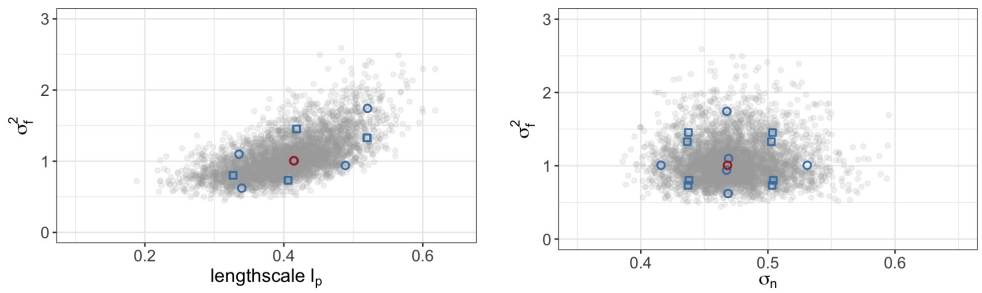


Figure 4.1. *Integration points generated by CCD for the motorcycle study, plotted on top of the MCMC samples. The star points are denoted with circular glyphs, while the design points are denoted with square glyphs. The modal point is highlighted in red.*

When the model is fitted on all 133 observations, the three integration techniques output very similar predictions for the locations of the model parameters. The numerical results of the experiment are summarized in Table 4.1 and predictive inference plots are provided in the appendix A.2. We can observe that the parameter predictions provided by the CCD integration method are closer to the MCMC results compared to the outputs of the MAP estimate strategy, while requiring only moderately greater number of log-density gradient evaluations for point generation. For the sake of comparison, simple log-density evaluations have been equated to log-density gradient evaluations in Table 4.1, providing a pessimistic evaluation of CCD computational demand. Nevertheless, the drastic difference in resource demand between the CCD approach and MCMC integration is evident.

Integration strategy	Log-density gradient evaluations	$E[l_p]$	$E[\sigma_f^2]$	$E[\sigma_n]$
Point estimate	15	0.394	0.910	0.466
CCD integration	59	0.412	1.023	0.469
MCMC sampling	23534*	0.406	1.067	0.472

Table 4.1. *Motorcycle accident experiment results on the full data set.*
$*$ log-density gradient evaluations that occurred during the warm-up stage of sampling are not included.

To further explore the trade-off between inference accuracy and computational cost associated with CCD when compared to MCMC, we can estimate the number of posterior gradient evaluations required for the MCMC method to achieve prediction accuracy matching that of CCD. Figure 4.2 demonstrates how the MCMC prediction error for the length scale estimate changes with the number of gradient evaluations. To provide a ground truth value of the length scale, an estimate from a large MCMC-generated sample was used.

As can be observed from Figure 4.2, the MCMC integration strategy has the potential to offer significantly better accuracy than the two studied deterministic approaches. However, it requires a noticeably larger number of posterior gradient evaluations to achieve low estimation errors, with the number of evaluations required to match CCD's accuracy being two orders of magnitude larger than that of CCD integration itself.

Reduced data set experiment

With the volume of training data restricted to 10% of its original volume, the differences in the performance of the three numerical integration

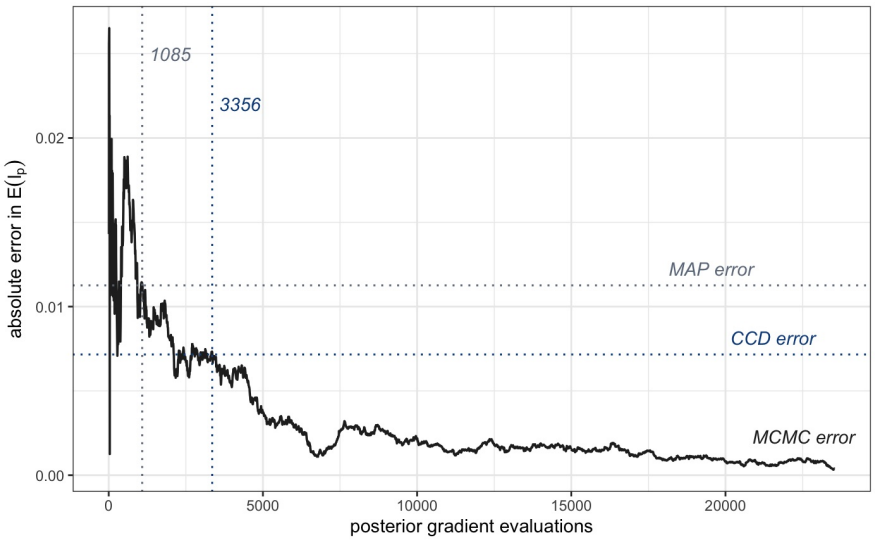


Figure 4.2. *Approximation error for $E[l_p]$ of the three studied algorithms.*

approaches becomes more apparent. Looking at the visualization of the proposed integration points presented in Figure 4.3, we can notice that the mode and the CCD integration points do not cover the posterior mass quite as well as in the previous experiment. When fitted on the reduced data set, the posterior shape gains thicker right tails for the length scale l_p and residual parameter σ_n. As a result, the mode becomes a less effective representative of the overall posterior, indicating that the MAP estimate model might not be suitable. The CCD points do capture some of the added uncertainty, but remain concentrated close to the mode, leaving the tails largely unexplored.

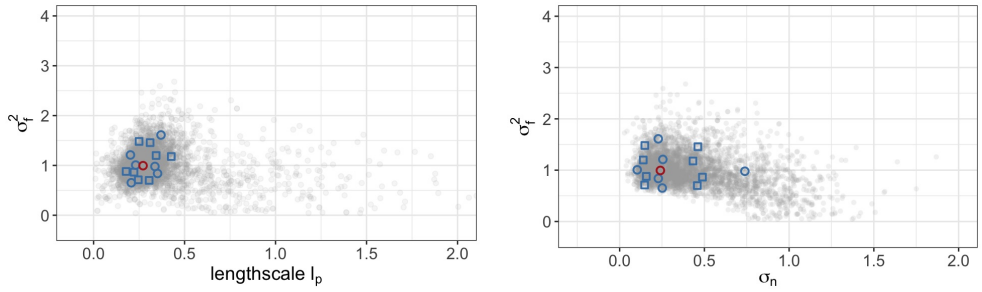


Figure 4.3. *Integration points generated by CCD for the truncated data set motorcycle study, plotted on top of the MCMC samples. The star points are denoted with circular glyphs, while the design points are denoted with square glyphs. The modal point is highlighted in red.*

The MAP estimate technique produces an overfitted and overconfident model, underestimating both the length scale l_p and the residual σ_n parameters. The resulting predictive interval covers the training data well, but misses a large portion of the observations what were omitted from the training set. The CCD integration estimates for the parameter locations improve upon the MAP estimate at modest added computational cost,

while still falling short of the accuracy of the more expensive MCMC approach. The obtained results are consistent with the posterior exploration shortcomings of the two deterministic models, as discussed above.

Integration strategy	Log-density gradient evaluations	$E[l_p]$	$E[\sigma_f^2]$	$E[\sigma_n]$
Point estimate	23	0.248	0.926	0.181
CCD integration	44	0.279	1.012	0.280
MCMC sampling	32296*	0.342	0.998	0.418

Table 4.2. *Motorcycle accident experiment results on the truncated data set.*
** log-density gradient evaluations that occurred during the warm-up stage of sampling are not included.*

5. Conclusions

5.1 Summary

In this thesis, we explored central composite design integration as a numerical integration method in the context of Bayesian statistics. We outlined the motivation behind the use of approximate integration in the Bayesian inference process and introduced two common approaches to selecting the integration points — MAP estimate and MCMC sampling. Following this, we discussed how the CCD strategy can address the shortcoming of the two existing methods, specifically the loss of uncertainty associated with the MAP estimate and the heavy computational load of MCMC sampling.

Next, we detailed the steps involved in generating the integration points using CCD. The algorithm was subsequently illustrated with an example based on a bivariate posterior obtained from a bioassay experiment. The provided visualizations demonstrated how the locations of the integration points are selected and then adjusted to better capture the shape of the posterior distribution.

Finally, we presented our implementation of central composite design integration by utilizing the method in an existing experimental study. Our results showed that CCD integration is capable of meaningfully improving upon the accuracy of the MAP estimate without elevating the computational cost of the inference to levels typically demanded by MCMC.

Overall, our study highlights the potential of the central composite design integration method as a valuable tool for Bayesian inference. The method can be particularly useful in scenarios where MCMC sampling is not feasible, such as when studying posteriors that are very computationally costly to evaluate.

5.2 Future work

While our experiments have shown CCD integration to be an effective numerical integration strategy, further research is necessary to fully understand its advantages and limitations. Evaluating the performance of CCD integration on models with a higher number of parameters is of particular interest, as the reduction in computational demand is expected to be more pronounced. Additionally, applying CCD integration to posteriors with heavily non-Gaussian shapes, such as funnel-shaped distributions, could help determine the limitations of this numerical integration approach's applicability.

Another avenue of future research that might be worth exploring is the use of points other than the mode as the central point in CCD integration. The mode has the advantage of being reasonably quick to locate with established optimisation methods, but for some posteriors, the mode can be located far away from most of the probability mass of the distribution. In such cases, the presented implementation of CCD would not be able to generate points that represent the posterior mass well, even when utilizing the directional shifting technique. It might then be advantageous to alter the point selection algorithm in such a way that it does not rely on the mode as the central integration point. One possible approach is outlined in [15].

Bibliography

[1] R. van de Schoot, S. D. Winter, O. Ryan, M. Zondervan-Zwijnenburg, and S. Depaoli, "A systematic review of bayesian articles in psychology: The last 25 years," *Psychological Methods*, vol. 22, p. 217–239, 2017.

[2] D. Ashby, "Bayesian statistics in medicine: a 25 year review," *Statistics in Medicine*, vol. 25, no. 21, p. 3589–3631, 2006.

[3] H. Rue, S. Martino, and N. Chopin, "Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 2, p. 319–392, 2009.

[4] I. Fornacon-Wood, H. Mistry, C. Johnson-Hart, C. Faivre-Finn, J. P. O'Connor, and G. J. Price, "Understanding the differences between bayesian and frequentist statistics," *International Journal of Radiation Oncology*Biology*Physics*, vol. 112, no. 5, p. 1076–1082, Apr 2022.

[5] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*. Courier Corporation, Jan 2007, google-Books-ID: gGCKdqka0HAC.

[6] J. Vanhatalo, V. Pietiläinen, and A. Vehtari, "Approximate inference for disease mapping with sparse gaussian processes," *Statistics in Medicine*, vol. 29, no. 15, pp. 1580–1607, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3895

[7] S. M. Sanchez and P. J. Sanchez, "Very large fractional factorial and central composite designs," *ACM Transactions on Modeling and Computer Simulation*, vol. 15, no. 4, p. 362–377, Oct 2005.

[8] S. Martino, "Approximate bayesian inference for latent gaussian models," Doctoral thesis, Fakultet for informasjonsteknologi, matematikk og elektroteknikk, 2007, accepted: 2014-12-19T13:57:02Z. [Online]. Available: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/258135

[9] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, "Bayesian data analysis third edition," Oct 2022.

[10] A. Racine, A. P. Grieve, H. Flühler, and A. F. M. Smith, "Bayesian methods in practice: Experiences in the pharmaceutical industry," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 35, no. 2, p. 93–120, Jun 1986.

[11] J. A. Nelder and R. Mead, “A simplex method for function minimization,” *The Computer Journal*, vol. 7, no. 4, p. 308–313, Jan 1965.

[12] B. W. Silverman, “Some aspects of the spline smoothing approach to non-parametric regression curve fitting,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 47, no. 1, p. 1–52, 1985.

[13] V. Pietiläinen, “Approximations for Integration over the Hyperparameters in Gaussian Processes,” Master’s thesis, 2010. [Online]. Available: http://urn.fi/URN:NBN:fi:aalto-201203131405

[14] A. Vehtari, “Gaussian process demonstration with stan,” 2023, accessed on: 2023-05-16. [Online]. Available: https://avehtari.github.io/casestudies/Motorcycle/motorcycle_gpcourse.html

[15] L. Zhang, B. Carpenter, A. Gelman, and A. Vehtari, “Pathfinder: Parallel quasi-newton variational inference,” *Journal of Machine Learning Research*, vol. 23, no. 306, p. 1–49, 2022.

A. Additional materials for the motorcycle accident case study

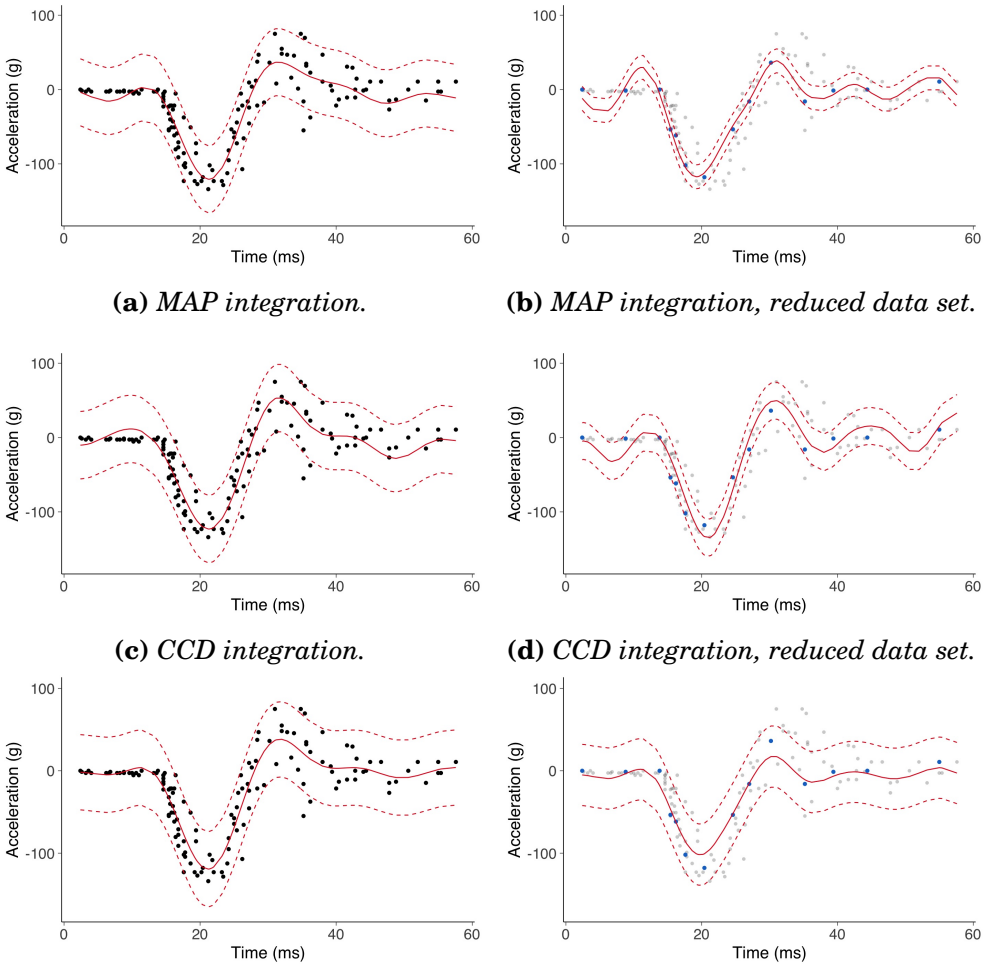
A.1 Stan code

```
functions {
  vector gp_pred_rng(array[] real x2,
                     vector y1,
                     array[] real x1,
                     real sigma_f,
                     real lengthscale_f,
                     real sigma,
                     real jitter) {
    int N1 = rows(y1);
    int N2 = size(x2);
    vector[N2] f2;
    {
      matrix[N1, N1] L_K;
      vector[N1] K_div_y1;
      matrix[N1, N2] k_x1_x2;
      matrix[N1, N2] v_pred;
      vector[N2] f2_mu;
      matrix[N2, N2] cov_f2;
      matrix[N1, N1] K;
      K = gp_exp_quad_cov(x1, sigma_f, lengthscale_f);
      for (n in 1:N1)
        K[n, n] = K[n,n] + square(sigma);
      L_K = cholesky_decompose(K);
      K_div_y1 = mdivide_left_tri_low(L_K, y1);
      K_div_y1 = mdivide_right_tri_low(K_div_y1', L_K)';
      k_x1_x2 = gp_exp_quad_cov(
        x1, x2, sigma_f, lengthscale_f
      );
      f2_mu = (k_x1_x2' * K_div_y1);
      v_pred = mdivide_left_tri_low(L_K, k_x1_x2);
      cov_f2 = gp_exp_quad_cov(
        x2, sigma_f, lengthscale_f
      ) - v_pred' * v_pred;

      f2 = multi_normal_rng(
        f2_mu, add_diag(cov_f2, rep_vector(jitter, N2))
```

```
      );
    }
    return f2;
  }
}
data {
  int<lower=1> N;       // number of observations
  vector[N] x;          // univariate covariate
  vector[N] y;          // target variable
  int<lower=1> N2;      // number of test points
  vector[N2] x2;        // univariate test points
}
transformed data {
  // Normalize data
  real xmean = mean(x);
  real ymean = mean(y);
  real xsd = sd(x);
  real ysd = sd(y);
  array[N] real xn = to_array_1d((x - xmean)/xsd);
  array[N2] real x2n = to_array_1d((x2 - xmean)/xsd);
  vector[N] yn = (y - ymean)/ysd;
  real sigma_intercept = 1;
  vector[N] zeros = rep_vector(0, N);
}
parameters {
  real<lower=0> lengthscale_f; // lengthscale of f
  real<lower=0> sigma_f;       // scale of f
  real<lower=0> sigman;         // noise sigma
}
model {
  // covariances and Cholesky decompositions
  matrix[N, N] K_f = gp_exp_quad_cov(
    xn, sigma_f, lengthscale_f
  ) + sigma_intercept^2;
  matrix[N, N] L_f = cholesky_decompose(
    add_diag(K_f, sigman^2)
  );
  // priors
  lengthscale_f ~ normal(0, 1);
  sigma_f ~ normal(0, 1);
  sigman ~ normal(0, 1);
  // model
  yn ~ multi_normal_cholesky(zeros, L_f);
}
generated quantities {
  // function scaled back to the original scale
  vector[N2] f = gp_pred_rng(
    x2n, yn, xn, sigma_f, lengthscale_f, sigman, 1e-9
  ) * ysd + ymean;
  real sigma = sigman*ysd;
}
```

A.2 Predictive plots



(a) *MAP integration.* **(b)** *MAP integration, reduced data set.*

(c) *CCD integration.* **(d)** *CCD integration, reduced data set.*

(e) *MCMC integration.* **(f)** *MCMC integration, reduced data set.*

Figure 1.1. *Predictive interval plots for the GP model plotted on top of the original simulated observations. Plots (a), (c), and (d) depict results for models fitted on the full data set, while plots (b), (d), and (f) depicts results for models fitted on a reduced data set. For the right column, the data points used for training are highlighted in blue.*