Delft University of Technology

Bachelor Project
AM3000

Estimating Normalizing Constants using Stochastic Simulation

Author:
Rinze Hallema (5280133)

Thesis comittee:
Dr. Ir. J. Bierkens
Dr. Y. van Gennip
June 18, 2023

Summary

The computation of normalizing constants often brings (higher-dimensional) integrals, which could not be computed analytically or are computationally expensive. Stochastic methods will be explored to find an estimate for normalizing constants. These stochastic methods will be used to estimate the Bayes factor. With this Bayes factor can be searched for a best fitting model to data.

Contents

Chapter 1

Introduction

A density function describes the likelihood of an outcome for a random variable. By integrating this density function over its entire support, we find the normalizing constant of the density function. Now, the density function divided by the normalizing constant integrates to 1, making it a probability density function. A density function is often higher dimensional, which results in integrals that are intractable to calculate analytically or computationally too expensive to calculate. A numerical approach is needed to find an estimate for normalizing constants. The numerical approaches to estimate normalizing constants in this thesis will be used to search for the best fitting model to known data x, such that we can do accurate predictions based on this model. Finding the best fitting model can be done using the Bayes factor.

The Bayes factor gives the strength of evidence in favor of one model compared to another model. The Bayes factor is defined as the likelihood of model 1, $l(M_1)$, where $l(\cdot)$ is thus the likelihood function and M_1 model 1, divided by the likelihood of model 2, $l(M_2)$. The likelihood of model 1 is equal to the probability of data x knowing model 1, so $l(M_1) = p(x|M_1)$ and $l(M_2) = p(x|M_2)$. Thus when the Bayes factor is greater than 1, model 1 fits better to the data and when the Bayes factor is smaller than 1, model 2 fits better to the data. By integrating over the whole parameter space Θ, the Bayes factor $BF(M_1, M_2)$ can be rewritten as follows

$$\begin{aligned} BF(M_1, M_2) &= \frac{l(M_1)}{l(M_2)} = \frac{p(x|M_1)}{p(x|M_2)} \\ &= \frac{\int p(x|\theta, M_1) p(\theta|x) d\theta}{\int p(x|\theta, M_2) p(\theta|x) d\theta}. \end{aligned}$$

The integrals gives a measure of the amount of evidence the data gives for the model, this could be interpreted as a normalizing constant.

When we are not able to calculate the normalizing constant z analytically by integrating over the unnormalized density function $q(\cdot)$, we can use a sampling method to find an estimate for z. The main idea of a sampling method is that it gives an estimate $\hat{z}$ of z based on random draws from the normalized density function $p(\cdot)$. However, when it is not possible to take draws from $p(\cdot)$, we have to search for a proposal density $\tilde{p}(\cdot)$ from which draws can be taken. This proposal density could be for example an analytical approximation of $p(\cdot)$ or be obtained from a Markov Chain Monte Carlo (MCMC) method (Chapter 2.3).

Chapter 3 explores the sampling methods described in Gelman and Meng (1998). In Chapter 4 a Gaussian field will be considered. A Gaussian field is a field, where the points on the field get an value from a Gaussian distribution. In our setting are the points on this Gaussian field censored, as all points with negative value are observed as zeros. Using a sampling method, an estimate will be searched for the probabilty that a specific arbitrary point on the field is below a threshold. The approach is taken over from Stein (1992). It is in practice often not known what model fits to such a field. In Section 4.6 the best fitting model to the observed data points will be searched using the Bayes factor, where the approach is again taken over from Stein (1992).

1.1 Notation

Throughout this thesis the following notation will be used:

Notation	**Description**
$q(\omega)$	unnormalized density function
$p(\omega)$	normalized density function
$\tilde{p}(\omega)$	proposal density
z	normalizing constant
r	ratio of normalizing constants ($r = \frac{z_1}{z_2}$)
λ	logarithmic ratio of normalizing constants ($\lambda = \log \frac{z_1}{z_2}$)
$-U(\omega, \theta)$	potential energy
$\alpha(\cdot)$	arbitrary function
$\Phi(\cdot)$	cumulative distribution function standard normal
$l(\cdot)$	likelihood function
$L(\cdot)$	Lagrangian
$BF(M_1, M_2)$	Bayes factor between model 1 and model 2
$T(\cdot\|\cdot)$	Transition kernel/proposal distribution

We use a hat above the variable to denote that the variable is an estimate. So $\hat{z}$ is an estimate of z.

Chapter 2

Preliminary Theory

In this Chapter some theory that we will use later on will be explained. In Section 2.1 the Euler-Lagrange equation will be proved. With the Euler-Lagrange equation the extreme value of a functional could be found. In Section 2.2 is the Cauchy-Schwartz inequality proved. This inequality will be usefull for finding a lowerbound to the variance of an estimator. In Section 2.3.1, a popular MCMC method will be introduced. This method is tried in an example in Section 2.3.2 to draw from a normal distribution.

2.1 Euler-Lagrange equations

The Euler-Lagrange equation is useful for finding the extreme values of a functional. The Euler-Lagrange equation will be proved following the steps of Boas (2005). Assume that we want to minimize $I[f(x)] = \int_a^b L(x, f(x), f'(x))dx$ with respect to $f(x)$. Here is L the Lagrangian and we will assume that it is twice continuously differentiable. Define $f_\epsilon(x) = f(x)+\epsilon\eta(x)$, where $\eta(x)$ is an arbitrary function with $\eta(a) = \eta(b) = 0$ and ϵ is a small perturbation such that $f_\epsilon(x)$ differs slightly from $f(x)$. Then $I_\epsilon = \int_a^b L(x, f_\epsilon(x), f'_\epsilon(x))dx$. For the minimum I, we should have that $\frac{dI_\epsilon}{d\epsilon} = 0$ for $\epsilon = 0$. Now the Euler-Lagrange equation can be derived

$$\begin{aligned}
\frac{dI_\epsilon}{d\epsilon} &= \frac{d}{d\epsilon}\int_a^b L(x, f_\epsilon(x), f'_\epsilon(x))dx \\
&\overset{(1)}{=} \int_a^b \frac{d}{d\epsilon} L(x, f_\epsilon(x), f'_\epsilon(x))dx \\
&\overset{(2)}{=} \int_a^b \frac{\partial L}{\partial f_\epsilon}\frac{df_\epsilon}{d\epsilon} + \frac{\partial L}{\partial f'_\epsilon}\frac{df'_\epsilon}{d\epsilon}dx \\
&\overset{(3)}{=} \int_a^b \frac{\partial L}{\partial f_\epsilon}\eta(x) + \frac{\partial L}{\partial f'_\epsilon}\eta'(x)dx \\
&\overset{(4)}{=} \int_a^b \frac{\partial L}{\partial f_\epsilon}\eta(x)dx - \int_a^b \frac{\partial L}{\partial f'_\epsilon}\eta'(x)dx \\
&\overset{(5)}{=} \int_a^b [\frac{\partial L}{\partial f} - \frac{d}{dx}\frac{\partial L}{\partial f'}]\eta(x)dx.
\end{aligned}$$

In the first step Leibniz' integral rule (Border (2016)) to interchange the differentiation and integration is used. In step (2), the chain rule is used and in step (3) the derivatives of $f_\epsilon(x), f'_\epsilon(x)$ are calculated with respect to ϵ. In step (4), integration by parts is used

$$\int_a^b \frac{\partial L}{\partial f'_\epsilon}\eta'(x)dx = [\frac{\partial L}{\partial f'_\epsilon}\eta(x)]_a^b - \int_a^b \frac{d}{dx}\frac{\partial L}{\partial f'_\epsilon}\eta(x)dx = -\int_a^b \frac{d}{dx}\frac{\partial L}{\partial f'_\epsilon}\eta(x)dx,$$

as $\eta(a) = \eta(b) = 0$. In the last step (5) are the functions evaluated in $\epsilon = 0$ and are the integrals taken together. Hence, it can be concluded $\frac{dI_\epsilon}{d\epsilon} = 0$ for $\epsilon = 0$ if and only if $\frac{\partial L}{\partial f} - \frac{d}{dx}\frac{\partial L}{\partial f'} = 0$ as $\eta(x)$ is an arbitrary function. This Euler-Lagrange equation extends for a problem that has more than one dependent variable f (Boas (2005)). For a Lagrangian with two dependent variables f, g, so $L(x, f(x), f'(x), g(x), g'(x))$ the following equations should be solved to make $\int_a^b L(x, f(x), f'(x), g(x), g'(x))dx$ stationary

$$\begin{cases} \frac{d}{dx}\frac{\partial L}{\partial f'} - \frac{\partial L}{\partial f} = 0 \\ \frac{d}{dx}\frac{\partial L}{\partial g'} - \frac{\partial L}{\partial g} = 0. \end{cases}$$

2.2 Cauchy-Schwarz Inequality

Let $f(x), g(x)$ be two integrable functions on interval $[a,b]$ with $f(x)$ is not the zero function. Define $h(t) = \int_a^b (tf(x) - g(x))^2 dx$. Now expanding the square and using linearity of the integral gives

$$h(t) = \int_a^b (tf(x) - g(x))^2 dx = t^2 \int_a^b f(x)^2 dx - 2t \int_a^b f(x)g(x)dx + \int_a^b g(x)^2 dx.$$

Denoting the three separate integrals respectively A, B and C gives $h(t) = At^2 - 2Bt + C$. From $(tf(x) - g(x))^2 \geq 0$, follows that $h(t) \geq 0$. We find $h'(t) = 2At - 2B$ and $h''(t) = 2A$, as $A > 0$ follows that $h(t)$ has a minimum in $h'(t) = 0$. Thus in $t = \frac{B}{A}$. Substituting this in $h(t)$, gives

$$h(\frac{B}{A}) = (\frac{B}{A})^2 A - 2\frac{B}{A}B + C = \frac{AC - B^2}{A}.$$

From $h(\frac{B}{A}) \geq 0$ and $A > 0$ follows that $AC - B^2 \geq 0 \iff B^2 \leq AC$, which is the Cauchy-Schwartz inequality

$$\Big(\int_a^b f(x)g(x)dx\Big)^2 \leq \int_a^b f(x)^2 dx \int_a^b g(x)^2 dx.$$

2.3 Markov Chain Monte Carlo

When it is not possible to take draws from a density or from an analytical approximation of this density, a MCMC method could be used to generate draws. With such a method samples are being drawn and based on an acceptance criterion, this sample will be accepted or rejected. As such a method proceeds, it generates a chain of samples.

2.3.1 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is one of the most popular MCMC methods. The algorithm generates a Markov chain of generated samples from $p(\omega)$, which we call the target distribution. In a Markov chain, each generated sample does only depend on the previous generated sample and not on all generated samples before. For the Metropolis-Hasting algorithm an initial guess ω_0 is needed. We need to define an proposal distribution (or transition kernel) $T(\omega'|\omega_t)$. Now given ω_t, sample an ω' from $T(\omega'|\omega_t)$. This ω' could be seen as a proposed sample for ω_t. Based on the acceptance probability is this proposed sample either accepted or rejected. The acceptance probability of this proposed sample for ω_t is in the Metropolis-Hastings algorithm defined such that the proposal distribution satisfies the detailed balance equation. When the detailed balance equation is satisfied, the generated chain of samples converges to the target distribution [4]. The detailed balance equation states

$$p(\omega_t)\mathbb{P}(\omega' \to \omega_t) = p(\omega')\mathbb{P}(\omega_t \to \omega'),$$

where $\mathbb{P}(\omega' \to \omega_t)$ is the probability that ω moves to ω_t in the algorithm. The acceptance probability of the Metropolis-Hastings algorithm is defined as

$$A(\omega', \omega_t) = \min\big(1, \frac{p(\omega')T(\omega_t|\omega')}{p(\omega_t)T(\omega'|\omega_t)}\big).$$

Now, as $\mathbb{P}(\omega' \to \omega_t) = T(\omega'|\omega_t)A(\omega', \omega_t)$ we find

$$
\begin{aligned}
p(\omega_t)\mathbb{P}(\omega' \to \omega_t) &= p(\omega_t)T(\omega'|\omega_t)A(\omega',\omega_t) \\
&= p(\omega_t)T(\omega'|\omega_t)\min\bigl(1, \frac{p(\omega')T(\omega_t|\omega')}{p(\omega_t)T(\omega'|\omega_t)}\bigr) \\
&= \min\bigl(p(\omega_t)T(\omega'|\omega_t), p(\omega')T(\omega_t|\omega')\bigr) \\
&= p(\omega')T(\omega_t|\omega')\min\bigl(1, \frac{p(\omega_t)T(\omega'|\omega_t)}{p(\omega')T(\omega_t|\omega')}\bigr) \\
&= p(\omega')T(\omega_t|\omega')A(\omega_t,\omega') \\
&= p(\omega')\mathbb{P}(\omega_t \to \omega')
\end{aligned}
$$

Hence, it can be concluded that the Metropolis-Hastings algorithm satisfies the detailed balance equation. The generated Markov chain with this algorithm will thus converge to the target distribution. The steps are summarized in Algorithm 1.

Algorithm 1: Metropolis-Hastings

```
Data: $\omega_0$                                                          Initial guess
Set t = 0
Generate $\omega' \sim T(\omega'|\omega_t)$
Calculate $A(\omega', \omega_t) = \min 1, \frac{p(\omega')T(\omega_t|\omega')}{p(\omega_t)T(\omega'|\omega_t)}$
Generate $u \sim \text{Unif}[0, 1]$
if $u \leq A(\omega', \omega_t)$ then
    $\omega_{t+1} = \omega'$
else
    $\omega_{t+1} = \omega_t$
end
Iterate t = t + 1
```

2.3.2 Example normal distributions

In this example is wanted to sample from a normal distribution. Define $X \sim \mathcal{N}(2, 3)$ and use as proposal distribution the random walk proposal. This random walk proposal is defined by adding a $\mathcal{N}(0, 1)$ random number to the current draw, so $\omega' = \omega_t + \mathcal{N}(0, 1)$. Note that the random walk proposal is symmetric as the probability of getting proposal ω' from ω_t is the same as getting ω_t as proposal from ω', as the normal distribution is symmetric around 0. Thus, $T(\omega'|\omega_t) = T(\omega_t|\omega')$ and the acceptance probability is $A(\omega', \omega_t) = \min(1, \frac{p(\omega')}{p(\omega_t)})$. Chosing a symmetrical proposal distribution is the algorithm presented by Metropolis in 1953, this is called the Metropolis algorithm. Later on in 1970 Hastings presented the algorithm for an non-symmetric proposal distribution, which follows by adding the term $\frac{T(\omega_t|\omega')}{T(\omega'|\omega_t)}$ in the acceptance probability. This is called the Metropolis-Hastings algorithm. Doing one million iterations of the described Metropolis algorithm with initial guess $\omega_0 = 2$ we find in that the draws of the Metropolis algorithm fit the $\mathcal{N}(2, 3)$-density [6].

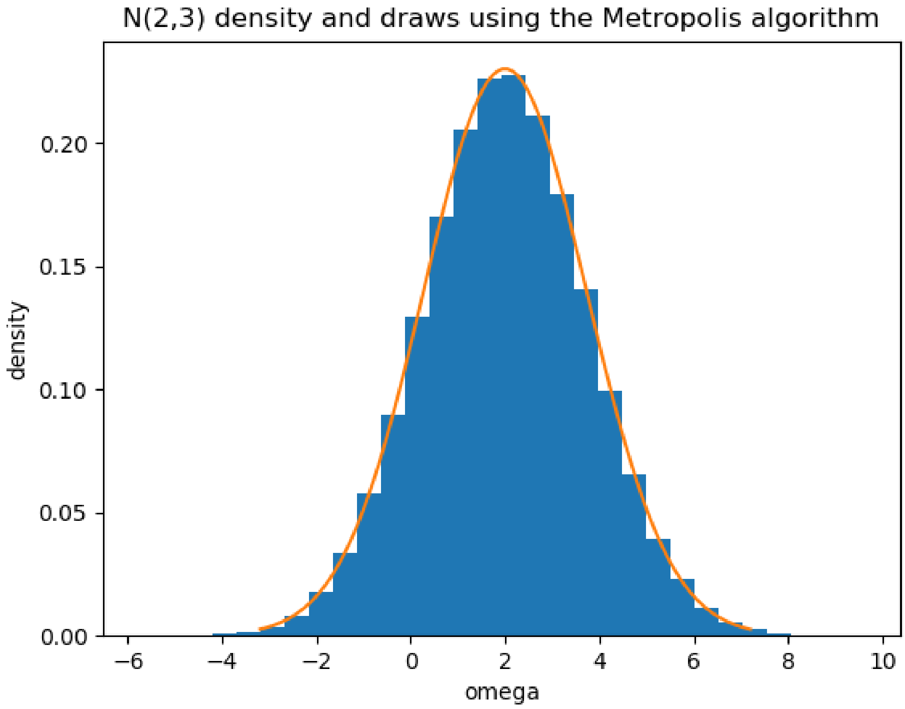


Figure 2.1: Draws Metropolis algorithm plotted on the normal density

Chapter 3

Sampling Methods

Sampling methods are techniques used when the normalizing constant of a distribution of interest cannot be calculated analytically or when the calculation is computationally expensive. These sampling methods use draws from the target distribution to estimate the normalizing constant. It is not always possible to draw from the target distribution, or from an analytical approximation of the target distribution. In this Chapter we will assume that there are no difficulties with taking draws from the target distribution. In Section 2.3 a method is discussed to generate draws using a MCMC method. The sampling methods work similarly with these generated draws. In this Chapter, we will introduce three specific sampling methods: Path Sampling (Section 3.1), Importance Sampling (Section 3.2) and Bridge Sampling (Section 3.3). These three sampling methods are found in Gelman and Meng (1998) and partly taken over.

3.1 Path Sampling

3.1.1 Path Sampling Estimate

The main idea of path sampling is that it generates a series of 'intermediate densities' between the two target densities. Based on samples from a path of these intermediate densities the ratio of normalizing constants of the function $q_0(\omega)$ and $q_1(\omega)$ will be estimated.
Assuming the legitimacy of the interchange of integration with differentiation, and using scalar parameter $\theta \in [0,1]$ we find

$$\begin{aligned} \frac{d}{d\theta}\log z(\theta) &\overset{\text{chain rule}}{=} \frac{1}{z(\theta)}\frac{d}{d\theta}z(\theta) \\ &= \int \frac{p(\omega|\theta)}{q(\omega|\theta)}\frac{d}{d\theta}q(\omega|\theta)d\omega \\ &= \mathbb{E}_\theta\left[\frac{1}{q(\omega|\theta)}\frac{d}{d\theta}q(\omega|\theta)\right] \\ &= \mathbb{E}_\theta\left[\frac{d}{d\theta}\log(q(\omega|\theta))\right], \end{aligned} \tag{3.1}$$

so we obtain $\lambda = \int_0^1 \mathbb{E}_\theta[\frac{d}{d\theta}\log(q(\omega|\theta))]d\theta$, where $\mathbb{E}_\theta$ denotes the expectation with respect to the target distribution. In statistical physics is $-\frac{d}{d\theta}\log(q(\omega|\theta))$ known as the potential energy. We will define $U(\omega,\theta) = \frac{d}{d\theta}\log(q(\omega|\theta))$ to simplify the equations. Introducing density $p(\theta) = \frac{p(\omega|\theta)}{p(\omega,\theta)}$ and considering θ as a random variable that is uniformly distributed on $[0,1]$, gives

$$\begin{aligned} \lambda = \int_0^1 \mathbb{E}_\theta[\frac{d}{d\theta}\log(q(\omega|\theta))]d\theta &= \int_0^1\int p(\omega|\theta)U(\omega,\theta)d\omega d\theta \\ &= \int_0^1 \mathbb{E}[\frac{U(\omega,\theta)}{p(\theta)}]d\theta \\ &= \mathbb{E}[\frac{U(\omega,\theta)}{p(\theta)}], \end{aligned}$$

where $\mathbb{E}$ is the expectation with respect to the joint density $p(\omega, \theta)$. Using n draws $(\omega_1, \theta_1), \ldots, (\omega_n, \theta_n)$ from the target distribution $p(\omega, \theta)$, the path sampling estimate is $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \frac{U(\omega_i, \theta_i)}{p(\theta_i)}$, with corresponding variance

$$\mathbb{V}\text{ar}(\hat{\lambda}) = \frac{1}{n}\left[\int_0^1 \frac{\mathbb{E}_\theta[U^2(\omega, \theta)]}{p(\theta)} d\theta - \lambda^2\right].$$

Extending this to the d-dimensional parameter space $\theta(t) = (\theta_1(t), \ldots, \theta_d(t))$, where $t \in [0, 1]$ an similar path sampling estimator could be found. Define $\theta_0 = \theta(0)$ and $\theta_1 = \theta(1)$, which are respectively the startpoint and the endpoint of the path. Define $U_k(\omega, \theta) = \frac{\partial \log q(\omega|\theta)}{\partial \theta_k}$ and $\dot{\theta}_k(t) = \frac{d\theta_k(t)}{dt}$. Using (3.1) and the chain rule to differentiate we find

$$\begin{aligned} \lambda &= \int_0^1 \mathbb{E}_{\theta(t)}[\frac{d}{dt} \log q(\omega|\theta(t))] dt \\ &= \int_0^1 \mathbb{E}_{\theta(t)}\left[\sum_{k=1}^{d} \dot{\theta}_k(t) U_k(\omega, \theta(t))\right] dt, \end{aligned}$$

where $\mathbb{E}_{\theta(t)}$ is the expectation with respect to $p(\omega|\theta(t))$. The corresponding path sampling estimator for λ is

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{k=1}^{d} \dot{\theta}_k(t_i) U_k(\omega_i, \theta(t_i))\right],$$

with variance

$$\mathbb{V}\text{ar}(\hat{\lambda}) = \frac{1}{n}\left[\int_0^1 \int \left(\sum_{k=1}^{d} \dot{\theta}_k(t_i) U_k(\omega_i, \theta(t_i))\right)^2 p(\omega|\theta) d\omega dt - \lambda^2\right]. \tag{3.2}$$

3.1.2 Choosing the best path

As one wants to use the optimal path sampling estimator, it is needed to minimize the variance of this estimator. $\mathbb{V}\text{ar}(\hat{\lambda}) = \mathbb{E}(\hat{\lambda}^2) - \mathbb{E}(\hat{\lambda})^2$, and $\mathbb{E}(\hat{\lambda})^2$ is independent of the path, so it suffices to minimize $\mathbb{E}(\hat{\lambda}^2)$.

$$\begin{aligned} \mathbb{E}(\hat{\lambda}^2) &= \frac{1}{n} \int_0^1 \int \frac{U^2(\omega, \theta)}{p^2(\theta)} p(\omega|\theta) p(\theta) d\omega d\theta \\ &= \frac{1}{n} \int_0^1 \frac{\mathbb{E}_\theta[U^2(\omega, \theta)]}{p(\theta)} d\theta \\ &= \frac{1}{n} \underbrace{\int_0^1 (\sqrt{p(\theta)})^2 d\theta}_{=1} \int_0^1 \left(\frac{\sqrt{\mathbb{E}_\theta[(U^2(\omega, \theta)]}}{\sqrt{p(\theta)}}\right)^2 d\theta \\ &\overset{\text{Cauchy-Schwartz}}{\geq} \left(\int_0^1 \sqrt{\mathbb{E}_\theta[U^2(\omega, \theta)] d\theta}\right)^2. \end{aligned}$$

The greater or equal sign becomes an equality when

$$p(\theta) = \frac{\sqrt{\mathbb{E}_\theta[U^2(\omega, \theta)]}}{\int_0^1 \sqrt{\mathbb{E}_{\theta'}[U^2(\omega, \theta')]} d\theta'}.$$

This is thus the optimal prior density and the corresponding optimal variance is

$$\mathbb{V}\text{ar}_{\text{opt}}(\hat{\lambda}) = \frac{1}{n}\left[\left(\int_0^1 \sqrt{\mathbb{E}_\theta[U^2(\omega, \theta)]} d\theta\right) - \lambda^2\right].$$

3.1.3 Example Path Sampling with optimal path in (μ, σ)-space

In this example an optimal path between two normal densities, $\mathcal{N}(\mu_0, \sigma_0)$ and $\mathcal{N}(\mu_1, \sigma_1)$ will be searched for. The path is denoted by $\theta(t) = (\mu(t), \sigma(t))$ with $t \in [0, 1]$, which is thus have a two-dimensional path. Define the startpoint and the endpoint respectively as $\theta_0 = (\mu_0, \sigma_0)$ and $\theta_1 = (\mu_1, \sigma_1)$. For any value of $\theta = (\mu, \sigma)$, the unnormalized density is given by

$$q(\omega|\theta) = \exp(-\frac{(\omega-\mu)^2}{2\sigma^2}).$$

So, the potential energy is given by

$$U_k(\omega, \theta) = \frac{\partial}{\partial \theta_k}\left(-\frac{(\omega-\mu)^2}{2\sigma^2}\right)$$

Using (3.2) gives

$$\begin{aligned}
\mathbb{V}\text{ar}(\hat{\lambda}) &= \frac{1}{n}\left[\int_0^1 \int \left(\sum_{k=1}^{2} \dot{\theta}_k(t) U_k(\omega, \theta(t))\right)^2 p(\omega|\theta) d\omega dt - \lambda^2\right] \\
&= \frac{1}{n}\left[\int_0^1 \int \left(\dot{\mu}\frac{\omega-\mu}{\sigma^2} + \dot{\sigma}\frac{(\omega-\mu)^2}{\sigma^3}\right)^2 p(\omega|\theta) d\omega dt - \lambda^2\right] \\
&= \frac{1}{n}\left[\int_0^1 \mathbb{E}_\theta\left[\left(\dot{\mu}\frac{\omega-\mu}{\sigma^2} + \dot{\sigma}\frac{(\omega-\mu)^2}{\sigma^3}\right)^2\right] dt - \lambda^2\right] \\
&\overset{(*)}{=} \frac{1}{n}\left[\int_0^1 \frac{\dot{\mu}^2}{\sigma^4}\mathbb{E}_\theta[(\omega-\mu)^2] + \frac{\dot{\sigma}^2}{\sigma^6}\mathbb{E}_\theta[(\omega-\mu)^4] dt - \lambda^2\right] \\
&= \frac{1}{n}\left[\int_0^1 \frac{\dot{\mu}^2 + 3\dot{\sigma}^2}{\sigma^2} dt - \lambda^2\right].
\end{aligned}$$

In (*) the cross-product is neglected, as the central moments of odd order are equal to zero for normal distributions. We minimize the variance with respect to θ $\int_0^1 \frac{\dot{\mu}^2 + 3\dot{\sigma}^2}{\sigma^2} dt$ and for this the Euler-Lagrange equations (see Section 2.1) with $L(\mu, \sigma, \dot{\mu}, \dot{\sigma}) = \frac{\dot{\mu}^2 + 3\dot{\sigma}^2}{\sigma^2} dt$ could be used, which gives

$$\begin{cases} \frac{\partial L}{\partial \mu} - \frac{d}{dt}\frac{\partial L}{\partial \dot{\mu}} = 0 \\ \frac{\partial L}{\partial \sigma} - \frac{d}{dt}\frac{\partial L}{\partial \dot{\sigma}} = 0. \end{cases}$$

Solving this system of equations and some tough calculations gives the optimal path $\theta(t) = (\mu(t), \sigma(t))$ with [5]:

$$\mu(t) = R\tanh(\phi_0(1-t) + \phi_1 t) + C$$

$$\sigma(t) = \frac{R}{\sqrt{3}}\text{sech}(\phi_0(1-t) + \phi_1 t),$$

where

$$R^2 = (\frac{\mu_0 - \mu_1}{2})^2 + \frac{3}{2}(\sigma_1^2 + \sigma_0^2) + \frac{9}{4}(\frac{\sigma_1^2 - \sigma_0^2}{\mu_1 - \mu_2})^2$$

$$C = \frac{\mu_0 + \mu_1}{2} + \frac{3}{2}\frac{\sigma_1^2 - \sigma_0^2}{\mu_1 - \mu_0}$$

$$\phi_t = \tanh^{-1}\left(\frac{\mu_t - C}{R}\right), \text{for t = 1,...,d.}$$

Now that the optimal path $\theta(t)$ is found, it is possible to write a program to find a path sampling estimator for λ. For each path sampling estimate $\hat{\lambda}_i$ draw firstly a t_i uniformly from [0,1], then draw ω_i from $N(\mu(t_i), \sigma(t_i)^2)$ and estimate $\hat{\lambda}$ as

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\hat{\lambda}_i = \frac{1}{n}\sum_{i=1}^{n}\left[\dot{\mu}(t_i)\frac{\omega_i - \mu(t_i)}{\sigma(t_i)^2} + \dot{\sigma}(t_i)\frac{(\omega_i - \mu(t_i))^2}{\sigma(t_i)^3}\right]$$

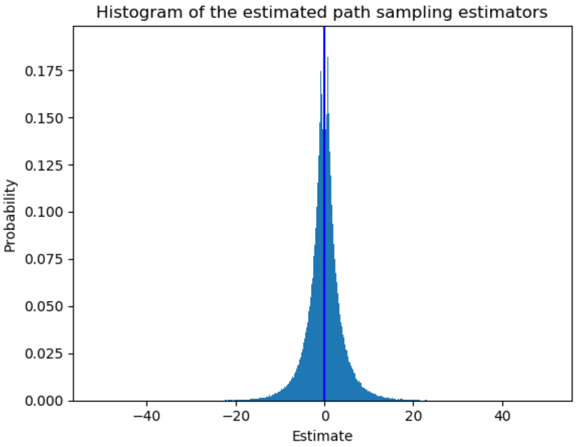


Figure 3.1: Histogram of the $\hat{\lambda}_i$'s

We use startpoint $\theta_0 = (0, 1)$ and endpoint $\theta_1 = (5, 1)$ are used. In Figure 3.1 one million path sampling estimates are plotted in the histogram (the script could be found at [6]). It could be noticed that the distribution of the $\hat{\lambda}_i$'s looks symmetric around 0. Also $\hat{\lambda}$ gives a value quite close to 0, namely -0.0036877. This agrees with the expected logarithmic ratio of the normalizing constants of two normal densities, as the theoretical normalizing constant of a normal is always $\sqrt{2\pi}$ and thus $\lambda = \log(\frac{\sqrt{2\pi}}{\sqrt{2\pi}}) = 0$. With the help of the Central Limit Theorem, we can state that the estimated value of -0.0036877 falls within the 95% confidence interval, which ranges from -0.0196 to 0.0196. The error, square root of the variance of the estimator, is given by [5]

$$\sqrt{n\,\mathbb{E}(\hat{\lambda} - \lambda)^2} = \sqrt{12}[\log(\frac{D}{\sqrt{12}} + \sqrt{1 + \frac{D^2}{12}})] = 4.0853,$$

where D is the difference between the mean values of the two normal distributions. In our code is found that the square root of the variance is $\hat{\lambda} = 4.0394$

3.1.4 Example Path Sampling for two t-distributions

Let's take a look at a similar example. Now will be searched for an estimate of the ratio of normalizing constants of two t-distributions. It is known that the probability density function of the t-distribution is given by

$$p(\omega|\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1 + \frac{\omega^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where ν is the degrees of freedom and $\Gamma(\cdot)$ is the gamma function. We will consider $\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}$ as the normalizing constant and look whether we can approximate the logarithmic ratio with a path sampling estimate. The unnormalized density function is here thus given by

$$q(\omega|\nu) = \left(1 + \frac{\omega^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

The path in this example is denoted as $\nu(t)$. As minimizing the variance of the path sampling estimator will cause a great calculation mess, a linear path will be used in this example, $\nu(t) = (1 - t)\nu_0 + t\nu_1$. The unbiased path sampling estimator for the logarithmic ratio of normalizing constants is here given as

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\frac{U(\omega_i, \nu_i)}{p(\nu_i)}.$$

We can calculate the potential energy

$$\begin{aligned} U(\omega_i,\nu_i) &= \frac{\partial}{\partial\nu}\log q(\omega_i,\nu_i) \\ &= \frac{\partial}{\partial\nu} - \frac{\nu+1}{2}\log(1+\frac{\omega^2}{\nu}) \\ &= \frac{\omega^2(\nu+1)}{2\nu^2(\frac{\omega^2}{\nu}+1)} - \frac{\log(\frac{\omega^2}{\nu}+1)}{2} \end{aligned}$$

As the prior distribution is a uniform distribution over the path, $p(\nu_i) = \frac{1}{\nu_1-\nu_0}$ and the path sampling estimator

$$\hat{\lambda}_i = \left(\frac{\omega_i^2(\nu(t_i)+1)}{2\nu(t_i)^2(\frac{\omega_i^2}{\nu(t_i)}+1)} - \frac{\log(\frac{\omega_i^2}{\nu(t_i)}+1)}{2}\right)(\nu_1-\nu_0)$$

For this example let the begin degrees of freedom $\nu_0 = 2$ and end degrees of freedom $\nu_1 = 3$. Generating 100000 path sampling estimates gave $\hat{\lambda} = -0.0390323$ [6], which is again close to the theoretical lambda $\lambda = -0.0388319$ and in the 95% confidence level interval. In Figure 3.2 note that the path sampling estimates are not symmetrically distributed around the theoretical lambda, which was the case for the normal densities example. This is consistent with the characteristics of the t-distribution, which is known for its non-symmetrical distribution around the mean and heavier tails.

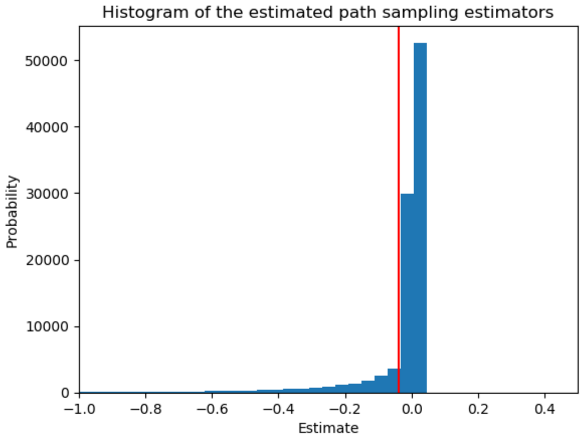


Figure 3.2: Histogram of the $\hat{\lambda}_i$'s

3.2 Importance Sampling

In the past Section is found that path sampling relies on a path of intermediate densities. Importance sampling relies on the target density, or a related density. In this Section, two importance sampling methods will be discussed. The first importance sampling method uses a trial density which is completely known and different from the target density. This could be for example an analytical approximation. The second importance sampling method does not use an trial density, but relies on draws of the unnormalized density.

3.2.1 Importance Sampling estimator

Using the proposal density $\tilde{p}(\omega) \approx p(\omega) = \dfrac{q(\omega)}{z}$, we find

$$\begin{aligned} z &= \int q(\omega)d\omega \\ &= \int \frac{q(\omega)}{\tilde{p}(\omega)}\tilde{p}(\omega)d\omega \\ &= \mathbb{E}_{\tilde{p}(\omega)}[\frac{q(\omega)}{\tilde{p}(\omega)}]. \end{aligned}$$

This gives importance sampling estimator $\hat{z} = \dfrac{1}{n}\sum_{i=1}^{n}\dfrac{q(\omega_i)}{\tilde{p}(\omega_i)}$, where $\omega_1, ..., \omega_n$ are draws of $\tilde{p}(\omega)$. This method depends heavily on the choice of the proposal density and it is thus important to find a good proposal density. The second importance sampling method searches an estimate for the ratio of normalizing constants of $q_0(\omega)$ and $q_1(\omega)$. It will be shown that

$$\frac{\mathbb{E}_0[q_1(\omega)\alpha(\omega)]}{\mathbb{E}_1[q_0(\omega)\alpha(\omega)]} = \frac{\int q_1(\omega)\alpha(\omega)\frac{q_0(\omega)}{z_0}d\omega}{\int q_0(\omega)\alpha(\omega)\frac{q_1(\omega)}{z_1}d\omega}$$

is equal to the ratio of normalizing constants. Here is $\mathbb{E}_0$ the expectation with respect to $p_0(\omega) = \dfrac{q_0(\omega)}{z_0}$ and $\mathbb{E}_1$ the expectation with respect to $p_1(\omega) = \dfrac{q_1(\omega)}{z_1}$.

$$\begin{aligned} \frac{\mathbb{E}_0[q_1(\omega)\alpha(\omega)]}{\mathbb{E}_1[q_0(\omega)\alpha(\omega)]} &= \frac{\int q_1(\omega)\alpha(\omega)\frac{q_0(\omega)}{z_0}d\omega}{\int q_0(\omega)\alpha(\omega)\frac{q_1(\omega)}{z_1}d\omega} \\ &= \frac{z_1}{z_0}\frac{\int q_0(\omega)q_1(\omega)\alpha(\omega)}{\int q_0(\omega)q_1(\omega)\alpha(\omega)} \\ &= \frac{z_1}{z_0}. \end{aligned} \tag{3.3}$$

Here is $\alpha(\omega)$ an arbitrary function which satisfies that $0 < \left|\int_{\Omega_0\cap\Omega_1}\alpha(\omega)p_0(\omega)p_1(\omega)d\omega\right| < \infty$, where respectively Ω_0 and Ω_1 are the support of $p_0(\omega)$ and $p_1(\omega)$. Now with the draws $(\omega_{0i}, i = 1, ..., n_0)$ from $p_0(\omega)$ and the draws $(\omega_{1i}, i = 1, ..., n_1)$ from $p_1(\omega)$ we obtain importance sampling estimator

$$\hat{r}_\alpha = \frac{\frac{1}{n_0}\sum_{i=1}^{n_0} q_1(\omega_{0i})\alpha(\omega_{0i})}{\frac{1}{n_1}\sum_{i=1}^{n_1} q_0(\omega_{1i})\alpha(\omega_{1i})}. \tag{3.4}$$

3.2.2 Finding the optimal $\alpha(\omega)$

As $\alpha(\omega)$ is an arbitrary function in (3.4), it is interesting to search for the function that minimizes the variance of this estimator. We search for the minimal relative square error. The numerator of (3.4) is denoted as $\bar{\eta}_0$ and the denominator as $\bar{\eta}_1$. Similarly, for the theoretic ratio of normalizing constants define η_0 and η_1 to be the numerator and denominator respectively. The relative square error is then given by

$$\begin{aligned} RE^2(\hat{r}_\alpha) &= \frac{\mathbb{E}[(\hat{r}_\alpha - r)^2]}{r^2} \\ &= \frac{\mathbb{E}\left[\left(\frac{\bar{\eta}_0}{\bar{\eta}_1} - \frac{\eta_0}{\eta_1}\right)^2\right]}{\left(\frac{\eta_0}{\eta_1}\right)^2} \end{aligned}$$

Using the independence of $\bar{\eta}_0$ and $\bar{\eta}_1$, this can be rewritten to

$$RE^2(\hat{r}_\alpha) = \frac{\mathbb{V}\mathrm{ar}(\bar{\eta}_0)}{\eta_0} + \frac{\mathbb{V}\mathrm{ar}(\bar{\eta}_1)}{\eta_1} + \mathcal{O}(\frac{1}{n^2}).$$

We find

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\bar{\eta}_0) &= \mathbb{V}\mathrm{ar}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} q_1(\omega_{0i})\alpha(\omega_{0i})\right) \\
&= \frac{1}{n_0}(\mathbb{E}_0(\bar{\eta}_1^2) - \mathbb{E}_0(\bar{\eta}_1)^2) \\
&= \frac{1}{n_0}\left[\int p_0(\omega)q_1^2(\omega)\alpha^2(\omega)d\omega - \left(\int p_0(\omega)q_1(\omega)\alpha(\omega)d\omega\right)^2\right] \\
&= \frac{z_1^2}{n_0}\left[\int p_0(\omega)p_1^2(\omega)\alpha^2(\omega)d\omega - \left(\int p_0(\omega)p_1(\omega)\alpha(\omega)d\omega\right)^2\right].
\end{aligned}$$

Similarly,

$$\mathbb{V}\mathrm{ar}(\bar{\eta}_1) = \frac{z_0^2}{n_1}\left[\int p_0^2(\omega)p_1(\omega)\alpha^2(\omega)d\omega - \left(\int p_0(\omega)p_1(\omega)\alpha(\omega)d\omega\right)^2\right].$$

Hence, combining these variances gives the following equation for the relative square error that needs to be minimised.

$$\begin{aligned}
RE^2(\hat{r}_\alpha) &= \frac{\frac{z_1^2}{n_0}\left[\int p_0(\omega)p_1^2(\omega)\alpha^2(\omega)d\omega - \left(\int p_0(\omega)p_1(\omega)\alpha(\omega)d\omega\right)^2\right]}{z_1^2\left(\int p_0(\omega)p_1(\omega)\alpha(\omega)d\omega\right)^2} + \frac{\frac{z_0^2}{n_1}\left[\int p_0^2(\omega)p_1(\omega)\alpha^2(\omega)d\omega - \left(\int p_0(\omega)p_1(\omega)\alpha(\omega)d\omega\right)^2\right]}{z_0^2\left(\int p_0(\omega)p_1(\omega)\alpha(\omega)d\omega\right)^2} \\
&= \frac{\int p_0(\omega)p_1^2(\omega)\alpha^2(\omega)d\omega}{n_0\left(\int p_0(\omega)p_1(\omega)\alpha(\omega)d\omega\right)^2} - \frac{1}{n_0} + \frac{\int p_0^2(\omega)p_1(\omega)\alpha^2(\omega)d\omega}{n_1\left(\int p_0(\omega)p_1(\omega)\alpha(\omega)d\omega\right)^2} - \frac{1}{n_1} \\
&= \frac{\int p_0(\omega)p_1(\omega)\alpha^2(\omega)\left(\frac{p_1(\omega)}{n_0} + \frac{p_0(\omega)}{n_1}\right)d\omega}{\left(\int p_0(\omega)p_1(\omega)\alpha(\omega)d\omega\right)^2} - \frac{1}{n_0} - \frac{1}{n_1},
\end{aligned}$$

where the error term of $\mathcal{O}(\frac{1}{n^2})$ is neglected. Note that for minimizing the relative square error, $\frac{1}{n_0}$ and $\frac{1}{n_1}$ can be left out as these are not dependent of $\alpha(\omega)$. Using Cauchy-Schwartz (Section 2.2), follows that the denominator

$$\left(\int p_0(\omega)p_1(\omega)\alpha(\omega)d\omega\right)^2 \leq \int \frac{p_0(\omega)p_1(\omega)}{\frac{p_1(\omega)}{n_0} + \frac{p_0(\omega)}{n_1}}d\omega \int p_0(\omega)p_1(\omega)\alpha^2(\omega)\left(\frac{p_0(\omega)}{n_1} + \frac{p_1(\omega)}{n_0}\right)d\omega.$$

An equality (and thus have minimized the relative square error with respect to $\alpha(\omega)$) is found when

$$p_0p_1\alpha^2\left(\frac{p_0}{n_1} + \frac{p_1}{n_0}\right) \propto \frac{p_0p_1}{\frac{p_0}{n_1} + \frac{p_1}{n_0}}$$

Thus

$$\begin{aligned}
\alpha_{\text{opt}} &\propto \frac{1}{\frac{p_1}{n_0} + \frac{p_0}{n_1}} \\
&\propto \frac{1}{rs_0q_0 + s_1q_1}.
\end{aligned} \tag{3.5}$$

Here is $s_0 = \dfrac{n_0}{n_0 + n_1}$ and $s_1 = \dfrac{n_1}{n_0 + n_1}$. Now the problem is that r is not known and α_{opt} cannot be used directly. Substituting $\alpha_{\text{opt}}(\omega)$ in the importance sampling estimator (3.4) gives

$$\begin{aligned}\hat{r}_{\alpha_{\text{opt}}} &= \frac{\frac{1}{n_0}\sum_{i=1}^{n_0} \frac{q_1(\omega_{0i})}{rs_0q_0(\omega_{0i}) + s_1q_1(\omega_{0i})}}{\frac{1}{n_1}\sum_{i=1}^{n_1} \frac{q_0(\omega_{1i})}{rs_0q_0(\omega_{1i}) + s_1q_1(\omega_{1i})}} \\ &= \frac{\frac{1}{n_0}\sum_{i=1}^{n_0} \frac{l_{0i}}{rs_0 + s_1l_{0i}}}{\frac{1}{n_1}\sum_{i=1}^{n_1} \frac{1}{rs_0 + s_1l_{1i}}},\end{aligned} \tag{3.6}$$

where $l_{mi} = \dfrac{q_1(\omega_{mi})}{q_0(\omega_{mi})}$. Using an initial guess $\hat{r}^{(0)} > 0$ and the fixed point iteration method, the importance sampling estimate can be calculated iteratively.
The fixed point iteration method gives

$$\hat{r}_{\alpha}^{(t+1)} = \frac{\frac{1}{n_0}\sum_{i=1}^{n_0} \frac{l_{0i}}{\hat{r}^{(t)}s_0 + s_1l_{0i}}}{\frac{1}{n_1}\sum_{i=1}^{n_1} \frac{1}{\hat{r}^{(t)}s_0 + s_1l_{1i}}}, \tag{3.7}$$

this will be called the fixed point estimator.
Now the question arises whether this fixed point estimator converges and whether it has the same error as the original predictor. Meng and Wong (1996) found that for any initial guess $\hat{r}^0$, the fixed point estimator converges to a unique limit $\hat{r}_F$ with the following property for $\hat{r}_F^{(t)} \neq \hat{r}_F$

$$|\hat{r}_F^{(t+1)} - \hat{r}_F| < |\hat{r}_F^{(t)} - \hat{r}_F| \tag{3.8}$$

To prove this identity, we will first look at the limit. When the limit exists, there should be some point, for a t large enough, that $\hat{r}^{(t+1)} = \hat{r}^{(t)}$. Setting these terms equal to r, define

$$S(r|\omega) = \sum_{i=1}^{n_0} \frac{s_1q_1(\omega_{0i})}{rs_0q_0(\omega_{0i}) + s_1q_1(\omega_{0i})} - \sum_{i=1}^{n_1} \frac{rs_0q_0(\omega_{1i})}{rs_0q_0(\omega_{1i}) + s_1q_1(\omega_{1i})}.$$

This formula could be obtained by rewriting (3.7) as $r = \dfrac{A(r)}{B(r)} \iff rA(r) - B(r) = 0$, where $A(r)$ and $B(r)$ denote respectively the numerator and denominator. Thus, the root of $S(r|\omega)$ is equal to the limit of $\hat{r}^{(t)}$, when this converges.

$$\frac{dS(r|\omega)}{dr} = -\sum_{i=1}^{n_0} \frac{s_0s_1q_0(\omega_{0i})q_1(\omega_{0i})}{(rs_0q_0(\omega_{0i}) + s_1q_1(\omega_{0i}))^2} - \sum_{i=1}^{n_1} \frac{s_0s_1q_0(\omega_{1i})q_1(\omega_{1i})}{(rs_0q_0(\omega_{1i}) + s_1q_1(\omega_{1i}))^2}$$

Without loss of generality, it can be assumed that $q_0(\omega), q_1(\omega) \geq 0$. Thus $\dfrac{dS(r|\omega)}{dr} < 0, \forall r$ and as $S(0|\omega) = n_2 > 0$ and $S(\infty|\omega) = -n_1 < 0$, $S(r|\omega)$ is thus a strictly increasing function of r. We can thus conclude that $S(r|\omega)$ has a unique root. To finally prove (3.8), we use following properties: $rM(r)$ is strictly decreasing and $\dfrac{M(r)}{r}$ is strictly increasing in r. To show this differentiate these functions and then $\dfrac{d(rM(r))}{r} < 0$ and $\dfrac{d(\frac{M(r)}{r})}{r} > 0$.
When $\hat{r}^{(t)} > r^*$, we find

$$\begin{aligned}\frac{\hat{r}^{(t+1)}}{\hat{r}^{(t)}} = \frac{M(\hat{r}^{(t)})}{\hat{r}^{(t)}} &< \frac{M(r^*)}{r^*} = 1 \\ \iff \hat{r}^{(t+1)} - r^* &< \hat{r}^{(t)} - r^*\end{aligned} \tag{3.9}$$

and

$$\begin{aligned}\hat{r}^{(t+1)}\hat{r}^{(t)} &= M(\hat{r}^{(t)})r^{(t)} > M(r^*)r^* = (r^*)^2 \\ \iff \hat{r}^{(t+1)} - r^* &> \frac{(r^*)^2}{\hat{r}^{(t)}} - r^* \\ &= \frac{r^*}{\hat{r}^{(t)}}(r^* - \hat{r}^{(t)}) > r^* - r^{(t)}\end{aligned} \tag{3.10}$$

Now with combining (3.9) and (3.10), could be concluded that

$$|\hat{r}_F^{(t+1)} - \hat{r}_F| < |\hat{r}_F^{(t)} - \hat{r}_F|.$$

Analogous arguments give the same result for $\hat{r}^{(t)} < r^*$. Now the Global Convergence Theorem implies that $\hat{r}^{(t)}$ converges to r^*.

$\mathbb{E}(\hat{r} - r)^2 = \dfrac{\mathbb{V}\mathrm{ar}(S(r|\omega))}{\mathbb{E}^2(S'(r|\omega))} + \mathcal{O}(\dfrac{1}{n^2})$

Meng and Wong (1992) also found that the asymptotic relative mean square error is

$$RE^2(\hat{r}_F) = \frac{1}{n}\left[\int \frac{1}{\frac{p_1}{n_0} + \frac{p_0}{n_1}} d\omega - \frac{1}{n_0} - \frac{1}{n_1}\right]$$

The asymptotic relative mean square error is equal to the one we found earlier with optimal $\alpha(\omega)$, but which was not usable in practice as the estimator was dependent of r.

3.2.3 Example iterative importance sampling for two t-distributions

In Section 3.1.4 a path sampling estimate for the logarithmic ratio of normalizing constants is found for two t-distributions. In the same setting we will search for an importance sampling estimate such that we can compare the two sampling methods.

Testing some simulations [6], we note that r converges really quick in just five iterations even if an initial guess is used which is far off. It is also noted that the estimate importance sampling estimate of r has some outliers and a single estimate is thus not reliable. For the final script is the iterative importance sampling method used 200 times with 15 iterations and initial guess of $r^{(0)} = 100$. After that is averaged over the founded $\hat{r}$'s. In the calculation of each of the 200 $\hat{r}$'s in (3.7), 1000 draws are taken from the t-distribution with 2 degrees of freedom and 1000 draws are taken from the t-distribution with 3 degrees of freedom. Averaging these 200 $\hat{r}$'s gives importance sampling estimate $\hat{r} = 0.96241096$. Now $\hat{\lambda} = \log(\hat{r}) \approx -0.03831$, which is a little bit further off the theoretical $\lambda = -0.03883$ than the path sampling estimate of Section 3.1.4.

To improve the accuracy of our estimate should be averaged over more than 200 $\hat{r}$'s.

3.3 Bridge Sampling

The last of the three sampling methods that will be discussed is bridge sampling. Bridge sampling can be seen as an combination of path sampling and importance sampling. The idea of bridge sampling is that it uses an arbitrary unnormalized density $q_{1/2}(\omega)$ 'between' the unnormalized densities $q_0(\omega)$ and $q_1(\omega)$. In the setting of (3.3), is $\alpha(\omega)$ for bridge sampling defined as

$$\alpha(\omega) = \frac{q_{1/2}(\omega)}{q_0(\omega)q_1(\omega)}.$$

In Section (3.2.2) is already found how to choose the optimal $\alpha(\omega)$. This implies that

$$\begin{aligned} q_{1/2}^{\text{opt}}(\omega) &= \alpha_{\text{opt}}(\omega)q_0(\omega)q_1(\omega) \\ &= \frac{p_0(\omega)p_1(\omega)z_0z_1}{s_0p_0(\omega) + s_1p_1(\omega)} \\ &\propto \frac{p_0(\omega)p_1(\omega)}{s_0p_0(\omega) + s_1p_1(\omega)} \end{aligned}$$

The optimal $q_{1/2}(\omega)$ is thus the harmonic mean of $p_0(\omega)$ and $p_1(\omega)$. Now using this optimal bridge in $\alpha(\omega)$, (3.4) gives that

$$\hat{r} = \frac{\frac{1}{n_0}\sum_{i=1}^{n_0} \frac{q_{1/2}(\omega_{0i})}{q_0(\omega_{0i})}}{\frac{1}{n_1}\sum_{i=1}^{n_1} \frac{q_{1/2}(\omega_{1i})}{q_1(\omega_{1i})}}$$

3.4 Overview of the sampling methods

Now that three different sampling methods are discussed arises the questions when does one use which method and which method is the best. Firstly, path sampling was discussed. For path sampling, a path needs to be defined. This choice of this path is crucial as it affects the accuracy of the path sampling estimate. By using multiple paths and averaging the estimates, the variance of the estimate could be reduced. With n draws (ω_i, θ_i) from the target distribution $p(\omega, \theta)$, the path sampling estimate of the logarithmic ratio of normalizing constants is given

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \frac{U(\omega_i, \theta_i)}{p(\theta_i)}.$$

This method is thus useful when an efficient path could be chosen without too much difficulties.

Another sampling method that is explored is importance sampling. When it is possible to sample from a proposal distribution that is closely related to the target distribution, but it is not possible to sample from the target distribution, importance sampling is useful. By sampling from the proposal distribution and reweighting the samples, importance sampling gives the estimate

$$\hat{z} = \frac{1}{n} \sum_{i=1}^{n} \frac{q(\omega)}{\tilde{p}(\omega)}.$$

The accuracy of importance sampling is highly dependent on the proposal distribution.

Lastly, bridge sampling is discussed. Bridge is a combination of path sampling and importance sampling. By using importance sampling, the variance of the estimate using only a path sampling approach is reduced. The accuracy of the estimate is thus less dependent on the choice of the path in comparison with path sampling. The estimate given with bridge sampling is

$$\hat{r} = \frac{\frac{1}{n_0} \sum_{i=1}^{n_0} \frac{q_{1/2}(\omega_{0i})}{q_0(\omega_{0i})}}{\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{q_{1/2}(\omega_{1i})}{q_1(\omega_{1i})}}.$$

Chapter 4

Censored Data

In this Chapter, a Gaussian field $W(x)$ will be considered. On this field all points are normally distributed. Conditioned on the observed points, an estimate could be found for an unobserved point. The observations are censored, only the positive values are observed and the negative values are set to 0. So the observed data is $Z(x) = W(x)^+ = \max(W(x), 0)$. The goal of this chapter is to determine numerically the conditional distribution of an unobserved point conditioned on the observed points. The same approach will be used as in Stein (1992).

In Section 4.1 all needed terminology will be defined for this Chapter. In Section 4.2 the conditional distribution will be calculated under the assumption that the model from which Gaussian field is generated is known. Two stochastic methods from Stein (1992) will be given to calculate this conditional distribution numerically. In Section 4.3 these two methods are being tested. From Section 4.4 on will be assumed that we do not have the knowledge from which model the data is generated. We use the Bayes factor to search for the best fitting model. In Section 4.5 a method from Stein (1992) will be introduced to numerically approach the Bayes factor. Lastly, in Section 4.6 are the results of this numerical method placed.

4.1 Setting

For the sake of notation suppose that $x_1, ..., x_n$ are ordered so that $z(x_1), ..., z(x_m)$ are positive and $z(x_{m+1}) = ... = z(x_n) = 0$. In this Chapter, n stands for the number of observations of the field, m is the number of positive observed values and thus is $n - m$ the number of censored values. Define $\mathbf{u} = (w(x_1), ..., w(x_m)) = (z(x_1), ..., z(x_m))$ and $\mathbf{U} = (W(x_1), ..., W(x_m))$. Similarly, $\mathbf{v} = (w(x_{m+1}), ..., w(x_n))$ and $\mathbf{V} = (W(x_{m+1}), ..., W(x_n))$. $\boldsymbol{U}$ is thus the positive part or known part of the data field and $\boldsymbol{V}$ is the censored part. The challenge is to estimate an arbitrary point on the field knowing that $\boldsymbol{U} = \boldsymbol{u}$ and $\boldsymbol{V} \leq \mathbf{0}$. A normal distribution conditioned on an other normal distributions is again a normal distribution (Bishop (2006)). We can calculate the mean and variance of $W(x)$ conditioned on $\mathbf{U} = \mathbf{u}$ and $\mathbf{V} = \mathbf{v}$ as follows

$$\mu_{W(x)|\mathbf{U}=\mathbf{u},\mathbf{V}=\mathbf{v}} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \tag{4.1}$$

and

$$\sigma^2_{W(x)|\mathbf{U}=\mathbf{u},\mathbf{V}=\mathbf{v}} = \Sigma_{aa} - \Sigma_{ba}\Sigma_{bb}^{-1}\Sigma_{ba}. \tag{4.2}$$

In these equations is the unknown part denoted by a and the known part is denoted by b. μ_a and μ_b are given by the law of the random field $W(\cdot)$. The Σ's are the covariance matrices. As example, we should have in this calculation that Σ_{bb} is an nxn covariance matrix filled with all covariance betwee all observed points on the field, as we have n known points.

4.2 Calculating the conditional distribution

Knowing the distribution of $w(x)$ conditioned on $\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} = \boldsymbol{v}$, we calculate the distribution of $W(x)$ conditioned on $\mathbf{U} = \mathbf{u}, \mathbf{V} \leq \mathbf{0}$, which is the restriction of the multivariate normal distribution to the negative orthant of $\mathbb{R}^{n-m}$.

$$\mathbb{P}(W(x) \leq t|\mathbf{U} = \mathbf{u}, \mathbf{V} \leq \mathbf{0}) = \frac{\int_{(-\infty,0]^{n-m}} \Phi\big(\frac{t - \mu_{W(x)|\mathbf{U}=\mathbf{u},\mathbf{V}=\mathbf{v}}}{\sigma_{W(x)|\mathbf{U}=\mathbf{u},\mathbf{V}=\mathbf{v}}}\big) p(\mathbf{v}|\mathbf{u}) d\nu}{\int_{(-\infty,0]^{n-m}} p(\mathbf{v}|\mathbf{u}) d\nu}, \tag{4.3}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. As these integrals are of dimension n-m, they are often intractable to calculate analytically and a numerical approach to approximate these integrals should be used. Stein (1992) proposes two methods.
Method 1:
Generate $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ from $N(\boldsymbol{\mu}, T)$, where $N(\boldsymbol{\mu}, T)$ is the conditional distribution of $W(x)$ given only $\boldsymbol{U} = \boldsymbol{u}$. The $\boldsymbol{v}_j$'s are thus simulations of the censored points. The naive estimate of (4.3) is then given by

$$\frac{\sum_{j=1}^{N} \Phi\big(\frac{t - a - \boldsymbol{b}'\boldsymbol{v}_j}{\tau}\big) \mathbb{1}_{\{\boldsymbol{v}_j \leq \mathbf{0}\}}}{\sum_{j=1}^{N} \mathbb{1}_{\{\boldsymbol{v}_j \leq \mathbf{0}\}}}. \tag{4.4}$$

The a and $\boldsymbol{b}'$ are respectively the dependencies on the conditional knowing $\boldsymbol{U} = \boldsymbol{u}$ and $\boldsymbol{V} = \boldsymbol{v}$ of $\boldsymbol{u}$ and $\boldsymbol{v}$. The expressions for them will be derived for simplicity in the case that we have a Gaussian field with expectation zero as we will also use this in the example of Section 4.3. The expressions follow by splitting the covariance matrices in block matrices with a '$\boldsymbol{u}$-part' and a '$\boldsymbol{v}$-part'. From (4.1) we find that

$$\begin{aligned}
\mu_{W(x)|\mathbf{U}=\mathbf{u},\mathbf{V}=\mathbf{v}} &= \Sigma_{ab}\Sigma_{bb}^{-1}\boldsymbol{x}_b \\
&= \begin{pmatrix} \Sigma_{tu} & \Sigma_{tv} \end{pmatrix} \begin{pmatrix} \Sigma_{uu} & \Sigma_{uv} \\ \Sigma_{vu} & \Sigma_{vv} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{x}_u \\ \boldsymbol{x}_v \end{pmatrix} \\
&= \begin{pmatrix} \Sigma_{tu} & \Sigma_{tv} \end{pmatrix} \begin{pmatrix} [\Sigma_{uu} - \Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu}]^{-1} & -[\Sigma_{uu} - \Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu}]^{-1}\Sigma_{uv}\Sigma_{vv}^{-1} \\ -[\Sigma_{vv} - \Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uv}]^{-1}\Sigma_{vu}\Sigma_{uu}^{-1} & [\Sigma_{vv} - \Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uv}]^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_u \\ \boldsymbol{x}_v \end{pmatrix} \\
&= \underbrace{(\Sigma_{tu}([\Sigma_{uu} - \Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu}]^{-1}) + \Sigma_{tv}(-[\Sigma_{vv} - \Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uv}]^{-1}\Sigma_{vu}\Sigma_{uu}^{-1}))\boldsymbol{x}_u}_{=a} \\
&+ \underbrace{(\Sigma_{tu}(-[\Sigma_{uu} - \Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu}]^{-1}\Sigma_{uv}\Sigma_{vv}^{-1}) + \Sigma_{tv}([\Sigma_{vv} - \Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uv}]^{-1}))}_{=\boldsymbol{b}}\boldsymbol{x}_v
\end{aligned} \tag{4.5}$$

Method 2:
The second method is an importance sampling method (see Section 3.2.). Now generate $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_N$ under some proposal density $h(\cdot)$, which has support on $(-\infty, 0]^{n-m}$. Then the importance estimate of (4.3) is given by

$$\frac{\sum_{j=1}^{N} \Phi\big(\frac{t - a - \boldsymbol{b}'\boldsymbol{v}_j}{\tau}\big) \frac{p(\mathbf{v}_j|\mathbf{u})}{h(\mathbf{v}_j)}}{\sum_{j=1}^{N} \frac{p(\mathbf{v}_j|\mathbf{u})}{h(\mathbf{v}_j)}} \tag{4.6}$$

The method to generate the $\boldsymbol{v}_j$'s is desribed in Section 4.2.1 and the choice of $h(\cdot)$ is given in Section 4.2.2.

4.2.1 Generating the $\mathbf{v}_j$'s

In method 2, we generate each $\mathbf{v}_j$ point for point under the truncated conditional univariate normal distribution given the positive observations and previously generated censored observations. Notationally, vector $\mathbf{v}_j^q = (v_{1j}, \ldots, v_{qj})$ is built given $\mathbf{U} = \mathbf{u}$ and $W(x_{m+k}) = v_{kj}$ for $1 \leq k \leq q-1$. We have $W(x_{m+q}) \sim N(\mu_{qj} = \alpha_q + \boldsymbol{\beta}_q' \boldsymbol{v}_j^{q-1}, \sigma_q^2)$, where α_q and $\boldsymbol{\beta}_q$ are again respectively the dependency of the positive observed points and the dependency of the censored points. They can be calculated in the same we as is done for a and $\boldsymbol{b}$ in (4.5). As it is known that all $\boldsymbol{v}_j$'s should be smaller or equal than $\mathbf{0}$, a method is needed which generates only negative values from $N(\mu_{qj}, \sigma_q^2)$. The following inverse CDF method is used

$$v_{qj} = \mu_{qj} + \sigma_q \Phi^{-1}\big(U_{qj}\Phi\big(-\frac{\mu_{qj}}{\sigma_q}\big)\big), \tag{4.7}$$

where the U_{qj}'s are arbitrary draws from Unif[0,1]. The standard inverse CDF method states that $\mu_{qj} + \sigma_q \Phi^{-1}(U_{qj})$ gives a draw from $N(\mu_{qj}, \sigma_q^2)$. Note that $U_{qj}\Phi(-\frac{\mu_{qj}}{\sigma_{qj}})$ is a random draw from Unif[0,$\Phi(-\frac{\mu_{qj}}{\sigma_{qj}})$].

Thus as Φ^{-1} is monotonically increasing we find that

$$U_{qj}\Phi(-\frac{\mu_{qj}}{\sigma_{qj}}) \leq \Phi(-\frac{\mu_{qj}}{\sigma_{qj}}) \iff \Phi^{-1}(U_{qj}\Phi(-\frac{\mu_{qj}}{\sigma_{qj}})) \leq \Phi^{-1}(\Phi(-\frac{\mu_{qj}}{\sigma_{qj}})) = -\frac{\mu_{qj}}{\sigma_{qj}}$$

Hence are all v_{qj} generated from (4.7) indeed smaller or equal than 0.

4.2.2 Importance Estimate

In Stein (1992) are the importance weights defined as follows

$$d_j = \frac{p(\boldsymbol{v}_j|\boldsymbol{u})}{h(\boldsymbol{v}_j)} = \prod_{q=1}^{n-m} \Phi(-\frac{\mu_{qj}}{\sigma_q}),$$

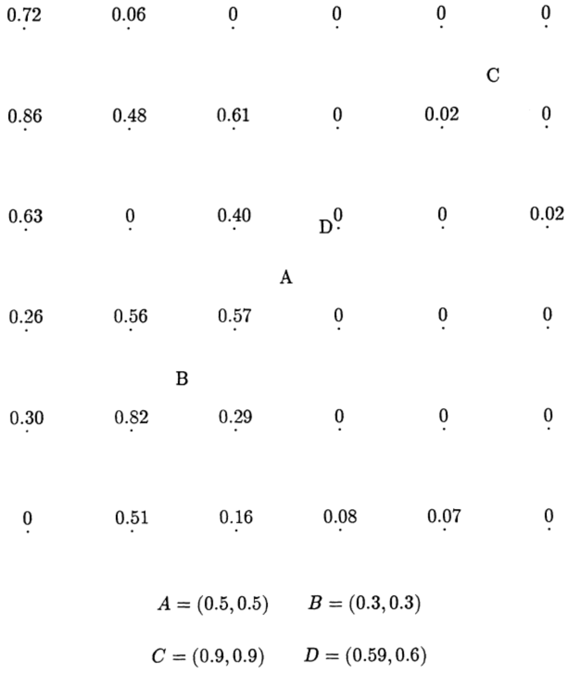
which gives the importance sampling estimate

$$\widehat{\mathbb{P}}_N(W(x) \leq t|\mathbf{U} = \mathbf{u}, \mathbf{V} \leq \mathbf{0}) = \frac{\sum_{j=1}^N d_j \Phi(\frac{t - \mu_{n+1,j}}{\tau})}{\sum_{j=1}^N d_j}, \tag{4.8}$$

where $\mu_{n+1,j} = a + \boldsymbol{b'}\boldsymbol{v}_j$.

4.3 Example Estimating Conditional Distribution

This example is taken over from Stein (1992) to look whether we can find the same results. The observed data from a truncated Gaussian random field $W(\cdot)$ is showed in Figure 4.1. This field is an $6x6$ square grid on $[0,1]^2$ with $\mathbb{E}(W(x)) = 0$ and $\mathbb{C}\text{ov}(W(x), W(x')) = e^{-|x-x'|}, \forall x, x' \in [0,1]^2$



0.72	0.06	0	0	0	0
0.86	0.48	0.61	0	0.02	0
0.63	0	0.40	0	0	0.02
0.26	0.56	0.57	0	0	0
0.30	0.82	0.29	0	0	0
0	0.51	0.16	0.08	0.07	0

$A = (0.5, 0.5)$ $B = (0.3, 0.3)$

$C = (0.9, 0.9)$ $D = (0.59, 0.6)$

Figure 4.1: Simulated Gaussian Field. Source: Stein (1992)

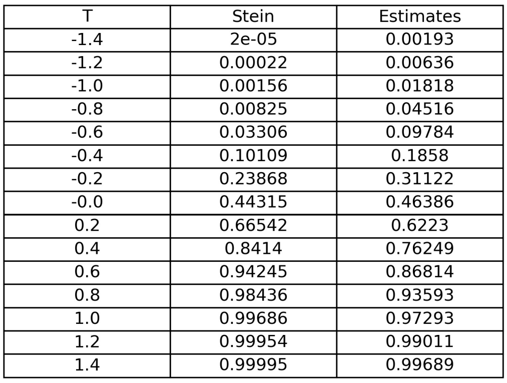
Using the techniques from Section 4.2, we can give a value to arbitrary points. We can for example calculate, whether point A has a value smaller than a specific bound.

4.3.1 Method 1

Generating $N = 10000$ random vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ from the conditional distribution knowing only $\boldsymbol{U} = \boldsymbol{u}$, we find that there is no vector $\boldsymbol{v}_j$ for which all components are less or equal than zero [6]. Also is found that approximately 36% of the components of all vectors $\boldsymbol{v}$ are smaller or equal than 0. This implies that the chances of getting a whole vector $\boldsymbol{v}$ of size $n - m = 17$ in our example smaller or equal than 0 are very small. Thus $\mathbb{1}_{\{\boldsymbol{v}_j \leq \boldsymbol{0}\}} = 0$ for all j in our example. Now we can not find an estimate for (4.3) and method 1 does thus not work in our example. When an example comes with less censored points $n - m$, is the probability of having some $\boldsymbol{v}_j$'s smaller or equal than zero higher and thus could this method work.

4.3.2 Method 2

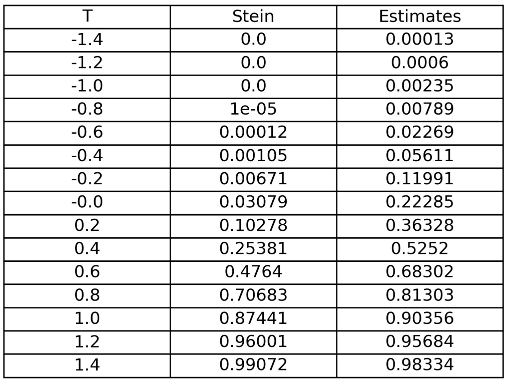
We implement method 2 with N = 10000 as in Stein (1992) and search for the estimate that $\mathbb{P}(W(A) \leq T | \boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} \leq \boldsymbol{0})$ for the T from which we have the estimates from Stein. To compare these estimates they are placed next to each other in Figure 4.2.



T	Stein	Estimates
-1.4	2e-05	0.00193
-1.2	0.00022	0.00636
-1.0	0.00156	0.01818
-0.8	0.00825	0.04516
-0.6	0.03306	0.09784
-0.4	0.10109	0.1858
-0.2	0.23868	0.31122
-0.0	0.44315	0.46386
0.2	0.66542	0.6223
0.4	0.8414	0.76249
0.6	0.94245	0.86814
0.8	0.98436	0.93593
1.0	0.99686	0.97293
1.2	0.99954	0.99011
1.4	0.99995	0.99689

Figure 4.2: Estimates of $W(A)$ in comparison to the estimates from Stein. [6]

For small values of T is found that our estimates give an higher probability that $W(A)$ is below this T and for high values of T is found that our estimates give an lower probability that $W(A)$ is below this T. Our $W(A)$ has thus a higher variance than the $W(A)$ from Stein. This is also noted in Figure 4.3, where the similar estimates for point B from Figure 4.1 are placed next to each other.



T	Stein	Estimates
-1.4	0.0	0.00013
-1.2	0.0	0.0006
-1.0	0.0	0.00235
-0.8	1e-05	0.00789
-0.6	0.00012	0.02269
-0.4	0.00105	0.05611
-0.2	0.00671	0.11991
-0.0	0.03079	0.22285
0.2	0.10278	0.36328
0.4	0.25381	0.5252
0.6	0.4764	0.68302
0.8	0.70683	0.81303
1.0	0.87441	0.90356
1.2	0.96001	0.95684
1.4	0.99072	0.98334

Figure 4.3: Estimates of $W(B)$ in comparison to the estimates from Stein. [6]

4.4 Bayes Factor

In the preceding Sections is used that the model from which the points on the field are generated is known. In a practical situation is this often not available in advance.

From now on we will consider the model where $\mathbb{E}(W(x)) = m$ and $\mathbb{C}\text{ov}(W(x), W(x')) = cK_a(x - x')$. $K_a(\cdot)$ is a class of covariance functions with possibly vector valued a. Parameter $\theta = (m, c, a)'$ is unknown. The goal is now to find an estimate for the likelihood of θ. The likelihood is given by

$$\begin{aligned} l(\theta; \boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} \leq \boldsymbol{0}) &= p_\theta(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} \leq \boldsymbol{0}) \\ &= p_\theta(\boldsymbol{u}) p_\theta(\boldsymbol{V} \leq 0 | \boldsymbol{U} = \boldsymbol{u}) \\ &= p_\theta(\boldsymbol{u}) \int_{(-\infty,0]^{n-m}} p_\theta(\boldsymbol{v}|\boldsymbol{u}) d\boldsymbol{v}, \end{aligned}$$

where the subscript θ means the distribution using model θ. With respect to an reference value θ_0, we can calculate the Bayes factor for model θ as follows

$$\text{Bayes Factor} = \frac{l(\theta; \boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} \leq \boldsymbol{0})}{l(\theta_0; \boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} \leq \boldsymbol{0})} = \frac{p_\theta(\boldsymbol{u})}{p_{\theta_0}(\boldsymbol{u})} \frac{\int_{(-\infty,0]^{n-m}} p_\theta(\boldsymbol{v}|\boldsymbol{u}) d\boldsymbol{v}}{\int_{(-\infty,0]^{n-m}} p_{\theta_0}(\boldsymbol{v}|\boldsymbol{u}) d\boldsymbol{v}}. \tag{4.9}$$

To choose a reference value θ_0, for example a maximum likelihood estimation could be used.

4.5 Importance Sampling Estimate Ratio of Likelihood Function

Firstly it is needed to generate the $\boldsymbol{v}_j$'s. For each θ for which an estimate for (4.9) is wanted, new μ_{qj}'s and σ_q's need to be calculated for each model, as the μ_{qj}'s and σ_q's are dependent on the model θ. To compare between the different models are the U_{qj}'s equal for all θ.
Defining the importance weights as in Stein (1992)

$$d_j(\theta) = \prod_{q=1}^{n-m} \Phi(-\frac{\mu_{qj}(\theta)}{\sigma_q(\theta)})$$

gives importance sampling estimate

$$\frac{p_\theta(\boldsymbol{u})}{p_{\theta_0}(\boldsymbol{u})} \frac{\sum_{j=1}^N d_j(\theta)}{\sum_{j=1}^N d_j(\theta_0)}. \tag{4.10}$$

4.6 Example Estimating Likelihood Function

In this example the setting $\mathbb{E}(W(x)) = m$, $\mathbb{C}\text{ov}(W(x), W(x')) = ce^{-|x-x'|}$ for the simulated field on Figure 4.1 will be used. The reference values $m_0 = -0.07$ and $c_0 = 0.631$ for m and c are chosen as approximate maximum likelihood estimates based on preliminary simulations in Stein (1992). We will take this reference values over. As done by Stein, we make an 21x21 grid for m and $\log(c)$. The gridpoints for m are from -2 to 2 with steps from 0.2 and the gridpoints for $\log(c)$ are from -2 to 1 with steps of 0.15. For each $\theta = (m, \log(c))$ on this grid, we calculate

$$\mu_{qj}(\theta) = m + \Sigma_{ab}\Sigma_{bb}^{-1}(\boldsymbol{x}_b - \boldsymbol{m})$$

and

$$\sigma_q(\theta) = c - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba},$$

where the a denotes the points for which we compute the mean and deviation and b denotes all known points. The covariance matrices are dependent on c. Using the same inverse CDF method as in Section 4.2.1 we generate the v_{qj}'s. So, $v_{qj}(\theta) = \mu_{qj}(\theta) + \sigma_q(\theta)\Phi^{-1}\big(U_{qj}\Phi(-\frac{\mu_{qj}(\theta)}{\sigma_q(\theta)})\big)$. Then for each θ on the grid, we calculate the $d_j's$ and the distribution of $\boldsymbol{u}$ under model θ could be calculated. Now from (4.10) the importance sampling estimate of the Bayes factor for each θ on the grid could be calculated. The Bayes factor is found to be the greatest on the grid for $\theta = (0.2, -1.4)$ [6]. We thus conclude that the model $\theta = (0.2, -1.4)$ that fits the best to the data considering all models from the grid.

Chapter 5

Conclusion and Further Research

The goal of this research was to search for the best fitting model to known data x, such that accurate predictions can be made based on this model. The Bayes factor is used to compare between models. We saw that it is often not possible to calculate the Bayes factor analytically. This Bayes factor was written as the ratio of two integrals such that a sampling method can be used to find an estimate of these integrals.

In Chapter 3, three sampling methods are explored, path sampling, importance sampling and bridge sampling. These sampling methods can be used to estimate an integral, when the integral can not be calculated analytically. To find the best estimator for each of these methods is searched to minimize the variance of the estimator obtained by the sampling method. For example for the path sampling method, we used the Cauchy-Schwartz inequality to find a lower bound for the variance of the estimator. Choosing the path in a convenient way, we can equal the variance of the estimator to the found lower bound. Choosing the path in this way, we have thus minimized the variance with respect to the path and have found the most accurate estimator.

In Chapter 4, a Gaussian field is considered. The observations on this field are censored, as all negative values are observed as zeros. The conditional distribution is determined for the value of an arbitrary position on the field. In Section 4.2 was used that the covariance structure and the Gaussian distribution from which the points on the field get a value are known. In Section 4.4 was assumed that the covariance structure and the Gaussian distribution from which points on the field get a value is not known. Now before we can find the conditional distribution for the arbitrary points, a model which fits the best to the data has to be found. Searching the best fitting model to the data was also the goal of the research. We used the Bayes factor to determine, which model is the most likely to fit to the data. This Bayes factor was calculated using importance sampling.

5.1 Further Research

Several things could be improved in this research. It would be interesting to analyse the choice of the sampling method. Importance sampling was chosen in Stein (1992) to approximate the ratio of integrals from the Bayes factor. A couple of questions arise from this choice. How is the used proposal function for importance sampling found and could this function be improved? How accurate is the estimate that is found for estimating the Bayes factor using importance sampling? It is also interesting to look at the computation time of importance sampling, we found that for $N = 10$ in Section 4.6 the computation time was around half an hour. Stein used $N = 1000$. Are there other possible sampling methods that could be chosen to lower the computation time?

In this thesis are models compared using the Bayes factor. There are a lot of other methods that could be used to search for the best fitting model. Examples are the likelihood ratio test and the Akaike information criterion. Using several methods could make the conclusion stronger.

Bibliography

[1] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[2] Mary L. Boas. *Mathematical Methods in the Physical Sciences.* John Wiley & Sons, Hoboken, NJ, 3rd edition edition, 2005.

[3] KC Border. Differentiating an integral: Leibniz'rule. *Caltech Division of the Humanities and Social Sciences*, 2016.

[4] Chain Monte Carlo. Markov chain monte carlo and gibbs sampling. *Lecture notes for EEB*, 581(540):3, 2004.

[5] Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

[6] Rinze M Hallema. https://github.com/rinzehallema/bachelor-project, 2023.

[7] Michael L Stein. Prediction and inference for truncated spatial data. *Journal of Computational and Graphical Statistics*, 1(1):91–110, 1992.