AMERICAN UNIVERSITY OF BEIRUT

Application of Higher-Order Approximations in Bayesian Inference

by
ESMAIL HARB ABDUL FATTAH

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science
to the Computational Science Program
of the Faculty of Arts and Sciences
at the American University of Beirut

Beirut, Lebanon
May 2018

AMERICAN UNIVERSITY OF BEIRUT

Application of Higher-Order Approximations in Bayesian Inference

by

ESMAIL HARB ABDUL FATTAH

Approved by:

Prof. Abbas Al Hakim, Associate Professor — Advisor

Mathematics

Prof. Samer Kharroubi, Associate Professor — Advisor

Nutrition and Food Sciences

Prof. Nabil Nassif, Professor — Member of Committee

Mathematics

Date of thesis defense: May 8, 2018

AMERICAN UNIVERSITY OF BEIRUT

THESIS, DISSERTATION, PROJECT RELEASE FORM

Student Name: Abdul Fattah Esmail Harb
Last First Middle

☑ Master's Thesis ◯ Master's Project ◯ Doctoral Dissertation

☑ I authorize the American University of Beirut to: (a) reproduce hard or electronic copies of my thesis, dissertation, or project; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

☐ I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes after:

One ___ year from the date of submission of my thesis, dissertation or project.
Two ___ years from the date of submission of my thesis , dissertation or project.
Three ___ years from the date of submission of my thesis , dissertation or project.

Signature May 11, 2018
Date

This form is signed when submitting the thesis, dissertation, or project to the University Libraries

Acknowledgements

First, I would like to express my sincere gratitude and thanks to Prof. Abbas Alhakim and Prof. Samer Kharroubi for their patience, motivation, and immense knowledge during my masters study.

I have really enjoyed the theory of Statistics with Prof. Abbas who mentored me from the very beginnings in my statistical foundations and overcame all the difficulties together, and the computational statistics with Prof. Kharroubi who gave me the opportunity to work with him on a research project which I really enjoyed as a part of the two years Masters program I have completed.

I would also like to thank Prof. Nabil Nassif for his insightful comments, encouragement, and valuable advice.

Last but not the least, I would like to thank my parents for providing me with unwavering support and continuous encouragement throughout my years of study.

An Abstract of the Thesis of

Esmail Abdul Fattah for Master of Computational Science
Major: Computational Science

Title: Application of Higher-Order Approximations in Bayesian Inference

In Bayesian methods, one almost is required to calculate certain characteristics of posterior and predictive distributions, including the mean, variance and density. When a conjugate prior likelihood pair is used, calculations of these tasks are usually immediate. However, in most useful applications, it is hard to find conjugate priors and so the posterior calculations cannot be obtained in closed form. In such cases analytic or numerical approximations are then needed. In these cases, it is often useful to have analytic approximations that are more accurate than the usual first order normal approximation but at the same time are not as computationally intensive as numerical integration, especially in cases with high dimensional parameter space. For several particular case studies including single and multi-parameter cases, we explored the use of higher order Laplace approximation in getting such estimates and compared the estimates with those obtained via Monte Carlo Methods. The methods will be illustrated by a genetic linkage model and a censored regression model.

Contents

Acknowledgements v

Abstract vi

1 Introduction 1

2 Preliminaries 2
- 2.1 Asymptotic Notations 2
- 2.2 Bayes Theorem 2
- 2.3 The Posterior Distribution 3
- 2.4 Newton-Raphson Method 3
- 2.5 Simple Monte Carlo 4
- 2.6 Gibbs Sampling 4
- 2.7 Metropolis Random Walk 5

3 Exponential Family and the Choice of Prior 6
- 3.1 Exponential Family 6
- 3.2 The Choice of Prior 7
  - 3.2.1 Non Informative Prior 7
  - 3.2.2 Conjugate Priors 8
  - 3.2.3 Prior for Normal Distribution 9

4 First Order Approximation For Univariate Case 10
- 4.1 Maximum Likelihood 10
- 4.2 Normal Approximation 10
- 4.3 Some Common Conjugate Priors and their First Order Approximations . 11
- 4.4 The Predictive Distribution 14

5 Higher Order Approximation to the Posterior Distribution for Univariate Case 15
- 5.1 Introduction 15
- 5.2 Preliminary Results 16
- 5.3 Transformation to Signed Roots 16
  - 5.3.1 Laplace's Approximation for the Normalizing Constant 16
  - 5.3.2 Standardized Signed Roots 17
- 5.4 Posterior Expectations 18
- 5.5 Some Common Conjugate Priors and their Higher Order Approximations 20
- 5.6 Type I Censored Data Example 22

5.6.1 The Likelihood Function 22
5.6.2 Exponential Lifetimes and Gamma Prior 23
5.6.3 Predictive Density for Censored Data 23
5.6.4 Normal Distribution Approximation 23
5.6.5 Higher Order Approximation 24

6 Higher Order Approximation to the Posterior Distribution for Multivariate Case 26
6.1 Posterior Approximation using Signed Root log-likelihood ratios 26
6.2 Example: Censored Regression 27

7 Conclusion 30

Bibliography 31

List of Figures

4.1 Distributions of the exact and first order approximation in exponential - gamma prior . . . 12
4.2 Distributions of the exact and first order approximation in binomial - beta prior . . . 13
4.3 Density plot of the exact and first order approximation (red line)in Poisson - gamma prior . . . 14

5.1 Comparison of the exact (black line) distribution and its approximations in exponential - gamma prior . . . 20
5.2 Comparison of the exact (black line) distribution and its approximations in binomial - beta prior . . . 21
5.3 Comparison of the exact (black line) distribution and its approximations in Poisson - gamma prior . . . 22
5.4 Comparison of the exact (black line) distribution and their approximations 25

List of Tables

4.1 Exponential - Gamma Prior 11
4.2 Binomial - Beta Prior 12
4.3 Poisson - Gamma Prior 13

5.1 First and Higher Order Approximation for different values of α and β - Censored Data 25

6.1 Life test data on mottorettes - Insulation life in hours at various test temperatures 28

Chapter 1

Introduction

In this report, different methods are used to approximate some characteristics of posterior and predictive distributions, especially their densities and means. One of these methods is the known first order approximation. However, it is useful in most applications to search for other methods to get better approximation such as the higher order approximation which is illustrated by the signed root log-likelihood or log-posterior density ratios.

In Bayesian inference, when the prior is not conjugate, it is hard to find the posterior in closed form. That's why analytical or numerical methods are needed. Tierney and Kadane (1986) and Tierney (1989) use Laplace methods to derive approximations for densities and expectations and have been shown to provide good approximations in many cases.

Sweeting (1995, 1996) and Sweeting and Kharroubi (2003, 2010), approach the problem of higher order approximations for various applications. They provide transformed signed roots, proposed by Brandorff-Nielson (1988, 1991). Also, Ventura and Reid (2014) discussed the approximate Bayesian computation based on the asymptotic theory of modified likelihood ratios. They outlined the role of computational tools for approximations in Bayesian inference, where high computational power allows the use of stochastic simulation to obtain exact answers. Ruli, Sartori and Ventura (2012) showed the advantage of MCMC methods where samples are drawn independently with lower computational time.

In this thesis, conjugate priors are used to compare the first order approximations, using Laplace methods, and the higher order approximations using the transformed log-likelihood ratios in univariate case. The higher order methods in multivariate case are compared to the estimates with those obtained via Monte Carlo Methods, such as the Metropolis random walk. The used approaches are illustrated by a genetic linkage model and a censored regression model.

Chapter 2

Preliminaries

2.1 Asymptotic Notations

Throughout the chapters, $\overset{k}{\propto}$, for general k, denotes the proportionality to $\mathcal{O}(n^{-k/2})$, $\overset{k}{=}$ denotes the equality to $\mathcal{O}(n^{-k/2})$, and $\doteq$ denotes the equality to fourth order respectively.

2.2 Bayes Theorem

In Bayesian approach, unlike the classical or frequentist approach, the parameters are viewed as random variables. Thomas Bayes figured out that the more balls are thrown, the better we should know the position of the first ball. Bayes theorem, the basis of statistical inference, relates the conditional and marginal probabilities of stochastic events A and B, see [17].

Bayes rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The proof of this theorem can be done by equating $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

For random variables X and Y with joint pdf $f(x, y)$,

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)} = \frac{f(y|x)f(x)}{\int f(x, y)dx}$$

Law of total probability:
Given pairwise disjoint events B_i, $i = 1, 2, ...$ whose union is the entire sample space Ω, then for any event A , we have

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$$

2.3 The Posterior Distribution

Consider the problem of finding a point estimator of the parameter $\theta = (\theta^1,\theta^2,...,\theta^d)$, and $X = (X_1, X_2, ...X_n)$ is a set of independent and identically distributed observations whose joint probability density function $X \sim f(.|\theta)$. Denote $\lambda(\theta)$ the prior density of the parameter θ, before the data is considered. The likelihood function of θ is defined by $L(\theta|X) = \prod_{i=1}^{n} f(X_i|\theta)$.

By reinterpreting the events in Bayes formula, the distribution of θ, given X, which is called the posterior distribution, is given by

$$\pi(\theta|X) = c^{-1}\lambda(\theta)L(\theta|X)$$

where, $c = \int_{\Theta} \lambda(\theta)L(\theta|X)d\theta$ is the marginal distribution of X.

It is useful to work with log-likelihood function,

$$l(\theta) = \log\ L(\theta|X) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

and the score function is

$$l'(\theta|X) = \frac{\partial l(\theta|X)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log f(X_i|\theta)}{\partial \theta}$$

and the second derivative

$$l''(\theta|X) = \frac{\partial^2 l(\theta|X)}{\partial \theta^2} = \sum_{i=1}^{n} \frac{\partial^2 \log f(X_i|\theta)}{\partial \theta^2}$$

2.4 Newton-Raphson Method

Newton's Method is used to solve the likelihood equation i.e $l'(\theta/X) = 0$. Parameters from the same data are chosen in a way that maximizes the probability of the sample that was actually observed. It is an iterative approach based on quadratic Taylor series approximation of $l(\theta/X)$
When finding the root of $l'(\theta|X) = 0$ the Newton-Raphson algorithm take the form

initialization θ_1, N, tolerance tol;
for *i:1,2,...N* **do**
$\quad \theta_{k+1} \leftarrow \theta_k + \left\{ - \frac{\partial^2 l(\theta|X)}{\partial \theta^2}\Big|_{\theta_k} \right\}^{-1} \left\{ \frac{\partial l(\theta|X)}{\partial \theta}\Big|_{\theta_k} \right\}$;
$\quad$ **if** $|\theta_{k+1} - \theta_{k+1}| < tol$ **then**
$\quad\quad$ return θ_{k+1} ;
$\quad$ **end**
end
Print('No Convergence');

Algorithm 1: Newtons-Raphson Algorithm

The choice of the initial guess is important and can lead to divergence, adapted from [2].

2.5 Simple Monte Carlo

When the prior $\lambda(\theta)$ is a density function from which a sample $\{\theta^i\}_{i=1}^n$ can be directly generated, then the simple Monte Carlo can be used to estimate

$$c = \int_\Theta \lambda(\theta)L(\theta|X)d\theta$$

by

$$\frac{1}{n}\sum_{i=1}^{n} L(\theta^i|X).$$

In case, the prior is non-informative and corresponds to uniform distribution, the estimation may be poor when the range is too narrow and an inefficient when the range is too wide. That means the range should be carefully defined.
On the other hand, when it is not simple to sample from the prior, rejection sampling can be used. Another distribution $q(\theta)$ is defined from which a sample can be generated under the restriction that $\lambda(\theta) < Mq(\theta)$ where M is an appropriate bond for $\dfrac{\lambda(\theta)}{q(\theta)}$. The rejection sampling algorithm can be summarized as follows,

initialization;
while *While* $i < N$ **do**
  $\theta^i \sim q(\theta)$;
  $u \sim U(0,1)$;
  if $u < \frac{\lambda(\theta^i)}{Mq(\theta^i)}$ **then**
    accept θ^i ;
    $i \leftarrow i + 1$;
  else
    reject θ^i;
  end
end

Algorithm 2: Rejection Sampling

See [4] for more computation methods.

2.6 Gibbs Sampling

Given a sequence of random variables $\theta_1, \theta_2, \theta_3...$, each variable is sampled from the distribution $Q(\theta^{t+1}, \theta^t|X)$ where the next sample depends on the current state only. This sequence is called Markov Chain.

Markov Chain Monte Carlo (MCMC) techniques are methods to construct sampled chain from probability distributions using Markov chains. One of these techniques is the Metropolis random walk.

One of the computational methods used to approximate the posterior distribution is Gibbs sampling. The main idea of Gibbs is to use the prior information to construct an ergodic Markov Chain whose limiting distribution is the posterior distribution. Samples from posterior are generated by sweeping through each variable to sample from its conditional distribution with the remaining variables fixed to their current values. See [8] for more details.

2.7 Metropolis Random Walk

Metropolis is a random walk that uses an acceptance/rejection rule (Hasting ratio) to converge to the target distribution, and as much as the sample becomes larger the better approximation to the desired distribution we get, assuming convergence exists. When it is hard to get conditional distributions for the variables, Metropolis sampling can be used as an option to approximate posterior distribution. The Metropolis algorithm as stated in [4] can be summarized as follows,

initialization of θ^t;
$\theta^{t+1} \sim Q$ the proposal density;
for *i:1,2,...N* **do**
  Set r $= \dfrac{\pi(\theta^{t+1}|X)Q(\theta^{t+1})}{\pi(\theta^t|X)Q(\theta^t)}$;
  $\theta^i \sim q(\theta)$;
  $u \sim U(0,1)$;
  if $u < min\{1, r\}$ **then**
    $\theta^t \leftarrow \theta^{t+1}$;
  end
end

Algorithm 3: Metropolis Sampling

Chapter 3

Exponential Family and the Choice of Prior

3.1 Exponential Family

In Bayesian inference, there are family of distributions that depends on the number and value of parameters that shapes them differently. As in [14], an exponential family distribution has the following form,

$$p(x|\eta) = h(x)\ exp\{\eta^T t(x) - a(\eta)\}$$

where η is a natural parameter, $t(x)$ is the sufficient statistic, $h(x)$ is the underlying measure and $a(\eta)$ is the log normalizer where we integrate the unnormalized density over the sample space. This ensures that the density integrates to one.

Example: The Gaussian distribution for one-parameter can be written as:

$$p(x/\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big\{\frac{-(x-\mu)^2}{2\sigma^2}\Big\}$$

After expanding the identity, we can see that $\eta = (\mu/\sigma^2, -1/2\sigma^2)$, $t(x) = (x, x^2)$, $a(\eta) = \mu^2/2\sigma^2 + \log\sigma$ and $h(x) = 1/\sqrt{2\pi}$

In a similar way, binomial, Poisson, uniform, gamma and beta distributions are examples of exponential family distributions.

Definition: We say P, class of prior distributions $p(\theta)$ for θ, is conjugate to S, class of sampling distributions $s(X/\theta)$ for θ, if $p(\theta/x) \in P\ \forall\ s(./\theta) \in S$

Conjugate Families arises when the likelihood times the prior produces a recognizable posterior kernel.

$$\pi(\theta|X) \propto \lambda(\theta)L(\theta|X)$$

where the kernel is the characteristic part of the distribution function that depends on the random variable(s), excluding the normalizing constant.

If the posterior distribution $\pi(\theta|X)$ is in the same family as the prior probability distribution $\lambda(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $L(\theta|X)$. Now, we discuss a few common conjugate family results with their first order approximations.

3.2 The Choice of Prior

3.2.1 Non Informative Prior

When we have no information about the prior, we call it non informative prior. In this case, the posterior distribution is approximately equal to the standarized likelihood.

Jeffreys [16] introduced an approach for choosing the prior. Suppose that $\theta = (\theta^1, \theta^2, ..., \theta^d)$, the Fisher information matrix

$$I(\theta) = -E\Big[\frac{\partial^2 l(\theta|X)}{\partial\theta_i \partial\theta_j}\Big]$$

and the Jeffreys non-informative prior is

$$\lambda(\theta) \propto det(I(\theta))^{1/2}$$

Jeffreys justified it on the ground of its invariance under any transformation on the parameter space. By considering 1-1 transformation of the parameter $\phi = t(\theta)$, if $\lambda(\theta)$ is the prior of θ, then the corresponding density of ϕ is

$$g(\phi) = \lambda(t^{-1}(\phi))|J(\phi)|$$

where J is the Jacobian of the transformation. Jeffreys claimed that after transforming θ to ϕ the prior of ϕ should be as follows

$$\lambda^*(\phi) = g(\phi), \forall\phi$$

In one-dimensional case, we have,

$$\lambda(\theta) \propto \Big\{ - E\Big[\frac{\partial^2 l(\theta|X)}{\partial\theta^2}\Big]\Big\}^{1/2}$$

and

$$g(\theta) = \lambda(t^{-1}(\phi))\Big|\frac{dt^{-1}(\phi)}{d\phi}\Big|$$

After the parameterization $\phi = t(\theta)$, the Fisher information is

$$-E\Big[\frac{\partial^2 l(t^{-1}(\phi)|X)}{\partial\phi^2}\Big]$$

and it can be written as

$$-E\Big[\frac{\partial^2 l(t^{-1}(\phi)|X)}{\partial t^{-1}(\phi)^2}\frac{\partial t^{-1}(\phi)^2}{\partial\phi^2}\Big] = -\Big(\frac{\partial t^{-1}(\phi)}{\partial\phi}\Big)^2 E\Big[\frac{\partial^2 l(t^{-1}(\phi)|X)}{\partial t^{-1}(\phi)^2}\Big]$$

which is proportional to $g^2(\phi)$, and hence Jeffreys invariance property holds.

As a notice, Jeffreys mentioned himself that in multidimensional case the chosen prior should be chosen with caution.

Example: The log-likelihood binomial function is

$$l(p) = \sum_{i=1}^{n} X_i \log p + (N - \sum_{i=1}^{n} X_i) \log(1-p)$$

So,

$$-E\Big[\frac{d^2(l(p))}{dp^2}\Big] = \frac{N}{p(1-p)}$$

So, Jeffreys non informative prior for p is proportional to $[p(1-p)]^{-1/2}$, and must be a Beta(1/2,1/2) density.

3.2.2 Conjugate Priors

As mentioned previously, when the posterior distribution follow the same parametric shape of the prior distribution, this leads to conjugate families. In this case, prior is called informative. For example, Beta is conjugate with Bernoulli and Binomial. The table below shows some common informative conjugate priors.

Likelihood	**Conjugate Prior Distribution**	**Prior Hyperparameters**	**Posterior Hyperparamters**
Binomial $\prod_{i=1}^{n} p^{x_i}(1-p)^{N_i - x_i}$	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^{n} x_i, \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$
Poisson $\prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x!}$	Gamma	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^{n} x_i, \beta + n$
Geometric $\prod_{i=1}^{n} p(1-p)^{x_i - 1}$	Beta	$\alpha,\ \beta$	$\alpha + n, \beta + \sum_{i=1}^{n} x_i + n$
Normal (Known σ^2)	Normal	$\mu_0,\ \sigma_0^2$	$\frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}$
Normal (Known μ)	Inverse gamma	$\alpha,\ \beta$	$\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n} (x_i - \mu)^2$
Exponential $\lambda^n \exp\{-\lambda \sum_{j=1}^{n} x_j\}$	Gamma	$\alpha,\ \beta$	$\alpha + n, \beta + \sum_{i=1}^{n} x_i$
IG (known α) $\prod_{i=1}^{n} x_i^{-\alpha-1} \exp\left(-\frac{\beta}{x_i}\right)$	Gamma	$\alpha_0,\ \beta_0$	$\alpha_0 + n\alpha,\ \beta_0 + \sum_{i=1}^{n} \frac{1}{x_i}$

3.2.3 Prior for Normal Distribution

The prior of a population is normally distributed with known mean μ and known variance σ_0. $\bar{x}$ is the mean of a random sample of size n from a normal population with known variance σ^2.

The density function of our sample is

$$L(x_1, x_2, ...x_n|\mu) = \frac{1}{(2\pi)^{2/n}\sigma^n} exp\Big\{ -\frac{1}{2}\sum_{i=1}^{n}\Big(\frac{x_i-\mu}{\sigma}\Big)^2\Big\}$$

for $-\infty < x_i < \infty$ and $i = 1, 2, ...n$, the prior is

$$\lambda(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} exp\Big\{ -\frac{1}{2}\Big(\frac{\mu-\mu_0}{\sigma_0}\Big)^2\Big\}$$

Then the posterior distribution of μ is

$$\begin{aligned}\pi(\mu) &\propto exp\Big\{ -\frac{1}{2}\Big[\Big(\frac{\mu-\mu_0}{\sigma_0}\Big)^2 + \sum_{i=1}^{n}\Big(\frac{x_i-\mu}{\sigma}\Big)^2\Big]\Big\} \\ &\propto exp\Big\{ -\frac{1}{2}\Big[\frac{(\mu-\mu_0)^2}{\sigma_0^2} + \frac{n(\bar{x}-\mu)^2}{\sigma^2}\Big]\Big\} \\ &\propto exp\Big\{ -\frac{1}{2}\Big(\frac{\mu-\mu^*}{\sigma^*}\Big)^2\Big\}\end{aligned}$$

where

$$\mu^* = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2+\sigma^2} \text{ and } \sigma^* = \sqrt{\frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}}$$

due to $\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2$ and completing the square of μ.

When the likelihood function and prior are normal, we get normal posterior distribution. The prior is a conjugate prior with the likelihood.

Chapter 4

First Order Approximation For Univariate Case

4.1 Maximum Likelihood

The maximizer likelihood estimator of θ is $\hat{\theta}$, the value of θ that maximizes the likelihood function, or equivalently the log-likelihood function.

If $l'(\theta|X)$ is differentiable with respect to θ, $\hat{\theta}$ is a solution of $l'(\theta|X) = 0$, and for a maximum $l''(\theta|X) < 0$.

4.2 Normal Approximation

The log-likelihood function $l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$ is asymptotically $\mathcal{O}(n)$ since $X = (X_1, X_2, ...X_n)$ are i.i.d. Similarly all the derivatives of $l(\theta)$ are $\mathcal{O}(n)$.

Furthermore, $\hat{\theta}$ is considered as a consistent estimator, that means as the number of data n increases, the resulting sequence of estimates converges to a true value θ_0, as $\hat{\theta} \to \theta_0$, n $\to \infty$.

Using Bernstein-von Mises Theorem and under regularity conditions and some fixed value θ, $\hat{\theta}$ is approximately normal with mean θ and variance $I(\theta)^{-1}$, the Fisher's Information. $I(\theta)$ is also $\mathcal{O}(n)$.

By definition, Standard normal distribution z is $\mathcal{O}(1)$.

$$z = \sqrt{I(\theta)}(\hat{\theta} - \theta) \overset{1}{\sim} N(0,1) \text{ and } (\hat{\theta} - \theta) \overset{1}{\sim} N(0, I(\theta)^{-1})$$

then $(\hat{\theta} - \theta)$ is $\mathcal{O}(n^{-1/2})$, and $(\hat{\theta} - \theta)^2$ is $\mathcal{O}(n^{-1})$.

Using Taylor series to approximate posterior density

$$l(\theta|X) \approx l(\hat{\theta}) + (\theta - \hat{\theta})l'(\theta|X)|_{(\theta=\hat{\theta})} + \tfrac{1}{2}(\theta - \hat{\theta})^2 l''(\theta|X)|_{(\theta=\hat{\theta})}$$

$l(\hat{\theta}) = constant$, and $(\theta - \hat{\theta})l'(\theta|X)|_{(\theta=\hat{\theta})} = 0$ because $\hat{\theta}$ is maximizer of the log-likelihood function.

$$l(\theta|X) \approx constant + \tfrac{1}{2}(\theta - \hat{\theta})^2 l''(\theta|X)|_{(\theta=\hat{\theta})}$$

Let $J = -l''(\theta|X)|_{(\theta=\hat{\theta})}$, J is called the observed information, and $\mu = \hat{\theta}$ and $\sigma^{-2} = J = -l''(\theta|X)|_{(\theta=\hat{\theta})}$

The likelihood function becomes:

$$L(\theta|X) \approx exp\{constant - \frac{1}{2}\frac{(\theta-\mu)^2}{\sigma^2}\} \overset{1}{\propto} exp\{-\frac{1}{2}\frac{(\theta-\mu)^2}{\sigma^2}\} = exp\{-\frac{1}{2}J(\theta-\hat{\theta})^2\}$$

Let $\nu(\theta) = log\lambda(\theta)$. Expand $\nu(\theta)$ about the fixed value $\hat{\theta}$, $\lambda(\theta) = e^{\nu(\theta)} = e^{\nu(\hat{\theta})+(\hat{\theta}-\theta)\nu'(\hat{\theta})+\dots}$

$$\lambda(\theta) \overset{1}{\propto} \text{constant.}$$

For $\mathcal{O}(n^{-1/2})$ approximation,

$$\pi(\theta|X) \propto \lambda,(\theta)L(\theta|X) \propto exp\{\nu(\theta) + l(\theta|X)\} \overset{1}{\propto} L(\theta|X) \overset{1}{\sim} N(\hat{\theta}, J^{-1}) \qquad (4.1)$$

So, posterior density is approximated proportional to normal density of mean $\hat{\theta}$ and variance J^{-1}, as shown in [1] and [2].

4.3 Some Common Conjugate Priors and their First Order Approximations

Exponential - Gamma Prior

Model	$X_i \sim Exponential(\lambda)$
Likelihood function $L(\alpha)$	$\alpha^n e^{-\alpha \sum_{i=1}^n X_i}$
Log-Likelihood function $l(\alpha) = logL(\alpha)$	$nlog(\alpha) - \alpha \sum_{i=1}^n X_i$
$l'(\alpha) = d(l(\alpha))/d\alpha$	$n/\alpha - \sum_{i=1}^n X_i$
$l''(\alpha) = d^2(l(\alpha))/d\alpha^2$	$-n/\alpha^2$
Solution of $l'(\alpha) = 0$	$\hat{\alpha} = n/\sum_{i=1}^n X_i = 1/\bar{X}$
Observed Information J	$n\bar{X}^2$
Gamma Prior	$Gamma(\alpha, \beta)$
Gamma Posterior	$Gamma(\mu_p, \sigma_p^2) = Gamma(\alpha + n, \beta + \sum_{i=1}^n X_i)$
First Order Normal Approximation	$N(\sum_{i=1}^n X_i/n, 1/(n\bar{X}^2))$

Table 4.1: Exponential - Gamma Prior

For $\alpha = 1$, $\beta = 2$ and $\sum_{i=1}^{10} X_i = 57.45281918$, we get the approximation in Figure 4.1 that shows the exact distribution and the first order approximation (Red line) of

θ. The exact posterior probability $P(\theta < \theta^1/X) = 0.5401112973$, $P(\theta < \theta^2/X) = 1.644022402e^{-08}$, and $P(\theta < \theta^3/X) = 0.9935720566$, where $\theta^1 = \alpha/\beta$, $\theta^2 = \theta^1 - 3\sqrt{\alpha}/\beta$, and $\theta^3 = \theta^1 + 3\sqrt{\alpha}/\beta$. For the first order approximation, $P(\theta < \theta^1/X) = 0.5789509273$, $P(\theta < \theta^2/X) = 0.00224600844$, and $P(\theta < \theta^3/X) = 0.9994019129$. It is clear that there is a noticeable discrepancy between the two distributions. Better approximation of theta distribution will be shown in the following section.

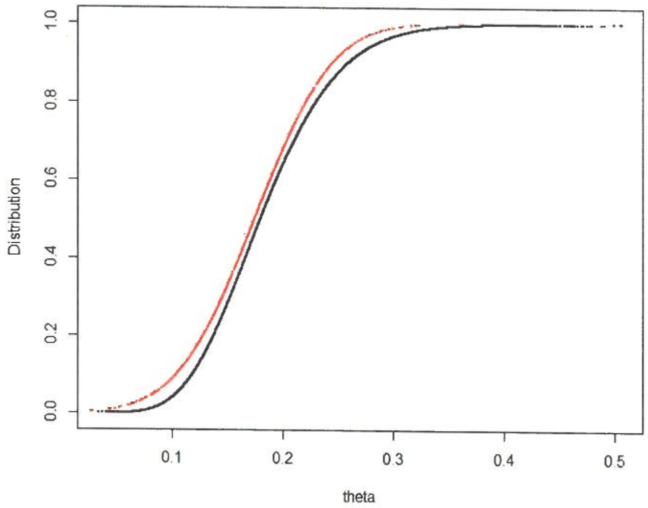

Figure 4.1: Distributions of the exact and first order approximation in exponential - gamma prior

Binomial - Beta Prior

Model	$X_i \sim Binomial(N_i, p)$
Likelihood function $L(p)$	$\prod_{i=1}^n p^{X_i}(1-p)^{N_i - X_i} = p^{\sum_{i=1}^n X_i}(1-p)^{N-\sum_{i=1}^n X_i}$
Log-Likelihood function $l(p) = logL(p)$	$\sum_{i=1}^n X_i \log p + (N - \sum_{i=1}^n X_i)\log(1-p)$
$l'(p) = d(l(p))/dp$	$\sum_{i=1}^n X_i/p - (N - \sum_{i=1}^n X_i)/(1-p)$
$l''(p) = d^2(l(p))/dp^2$	$-\sum_{i=1}^n X_i/p^2 - (N - \sum_{i=1}^n X_i)/(1-p)^2$
Solution of $l'(p) = 0$	$\hat{p} = \sum_{i=1}^n X_i/N$
Observed Information J	$\sum_{i=1}^n X_i/\hat{p}^2 - (N - \sum_{i=1}^n X_i)/(1-\hat{p})^2$
Beta Prior	$Beta(\alpha, \beta)$
Beta Posterior	$Beta(\alpha + \sum_{i=1}^n X_i, \beta + N - \sum_{i=1}^n X_i)$
First Order Normal Approximation	$N(\mu_{NA}, \sigma_{NA}^2) = N(\sum_{i=1}^n X_i/N, J^{-1})$

Table 4.2: Binomial - Beta Prior

For $\alpha = 3$, $\beta = 2$, $N = 100$ and $\sum_{i=1}^{10} X_i = 50$, we get the approximation in Fig 4.2 that shows the exact distribution and the first order approximation (red line) of θ. The exact posterior probability $P(\theta < \theta^1/X) = 0.5047619048$, where $\theta^1 = \alpha/(\alpha+\beta)$, . For the first order approximation, $P(\theta < \theta^1/X) = 0.5379371441$.

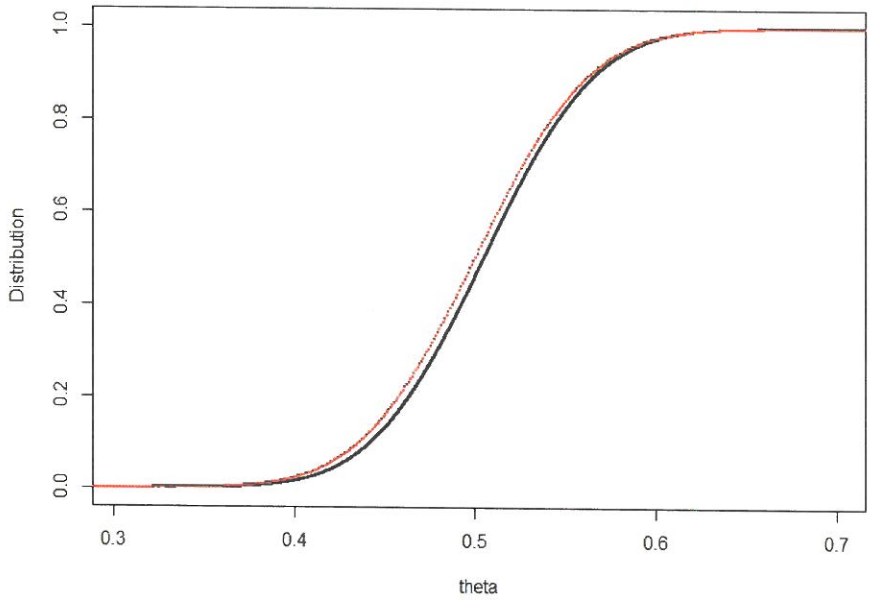

Figure 4.2: Distributions of the exact and first order approximation in binomial - beta prior

Poisson - Gamma Prior

Model	$X_i \sim Poisson(\lambda)$
Likelihood function $L(\lambda)$	$(1/\prod_{i=1}^{n} X_i!)\lambda^{\sum_{i=1}^{n} X_i} e^{-n\lambda}$
Log-Likelihood function $l(\lambda) = logL(\lambda)$	$\sum_{i=1}^{n} X_i \log\lambda - n\lambda - \log\prod_{i=1}^{n} X_i!$
$l'(\lambda) = d(l(\lambda))/d\lambda$	$(1/\lambda)\sum_{i=1}^{n} X_i - n$
$l''(\lambda) = d^2(l(\lambda))/d\lambda^2$	$(-1/\lambda^2)\sum_{i=1}^{n} X_i$
Solution of $l'(\lambda) = 0$	$\hat{\lambda} = \sum_{i=1}^{n} X_i/n$
Observed Information J	$n^2/\sum_{i=1}^{n} X_i$
Gamma Prior	$Gamma(\alpha, \beta)$
Gamma Posterior	$Gamma(\alpha + \sum_{i=1}^{n} X_i, \beta + n)$
First Order Normal Approximation	$N(\mu_{NA}, \sigma_{NA}^2) = N(\sum_{i=1}^{n} X_i/n, J^{-1})$

Table 4.3: Poisson - Gamma Prior

For $\alpha = 2$, $\beta = 4$ and $\sum_{i=1}^{10} X_i = 24.34864851$, we get the following density approximation

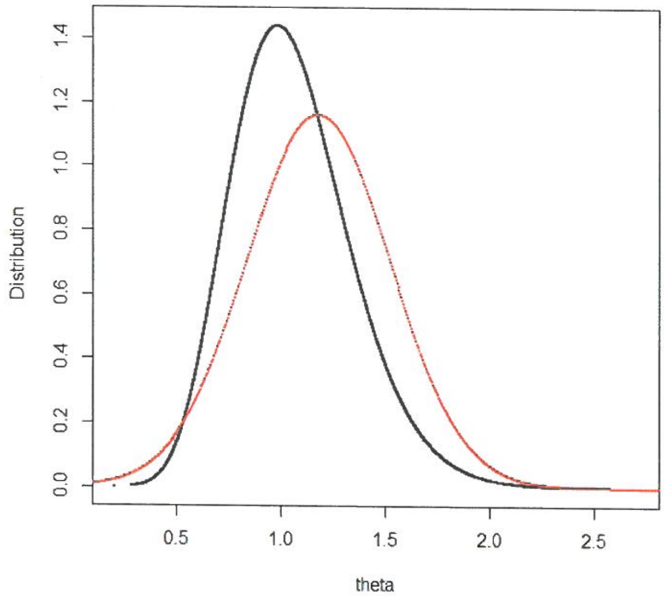

Figure 4.3: Density plot of the exact and first order approximation (red line)in Poisson - gamma prior

4.4 The Predictive Distribution

For a fixed value of θ, data X follows the distribution $p(X|\theta)$. The uncertainty of θ is represented by the prior distribution $p(\theta)$. For a new data y and before having data X, we get the prior predictive distribution:

$$p(y) = \int_{\Theta} p(y, \theta) d\theta = \int_{\Theta} p(y|\theta)\lambda(\theta) d\theta \tag{4.2}$$

After taking data X, the posterior predictive distribution for a new data point Y is:

$$p(y|X) = \int_{\Theta} p(y|\theta, X) p(\theta|X) d\theta = \int_{\Theta} p(y|\theta)\pi(\theta|X) d\theta \tag{4.3}$$

Expression 4.3 displays the distribution of Y as an average over the posterior distribution of θ.

Chapter 5

Higher Order Approximation to the Posterior Distribution for Univariate Case

5.1 Introduction

Suppose the likelihood function $L(\theta|X)$ is continuous and unimodal, and θ is a scalar parameter. Knowing that $l(\theta) = log(L(\theta|X))$ and $\hat{\theta} = argmax\{L(\theta|X)\}$. Consider the following transformation:

$$w(\theta) = 2\log\frac{L(\theta)}{L(\hat{\theta})} = 2[l(\hat{\theta}) - l(\theta)] > 0$$

The likelihood ratio statistic $w(\theta)$ has the asymptotic χ^2 distribution with one degree of freedom. It may be replaced by

$$r(\theta) = sign(\theta - \hat{\theta})\sqrt{2(l(\hat{\theta}) - l(\theta))}$$

Then,

$$L(\theta) \propto \frac{e^{l(\theta)}}{e^{l(\hat{\theta})}} = e^{(1/2)w(\theta)} = e^{-(1/2)r^2}$$

Since $e^{-(1/2)r^2}$ is the kernel of the standard normal density, it follows that the likelihood function $L(\theta)$ is standard normal in r.
That is,

$$L(\theta) \propto \phi(r(\theta))$$

It is shown in Barndorff-Nielsen and Cox (1989) that the quantity $R = r(\theta)$ is asymptotically standard normally distributed and R is referred to as directed likelihood ratio. This quantity is also referred to as Signed Root Log-Likelihood Ratio (SRLLR) statistic.

5.2 Preliminary Results

Let F_n be a sequence of univariate distributions. We say that f_n is an effective density sequence of F_n if

$$F_n(r) \doteq \int_{-\infty}^{r} f_n dx$$

for all r. Let q_n be a sequence of a real-valued functions on R. We shall say $(F_n) \in \Phi[q_n]$ if it has an effective density sequence f_n satisfying

$$f_n(r) \propto \phi(r) q_n(r)(1 + \epsilon_n r) \tag{5.1}$$

where ϵ_n denotes a sequence of $\mathcal{O}(n^{-3/2})$ independent of r, and $q_n(r)$ is of the form:

$$q_n(r) = 1 + a_n r + b_n r^2 + c_n r^3 + d_n r^4 \tag{5.2}$$

where $a_n = \mathcal{O}(n^{-1/2}), b_n = \mathcal{O}(n^{-1/2}), c_n = \mathcal{O}(n^{-3/2})$ and $d_n = \mathcal{O}(n^{-2})$. It is shown in Sweetings (1995) that this class $\Phi[.]$ has a number of attractive properties. For example, if $F_n \in \Phi[q_n]$ then,

$$F_n(r) \doteq \Phi[r] - \phi(r)\Big(\frac{q_n(r) - 1}{r} + \epsilon_n\Big)$$

$$f_n(r) = \frac{\phi(r) q_n (1 + \epsilon_n r)}{1 + b_n}$$

where $(1 + b_n)$ is the proportionality constant in 5.1 and b_n is the second coefficient of q_n in the expansion 5.2, see Sweeting (2003) and Kharroubi and Sweeting (2010).

5.3 Transformation to Signed Roots

5.3.1 Laplace's Approximation for the Normalizing Constant

The basic idea of Laplace's approximation is to find the maximum of the function to be integrated and apply a second order Taylor series approximation for the logarithm of that function.

Assume that an unnormalized probability density $\lambda(\theta)L(\theta|X)$, whose normalizing constant is c^{-1} such that

$$c = \int_\Theta \lambda(\theta) L(\theta|X) d\theta$$

where $\hat{\theta}$ is a maximizer of $L(\theta|X)$. We Taylor-expand the logarithm of $\lambda(\theta)L(\theta|X)$ around $\hat{\theta}$:

$$\log \lambda(\theta) L(\theta|X) \approx log \lambda(\hat{\theta}) L(\hat{\theta}|X) - \frac{k}{2}(\theta - \hat{\theta})^2 + ...$$

where

$$k = -\frac{\partial^2}{\partial \theta^2} \log \lambda(\theta) - l''(\theta|X)|_{\theta=\hat{\theta}} = -\frac{\partial^2}{\partial \theta^2} \log \lambda(\theta)|_{\theta=\hat{\theta}} + J = \mathcal{O}(n) + J$$

Then we can approximate

$$\log \lambda(\theta)L(\theta|X) \approx log\lambda(\hat{\theta})L(\hat{\theta}|X) - \frac{J}{2}(\theta - \hat{\theta})^2 + ...$$

We then approximate $\lambda(\theta)L(\theta|X)$ by an unnormalized Gaussian

$$Q(\theta) = \lambda(\hat{\theta})L(\hat{\theta}|X)exp\{-\frac{J}{2}(\theta - \hat{\theta})^2)\}$$

and we approximate the normalizing constant c by the normalizing constant of this Gaussian,

$$c \approx \lambda(\hat{\theta})L(\hat{\theta}|X)\sqrt{\frac{2\pi}{J}}$$

Posterior Approximation using Signed Roots Transformation

Using the above approximated normalizing constant and as shown in [1] and [2], the posterior density $\pi(\theta|X) = c^{-1}\lambda(\theta)L(\theta|X)$ becomes:

$$\pi(\theta|X) = c^{-1}\lambda(\theta)L(\theta|X) \stackrel{2}{=} \frac{1}{\sqrt{2\pi}}|J|^{1/2}\frac{\lambda(\theta)}{\lambda(\hat{\theta})}\frac{L(\theta)}{L(\hat{\theta})} = \frac{1}{\sqrt{2\pi}}|J|^{1/2}\frac{\lambda(\theta)}{\lambda(\hat{\theta})}\frac{e^{l(\theta)}}{e^{l(\hat{\theta})}}$$

Then the posterior function for θ is then approximated to the same order by

$$\int_{\theta_0}^{\infty} \pi(\theta|X)d\theta \stackrel{2}{=} \int_{\theta_0}^{\infty} \frac{1}{\sqrt{2\pi}}|J|^{1/2}\frac{\lambda(\theta)}{\lambda(\hat{\theta})}e^{-(1/2)r(\theta)^2}d\theta$$

where $r(\theta) = sign(\theta - \hat{\theta})\sqrt{w(\theta)}$ and $w(\theta) = 2[l(\hat{\theta}) - l(\theta)]$.

The next step is to change the variable of integration from θ to $r = r(\theta)$. The quantity $e^{-r(\theta)^2}$ is the kernel of the standard normal density. The Jacobian of the transformation is $dr(\theta)/d\theta = -l'(\theta)/r(\theta)$ where $l'(\theta)$ is the score function. Let $b(r) = |J(\hat{\theta})|^{1/2}\{\lambda(\theta)/\lambda(\hat{\theta})\}\{r(\theta)/l'(\theta)\}$ and $r_0 = r(\theta_0)$, then the posterior function becomes

$$\int_{\theta_0}^{\infty} \pi(\theta|X)d\theta \stackrel{2}{=} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{r_0} e^{(-(1/2)r(\theta)^2 + \log b(r))}dr$$

The posterior density of r can be expressed by

$$\pi(\theta|X) \stackrel{2}{=} \frac{1}{\sqrt{2\pi}}e^{(-(1/2)r(\theta)^2 + \log b(r))}$$

5.3.2 Standardized Signed Roots

Changing the variable r to $\tilde{r}$ by the following transformation

$$\tilde{r} = \tilde{r}(\theta) = r - r^{-1}\log b(r)$$

so that $-(\tilde{r})^2 = -r^2 + 2\log b(r) - (r^{-1}\log b(r))^2$. The Jacobian of the transformation and the third term in $-(\tilde{r})^2$ contribute to the error of the posterior approximation using r. Using the following equalities:

$$\begin{aligned}-\frac{1}{2}(r^2 - 2logb(r)) &= -\frac{1}{2}(r^2 - 2logb(r) + r^{-2}\log^2 b(r)) \\ &= -\frac{1}{2}\tilde{r}^2 + \frac{1}{2}r^{-2}\log^2 b(r)\end{aligned}$$

and

$$d\tilde{r}/dr = -r^{-2}\log^2 b(r)$$

The posterior function becomes

$$\int_{\theta_0}^{\infty} \pi(\theta|X)d\theta \overset{2}{=} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{r_0} exp\{-\frac{1}{2}\tilde{r}^2\}exp\{\frac{1}{2}r^{-2}\log^2 b(r)\}(-r^{-2}\log^2 b(r))^{-1}d\tilde{r}$$

Since the transformed variable $\tilde{r}$ has a normal distribution to $\mathcal{O}(n^{-3/2})$ then the posterior function is approximated by

$$\int_{\theta_0}^{\infty} \pi(\theta|X)d\theta \overset{3}{=} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{r_0} exp\{-\frac{1}{2}\tilde{r}^2\}d\tilde{r} = \Phi(\tilde{r}) \tag{5.3}$$

where $\Phi(.)$ is the standard normal distribution function, see [1] and [5].

5.4 Posterior Expectations

As is [?], to compute approximate posterior expectations of a general function $v(\theta)$ in the following fraction:

$$E[v(\theta)|X] = \frac{\int L(\theta)\lambda(\theta)v(\theta)\ d\theta}{\int L(\theta)\lambda(\theta)d\theta} \tag{5.4}$$

The normalizing constant in equation 5.1 is $s_n = 1 + b_n$, where b_n is the coefficient of r^2 in 5.2. That is

$$\int \phi(r)q(r)(1+\epsilon_n r)dr \doteq s$$

This can be transformed back to θ in order to obtain an approximation to the normalizing constant in 5.1 and we obtain,

$$s \doteq \int \frac{1}{\sqrt{2\pi}}|J|^{1/2}\frac{\lambda(\theta)}{\lambda(\hat{\theta})}\frac{L(\theta)}{L(\hat{\theta})}d\theta$$

which can be written as,

$$s \doteq \frac{|J|^{1/2}}{\sqrt{2\pi}\lambda(\hat{\theta})L(\hat{\theta})}\int \lambda(\theta)L(\theta)d\theta$$

Where $q_n = \frac{\lambda(\theta)}{\lambda(\hat{\theta})}\frac{d\theta}{dr} = -\frac{\lambda(\theta)}{\lambda(\hat{\theta})}\frac{r|J|^{1/2}}{l'(\theta)}$

and in particular, that

$$\int \lambda(\theta)L(\theta)d\theta \doteq s\frac{\sqrt{2\pi}}{|J|^{1/2}}\lambda(\hat{\theta})L(\hat{\theta})$$

Similarly, the numerator of 5.4 can be written as,

$$\sqrt{\frac{2\pi}{J}}s^* L(\hat{\theta})\lambda(\hat{\theta})\upsilon(\hat{\theta})$$

where $s^* = 1 + b^* = \int \phi(r)q^*(r)(1+\epsilon_n r)dr$, and $q^*(r) = \frac{\upsilon(\theta)}{\upsilon(\hat{\theta})}q(r)$, and b^* is the second coefficient in the expansion $q^*(r)$.

Taking the ratio of the above two approximations, we get

$$\begin{aligned} E[\upsilon(\theta)|X] &\doteq \frac{\sqrt{\frac{2\pi}{J}}s^* L(\hat{\theta})\lambda(\hat{\theta})\upsilon(\hat{\theta})}{\sqrt{\frac{2\pi}{J}}sL(\hat{\theta})\lambda(\hat{\theta})} \\ &= \upsilon(\hat{\theta})\frac{s^*}{s} \\ &= \upsilon(\hat{\theta})\frac{1+b^*}{1+b} \end{aligned} \tag{5.5}$$

Expansions are not necessary for the calculations of s_n. This is achieved by noting that

$$s_n = 1 + b_n \doteq \frac{1}{2}(q_n(-1) + q_n(1))$$

Define now θ^- by $r_n(\theta^-) = -1$ and θ^+ by $r_n(\theta^+) = +1$. Suppressing n from now on, we find that

$$\begin{aligned} s &\doteq \frac{1}{2}\Big(\frac{J^{1/2}}{l'(\theta^-)}\frac{\lambda(\theta^-)}{\lambda(\hat{\theta})} + \frac{-J^{1/2}}{l'(\theta^+)}\frac{\lambda(\theta^+)}{\lambda(\hat{\theta})}\Big) \\ &= \frac{1}{2}J^(1/2)(\lambda(\hat{\theta}))^{-1}\tau \end{aligned}$$

where $\tau = \Big(\frac{\lambda(\theta^-)}{l'(\theta^-)}\Big) + \Big(\frac{-\lambda(\theta^+)}{l'(\theta^+)}\Big)$.

Similarly,

$$\begin{aligned} s^* &= 1 + b^* \\ &\doteq \frac{1}{2}(q_n^*(-1) + q_n^*(1)) \\ &= \frac{1}{2}J^{1/2}(\lambda(\hat{\theta})\upsilon(\hat{\theta}))^{-1}\tau^* \end{aligned}$$

where $\tau^* = \Big(\frac{\lambda(\theta^-)\upsilon(\theta^-)}{l'(\theta^-)}\Big) + \Big(\frac{-\lambda(\theta^+)\upsilon(\theta^+)}{l'(\theta^+)}\Big)$.

Substituting in the approximation 5.5 for s and s^*, we obtain

$$
\begin{aligned}
E[\upsilon(\theta)|X] &\doteq \upsilon(\hat{\theta})\frac{\frac{1}{2}J^{1/2}(\lambda(\hat{\theta})\upsilon(\theta^-))^{-1}\tau^*}{(\lambda(\hat{\theta}))^{-1}\tau} \\
&= \frac{\tau^*}{\tau} \\
&= \Big(\frac{\lambda(\theta^-)\upsilon(\theta^-)}{l'(\theta^-)} + \frac{-\lambda(\theta^+)\upsilon(\theta^+)}{l'(\theta^+)}\Big)/\tau \\
&= \Big(\frac{\lambda(\theta^-)/l'(\theta^-)}{\tau}\Big)\upsilon(\theta^-) + \Big(\frac{-\lambda(\theta^+)/l'(\theta^+)}{\tau}\Big)\upsilon(\theta^+) \\
&= \alpha^-\upsilon(\theta^-) + \alpha^+\upsilon(\theta^+)
\end{aligned}
$$

where $\alpha^- = \tau^{-1}\Big(\frac{\lambda(\theta^-)}{l'(\theta^-)}\Big)$, $\alpha^+ = \tau^{-1}\Big(\frac{\lambda(\theta^+)}{l'(\theta^+)}\Big) = 1 - \alpha^-$. Note that $\theta^+ = \hat{\theta} + J^{-1/2}$, $\theta^- = \hat{\theta} - J^{-1/2}$ and $\alpha^- = \alpha^+ = \frac{1}{2} + \mathcal{O}(n^{-1/2})$, for more details see [1] and [2].

5.5 Some Common Conjugate Priors and their Higher Order Approximations

First order approximations in part 4.3 show a clear discrepancy between the exact distributions and the approximated one. Here below, plotted are showed again to show how the discrepancy decreases and approximations become better.

For Exponential - Gamma prior case,

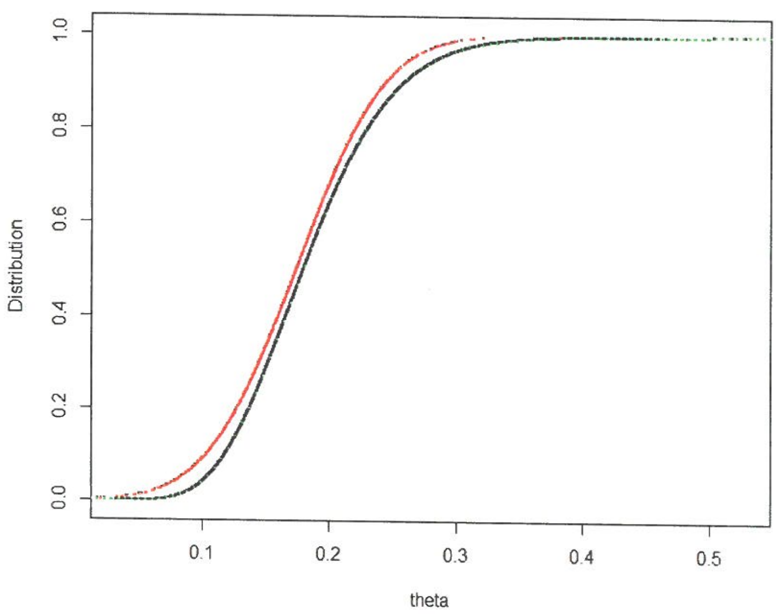

Figure 5.1: Comparison of the exact (black line) distribution and its approximations in exponential - gamma prior

Figure 5.1 shows the exact distribution, the first order approximation (red line) of θ and the higher order approximation (green dotted points). The exact posterior probability $P(\theta < \theta^1/X) = 0.5401112973$, $P(\theta < \theta^2/X) = 1.644022402e^{-08}$, and $P(\theta < \theta^3/X) = 0.9935720566$. For the higher order approximation, $P(\theta < \theta^1/X) = 0.5395864305$, $P(\theta < \theta^2/X) = 1.644022402e^{-08}$, and $P(\theta < \theta^3/X) = 0.9935586228$. It is clear that discrepancy has improved between the two distributions.

For Binomial - Beta prior case,

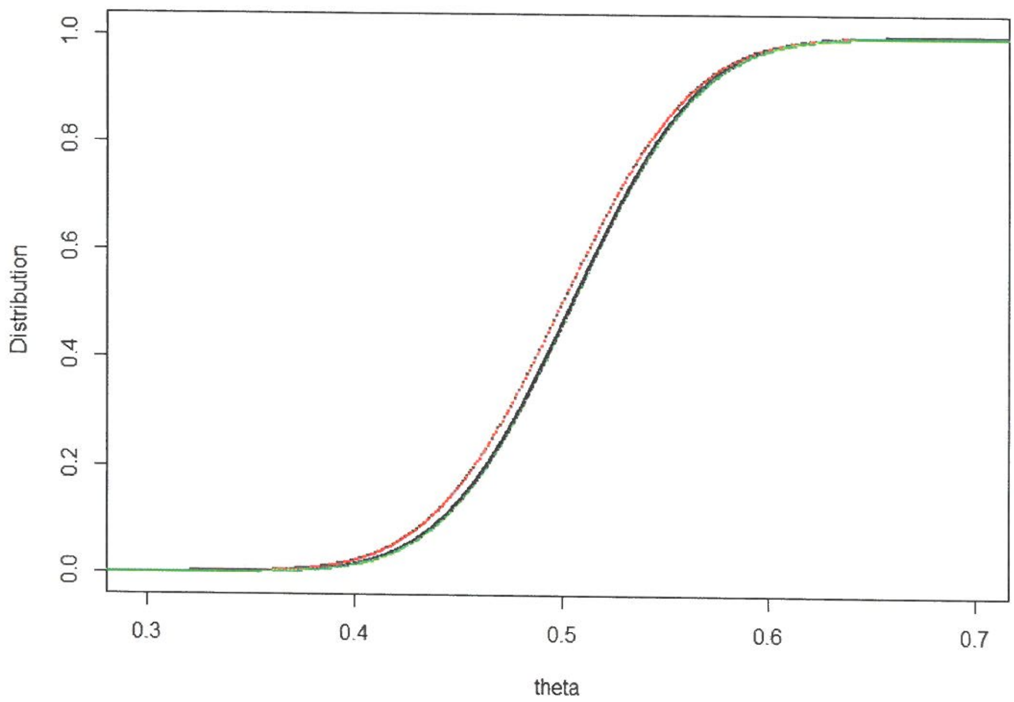

Figure 5.2: Comparison of the exact (black line) distribution and its approximations in binomial - beta prior

In Figure 5.2 the higher order approximation (green line) is more accurate compared to the first order approximation (red line), and the $P(\theta < \theta^1/X) = 0.4989557325$ for the case where $\alpha = 3$ and $\beta = 2$.

For Poisson - Gamma prior case

In Figure 5.3, for the same α, β and $\sum_{i=1}^{10} X_i$ as in sec 4.3, higher order approximation of the density in the green line is plotted (green line) and shows an accurate result compared to the first order (red line).

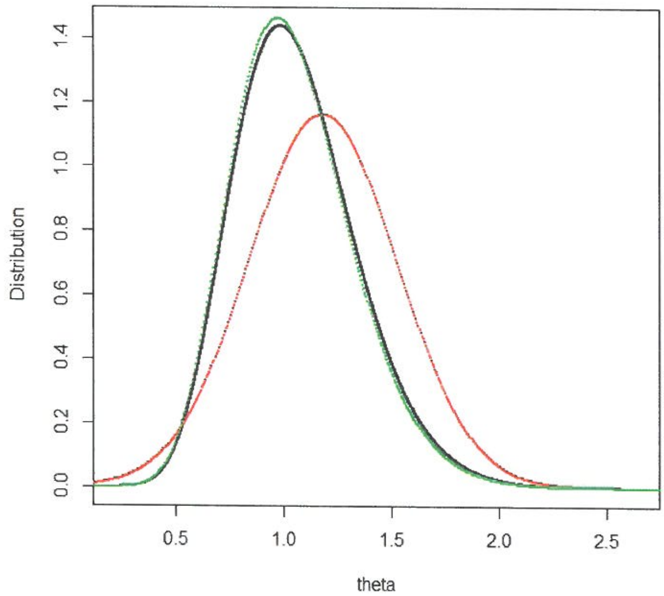

Figure 5.3: Comparison of the exact (black line) distribution and its approximations in Poisson - gamma prior

5.6 Type I Censored Data Example

5.6.1 The Likelihood Function

Censoring is when we know an incomplete information about the observation. Type-I censoring occurs when a failure time T_i, which denotes the response for i^{th} object, exceeds a constant c_i, which denotes the censoring time for i^{th} object. T_i's are assumed to be i.i.d. with density f, cdf F and survivor function S = 1 - F.

Define the observed response $X_i = min\{T_i, C_i\}$, and let δ_i denotes the indicator,

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i, \text{ data is uncensored} \\ 0 & \text{if } T_i > C_i, \text{ data is censored} \end{cases}$$

When X_i is uncensored, $f(X_i)$ contributes to the likelihood, and when X_i is censored, $P(x > X_i)$ contributes to the likelihood. The joint p.d.f of X_i and δ_i is

$$f(X_i)^{\delta_i} S(X_i)^{1-\delta_i}$$

Hence, the likelihood function is:

$$\prod_u f(X_i) \prod_c S(X_i) \tag{5.6}$$

where "u" and "c" denote the uncensored and censored observations respectively.

5.6.2 Exponential Lifetimes and Gamma Prior

Given a random sample T_i that follows an exponential distribution with p.d.f $f(t) = \theta e^{-\theta t}$ and the survivor function is $S(t) = e^{-\theta t}$, $t > 0$. Assume that T_i are i.i.d. Let n_u denote the number of uncensored observations in the sample and $s = \sum_{i=1}^{n} x_i$ where x_i is the observed event time. Using equation 5.6, the likelihood function is

$$L(\theta|X) = \theta^{n_u} e^{-\theta s}$$

By using the standard form of the gamma density as a prior distribution which is denoted by $Ga(a, b)$

$$\lambda(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}, \theta > 0$$

It follows that the posterior density is

$$\pi(\theta|X) = \frac{(b+s)^{a+n_u}}{\Gamma(a+n_u)} \theta^{a+n_u-1} e^{-(b+s)\theta}$$

That is $Ga(a + n_u, b + s)$

For illustration we use *rexp* in R to generate a vector of 10 observations from an exponential distribution, in which 8 observations are uncensored. With $n = 10$, $n_u = 9$ and $s = \sum_{i=1}^{10} x_i = 10.02414223$, we can the exact posterior probability that θ is less than 1 under a flat prior $Ga(1, 0)$

$$p(\Theta < 1|X) = 0.5450870519$$

5.6.3 Predictive Density for Censored Data

Using equation 4.3, the predictive density for a new observation Y that follows an exponential distribution with density $p(y|\theta) = \theta e^{-\theta y}$, $y > 0$ is:

$$\begin{aligned} p(y|\theta) &= \frac{(b+s)^{(a+n_u)}}{\Gamma(a+n_u)} \int_0^\infty \theta^{(a+n_u)} e^{-(b+s+y)\theta} d\theta \\ &= (b+s)^{(a+n_u)} (a+n_u)(b+s+y)^{-(a+n+1)}, y > 0 \end{aligned}$$

Using the same simulated data used in section 5.6.2, the exact predictive density under a uniform prior is

$$p(y = 1|X) = 0.6308424869$$

and a mean = 1.16412287.

5.6.4 Normal Distribution Approximation

The log-likelihood function in the censored data example is:

$$l(\theta|X) = n_u \log\theta - \theta s$$

and the score function can be expressed as:

$$l'(\theta|X) = \frac{\partial l(\theta|X)}{\partial \theta} = \frac{n_u}{\theta} - s$$

from which we get the maximum likelihood estimator $\hat{\theta} = \frac{n_u}{s}$.

The second derivative $l'(\theta|X) = \frac{\partial^2 l(\theta|X)}{\partial \theta^2} = -\frac{n_u}{\theta^2}$ is used to calculate the observed information $J = -l''(\hat{\theta}|X) = \frac{n_u}{\hat{\theta}^2}$. Therefore, using equation 4.1, the normalized likelihood can be approximated as a normal with mean $\hat{\theta}$ and variance $\frac{\hat{\theta}^2}{n_u}$ as

$$\pi(\theta|X) \overset{1}{\sim} N(\hat{\theta}, \frac{\hat{\theta}^2}{n_u})$$

To compare the first order approximation with the exact, we illustrate the previous result using the same data used in section 5.6.2, $n = 10$, $n_u = 9$, and $s = \sum_{i=1}^{10} x_i = 10.02414223$. We get $\hat{\theta}) = 0.8978$ and standard deviation $= 0.2993$. Based on the uniform prior of θ i.e $\lambda(\theta) \propto 1$. The approximate posterior probability is

$$p(\Theta < 1|X) = 0.6335915155$$

and a mean $= 1.058293518$.

5.6.5 Higher Order Approximation

Using equation 5.3, the posterior distribution can be approximated by

$$\int_{\theta_0}^{\infty} \pi(\theta|X) d\theta \overset{3}{=} \Phi(\tilde{r})$$

The approximate posterior probability $F(\theta) = p(\Theta < 1|X)$ based on the asymptotic normal distribution of $\tilde{r}$

$$p(\Theta < 1|X) = 0.544578488$$

We get $\theta^{+} = 1.392955314$, $\theta^{-} = 0.7236317233$ and a mean $= 1.16412287$.
This compares well compared to the first order approximation, where there is clear discrepancy, and this is shown in figure 5.4 where the exact distribution and the corresponding asymptotic approximation yield virtually identical curves.

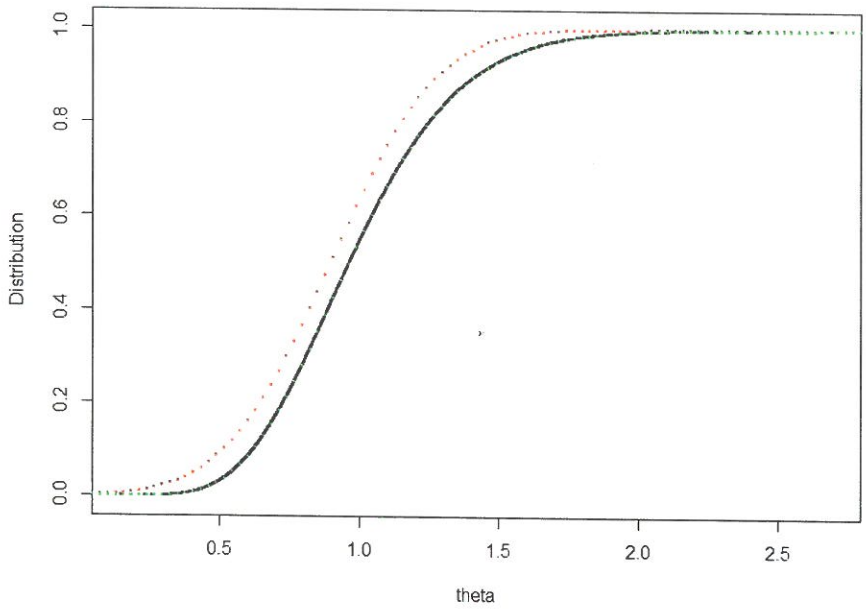

Figure 5.4: Comparison of the exact (black line) distribution and their approximations

For Gamma prior as informative prior and for different values of $\alpha's$ and $\beta's$ we get the following approximations as presented in the table

α	β	n	n_u	s	**App.**	$p(\Theta < \theta^1 \| X)$	$p(\Theta < \theta^2 \| X)$	$p(\Theta < \theta^3 \| X)$	**Mean**
1	2	10	9	9.9493	E	0.5421	0.9933	2.191e-10	0.8369
					FO	0.4111	0.9920	0.0021	0.9046
					HO	0.5523	0.9939	2.303e-10	0.8289
1	6	10	8	11.2640	E	0.5443	0.9929	0.0000	0.5213
					FO	0.2259	0.9072	0.0023	0.7102
					HO	0.6102	0.9961	0.0000	0.5060
4	2	10	10	10.9548	E	0.5356	0.9942	1.405e-06	1.0807
					FO	0.7195	0.9999	0.0078	0.9128
					HO	0.5139	0.9936	9.012e-07	1.0931

Table 5.1: First and Higher Order Approximation for different values of α and β - Censored Data

App.: Approximation, FO: First Order, HO: Higher Order, $\theta^1 = \alpha/\beta$, $\theta^2 = \theta^1 - 3\sqrt{\alpha}/\beta$, and $\theta^3 = \theta^1 + 3\sqrt{\alpha}/\beta$

Chapter 6

Higher Order Approximation to the Posterior Distribution for Multivariate Case

6.1 Posterior Approximation using Signed Root log-likelihood ratios

The notation in Sweeting (1996) and Sweeting and Kharroubi (2003, 2010) is used. For multivariate case. θ is a vector of parameter value where $\theta = (\theta^1,\theta^2,...,\theta^d) \in \Omega \subset \mathbb{R}^d$, $d \geq 1$. Let $\theta_i = (\theta^1,\theta^2,...,\theta^i)$ be the first i components of θ vector, and $\theta^{(i)} = (\theta^i,\theta^{i+1},...,\theta^d)$, the last $d-i+1$ components.

In addition to $\hat{\theta}_n = (\hat{\theta}_n^1, \hat{\theta}_n^2, ..., \hat{\theta}_n^d) = argmax\{L_n(\theta)\}$, define $\hat{\theta}_n^{i+1}(\theta_i)$ to be the maximizer of L_n conditional on θ_i. For $j > i$, $\hat{\theta}_i^j$ denotes the jth component of $(\theta_i, \hat{\theta}_n^{i+1}(\theta_i))$. For a function $g(\theta)$, when $1 \leq i < d$ we use $g(\theta_i)$ to denote $g(\theta_i, \hat{\theta}_n^{i+1}(\theta_i)) = g(\theta^1,\theta^2,...,\theta^i, \hat{\theta}_n^{i+1}, \hat{\theta}_n^{i+2}, ...,\hat{\theta}_n^d)$ and $\hat{\theta}_n^i(\theta_{i-1})$ is the unique solution of the conditional likelihood equation $l_i(\theta) = 0$, where $l(\theta) = logL(\theta)$ is the log-likelihood function and $l_i(\theta) = \partial l(\theta)/\partial\theta^i$.

Now define, $l'(\theta) = dl(\theta)/d\theta = (l_1(\theta), l_2(\theta), ...l_d(\theta))^T$, $j(\theta) = -d^2 l(\theta)/d\theta^2$ and $J = j(\hat{\theta})$, the observed information. For $i = 1, ..., d$ the log-likelihood ratios

$$w_n^i = w_n^i(\theta_i) = 2\{l_n(\theta_{i-1}) - l_n(\theta_i)\}$$

and the signed root transformation

$$r_n^i = r_n^i(\theta_i) = sign\{\theta^i - \hat{\theta}_n^i(\theta_{i-1})\}\{w_n^i\}^{1/2}$$

Note that $w_n = \sum_i w_n^i = 2\{l_n(\hat{\theta_n}) - l_n(\theta_n)\}$, and r^i is a function of the first i components $\theta_i = (\theta^1, ..., \theta^i)$ of θ

Writing $W_n = w_n(\theta)$ and $R_n = (r_n^1(\theta_1), ..., r_n^d(\theta_d))$. Suppressing n from now on, as in Sweeting (2010) the density $f(r)$ of R satisfies

$$f(r) \propto \phi(r) \prod_{i=1}^{d} q^i(r_i)(1 + \epsilon^T r) \tag{6.1}$$

where now $\phi(.)$ is the d-dimensional standard normal density, ϵ is an $\mathcal{O}(n^{-3/2})$ sequence independent of θ and

$$q^i(r_i) = \{-r^i/l_i(\theta_i)\}\{|j^{(i)}(\theta)|^{1/2}/|j^{(i+1)}(\theta)|^{1/2}\} \tag{6.2}$$

where $j^{(i)}$ is the submatrix of j corresponding to $\theta^{(i)}$ (Setting $|j^{(d+1)}(\theta)| = 1$). As in equation 6.2, q^i is of a function of r^i and is assumed to be in this form

$$q^i(r_i) \doteq 1 + a^i(r_{i-1})r^i + b^i(r_{i-1})(r^i)^2 + c^i(r_{i-1})(r^i)^3$$

where $a^i(r_{i-1}) = \mathcal{O}(n^{-1/2})$, $b^i(r_{i-1}) = \mathcal{O}(n^{-1})$, $c^i(r_{i-1}) = \mathcal{O}(n^{-3/2})$.

To $\mathcal{O}(n^{-2})$, the constant of proportionality in 6.1 is $\prod_{i=1} s^i$, where $s^i = 1 + b^i$. As in single parameter case, s^i may be calculated without expansion by noting that:

$$s^i \doteq \frac{1}{2}(q^i(-e_i) + q^i(-e_i)) \tag{6.3}$$

where e_i is the i-dimensional vector $(0, ..., 0, 1)$. Let θ^{i+} and θ^{i-} be the solutions to the equations $r^i(\hat{\theta}_{i-1}, \theta^i) = +1$ and $r^i(\hat{\theta}_{i-1}, \theta^i) = -1$, and write $\theta_i^+ = (\hat{\theta}_{i-1}, \theta^{i+})$ and $\theta_i^- = (\hat{\theta}_{i-1}, \theta^{i-})$. Then from 6.3 we have:

$$s^i \doteq \frac{1}{2}|J^{(i)}|^{1/2}(\lambda(\hat{\theta}))^{-1}\tau^i$$

where $\tau^i = (\nu_i(\theta_i^-)/l_i(\theta_i^-)) + (-\nu_i(\theta_i^+)/l_i(\theta_i^+))$, and $\nu_i(\theta) = \lambda(\theta)|j^{(i+1)}|^{-1/2}$ Then we obtain the following approximation,

$$\int L(\theta)\lambda(\theta)d\theta \doteq (2\pi)^{d/2}|J|^{-1/2}L(\hat{\theta})\lambda(\hat{\theta})\prod_{i=1}^{d} s^i$$

As in univariate case, formula 5.5 can be used to compute an approximation to the posterior expectation of a general formula $\upsilon(\theta)$. This leads to the formula

$$E(\upsilon(\theta)|X) \doteq \upsilon(\hat{\theta})\prod_{i=1}^{d}\frac{s^{*i}}{s^i}$$

where $\upsilon_i^- = \upsilon(\theta_i^-), \upsilon_i^+ = \upsilon(\theta_i^+), \hat{\upsilon} = \upsilon(\hat{\theta})$. Since $s^{*i}/s^i = 1 + \mathcal{O}(n^{-1})$, we can deduce that the alternative summation form

$$E(\upsilon(\theta)|X) \doteq \hat{\upsilon} + \sum_{i=1}^{d}\left\{\alpha_i^-\upsilon_i^- + \alpha_i^+\upsilon_i^+ - \hat{\upsilon}\right\}$$

6.2 Example: Censored Regression

We consider the censored failure data given by Crawford (1970) presented below in table 6.1. These data arise from temperature accelerated life tests on electrical insulation in n = 40 motorettes. Ten motorettes were tested at each of four temperatures in degrees Centigrade, resulting in l = 17 failed (i.e uncensored) units and n-l = 23 unfailed (i.e censored) units.

150°C	170°C	190°C	220°C
8064*	1764	408	408
8064*	2772	408	408
8064*	3444	1344	504
8064*	3542	1344	504
8064*	3780	1440	504
8064*	4860	1680*	528*
8064*	5196	1680*	528*
8064*	5448*	1680*	528*
8064*	5448*	1680*	528*
8064*	5448*	1680*	528*

(* denotes a censored time)

Table 6.1: Life test data on mottorettes - Insulation life in hours at various test temperatures

As in Schmee and Hahn (1979), we fit a model of the form

$$y_i = \beta_0 + \beta_1 v_i + \sigma\epsilon_i$$

where y_i is $\log_{10}$(failure time), with time in hours,v_i = 1000/(temperature + 273.2) and ϵ_i are independent standard normal errors. Reordering the data so that the first l observations are uncensored, with observed log-failure times y_i, and the remaining nl are censored at times c_i. The log-likelihood function is

$$l(\theta) = l(\beta_0, \beta_1, \sigma) = -\log\sigma - \frac{1}{2}\sum_{i=1}^{l}\left(\frac{y_i - \beta_0 - \beta_1 v_i}{\sigma}\right)^2 + \sum_{i=l+1}^{n}\log\left\{1 - \Phi\left(\frac{c_i - \beta_0 - \beta_1 v_i}{\sigma}\right)\right\}$$

where θ parameter is a vector of three dimensions $\theta^1 = \beta_0$, $\theta^2 = \beta_1$ and $\theta^3 = \sigma$ and Φ is the standard normal distribution function.

The score function $l'(\theta) = dl(\theta)/d\theta = (l_1(\theta), l_2(\theta), l_3(\theta))^T$ where

$$l_1(\theta) = \partial l(\theta)/\partial\theta^1 = \sum_{i=1}^{l} - \frac{\phi(e)}{\sigma(1-\Phi(e))} - \sum_{i=l+1}^{n}\frac{-y_i + \beta_0 + \beta_1 v_i}{\sigma^2}$$

$$l_2(\theta) = \partial l(\theta)/\partial\theta^2 = \sum_{i=1}^{l} - \frac{v_i\phi(e)}{\sigma(1-\Phi(e))} - \sum_{i=l+1}^{n}\frac{-v_i(y_i - (\beta_0 + \beta_1 v_i))}{\sigma^2}$$

$$l_3(\theta) = \partial l(\theta)/\partial\theta^3 = \sum_{i=1}^{l} - \frac{e\phi(e/\sigma)}{\sigma^2(1-\Phi(e/\sigma))} - \sum_{i=l+1}^{n}\frac{-y_i + (\beta_0 + \beta_1 v_i)^2}{\sigma^3} + l/\sigma$$

and e = $(c_i - (\beta_0 + \beta_1 v_i))/\sigma$

Using non-linear optimization function *nlm* in *R* that carries Newton-type Method, we find $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}) = (-6.0193, 4.3112, 0.2592)$ and the Hessian matrix

$$H = \begin{bmatrix} 427.8675 & 931.9110 & -251.3856 \\ 931.9110 & 2035.2276 & -558.2911 \\ -251.3856 & -558.2911 & 614.8931 \end{bmatrix}$$

$$\theta^- = \begin{bmatrix} -6.9661 & 4.7498 & 0.2751 \\ -6.0192 & 4.2892 & 0.2437 \\ -6.0192 & 4.3112 & 0.2236 \end{bmatrix} \; and \; \theta^+ = \begin{bmatrix} -5.0724 & 3.8784 & 0.2592 \\ -6.0192 & 4.3336 & 0.2829 \\ -6.0192 & 4.3112 & 0.3056 \end{bmatrix}$$

In case of non informative prior $\lambda(\theta) \propto 1/\sigma$, the posterior expectation of the function $v(\theta) = \beta_0 + \beta_1 + \sigma$, obtained via methods used in Tanner and Wong (1987), is -1.4989, while the first order approximation based on maximum likelihood estimation is -1.4488. The values of α^+ and α^- are (0.4555, 0.5742, 0.55434) and (0.5445, 0.4258, 0.4456) respectively. The approximate posterior expectation of $v(\theta)$ using 5.3 is -1.4625.

Assuming β_0,β_1 and σ are independent with the following prior distributions $N(\hat{\beta}_0, 1/J[1,1]^{-1/2})$, $N(\hat{\beta}_1, 1/J[2,2]^{-1/2})$, and $X^2(1)$, the approximate posterior expectation using metropolis random walk with 10000 iterations and 5000 burn in is -1.4489.

Chapter 7

Conclusion

This report starts with some elementary concepts in Bayesian inference that are developed later to show a comparison between first and higher order approximations for different characteristics including densities and expected values. The higher order approximations that are represented by log-likelihood, or log-posterior density ratios show good results up to $\mathcal{O}(n^{-3/2})$.

Some good features of log-likelihood approximations are that they are easy to obtain, the need only for the second order derivatives and not beyond, and for implementation for θ^{+} and θ^{-}, only $\hat{\theta}$ and J are required. Generally, they can be considered as a good start to reach exact computations.

Also, there are many stochastic simulation techniques can be used to obtain approximations such as Metropolis and Gibbs sampling. The use of metropolis sampling in the censored regression example for multivariate case shows sufficient accuracy, and for low computational power it gives a similar result as the first order normal approximation.

The presented chapters here can be considered as a good start for further research and to investigate more computational tools in approximations such as the different methods in computing an integrated likelihood and Bayesian computation presented in Zhenyu and Severn (2017) and the hybrid methods presented in Kharoubi and Sweetings (2010).

Bibliography

[1] Kharroubi, S.A. and Sweeting, T.J. (2010), Bayesian Analysis, 5, 787-816.

[2] Sweeting, T.J. and Kharroubi, S.A. (2003), Test, 12, 497-521.

[3] Sweeting, T.J. (1996), Approximate Bayesian computation based on signed roots of log-density ratios. Bayesian Statistics, 5, 427-444.

[4] Zhenyu and Severn (2017), Integrated likelihood computation methods, Comput Stat, 32, 281 - 313.

[5] Ruli E. and Ventura Laura (2014), Higher-order Bayesian Approximations for Pseudo-posterior Distributions, Journal Communications in Statistics - Simulation and Computation, 45, 2863-2873.

[6] Ruli, E., Sartori N., Ventura L. (2012), A note on marginal posterior simulation via higher-order tail area approximations, Bayesian Analysis, arXiv:1212.1038v1 [stat.CO].

[7] Tierney, L. and B. Kadane, J. (1986), Fully exponential Laplace approximations for posterior moments and marginal densities, Journal of the American Statistical Association 81, 82-86.

[8] Suess, E. A., and Trumbo, B. E. (2010), Introduction to Probability Simulation and Gibbs Sampling with R, New York : Springer.

[9] Tierney, L., Kass, R. and B. Kadane, J. (1989), Fully Exponential Laplace Approximations of Expectations and Variances of Non-Positive Functions, Journal of the American Statistical Association 84, 710-716.

[10] Tierney, L., Kass, R. and B. Kadane, J. (1989), Fully Exponential Laplace Approximations of Expectations and Variances of Non-Positive Functions, Journal of the American Statistical Association 84, 710-716.

[11] Barndorff-Nielsen, O.E. and Hall, P. (1988). On the level-error after Bartlett adjustment of the likelihood ratio statistic. Biometrika, 75, 374 - 378.

[12] Barndorff-Nielsen, O.E. and Cox, D. R (1989), Asymptotic Techniques for Use in Statistics. London: Champan and Hall.

[13] Barndorff-Nielsen, O.E. (1991), Modified signed log likelihood ratio Biometrika, 78, 557563.

[14] Bishop. C. M., (2006) Pattern Recognition and Machine Learning. Springer New York.

[15] Tanner, M.A. and Wong, W.H. (1987), The calculations of posterior distributions by data augmentations, 82, 528-540.

[16] Jeffreys, H.S., (1961), Theory of Probability. Oxford: Oxford University Press.

[17] Statisticat, L., 2014a. Bayesian inference. URL.http://www.bayesian-inference.com/bayesian.