UNIVERSITY OF NAIROBI

COLLEGE OF BIOLOGICAL AND PHYSICAL SCIENCES
SCHOOL OF MATHEMATICS

BAYESIAN SPATIAL AND SPATIOTEMPORAL MODELLING (Applied to precipitation dataset)

Erick Otieno Arieda Okuto

I56/68975/2011

A RESEARCH PROJECT REPORT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE OF MASTER OF SCIENCE IN BIOMETRY OF THE UNIVERSITY OF NAIROBI.

JULY 2013

DECLARATION

This research project is original and has never been presented for examination at any of the learning institution/University whether in Kenya or elsewhere.

STUDENT

OKUTO ERICK OTIENO ARIEDA

SIGN: ------------------------------- DATE: -------------------

This project has been submitted for examination with the approval of the following as University supervisor.

SUPERVISOR

1. Prof. J.A.M OTTIENO

   SIGN: ------------------------------- DATE: -------------------

2. Dr. NELSON OWUOR

   SIGN: ------------------------------- DATE: -------------------

ACKNOWLEDGEMENT

Thanks to the almighty God for bringing me up to this far. He has truly been my Ebenezer. I wish to thank my supervisors Prof. J.A.M Ottieno and Dr. Nelson Owuor, School of Mathematics for their priceless support without which this project would have not been successful.

DEDICATION

This project is dedicated to my family members for their infinity support and guidance.

To Paul “For I know the plans I have for you say the Lord plans to prosper and not to perish.’’ Jeremiah 29:11.

Contents

EXECUTIVE SUMMARY

The daily rainfall data is one type of data that is not easy to fit a simple model. This is due to the fact that there may be several days when it rains and it doesn't rain in other days. Besides, this special feature of rainfall, it has spatial dependence which complicates its predictions under the classical paradigm. In addition, if the daily amount of rainfall is taken over time say one year or so together with appropriately chosen covariates, temporal effects may also affect the reliability of prediction rule. Furthermore, we may need the average amount of rainfall on a particular day at given sites where we have ongoing experiments with unknown amount of rainfall. We propose to fit a semi-continuous model using Bayesian method and take advantage of Stochastic Partial Differential Equation (SPDE) approach as provided in the Integrated Nested Laplace Approximation (INLA).

Due to scares resources, we may not be able to observe the amount or rainfall recorded within the entire statistical population. Sampled data can therefore be obtained from such sources including satellite images and weather stations. Irrespective of the data source, we need to determine with certainty; the average daily rainfall at specified points for the days of interest , distribution of rainfall within and outside the practical range over time and if possible interpolate and extrapolate taking into account effects of the available covariates, spatial dependence and temporal effects.

Surely, this problem is beyond classical interpolation methods like simple or ordinary kriging where; trend surface function is constant, variogram is constant in the whole area of interest and the target variable is assumed to follow approximately Gaussian distribution which at least is untrue in this case.

Since, we always try to go for the most flexible, most comprehensive and the most robust technique, we have adopted a Bayesian method in this work. Both covariates and random effects will be treated as random. We know that covariates are fixed effects; therefore the randomness here is on the uncertainty about their true values. We will create two outcome variables; one for the occurrence of rainfall and the other for the amount of rainfall. The occurrence which depict either success or failure probability will be assume a binomial likelihood while the amount which is strictly positive will assume a gamma likelihood. On the other hand, fixed effects which include elevation and distance from the

sea will assume $N(0, 0.001)$ where 0.001 is the precision parameter equivalent to variance of 1000 while the random effects; that is temporal effects will assume an autoregressive process of order one (ar1) and the spatial dependence will assume stochastic partial differential equation (SPDE). Due to the challenges that MCMC schemes have faced over the years, that is convergence and running time, we adopt the Bayesian technique using numerical methods as given by the Integrated Nested Laplace Approximation (INLA).

CHAPTER ONE

1.0 Background of the study

The aim of this project was to build robust models that may address spatial and temporal variation in data that may lead to misinterpretation of analysis results. We may want to predict agricultural produce based on a few sampled known farms in a given country. The produce is known to depend on a number of covariates, among them amount of rainfall which may largely have both space (spatial) effect and time (temporal) effect (or variation). What would be the best way to handle this situation?

I suggest building two Bayesian spatial models where amount of rainfall is the dependent variable and its predictors are elevation and distance from the sea among other such covariates. We interpolate amount of rainfall and obtain the predictive means on specific points where the target farms are located. If the target farms are everywhere (the statistical population) then we obtain the corresponding predictive means for very point on the surface. Having obtained the required estimates we can then build the second regression model with crop yield and the independent variables(predictor variables) including rainfall (the previously obtained predictive means will be used as the rainfall values) and other predictors of interest included in the regression. If our interest was only rainfall interpolation, then the task ends at first model. Data sources for rainfall might include satellite imagery, weather stations among other such sources. Whichever source, we can obtain the posterior estimates (updated means) for the parameters of interest as well as predictive means for the target locations whose amount of rainfall were unpredictable.

When all the parameters of interest are represented by an appropriately chosen probability distribution we can obtain an exact posterior distributions (the estimate given data, assumption which might include the assumed prior distribution and likelihood) using R-INLA and subsequently the predictive means. **Predictive means** are not similar to **posterior means**, of course they are related. Whereas, posterior means is the estimate given the data and other sets of assumptions, predictive means normally done after estimation is the posterior mean given the updated mean (previously obtained posterior mean) and information from the data excluding the one we are predicting. If we have say

100 observations in the data, posterior mean is the mean of parameter estimate given set of assumptions. Take for instance we wanted to know the value of 101th observation from another location. Here we take into consideration the fact that our belief has been already updated and find the resultant posterior mean given the updated mean and 100 observations. This can subsequently be done for all the possible unvisited locations which turn out to be interpolation if our target locations are equivalent to statistical population.

One major breakthrough in Bayesian estimation came with invention of software called INLA which has been implemented as an R package (http://www.r-inla.org/home). With INLA, a number of statistical analyses are possible.

The Bayesian methods in INLA are quite similar to Markov Chain Monte Carlo (MCMC) methods, commonly used in Bayesian analysis preferably in WINBUGS software, except:

MCMC method requires that the simulation converges for one to be pretty sure that we have a reliable posterior distribution; otherwise we may obtain false positive estimates. Convergence is not guaranteed. Also, even if we have convergence, what is normally referred to as posterior distribution is always almost an approximation to the posterior distribution but not the posterior distribution itself. We have worked in this scenario because this has provided a major breakthrough to Bayesian methods from around 1990. However, reducing the challenges of estimation such as long computing time and the difficulty of reaching convergence that affected MCMC based methods have always been intensified. INLA method though is still under development has since addressed this challenge.

INLA uses numerical methods to obtain "exact posterior distributions". Where MCMC algorithm needs hours or days to run, INLA method provide more precise estimates in seconds or minutes (http://www.math.ntnu.no/~hrue/r-inla.org/papers/inla-rss.pdf)

As in most analysis cases, we can adjust our assumptions when we wish to meet certain goals. Consider the semi-continuous case, where rainfall amount can only be represented logically as a positive value. As is customary in INLA, we use a Gaussian likelihood (the normal distribution), that assumes the variable rainfall can take any value from negative infinity to positive infinity. In such circumstances, we transform the variable by taking the logarithm, and later transform back the estimates we obtained for purposes of interpretation. In addition, we could use a gamma likelihood in which $0 \leq x \leq \infty$ where x is amount of rainfall. In both cases, only a single likelihood is used rather than the previous case. Gaussian likelihood is computationally more effective than otherwise. Perhaps, due to the fact that every scenario tends to be Gaussian as sample size grows large. This could be one of the other advantages of using R-INLA. Applicability of likelihoods like Gamma, beta, zero-inflated discrete related distributions among many others have since been useful.

In my analysis, I attached prior distribution to each of the parameters of interest. By default fixed effects (elevation and sea distance) are assumed to be from the normal distribution with mean zero and a small precision of 0.001, equivalent to variance of 1000, in notation form, $\beta \sim N(0, 0.001)$. The temporal correlation of rainfall values was assumed to take the form of Auto-regression order 1 (AR1) or order 2 (AR2), beyond which may not be computationally effective. Whereas in temporal effect; Markov property observations are related such that the ($present\ value \sim immediate\ past\ value$), in spatial effect; Markov property ($value\ at\ point\ say\ i \sim value\ at\ \ point\ say\ j$) are related by Mat$e'$rn correlation function which incorporates distance between each point from the others. This leads to a sparse matrix rather than the traditional variance covariance matrix which is associated with the "big n" problem while using MCMC methods. This is the basis of stochastic Partial Differential Equations (SPDE). Note: SPDE is used here as a probability distribution (random effect) that best approximates the space effect in a continuous domain setting.

1.1 Problem statement

Climate data provides a common example of data structure that requires spatial statistical methods for analysis. The daily rainfall data is a typical example of such datasets that is not easy to fit a simple model. This is partly because there are days when it rains and days when it doesn't rain. The rainfall distribution is very asymmetric. Similarly, taking a point at which no rain was recorded, then it is likely that there is no rain at neighboring sites. In addition, choosing a site where larger rainfall was recorded then we are likely to find rainfall at neighboring sites. If daily rainfall is recorded say, for a year or so, then we have an outcome with both spatial and temporal effects. Besides this, rainfall is a mixture of both occurrence and amount variables and as such no form of data transformation would be appropriate for such scenarios. The principal problem however, is that we have ongoing experiments at locations with unknown rainfall measurement.

So we build a jointly model for rainfall, taking into account these facts. We build a semi-continuous model that is a jointly model to occurrence and to positive values and interpolate accordingly. We will use a simple dataset from Kenya sampled weather stations to illustrate the method.

1.2 Study objective

The main objective of this study is to present the basic ingredients of the link between continuous domain and Markov models and show how to perform Bayesian spatial and spatiotemporal inference on a semi-continuous outcome using the R-INLA software package (http://www.r-inla.org).

1.3 Significance of the study

We might want to predict the amount of say maize production in the country and identify places with the highest production and more interestingly predict steps forward where n is large enough. This is a complex problem to focus on given that production depends on climate factors like rainfall with high variability. We might not know the average amount of rainfall at the known farms hence the need to develop a model using the data collected from the few available weather stations and predict based on the Bayesian model, the average amount of rainfall from the unvisited locations (non-weather stations including targeted farms). The predictive posterior means can then be used as covariate in the second model which in our view is more reliable and the second model may be more robust to predict the production with certainty.

CHAPTER TWO

2.0 REVIEW OF CLASSICAL SPATIAL STATISTICS

2.1 Introduction

The daily rainfall data is one type of data that is not easy to fit a simple model. This is due to the fact that there are not days when it rains and it rains in other days, the distribution of rainfall is logically asymmetric. Besides, this special feature of rainfall, it also have spatial dependence which complicates its predictions under the classical paradigm. In addition, if the daily amount of rainfall is taken over time say one year together with appropriately chosen covariates, temporal effects may also affect the reliability of prediction rule. In some cases, researchers have ignored the spatial effect and concentrated on a uni-dimensional analysis of time series which has then been used to predict the future.

Due to scares resources, we may not be able to observe the amount or rainfall recorded within the entire statistical population. Sampled data can therefore be obtained from such sources including satellite images and weather stations. Irrespective of the data source, we need to determine with certainty; the mean daily rainfall at specified points for the days of interest , distribution of rainfall within and outside the practical range over time and if possible interpolate and extrapolate taking into account effects of the available covariates, spatial dependence and temporal effects.

Due to the climatic variations, farmers for instance cannot plan for the beginning of growing season with some degree of certainty. Other factors have since contributed to climate change and it has been difficult to predict future scenarios with certainty. A robust model will enhance climate adaptation and mitigation. In East Africa for instance, the summer beginning of growing season used to be around January to early February. This has since been fluctuating between March and April and in our view, a robust model that account for the uncertainty may be necessary to predict future scenarios.

Surely, this problem is beyond simple or ordinary kriging where; trend surface function is constant, variogram is constant in the whole area of interest and the target variable is assumed to follow approximately Gaussian distribution which at least is untrue in this case. The constant mean is usually subtracted from each of the data points and the residuals interpolated.

Similarly, the widely used universal kriging (kriging with trend) which has been called names depending on an individual's interest including co-kriging, regression kriging and such other names. Here, a trend is modeled across the domain. The fitted model may contain both fixed effects (covariates) and random effects in the classical meaning which might include the temporal effects. The trend (fitted) values are subtracted from the point data values and the residuals interpolated. But, as in ordinary kriging, the variogram is constant in the whole area of interest and the outcome expected to be Gaussian so as to be computationally effective (Hengl, 2009).

Indicator and probability kriging in which a threshold is established and the probability that each data point is greater than the established threshold interpolated might be irrelevant in this case where we have covariates and other effects . In our view, this can be a more improved version of simple or ordinary kriging.

A model may have to represent a reality, but it is not itself a reality. According to (Box, 1976) "All models are wrong but some are useful". While a model can never be truth, a model might be ranked from very useful, to useful, to somewhat useful, to finally essentially useless.

As we saw, there are many interpolation techniques that can be used to map rainfall and other climatic phenomena. In fact, we could not exhaust them. Researchers have sometimes adopted modified versions of the above techniques for instance weighted least squares have been used to obtain fitted values in regression kriging with the hope that the magnitude of noise will be minimized.

In reality, we always try to go for the most flexible, most comprehensive and the most robust technique. In fact, many (geo) statisticians hence researchers believe that there is only one Best Linear Unbiased Prediction (BLUP) model for spatial data.

Diggle and Ribeiro, 2007 (Model based geostatistics (Interpolation in a continuous space), page 157) after a six book chapters on classical interpolation techniques wrote in their final chapter (Bayesian method) "An obvious concern with the two phase approach in classical interpolation is that of ignoring uncertainty in the parameter estimates which may lead to optimistic assessments of predictive accuracy. It is possible to address this concern without being Bayesian, but in our view the Bayesian approach gives a more elegant solution and it is the one which we have adopted in our own work”.

(Cressie, 1993)asserts that one way to overcome the limitations of assumed constant variogram across the study area is by adoption of Bayesian methods which at the time had well-grounded theory in literature without computing.

(Votano, Parham, & Hall, 2004)also emphasized this in his introduction to a 432 page book on Spatial data analysis theory and practice by saying “Bayesian approaches attracted attention in the 1990’s, in part because of availability of numerical methods within new software for fitting a wide range of models. Prior to 1990’s much spatial modeling was based on spatial modification to the linear regression model in which, for example, spatial dependence was modeled through the response variable and semi-variograms. There were few applications of Bayesian methods (Hepple, 1979). Bayesian methods despite having challenges of computational efficiency, have introduced more interesting ways for modeling the effects of spatial dependence”

The invention of WinBugs for Bayesians using Monte-Carlo methods was therefore a major breakthrough in early 1990’s (Congdon, 2006) and (Ntzoufras, 2009). Since it uses the principle of Markov chain (present outcome depend on its immediate past), Metropolis and Gibbs algorithms were invented to cater for both dependent and independent outcomes. Despite this development, MCMC methods still faced computational and convergence challenges and many researchers would prefer the traditional classical methods than one that would run for hours or days. Also, Mathematicians and statisticians concentrated on the theory of Bayesian methods which indeed is an applied mathematics problem. Instead, researchers were shy getting involved mathematical jargons and preferred the said traditional methods for convenience (P. J. Diggle, 2011).

INLA software which is also available as an R package (R-INLA) has since addressed these challenges. We might not have any valid reason for assuming that the coefficients of independent variables are constants and unknown any more especially now that a researcher friendly software is available that runs in seconds or minutes where MCMC could take hours or days besides being convergence problem free. That is Integrated Nested Laplace Approximation (INLA).

2.2 History of Spatial statistics

It is past of human nature to discover patterns from a seemingly arbitrary set of events. According to applied spatial statistics for public health data, a book by Walter and Samuel Wilks, 1965, we are taught from an early age to `correct the dot, learning that if we correct the right dots in the right way, a meaningful picture will emerge. People around the world look to the right sky and create patterns among the stars.

Although best known among spatial analysts for the broad street maps, it was Dr. Snow careful case definition and analysis of cholera deaths in a wider area of London that placed him among the founders of epidemiology rather than from his maps (Lilienfeld and stolley, 1984) central to this was Snows. Natural experiment, where he categorized cholera deaths by two water companies, one drawing waters upstream from London (and its sewage) the other downstream. The water company service was so intermingled that in many cases a single house, has a supply different from that on either side (Snow 1936 p.75)

This in addition to maps, study design and simple spatial statistics were important tools in Snows analysis.

According to Walter and Samuel Wilks, 2004, applying statistical methods in a spatial setting raises several challenges. Geographer and statistician (Tobler, 1970) summarized a key component affecting any analysis of spatially referenced data through his widely quoted and paraphrased first law of geography

Everything is related to everything else, but near things are more related than far things, (Tobler, 1970).This law succinctly defines the statistical notion of(positive) spatial autocorrelation, in which .pairs of observations taken nearby are more alike than those taken farther apart .Weakening the usual assumption of the independent observations in spatial trend in the probabilistic expected values of each observations.

By allowing spatial correlation between observations, observed spatial similarities in observation may be due to a spatial trend, spatial autocorrelation, or both. Second, a set of correlated observations contains less statistical information than the same number of independent observations. (Cressie, 1993) provides an example of the reduction in effective sample size induced by increasing autocorrelation. The result is a reduction in statistical precision in estimation and prediction from a given sample size of a correlated data compared to what we would see in the same sample size of independent observations e.g. confidence intervals based on independent observations are too narrow to reflect appropriate uncertainly associated with positive correlated data. This is well explained in the book applied spatial statistics for public health by Walter and Samuel Wilks 2004.

In reference to model based Geostatistics by (P. Diggle & Ribeiro, 2007), geostatistics refers to the sub branch of spatial statistics in which the data consist of a definite sample of measured values relating to an underlying spatially continuous phenomenon (P. J. Diggle, Menezes, & Su, 2010). Example they gave include; height above sea level in topographical survey; determination of soil properties from care sample e.t.c. According to the clue, the subject has an interesting history. Originally, the term geostatistics was coined by Georges Malheron and colleagues at Fontainainebleau, France, to describe their work addressing problems of spatial 'prediction arising in the mining industry. See, also (Malheron, 1963 and Malhenron, 1971b).The ideas of the Fontainebleau school were developed largely independently of the mainstream of spatial statistics.

These parallel development included work by Kolmogorov, (1941) and Matérn, (1960), reprinted as Matérn, (1986) whittle (1954, 1962, 1963), Bartlelt (1964, 1967) and others. For instance, according to (P. Diggle & Ribeiro, 2007), the case geostatistical method known as simple krigging is equivalent to minimum mean squared error prediction under a linear Gaussian model with known parameter values. Papers by Wilson (1971, 1972) and the book by Ripley (1981) made this connection explicit.

(Cressie, 1993) considered geostatistics to be one of three main branches of spatial statistics, the others being discrete spatial variation (covering distribution on lattices and mark or random fields) and spatial point processes. Fortunately, geostatistical methods are now used to in many areas of application and far beyond the mining context in which they were originally developed.

Diggle and Ribeiro, 2007, despite the apparent integration with spatial statistics, much geostatistical practice still reflects its independent origins and from a main stream statistical perspective, this has some undesirable consequences.

In particular, explicit stochastic models are not always declared and adhoc methods of inference are often used, rather than the likelihood based methods of inference which are central; to modern statistics.

The potential advantages of using likelihood–based methods of inference according to them are twofold. They generally lead more efficient estimation of unknown model parameters, and they allow for the property assessment of the uncertainty in spatial predictions including an allowance for the effects of uncertainty with the estimation of model parameters.

Diggle, Town and Moyeed, (1998) coined the phrased model based geostatistics to describe an approach to geostatistical problems based on the application of formal statistical methods under an explicitly assumed stochastic model. The book by Diggle and Ribeiro, 2007 takes this approach.

According to a book titled "practical guide to geostatistical mapping by (Hengl, 2009), in which he clarifies that spatial prediction model(algorithms) can be classified according to the amount of statistical analysis i.e. amount of expert knowledge included in the analysis.

He classified these models into three groups. First, mechanical (deterministic) models are models where arbitrary or empirical parameters are used. No estimate of the model error is available and usually no strict assumptions about the variability of a feature exist.

The most common techniques that belong to this group include; Thiessen polygons, inverse distance interpolation, regression on coordinate natural neighbor's splines e.t.c.

The second group, linear statistical (probability) models, the model parameters are commonly estimated in an objective way, following probability theory. The predictions are accompanied with an estimate of the prediction error. The drawback is that the input data set usually needs to satisfy strict statistical assumptions. There are at least four groups of linear statistical models; Krigging (plain geostatistics), environmental correlation (e.g. regression based) Bayesian based models (e.g. Bayesian maximum entropy) hybrid models (e.g. regression krigging) etc.

2.3 How classical kriging works

The Inverse Distance Weighted (IDW) and Spline interpolation tools are referred to as deterministic interpolation methods because they are directly based on the surrounding measured values or on specified mathematical formulas that determine the smoothness of the resulting surface. A second family of interpolation methods consists of geostatistical methods, such as kriging, which are based on statistical models that include autocorrelation—that is, the statistical relationships among the measured points. Because of this, geostatistical techniques not only have the capability of producing a prediction surface but also provide some measure of the certainty or accuracy of the predictions.

Kriging assumes that the distance or direction between sample points reflects a spatial correlation that can be used to explain variation in the surface. The Kriging tool fits a mathematical function to a specified number of points, or all points within a specified radius, to determine the output value for each location. Kriging is a multistep process; it includes exploratory statistical analysis of the data, variogram modeling, creating the surface, and (optionally) exploring a variance surface. Kriging is most appropriate when you know there is a spatially correlated distance or directional bias in the data. It is often used in soil science and geology.

Kriging formula

Kriging is similar to IDW in that it weights the surrounding measured values to derive a prediction for an unmeasured location. The general formula for both interpolators is formed as a weighted sum of the data:

$$\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i \; Z(s_i) \qquad (2a)$$

- where:

  $Z(s_i)$ = the measured value at the *i*th location

  λ_i = an unknown weight for the measured value at the *i*th location

  s_0 = the prediction location

  N = the number of measured values

In IDW, the weight, λ_i, depends solely on the distance to the prediction location. However, with the kriging method, the weights are based not only on the distance between the measured points and the prediction location but also on the overall spatial arrangement of the measured points. To use the spatial arrangement in the weights, the spatial autocorrelation must be quantified. Thus, in ordinary kriging, the weight, λ_i, depends on a fitted model to the measured points, the distance to the prediction location, and the spatial relationships among the measured values around the prediction location. The following sections discuss how the general kriging formula is used to create a map of the prediction surface and a map of the accuracy of the predictions.

Creating a prediction surface with kriging

To make a prediction with the kriging interpolation method, two tasks are necessary:

- Uncover the dependency rules.
- Make the predictions.

To realize these two tasks, kriging goes through a two-step process:

1. It creates the variograms and covariance functions to estimate the statistical dependence (called spatial autocorrelation) values that depend on the model of autocorrelation (fitting a model).
2. It predicts the unknown values (making a prediction).

It is because of these two distinct tasks that it has been said that kriging uses the data twice: the first time to estimate the spatial autocorrelation of the data and the second to make the predictions.

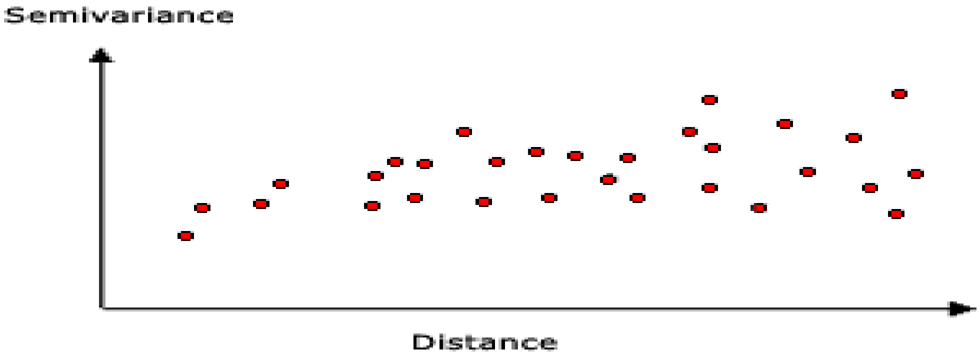
2.4 Variography

Fitting a model, or spatial modeling, is also known as structural analysis, or variography. In spatial modeling of the structure of the measured points, you begin with a graph of the empirical semivariogram, computed with the following equation for all pairs of locations separated by distance h:

Semi-variogram $(distance_h)$ = 0.5 * average $\{(value_i - value_j)^2\}$ $\quad (2b)$

The formula involves calculating the difference squared between the values of the paired locations.

Often, each pair of locations has a unique distance, and there are often many pairs of points. To plot all pairs quickly becomes unmanageable. Instead of plotting each pair, the pairs are grouped into lag bins. For example, compute the average semivariance for all pairs of points that are greater than 40 meters apart but less than 50 meters. The empirical semivariogram is a graph of the averaged semivariogram values on the y-axis and the distance (or lag) on the x-axis (see diagram below).



Empirical semivariogram graph example

Spatial autocorrelation quantifies a basic principle of geography: things that are closer are more alike than things farther apart. Thus, pairs of locations that are closer (far left on the x-axis of the semivariogram cloud) should have more similar values (low on the y-axis of the semivariogram cloud). As pairs of locations become farther apart (moving to the right on the x-axis of the semivariogram cloud), they should become more dissimilar and have a higher squared difference (moving up on the y-axis of the semivariogram cloud).

Fitting a model to the empirical semi-variogram

The next step is to fit a model to the points forming the empirical semivariogram. Semivariogram modeling is a key step between spatial description and spatial prediction. The main application of kriging is the prediction of attribute values at unsampled locations. The empirical semivariogram provides information on the spatial autocorrelation of datasets. However, it does not provide information for all possible directions and distances. For this reason, and to ensure that kriging predictions have positive kriging variances, it is necessary to fit a model—that is, a continuous function or curve—to the empirical semivariogram. Abstractly, this is similar to regression analysis, in which a continuous line or curve is fitted to the data points.

To fit a model to the empirical semivariogram, select a function that serves as your model—for example, a spherical type that rises and levels off for larger distances beyond a certain range (see the spherical model example below). There are deviations of the points on the empirical semivariogram from the model; some points are above the model curve, and some points are below. However, if you add the distance each point is above the line and add the distance each point is below the line, the two values should be similar. There are many semivariogram models from which to choose.

2.5 Semi variogram models

ArcGIS Spatial Analyst provides the following functions from which to choose for modeling the empirical semivariogram:

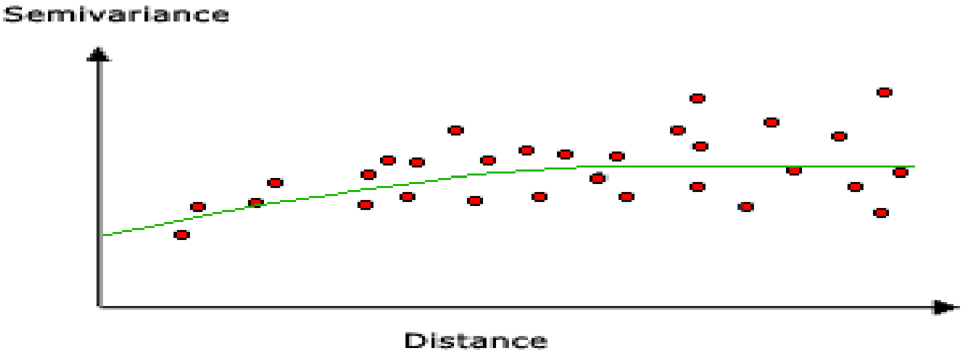
- Circular
- Spherical
- Exponential
- Gaussian
- Linear

The selected model influences the prediction of the unknown values, particularly when the shape of the curve near the origin differs significantly. The steeper the curve near the origin, the more influence the closest neighbors will have on the prediction. As a result, the output surface will be less smooth. Each model is designed to fit different types of phenomena more accurately.

The diagrams below show two common models and identify how the functions differ:

A spherical model

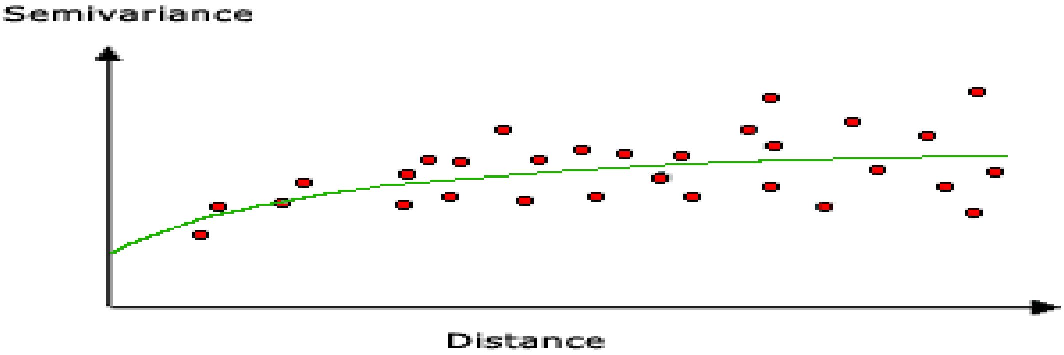
This model shows a progressive decrease of spatial autocorrelation (equivalently, an increase of semivariance) until some distance, beyond which autocorrelation is zero. The spherical model is one of the most commonly used models.



Spherical model example

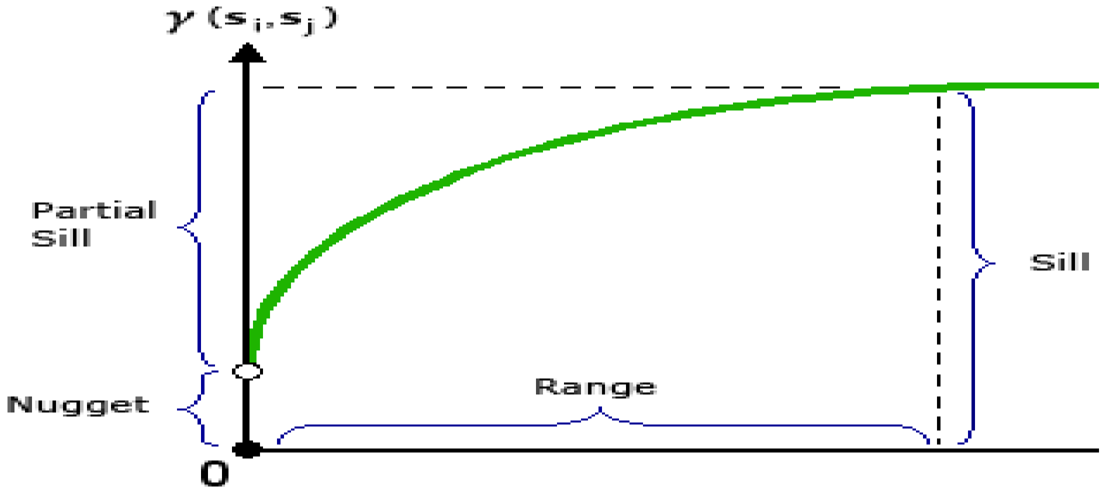
An exponential model example

This model is applied when spatial autocorrelation decreases exponentially with increasing distance. Here, the autocorrelation disappears completely only at an infinite distance. The exponential model is also a commonly used model. The choice of which model to use is based on the spatial autocorrelation of the data and on prior knowledge of the phenomenon.



Understanding semi-variogram

As previously discussed, the semi-variogram depicts the spatial autocorrelation of the measured sample points. Because of a basic principle of geography (things that are closer are more alike), measured points that are close will generally have a smaller difference squared than those farther apart. Once each pair of locations is plotted after being binned, a model is fit through them. Range, sill, and nugget are commonly used to describe these models.



Range and sill

When you look at the model of a semivariogram, you will notice that at a certain distance the model levels out. The distance where the model first flattens is known as the range. Sample locations separated by distances closer than the range are spatially auto correlated, whereas locations farther apart than the range are not.

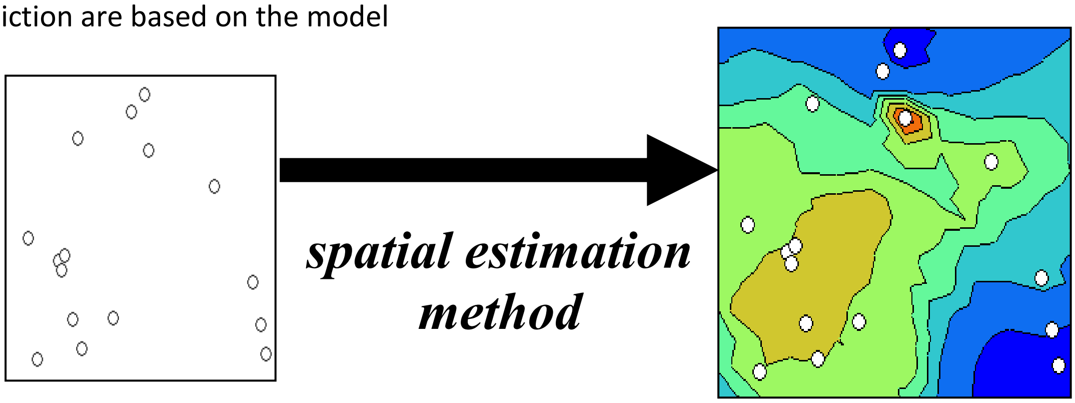
The value at which the semivariogram model attains the range (the value on the y-axis) is called the sill. A partial sill is the sill minus the nugget. The nugget is described in the following section.

Nugget

Theoretically, at zero separation distance (for example, lag = 0), the semivariogram value is 0. However, at an infinitely small separation distance, the semivariogram often exhibits a nugget effect, which is a value greater than 0. If the semivariogram model intercepts the y-axis at 2, then the nugget is 2.

The nugget effect can be attributed to measurement errors or spatial sources of variation at distances smaller than the sampling interval (or both). Measurement error occurs because of the error inherent in measuring devices. Natural phenomena can vary spatially over a range of scales. Variation at microscales smaller than the sampling distances will appear as part of the nugget effect. Before collecting data, it is important to gain an understanding of the scales of spatial variation in which you are interested.

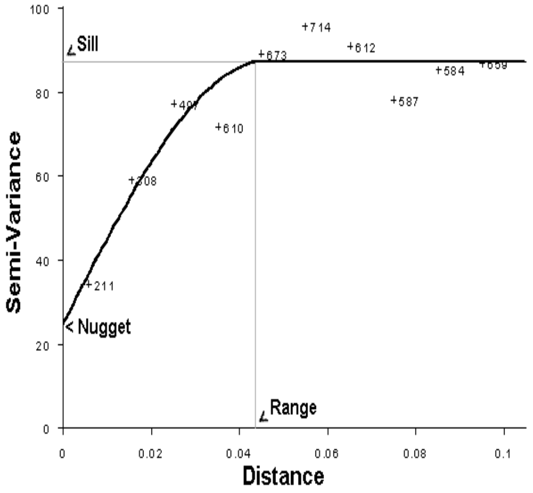
The standard version of krigging is called ordinary and simple krigging (O.K). Here the prediction are based on the model



spatial estimation method

point monitorin g data

continuous surface of estimates (map)



$$\begin{array}{ccccc} C & \times & w & = & D \\ \begin{bmatrix} C_{11} & \cdot & \cdot & \cdot & C_{1n} & 1 \\ \cdot & \cdot & & & \cdot & \cdot \\ \cdot & & \cdot & & \cdot & \cdot \\ \cdot & & & \cdot & \cdot & \cdot \\ C_{n1} & \cdot & \cdot & \cdot & C_{nn} & 1 \\ 1 & \cdot & \cdot & \cdot & 1 & 0 \end{bmatrix} & \times & \begin{bmatrix} w_{1j} \\ \cdot \\ \cdot \\ \cdot \\ w_{nj} \\ \mu \end{bmatrix} & = & \begin{bmatrix} C_{1j} \\ \cdot \\ \cdot \\ \cdot \\ C_{nj} \\ 1 \end{bmatrix} \end{array}$$

However, if

$y(s_i)$ Observation at site i depends on $y(s_j)$ observation in its neighbourhood, then the above model is likely to inflate type I error which may lead to a good fit when indeed the model is poor fit. Such a model cannot be used to predict the future with certainty.

Spatial lag model

$$y(s_i) = XB + \sum \rho\omega_{ij}\, y(s_j) + \mathcal{E}$$

With $\omega_{ij} = \begin{cases} 1 & \textit{if i and j are neighbours} \\ 0 & \textit{otherwise} \end{cases}$

In matrix form

$$Y = XB + \rho WY + \mathcal{E}$$

$$Y - \rho WY = XB + \mathcal{E}$$

$$Y(I - \rho W) = XB + \mathcal{E}$$

So that $$Y = (\boldsymbol{I} - \rho W)^{-1}XB + (\boldsymbol{I} - \rho W)^{-1}\mathcal{E}$$

With $$\mathcal{E} \sim MVN(0, \partial^2\boldsymbol{I})$$

$$Y \sim MVN((\boldsymbol{I} - \rho W)^{-1}X\beta, \Sigma)$$

Cov(y_i, y_j) may or may not be Zero depending on whether or not they are neighbours.

2.6 Simple/Ordinary kriging

Use a known (at least assumed) a constant mean across the domain. Usually used by subtracting point data values from mean, interpolating the residuals, and then adding back to the mean.

2.7 Universal kriging (kriging with a trend)

Uses a modeled trend across the domain. The trend is subtracted from the point data values and the residuals interpolated.

2.8 Indicator kriging

Point data are transformed to indicator variables (usually binary). For example, a threshold value is defined and data points that are below the threshold are assigned a value of 1, otherwise are assigned the value 0. The indicator values are interpolated and the resulting surface shows the probabilities (0-1) of exceeding (or being below) the threshold.

2.9 Probability

Similar to indicator kriging method but also incorporate the difference between a data point value and the defined threshold. Using this “proximity” to threshold information can result in more accurate probabilities.

Where $\boldsymbol{\mu}$ in the constant stationery function (global mean) and $\varepsilon(s)$ is the spatially correlated stochastic part.

The target variable is said to be stationary if several sample variograms are ‘similar’ (if they do not differ statistically) which is referred to as covariance stationary or second order stationary. In summary, three important requirements for ordinary and simple kriging are

Trend function is constant ($\boldsymbol{\mu}$=constant). The variogram is constant in the whole area of interest. The target variable follows (approximately a normal distribution). In practice, these requirements are often not met which is a serious limitation of this form of interpolation.

According to (Hengl, 2009) we always try to go for the most comprehensive and most robust technique (Preferably implemented in a software with a user friendly GUI). In fact, many (geo) statisticians believe that there is only one, best linear unbiased prediction (BLUP) model for spatial data, from which all other techniques can be derived. This is echoed by (Gotway and Stroup, 1997, Stein 1999, Christensen 2001).He introduces the concept of regression kriging, a superior version to .ordinary and simple kriging .It comprises of methods including universal kriging, kriging with external drift and co-kriging.

Despite the above development, all the geo (statistician) have a common believe that approximate Bayesian method will provide the most robust prediction of the outcome.

(Peter Diggle, 2010) in his support for a Bayesian approach to spatial statistics introduces preferential sampling once more in Bayesian paradigm. This method extensively been used under classical approach in their book. Model based geostatistics, 2007.

According to Diggle, geostatistics involves the fitting of spatially continuous models to spatially discrete data. Preferential sampling arises when the process being modeled are stochastically dependent. Conventional geostatistical method assume if only implicitly, that sampling is non-preferential. However these methods are often used in situations where sampling is likely to be preferential.

He would then give a general expression for the likelihood function of preferentially sampled geostatistical data and describe how this can be evaluated approximately by using Monte Carlo methods (MCMC) Diggle then proceeded to develop a geoR and geoRglm packages to facilitate these applications. However, as the tradition, this method was painfully slow even though its' achievement motivated Bayes' believers across the world. Rue et al, 2009 would two years later develop an Integrated Nested Laplace Approximation (INLA) which would give a more precise Bayesian estimates using numerical methods in seconds or minutes where the traditional MCMC methods would require hours or days.

The development of R-INLA enhanced even more development of Bayesian spatial statistics. Lindgren et al (2011).Continuous indexed Gaussian fields (GFS) is the most important ingredients in spatial statistical modeling and geostatistics. The specification through the covariance function gives an intuitive interpretation of the field properties. On the computational side, GFS are hampered with the big n problem, since the cost of factorizing dense matrices is cubic in the dimension. Although computational power is all time high, this fact seems still to be a computational bottleneck in many applications. Along with GFs, there is the class of Gaussian Markov random fields (GMRFs) which are discretely indexed. The Markov property makes the precision matrix involved sparse matrices for fields in R2 only use the square root of the time required by general algorithms. Lindgren showed that using an approximate stochastic weak solution to (linear) stochastic partial differential equations, we can, for some GFs in the marten class, provide an explicit link for any triangulation of between GFs and GMRFs, formulated as a basic function representation.

Cameletti et al (2012) then applied the concept developed in Lindgren et al 2011 to a hierarchical Spatio-temporal model in which R-INLA was used to developed approximate

Bayesian analysis giving more precise estimate than would be achieved using the only alternative traditional MCMC method that in addition comes at a high cost of computational time. The model for particulate matter (PM) concentration in the North Italian region Piemorite, involved a Gaussian field(GF), affected by a measurement error and a state process characterized by a first order autoregressive dynamic model .the model is well discussed in the paper published December 2012.

However, according to the paper, Bayesian inference though Markov chain Monte-Carlo (MCMC) techniques can be a challenge due to convergence problems and heavy computational loads. In particular, the computational issue refers to the infeasibility of linear algebra operations involving the big dense covariance matrices which occur when large spatio –temporal data sets are present. The main goal .of this paper was to present the most effective estimating and spatial prediction strategy for the considered spatio-temporal model. The model consists of GF with marten covariance function as a Gaussian Markov Random field. (GMRF) through the stochastic partial differential Equations (SPDE) approach. The main advantage of moving from a GF to a GMRF stems from good computational properties that the latter enjoys using R-INLA

Lindgren, 2012 has considered continuous domain spatial models giving step by step frameworks on how what traditionally has been called geostatistics can now be well using R-INLA and get the best prediction rule than ever before. Infant, Lindgren in his own words says ‘think continuous’

Finally in the (Krainski, 2013), he presents how we fit models to spatial point referenced data, the so called geostatistical models, using INLA and SPDE. He uses the try datasets and rainfall data to explicitly illustrate how to handle more complex model in this paradigm.

He considers a case for a non-Gaussian response, Semi-continuous model to daily rainfall, joint modeling a covariate with misalignment and non-stationery models using INLA. Non stationary models and many more expected to address challenges that researchers may face when dealing with spatial statistics.

CHAPTER THREE

REVIEW OF BAYESIAN STATISTICS

3.1 INTRODUCTION

Bayesian is a branch of statistics that concerns how we deal with evidence, how we deal with data, how we evaluate the evidence and measure the uncertainty involved. Update it as new knowledge arises and hopefully change minds in light of new data.

The classical statistics which is often referred to as Null hypothesis significant testing (NHST) (Kruschke, 2010) has many problems that should not be allowed in the 21st century. For instance,   In NHST, the data collector must pretend to plan the sample size in advance and pretend not to let preliminary looks at the data influence. Bayesian design, on the contrary, has no such pretenses because inferences are not based on the p values which depend on the sample size which on the other hand depend on the intension of the researcher.

In summary, the NHST analysis and conclusion depend on covert intensions of the experimenter, because those intentions define the space of all possible (unobserved) data. This dependence of the analysis on the experimenter's intentions conflicts with the opposite assumption that the experimenter's intentions have no effect on the observed data. The Bayesian analysis does not depend on the space of possible unobserved data. The Bayesian analysis operates only with the actual data obtained.

Moreover, in NHST, analysis of variance (ANOVA) has elaborate corrections for multiple comparisons based on the intentions of the analyst. Hierarchical Bayesian ANOVA uses no such corrections, instead rationality mitigating false alarms based on the data.

In many NHST analyses, missing data or otherwise unbalanced designs can produce computational problems. Bayesian models seamlessly handle unbalanced and small-sample designs.

Similarly, in multiple regression analysis, traditional analysis breakdown when the predictors are perfectly (or very strongly) correlated or if the number of predictors is more than the sample size this has always been partly resolved by using Partial least square regression (Yu et al, 2012) which borrows the concept of principle component analysis hence is not purely classical regression analysis, but the Bayesian analysis proceeds as usual and reveals that the estimated regression coefficients are anti-correlated (Kruschke, 2010).

The crucial problem with NHST is that the P-value is defined in terms of repeating the experiment, and what constitute the experiment is determinant by the experimenter's intentions. The single set of data could have a risen from many different experiments, and therefore the single set of data has many different P values. In all the conventional statistical tests, it is assumed that the experimenter intentionally fixed the sample size.

This dependence of P on the intended stopping rule for the data collection is well known, but rarely if ever acknowledged in applied textbooks on NHST.

The only situation in which standard NHST textbooks explicitly confront the dependence of P-value on experimenter intentions is when multiple comparisons are made. When there are several conditions are compared, each comparison inflates the probability of spuriously declaring a difference to be non-zero. To compensate for this inflation of false alarms, different "corrections" can be made on the P-value criterion used to declare significance. These corrections go by the names of Bonferroni, Scheff, Tukey, Dunnett, HSU or a variation called the false discovery rate (FDR) (Kruschke, 2010).

McNemar's test which is a normal approximation used on nominal data on a 2x2 contingency tables with dichotomous trait, with matched pair of subjects, to determine whether the row and column marginal frequencies are equal (Marginal homogeneity), is marred with corrections (Yates' correction) ranging from 0.1 to anything depending on the interest of the analyst. Bayesian methods are free of such influences.

NHST summarizes a data set with a value as t or f, which in-turn is based on a point estimate from the data, such as the mean and standard deviation for each group. The point estimate is the value for the parameter that makes the model most consistent with the data in the sense of minimizing the sum squared deviation or maximizing the sum squared deviation or maximizing the likelihood (or some other measure of consistency). Unfortunately, the point estimate provides no information about the range of other parameter values that are reasonably consistent with the data. Some researchers use confidence intervals for this purpose. But some NHST analyses do not easily provide confidence intervals, such as x^2 analyses of contingency table all probabilities. More fundamentally, confidence intervals are as fickle as P values because a confidence interval is simply the range of parameter values that would not be rejected by significance test. (And significance test depend on the intension of the analysts including the range of these possible values). Point estimates of parameters also provide no indication of correlation between plausible parameter values (Kruschke, 2010).

Statistical power in NHST is the probability of rejecting the null hypothesis when an alternative hypothesis is true. Because power increases with sample size, estimate of power are often used in research planning to anticipate the amount of data that should be collected, closely related in power is replication probability, which is the probability that a result found to be significant in one experiment will also be significant in another experiment. Replication probability can be used to assess the reliability of a finding. To estimate power and replication probability, the point estimate from a first experiment is used as an alternative hypothesis to contrast with the null hypothesis. Unfortunately, a point estimate yields little information about other alternative hypothesis that is reasonably consistent with the initial data.

The other hypothesis can span a very wide range, with each one yielding very different estimates of power and replications probability.

Therefore, the replication probability has been determined to be "virtually unknowable". Thus, NHST in combination with point estimation leaves the scientists with unclear Estimates of power and replication probability, and hence provides a very weak basis for assessing the reliability of an outcome (Kruschke, 2012).

As outlined above, NHST provides a palicity of dubious information. To obtain this, the analyst is also subject to many computational constraints for example, in analysis of variance (ANOVA), computations are much easier to conduct and interpret if all conditions have the same number of data points (i.e. so called balanced designs). Standard ANOVA also demands homogeneity of variances across the condition which is rarely practical (Kruschke, 2010).

In real life situation, assumption is rarely satisfied bringing focus on existing gap between real life and academia.

According to Benjamin Hobbs (1997), there is ample empirical evidence that people act contrary to the assumptions of Bayesians analysis. They fail to update prior beliefs using Bayes' law, and they have great difficulty specifying utility functions because their preferences are incoherent and consistent (Kahneman et al, 1982).

A Bayesian response to this criticism might be "so what?" The point of Bayesian analysis is to improve of upon unaided human judgment, upon to imitate it. Bayes' Law is not meant to be a psychological theory that can be used to predict behavior; rather it is supposed to be a guide to making more rational, consistent, and defensible decisions.

Bayesian analysis is the only integrated approach to inference and decision making that is fully consistent with a set of assumptions that have a normative appeal. Bayesian and application should complement each other otherwise the existing gap between academic and real world will continue to widen.

3.2 Bayes' theorem

One thing we do know about, reverend Thomas Bayes (1702-1761) (Bellhouse, 2004) is that he was an English Mathematician and Presbyterian minister, known for having formulated a specific case of the theorem that bears his name: Bayes' theorem, (1940). Bayes never published what would eventually become his most famous accomplishment; his notes were edited 10-15 years later after his death and published by Richard Price. We know very little about his lifetime and the picture we see is in the Wikipedia allegedly belong to one named after him. However, the Google has his handwriting from the institute of actuaries in London. During the time, there was a religion controversy on the existence of God the cause (primary/first cause). We do not know whether Thomas Bayes wanted to prove the existence of God but we do know that Bayes tried to deal with cause and effect mathematically and in so doing of course he produced a one line theorem that will never die (Sharon Bertsch, 2010).

According to Bayes', we modify our initial belief and he actually called it "initial guess" and if nothing seems to work, starts with "50-50" probability that it works and modify this guess with object new information and get an improved belief which we then carry with a commitment to update whenever new piece of information arrives. But as stated above, Bayes' did not believe in his theorem enough to publish it and he dies about 20 years later with this theorem in his notebook. Going through Bayes' papers, Richard Price decides that the theorem will help prove the existence of God. He would then spend the next two years off and on editing Bayes' theorem get it posthumously published in 1764 unfortunately, in a British journal that was neither read by Statisticians nor Mathematicians hence not get publicity. By today's standard, Richard Price would be considered as a Thomas Bayes' co-author.

However, we see two patterns emerging between the Bayesians and the then frequentist. Bayes' became an extreme example of the existing gap between academia and the real world (Sharon Bertsch, 2010).

By rights and as they did until 50 years later after this discovery, the theorem was entirely called Laplace theorem. As a young man of about 25 in the year 1774 discovered independently Bayes' theorem (Sharon Bertsch, 2010). According to Wikipedia, "in statistics, the so-called Bayesian interpretation of probability was developed mainly by Laplace." So until about 50 years later, Bayes rule was called Laplace theorem (Sharon Bertsch, 2010).

According to Laplace:

This is the foundation of what today we call Bayesian statistics. As illustrated this is entirely the work of Simeon Laplace commonly remembered only through the Laplace transform. We also agree that the theorem should be renamed entirely as Laplace theorem as they did up to 1827.

In a (very small!) nutshell, Bayesian inference boils down to the computation of posterior/predictive distributions.

Note: $p(\theta|y) = p(\theta)p(y)$. *The naïve version of probability that we learn of events that are independent.*

$$p\ (\theta/y) = \frac{P\ (y|\theta)P(\theta)}{\int p(y|\theta)P(\theta)d\theta}$$

where: $p(\theta|y)$ = posterior distribution;

$p(\theta) =$ prior distribution and

$p(y|\theta)$=likelihood distribution

Laplace theorem (Bayes' theorem). As a matter of fact from 1774 up to which Laplace died in 1827, applied statistics was entirely Bayesian (Sharon Bertsch, 2010). The normalizing constant at times would be a complex integral and Laplace transform a method of approximation was developed to solve it explicitly. Unfortunately, after his death in 1827, his method of approximation was forgotten almost overnight & sophisticated statisticians preferred to judge the probability of an event happening by how frequently it occurred. They would then be called “frequentists” and they became the chief opponents of Bayes’ rule that will never die. For them modern science requires both objectively & precision and Bayes’ of course, start with the measure of your belief a situation makes approximation and frequentists call this “ignorance coined into science” and sometimes say “they use Bayes rule with sigh as the only thing available under the circumstance”.

Despite Bayes’ usefulness, Ronald Fisher (1890/1962) started attacking Bayesians in 1920-1930s (Sharon Bertsch, 2010) and theoreticians were then sigh applying Bayes’ rule having been opposed by a man well known to have single-handed created the foundations of modern statistical science including maximum likelihood methods, sampling theory, non-parametric statistics, randomization methods, fisher’s information, analysis of variance, Fisher-Kolmogorov equation, Fisher’s geometric model coining the word “null hypothesis” ,Fisher’s exact test, F-distribution, he created Biometry as a potential way to reconcile the discontinuous nature of Mendelian inheritance with continuous variation and gradual evolution (Sharon Bertsch, 2010).

As we will see later Fisher is an example of how personalizing a concept can be destructive to the growth and development of field especially a small field or an emerging area. Fisher publicly oppose Bayes’ rule fortunately, without his knowledge, his classical discoveries made significant contributions in the development of Bayesian paradigm especially maximum likelihood method which is the likelihood function in Bayesian.

He kept on with the bad fight which was publicly demonstrated around 1950 when Richard Doll and A.B Hill came to a conclusion that smoking caused lung cancer in which Bayesian method was used (Sharon Bertsch, 2010). Instead, he compared correlation in their papers to a correlation in the import of apples and the rise of divorce in order to show that correlation indeed does not imply causation. He went further to smoking for the first time at 60 and in the public to contradict the findings (Sharon Bertsch, 2010). Fortunately, this result was supported by a study in 1938 at the John Hopkins University in which scientists suggested a strong negative correlation between smoking and lifespan. Moreover, five studies were published in 1950 in which smoking was powerfully implicated in the causation of lung cancer. Furthermore, four years later, in 1954 the British Doctors study, a study of some 40 thousand doctors over 20 years, confirmed the suggestion, based on which the government issued advice that smoking and lung cancer rates were related. The British doctor's study lasted till 2001, with results published every 10 years and final results published in 2004 by Doll and Richard Peto. So it would then be clear later after 1950 that Fisher's problem was entirely Bayesian method that was used in the paper but rather not the findings.

Despite this struggle, 30-40 years after the World War II, a group of 100 Bayes' believers kept on with the bad fight. For many years, the Bayesians took a lot of time to develop a lot of theories to show how Bayesian statistics work. During this period both Bayesians and frequentists' were busy trying to outdo each other in the public. A cold war that was too heavy to bear by Bayesians who were just a minority as far as science is concerned.

By 1984, there was a host of techniques floating around, Bayes, Laplace transform, random sampling, Monte Carlo, Markov chain, iteration etc. & two men realized how they could work together Valan Gelfand and Adian Smith a student of Prof. Lindley at the University of Sidney (Sharon Bertsch, 2010).

They wrote a paper in which they used the "b" word only 5 times in a 5 page paper so as to persuade statisticians (the then Fisher's believers) who were the chief opponents of Bayes' rule to read the paper. In which MCMC method was re (discovered).

Physicists were familiar with MCMC methodology from the 1950s Nick Metropolis and his associates had developed one of the first electronic supercomputers (for those days) and had been testing their theories in physics using Monte Carlo techniques.

Implementation of the MCMC methods in combination with the rapid evolution of personal computers made the computational tool popular within a few years. Bayesian statistics suddenly re (became) fashionable. Using MCMC we could then set up and estimate complicated models that describe and solve problem that could not be solved with traditional methods.

Fortunately, the paper came out at same time powerful desktops workstation become available. A couple of years later around 1990's there is off the shelf a software called "Bugs" become available for solving Bayesian problems through Lindley's academic

Grandson David Peterson. It could fit complicated models in a relatively easier manner using standard MCMC methods.

This revolution brings together computer Scientists, Mathematicians, Physicists, and Statisticians and got adopted almost overnight. Researchers could then adopt the most suitable method of analysis for their work. Fortunately, even the famous frequentists (Fisher's believers) in the town could be heard saying. "I'm always a Bayesian".

Since 1998, WinBugs the window version of BUGS, earned great popularity among researchers of diverse scientific fields (Ioannidis, 2005) .

Given that the one line Bayes theorem that will never die is undisputed by everyone including the Fisher's believers, the main challenge then focuses on computation since a Standard Computation can run MCMC simulations for hours or days before convergence to a posterior distribution or sometimes convergence may not be guaranteed after hours or days of running depending on the complexity of the model.

However, recently (Rue, Martino, & Chopin, 2009) introduced the use of Gaussian random field which has become increasingly popular among scientists. The integrated Nested Laplace Approximation (INLA) is an approach proposed by Rue et al (2009) to perform approximates fully Bayesian inference on the class of latent Gaussian models (LGM). INLA makes use of determinist nested Laplace approximations and, as an algorithm tailored to the class of (LGMS), it provides a faster and more accurate alternative to simulation based MCMC schemes. This is demonstrated in a series of examples ranging from simple to complex models in Rue et al, (2009). Although the theory behind INLA has been well established in Rue et al. (2009), the INLA method continues to be a research and in active research and development.

3.3 Introduction to Bayesian Analysis

As opposed to the point estimators (means, variances) used by **classical statistics, Bayesian statistics** is concerned with generating the posterior distribution of the unknown parameters given both the data and some prior density for these parameters. As such, Bayesian statistics provides a much more complete picture of uncertainty in the estimation of the unknown parameters, especially after the confounding effects of nuisance parameters are removed.

Our treatment here is intentionally quite brief and much information can be obtained in Lee(1997) and Draper(2000) for a complete introduction to Bayesian analysis, and the introductory chapters of Tanner (1996) for a more condensed treatment. While very deep differences in philosophy separate hard-core Bayesians from hard-core frequentists (Efron 1986, Glymour 1981), our treatment here of Bayesian methods is motivated simply by their use as a powerful statistical tool.

3.4 probability and statistics

Statistics is the study of uncertainly, how to measure it and how to make choices in the face of it.

Since uncertainty is an inescapable part of the human condition, statistics has the potential to be helpful in almost every aspect of daily life, including science (the acquisition of knowledge for its own sake) and decision-making (how to use that knowledge to make a choice among the available actions).

When you notice you're uncertain about something-for example, the truth status of a true-false proposition such as “This patient is HIV-positive” or “Obama will win a second term as U.S president in 2012”- it's natural to want.

a) To quantify how much uncertainty you have and
b) To figure out how to reduce your uncertainty if the answer to (a) is higher than the level necessary to achieve your goals.

Probability- Is the part of mathematics devoted to quantifying uncertainty, so it plays a fundamental role in statistics and so does data-gathering, because the best way to reduced your uncertainty is to get some relevant new information (data).

3.5 Description, Influence and prediction

Inference: about the nature of the underlying process generating the data.

This is the statistical version of what the 18th century philosopher Hume referred to as the problem of induction; it included as special cases

a) Answering question about causality and
b) Generalizing outward from sample of data values to a population (a broader universe of discourse).

Prediction of future data on the basis of past data, including quantifying how much uncertainty you have about your predictions.

This is important in science, because good (bad) scientific theories make good (bad) predictions and it is important to do the best we can do to use Good models (Theories).

Decision–Making; frequentist and Bayesian Probability

Decision-Making: Predicting the future under all possible actions open to you, and choosing your favorite future on that basis.

The systematic study of probability can be traced back to an exchange of letters between Pascal and Fermat in the 1650's, but the version of probability they developed turns out to be to two simplistic to help in 21st century problems of realistic complexity.

Instead, two other ways to give meaning to the concept of probability are in current use today:

The frequentist (relative frequency) approach in which you restrict attention to phenomenon that are inherently repeatable under " identical" conditions and define P (A) to be the limiting relative frequency with which A would occur in a repetitions, as $n \to \infty$ (this approach was developed around 1870 by Venn Boole and others and was refined in the 1930s by Von Mises): and the Bayesian approach, in which the argument B of the probability operation P (B/A) is a true-false proposition whose truth status is unknown to you and P (B/A) represents the weights of evidence in favor of the truth of B, given the information in A (this approach was first developed by Bayes and Laplace in the 18th Century and was refined by Keynes, de Finetti, Hamsay. Jeffrey. Turung, Good Savage, Jaynes and others in the 20th century).

Internal and External Information

The Bayesian approach includes the frequentist Paradigm as a special case, so you might think it would be the only version of probability used in statistical work today, but in quantifying your uncertainty about something unknown to you, the Bayesian paradigm requires you to bring all relevant information to bear on the calculation; this involves combining information both internal and external to the dataset you're gathered and (somewhat strangely) the external-information part of this approach was controversial in the 20th century, and Bayesian calculation require approximating high –dimensional integrals.

The Laplace approximation method commonly referred to as Laplace transform was forgotten almost immediately after the death of Laplace in the year. (Whereas the frequent approach mainly relies on maximization rather than integration), and this was a severe limitation to the Bayesian paradigm for a long time (from the 1750s to the 1980s).

Metropolis Algorithm; Bayesian + frequentist

Bayesian statisticians belatedly discovered that applied mathematicians (led by metropolis and Ulam, working at the intersection between chemistry and physics in the 1940s, had used Markov chains to develop a clever algorithm, for approximating integrals arising in thermodynamics that are similar to the kinds of integrals that come up in Bayesian statistics and Desk-top computers finally became fast enough to implement the metropolis algorithm in a feasible start amount of time.

The 20th century was definitely a frequentist century, in large part because maximization was an excellent technology for that moment in history from the 1920s (when the statisticians and geneticist emphasized it) through the 1980s; but a consensus is now emerging around the idea that in the 21st century it's important for statisticians to be fluent in both frequentist and Bayesian ways of thinking.

In the 20th century many people acted as if you had to choose one of these paradigms and defend it against attacks from people who favored the other one, but it turns out that both approaches have strengths, and weaknesses so that can't be the right way to frame the issue; it seems to us instead that our job as a statisticians in this century is to develop a fusion of the two approaches that emphasizes the strength and de-emphasizes the weaknesses.

Our passion fusion involves

Reasoning in a Bayesian way when formulating my inferences, predictions and decisions because Bayesian paradigm is the most flexible approach so far developed for incorporating all relevant sources of uncertainty.

Reasoning in a frequentist way when paying attention to how often I get the right answer, which is an inherently frequentist activity that’s central to good science and decision-making.

According to the useful history of mathematics website www-history.mcs.at-and.ac.uk, mathematics began in Babylonia in approximately 2000 BCE, with the development of a systematic way to record and manipulates numbers (both integers and fractions)).

Gambling, which you would think might prompt the creation of mathematics based on what we now call randomness, is even older; che-like objects made from animal bones have been traced back to at least 4500 BCE.

Thus we have been thinking mathematically as a species for about 4000 years and gambling for far longer than that, and yet no one seem to have laid down the foundations of probability until around 350 years ago.

Some specialized problems in games of chance had been solved by Italians mathematicians going back to the 1400s, and Galilei (1564-1642) worked in a fragmentary way on probability concepts in the early 17th century, but the subject was not properly launched as a branch of mathematician until an exchange of letters between the French mathematicians Blaise Pascal (1623-1662) and Pierre de Fermat (1601-1665) in 1654.

Conditional probability

(Pascal and Fermat, 1654) invented what we now call the classical approach to probability: It enumerates the elementary outcomes ($observations$) (the fundamental possibilities in the process under study). In a way that makes them equi-possible (i.e., so that none would be favored over any other in hypothetical repetitions of the process) and compute the classical probability P (A) of an outcome A as the ratio $n_A = number\ of\ observations$ favoured to A to n=total number of $observations$.

This works for assigning probabilities to outcomes of idealized games of chance (dice coins, roulette, cards) but fails in complicated problems like those people think about routinely today (e.g., what are the $observations$in a regression setting with 100,000 observation and 1000 predictor variables?).

The Dutch scientist Christian Huygens (1629-1695) published the first book on probability in 1657.

Another important early probability book was written by the Swiss Mathematician Jacob Bernoulli (1654-1705) and published in 1713, after his death; in it Bernoulli stated and proved the first (weak) law of large number (p of a sequence of random variables y_n to a non-random limit $\mu = E\ (y)\}$.

The Pascal-Fermat classical approach had no notion of conditional probability; it was remembered by Thomas Bayes (1702-1761), who gave the first definition of

$$p\ (B|A) = \frac{p\ (B\ and\ A)}{p\ (A)}\ , from\ which \qquad (3c)$$

$p\ (B\ and\ A) = p(A)P\ (B|A)$. For (true-false) propositions A and B, in a posthumous publication in 1764. Bayes was interested in causal relationships: you see an effect in the world. (e.g people dying of a disease) and you wonder what was its cause (e.g., drinking the water? eating something? breathing the air? etc).

He had the bravery/imagination to consider this probabilistically, and he noticed that P (effect caused was a lot easier to think about than $p\ (Cause|effect)$. So he wondered how $p\ (B|A)$ depended on $p(A|B)$ (he wanted to reverse the order of conditioning).

Bayes' Theorem for propositions

To find out he wrote down his definition in the other order.

$$p\ (A|B) = \frac{p(A\ and\ B)}{p\ (B)} from\ which \qquad (3d)$$

$p(A\ and\ B) = p(B)p$ (A|B). $So\ now\ he\ has\ P\ (B\ and\ A) = P\ (A)P\ (B|A) and$

$$p(A\ and\ B) = p\ (B)P\ (A|B) \qquad (3e)$$

$and\ he\ can\ equate\ the\ two\ equations, since$

$p(B\ and\ A)\ = p\ (A\ and\ B)$, and solve for what he wants to get Bayes Theorem for propositions:

$$p\ (B|A) = \frac{p\ (B)P\ (A|B)}{p\ (A)} \qquad (3f)$$

The main application he had in mind was more ambitious: B stood for an unknown rate at which something happens (today we might use the symbol (0< θ <1) and stood for some data relevant to on (in today's notation his data set was. y =(y_1,... , y$_n$), $where\ each\ y\ was\ a$ 1 or 0 variable with success rate θ).

Bayes' Theorem for real numbers

In words he thought of his result as having the following meaning:

$$p\ (unknown|data) = \frac{p\ (Unknown)\ p(data|unknown)}{p\ (data)} \qquad (3g)$$

He conjectured (correctly) that his Theorem still applies when B is a real number ($\emptyset$) and A is a vector of real numbers (y): in contemporary notation.

$$p\ (\theta|y) = \frac{P\ (\theta)P(y|\theta)}{P(y)} \qquad (3h)$$

Where

a) $p\ (\theta|y) and\ p(y|\theta) are$ conditional probability densities for θ given y and y given θ and (respectively) and

b) $p\ (\theta)$ are (unconditional) probability densities for θ and y (respectively).

This requires some interpreting: I want to use after the dataset y had arrived, to quantify my uncertainty about θ in light of the new information, so I want to condition on the data, i.e to treat the entire equation as a function of θ for fixed y: this has two implications:

3.6 Diachronic interpretation of Bayes' theorem

$$p(Hypothesis|Data, Assumptions, Information) = \frac{p(Hypothesis)p(Data|Hypothesis)}{p(Data)} \qquad (3i)$$

Likelihood function

a) $p(y)$ is just a constant-in fact I can think of it as the normalizing constant, put into the equation to make the product $p(\theta)\, p(y|\theta)$ integrate to 1 (as all densities, e.g. , $p(\theta|y)$, must); and.

b) $p(y|\theta)$ may look like the sampling distribution for y given $\theta,$ but, have to think of it as a function of θ for fixed y.

Much later, fisher (1922) popularized this same idea and called it the **likelihood function.**

$p\ (\theta|y) = k\ \mathrm{I}(y|\theta)$ where I=information

$(3j)$

Where k is an arbitrary positive constant commonly referred to as the normalizing constant but Bayes (1764) saw its importance first we this new notation and terminology Bayes' Theorem becomes

$$p\ = k\ p\ (\theta|y) I(\theta|y). \qquad (3k)$$

$p\ (\theta|y)$ represent s the information about the unknown θ internal to the dataset y, but this is only one ingredient in the process of drawing together all of the evidence about θ.

Synthesis of knowledge

As Bayes (1764) understood, there will typically also be information about θ external to y, and p (θ) is where this other information comes into the Synthesis of knowledge.

On the log scale, and ignoring irrelevant constants, Bayes' Theorem says.

$$I\ (\theta|y) = I\ \ (\theta) +\ I\ (\theta|y) \qquad (3l)$$

Which, in words, could be interpreted as?

$$\begin{pmatrix} total\ information \\ about \\ \theta \end{pmatrix} = \begin{pmatrix} Information \\ external \\ to\ y \end{pmatrix} + \begin{pmatrix} Information \\ Internal \\ to\ y \end{pmatrix} \qquad (3m)$$

One way (but not the only way) you could think about the information about ∅ external to Y is to recall the sequential nature of learning; the temporal nature of learning; the temporal events of observing the data set y divided the time line into the period before y (a priori) and the period after y (a posteriori).

Centuries after Bayes, researchers in the 1950s used this to suggest a different way to express∅.

$$I(\theta|y) = I\ (\theta) + I\ (\theta|y)$$

$$\begin{pmatrix} Posterior \\ Information \\ about \\ \theta \end{pmatrix} = \begin{pmatrix} Prioric \\ information \\ about \\ \theta \end{pmatrix} + \begin{pmatrix} Likelihood \\ information \\ about\ \theta \end{pmatrix} \qquad (3n)$$

With this in mind people called P $(\theta|y)$ the posterior distribution and P (θ) the prior distribution for θ respectively.

These are actually not very good names, because (as noted above) P $(\theta|y)$ is meant to quantify all information about θ external to y (whether that information arrives before or after y is irrelevant but through widespread usage we're stuck with them now.

With the notation and terminology Bayes' Theorem says.

$$P(\theta|y)\ = k\ (\theta)\ \ I\ (\theta)\ \ I\ (\theta|y)$$

$$Posterior\ = k\ (Prior)\ (Likelihood)$$

Bayes' Theorem

The foundation of Bayesian statistics is **Bayes' theorem.** Suppose we observe a random variable y and wish to make infrences about another random variable θ, where θ is drawn from some distribution $p(\theta)$. From the above definition,

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} \tag{3o}$$

Again from the definition of conditional probability, we can express the joint probability by conditional probability by conditioning on θ to give

$$p(y,\theta) \ = p(y|\theta)\ p(\theta)$$

Putting these together gives Bayes' Theorem:

$$p(\theta|y) = \frac{p(y|\theta)\ p(\theta)}{p(y)}$$

With n possible outcomes $(\theta_1, \theta_2, \ldots, \ \theta_n)$,

$$p(\theta_j|y) = \frac{p(y|\theta_j)p(\theta_j)}{p(y)} = \frac{p(y|\theta_j)}{\sum_{i-1}^{n} p(b_i)p(y|\theta_i)} \tag{3p}$$

$p(\theta)$ is the prior distribution of the possible θ values, while $p(\theta|y)$ is the posterior distribution of θ give the observed data y. The origin of Bayes' theorem has a fascinating history (Stigler 1983).It is named after the Rev. Thomas Bayes, a priest who never published a mathematical paper in his lifetime. The paper in which the theorem appears was posthumously read before the royal society by his friend Richard Price in 1764. Stigler suggests it was first discovered by Nicholas Saunderson, a blind mathematician/optician who, at age 29, became Lucasian professor of Mathematics at Cambridge (the position held earlier by Isaac Newton).

Bayesian Perspective on Probability

Probability is a measure of uncertainty.

- Probabilities can be assigned to future events or to unobservable quantities, as well as sampling scenarios.
- Subjective probability concerns the judgment of an individual about uncertain events or propositions
- Such probability cannot be thought of as frequency (long-run) probabilities.
- Subjective probabilities should obey the axioms of probability and a single person's probabilities should not be inconsistent.
- Arguments based on decision theory and theory of betting provides support for using probability as the right measure of uncertainty.
- Rational probability assignments are equivalent to betting in a coherent way; in a way that your opponent cannot be guaranteed to win based on inconsistency in your probability assignments.
- Classical statistical procedures can be incoherent (informally: non-sensical; formally: can result in bets that your opponent is guaranteed to win)
- Bayesian inference emerges naturally from subjective probability. i.e., if you are comfortable with using probability to measure uncertainty, then the Bayesian machinery is the principled approach to inference.

Bayesian perspective on subjectivity

- Both the likelihood and the prior are subjective and both must be specified as part of the model.
- Bayesian statistics formalizes the scientific process in which beliefs are updated based on the data. No single experiment will determine one's beliefs (unless overwhelmingly compelling or there is no other information); prior distribution captures this.

Advantage of using Bayes' method

- Bayesian logic and interpretation are simple; scientific questions can often be easily framed as inferential questions.
- Bayesian inference is simple in principle and provides a single recipe for coherent inference, all based on the posterior. Inference is conditional on the observed, and not on data that were possible but not observed, obeying the likelihood principle. In addition, it tells how to update prior beliefs and how to add additional information.
- Utility of using prior information, allowing one to combine various sources of information, including constraints.
- Inference for small samples is exact (but sensitive to the prior)
- Interpretation: Uncertainty is naturally framed as probability statements based on the posterior in a way that non-statisticians easily relate to (what else could 'statistical inference about θ' mean?)
- Bayesian inference naturally deals with conditioning, marginalization, and nuisance parameters
- Parameter uncertainty is naturally accounted for.
- Bayesian inference naturally meshes with decision theory
- Modern computational techniques allow models to be fit in other ways.
- Bayesian results often have good frequentist properties and frequentist inference is sometimes a special case of Bayesian results under a particular prior.
- Complicated hierarchical models can be naturally constructed in a Bayesian framework.
- Bayesian inference naturally penalizes complex models
- Bayesian inference can deal with multiple testing inherently if set up properly as a joint inference problem.

Disadvantage of using Bayes' method

- Computing the posterior, while simple in theory, is often difficult and time consuming in practice.
- Frequentist inference is often simpler in simple or standard statistical settings, in part because of standardized software.
- When it becomes more difficult to fit models, this may discourage full model exploration, assessment and comparison, which is key to good applied statistical work.
- Bayesian inference is model-based and classical methods may not generalize (partial likelihood, non-parametric testing, robust estimation, marginal models).
- Sensitivity to prior distribution and difficulty of specifying sensible priors in some cases, particularly for complicated models.

Comparison of Bayesian and Non-Bayesian Inference

Topics	Non-Bayesian	Bayesian
Probability	Limit of empirical frequencies; θ is fixed	Subjective believe; θ is random
Estimation	Likelihood based (MLE) and other criteria (e.g. UMVUE)	Baesd on posterior; often $E(\theta\|y)$
Sources of information	Data only	Data and prior beliefs
Phylosophy	Incoherent; criteria can lead to nonsensical procedures	Coherent approach
Biase	Unbiasedness is often a criterion	Biased; shrinkage and biase-variance tradeoff
Interval estimation	Interpreted in terms of long-run behavior of y	Interpreted as probability statements about θ

Likelihood principle	May violate	Obeys (except Jeffreys' prior)
Invariance	MLE is invariant to transformation	Posterior mean is not ; but posterior for any derived quantity is easily obtained
Computation	Optimization	Integration
Uncertainty	Often based on asymptotic	exact
Nuisance parameters	A nuisance	Integrated over

Current state of affairs

Bayesian methods are widespread in statistics and some applied areas often used for practical, computational reasons, rather that philosophical reasons. However, some consensus on using Bayesian inferential techniques and evaluating them in a frequentist way.

Proper posteriors

If prior is 'proper', then posterior

- Is proper
- Converse is not necessarily true.

Recall, **a distribution is proper** if:

i. $\int_{\theta}^{\omega} p(\theta) d\theta = 1$ if continuous

ii. $\sum_{\theta} p(\theta) = 1$ if discrete

Improper priors are often used but require care (and have some interpretational issues). If the posterior is not proper, it makes no sense to use it (hence to summarize it).

Summaries of the posterior distribution

i. Point estimate: $\hat{\theta} = E(\theta|y) = \int_{\theta}^{\omega} p(\theta|y)\,d\theta$

ii. Posterior median: Choose $\hat{\theta}$ such that $\int_{-\infty}^{\hat{\theta}} p(\theta|y) = \frac{1}{2}$

iii. Posterior mode: Choose $\hat{\theta}$ such that $p(\hat{\theta}|y) \geq p(\theta|y) \,\forall\, \theta$

When $p(\theta|y)$ is symmetric and unimodal, the 3 coincide.

Note

However, using any of the above estimators or even all the three simultaneously (mean mode and median), loses **the full power of a Bayesian analysis, as the full estimator is the entire posterior density itself**. If we cannot obtain the full form of the posterior distribution, it may still be possible to obtain one of the three estimators. However, as we will see later, we can generally obtain the posterior by numerical methods using Integrated Nested Laplace Approximations, and hence the Bayes estimate of a parameter is frequently presented as a frequency histogram from INLA samples of the posterior distribution.

Summary of Interval estimation (Credibility Interval)

The posterior credible intervals $100(1-\alpha)\%$. “equal tailed” posterior credible intervals (central intervals). The range of values above and below which lies exactly $100\left({}^{\alpha}/_{2}\right)\%$ of the posterior probability.

“Highest Posterior Density” (HPD) region Def: A region R is an HPD region of constent $(1-\alpha)$ for θ if: $p(\theta_0|y) \geq p(\theta_1|y) \;\; \forall\, \theta_0 \in R \; and \; \theta_1 \notin R$With $\int_{R}^{w} p(\theta|y)\,d\theta = 1-\alpha$

Therefore HPD is the region of the values containing $100(1-\alpha)\%$ of the posterior probability and the density within the region is never lower than outside. Central posterior interval ≡ HPD interval if the posterior is symmetric and unimodal. If ${\sigma^2}_0$ is large, then the interval is:

$\bar{y} + \sqrt{\frac{\sigma^2}{n}} \; Z{\alpha}/_{2}$ (Classical sampling theory). HPD method will lead to the posterior region of the shortest length for a given α.

We can also plot marginal posterior distributions in one and two dimensions, avoiding the problem of having to summarize the posterior using posterior moments.

If we have a closed form for the posterior, even without the normalizing constant, we can plot the density.

More often we have a sample of values from the posterior. We can plot a histogram or (often better) a smoothed density estimate.

Highest Density Regions (HDRs)

Given the posterior distribution, construction of confidence intervals is obvious. For example, a 100(1−α) confidence is given by any $(L\alpha/_{2}, H\alpha/_{2})$ satisfying

$$\int_{L\alpha/_{2},}^{H\alpha/_{2}} p(\theta|x)d\theta = 1 - \alpha$$

To reduce possible candidates, one typically uses **highest density regions,** or HDRs, where for a single parameter the HDR 100(1−α) **region** (s) are the shortest intervals giving an area of (1−α). More generally, if multiple parameters are being estimated, the HDR region (s) is those with the shortest volume in the parameter space. HDRs are also referred to as Bayesian **Confidence Intervals or Credibility Intervals.**

It is critical to note that there is a profound difference between a confidence interval (CI) from classical (frequentist) statistics and a Bayesian credibility interval. The interpretation of a classical confidence interval is that we repeat the experiment large number of times, and construct CIs in the same fashion, that (1−α) of the time the confidence interval with enclose the (unknown) parameter. With a Bayesian HDR, there is a (1−α) probability that the interval contains the true value of the unknown parameter. Often, the CI and Bayesian intervals have essentially the same value, but again the interpretational difference remains. The key point is that the Bayesian prior allows us to make direct probability statements about θ , while under classical statistics we can only make statements about the behavior of the statistic if we repeat an experiment a large number of times. Given the important conceptual difference between classical and Bayesian intervals, Bayesians often avoid using the term confidence interval.

3.7 Why prior in Bayesian Statistics

Some people may have the mistaken impression that the advantages of Bayesian methods are negated by the need to specify a prior distribution. In fact the use of prior is both appropriate for rational inference and advantages in practical applications.

It is inappropriate not to use a prior consider the well-known example of random disease screening. A person is selected at random to be tasted for a rare disease. The test result is positive, what is the probability that the person actually has the disease? it turn out, even if the test is highly accurate, the posterior probability of actually having the disease is surprisingly small. Why? Because the prior probability of the disease was so small. Thus, incorporating the prior is crucial for coming to the right conclusion.

Priors are explicitly specified and must be agreeable to a skeptical scientific audience. Priors are not capricious and cannot be covertly manipulated to predetermine a conclusion. If skeptics disagree with the specification of the prior, then the robustness of the conclusion can be explicitly examined by considering other reasonable priors. In most applications, with moderately large data sets and reasonably informed priors, the conclusions are quite robust to accommodate all the divergent views concerning the estimate.

We emphasize on priors because it is the power of Bayesian statistics. Priors are useful for cumulative scientific knowledge and for leveraging inference from small-sample research.

As an empirical domain matures, more and more data accumulate regarding particular procedures and outcomes. The accumulated result can inform the priors of subsequent research, yielding greater precision and firmer conclusions.

When different groups of scientists have different priors stemming from different theories and empirical emphases then Bayesian methods provide rational means for comparing the conclusions from different priors.

To summarize, priors are not problematic nuisance to be avoided. Instead, priors should be embraced as appropriate in rational inference and advantageous in real research.

If those advantages of Bayesian methods are not enough to attract change, there is a major reason to be repelled from the dominant methods of the 20th century.

3.8 Prior and likelihood specification; parametric modeling

This creates a specification problem: how do you quantify “information about θ $internal\ to$ y" in the likelihood distribution $I\ (\theta/y)\ and$ “information about θ external to y” in the prior distribution $P\ (\theta)$

I’ll give an example later of prior specification; what about specifying.

$$I\ (\theta/y) = P\ (y/\theta)$$

From a Bayesian perspective P (Y/∅) is the predictive distribution for how the data will come out before any data have arrived; how do you specify this?

Typical solution from 1764 through 1937; try to find a standard parametric family of probability distributions (indexed by r= (θ, n) that captures what you expect to seek in the data (base on previous experience with similar problems); for example with binary outcomes, you would firs try the Bernoulli (4) distribution, with count data you would first think of the Poisson (4) distribution, and with continuous outcomes you might well start with the Normal distributions.

This- parametric statistic modeling- was the standard approach for centuries, but there is a problem with it. Use the data to choose the model and then use the same data to draw conclusion on the basis of the same model.

The choice of a prior

Obviously, a critical feature of any Bayesian analysis is the choice of a prior. The key here is that when the data have sufficient signal, even a bad prior will still not greatly influence the posterior. In a sense, this is an asymptotic property of Bayesian analysis in that all but pathological priors will be overcome by sufficient amount of data. As mentioned above, one can check the impact of the prior by seeing how stable to posterior distribution is to different choices of priors. If the posterior is highly dependent on the prior, then the data (the likelihood function) may not contain sufficient information. However, if the posterior is relatively stable over a choice of priors, then the data indeed contain significant information.

The **location** of a parameter (mean or mode) and its **precision** (the reciprocal of the variance) of the prior is usually more critical than its actual shape in terms of conveying prior information. The shape (family) of the prior distribution is often chosen to facilitate calculation of prior, especially through the use of conjugate priors that, for a given likelihood function, return a posterior in the same distribution family as the prior (i.e., a gamma prior returning a gamma posterior when the likelihood is Poisson). We will return to conjugate priors towards the end of this section, but we first discuss other standard approaches for construction of priors.

Diffuse priors

Bayes rule only provides mathematically correct re-allocation of credibility across the candidate parameter values (Kruschke, 2012). The result reveals how strongly we should believe in each candidate parameter value given the data.

One of the most common priors is the flat, uninformative, or diffuse priors where the prior is simply a constant.,

$$p(\theta) = k = \begin{cases} \frac{1}{b-a}, & a \leq \theta \leq b \\ 0, & elsewhere \end{cases} \qquad (3s)$$

This conveys that we have no priori reason to favor any particular parameter value over another. With a flat prior, the posterior just a constant times the likelihood.,

$$p(\theta|x) = Cl(\theta|x)$$

And we typically write that $p(\theta|x) \propto \boldsymbol{\ell}(\theta|x)$. In many cases, classical expressions from frequentists statistics are obtained by Bayesian analysis by assuming a flat prior.

If the variable of interest ranges over $(0, \infty)$ or $(-\infty, +\infty)$, then strictly speaking a flat prior does not exist, as if the constant take on any non-zero value, the integral does not exists. In such cases a flat prior (assuming $p(\theta|x) \propto \boldsymbol{\ell}(\theta|x)$) is referred to as an **improper prior.**

Sufficient Statistics and Data-Transformed Likelihoods

Suppose we can write the likelihood for a given parameter θ and data vector x as

$$\ell(\theta|x) = g[\theta - t(x)] \qquad (3t)$$

Here the likelihood is a function $l = g(z)$, where $z = \theta - t(x)$. If the likelihood is of this form, the data x only influences θ by a translation on the scale of the function g, i.e., from $g(z)$ to $g(z + a)$. Further, note that $t(x)$ is the only value of the data that appears , and we call the function t a sufficient statistic. Other data sets with different values of x, but the same value of the sufficient statistic $t(x)$, have the same likelihood.

When the likelihood can be placed in the form

$\ell(\theta|x) = g[\theta - t(x)]$, a shift in the data gives rise to the same functional form of the likelihood function except for a shift in the location, from $(\theta + t[x_1])$ to $(\theta + t[x_2])$. Hence, this is a **natural** scale upon which to measure likelihoods, and on such a scale, a flat /diffuse prior seems natural.

Example

Consider n independent samples from a normal with unknown mean μ and known variance σ^2. Here

$$\ell(\mu|x) \propto exp\left(\frac{-(\mu - \bar{x})^2}{2\left({\sigma^2}/{n}\right)}\right)$$

Note immediately that $\bar{x}$ is a sufficient statistic for the mean, so that different datsets with the same mean (for n draws) have the same likelihood function for the unknown mean μ. Further, note that

$$g(z) = exp\left(\frac{-z^2}{2({\sigma^2}/{n})}\right)$$

Hence a flat prior for μ seems appropriate.

What is the natural scale for a likelihood function that does not satisfy

$\boldsymbol{\ell}(\theta|x) = g[\theta - t(x)]$? Suppose that the likelihood function can be written in data-translated format as

$$\boldsymbol{\ell}(\theta|x) = g[h(\theta) - t(x)]$$

When the likelihood function has this format, the natural scale for the unknown parameter is $h(\theta)$. Hence, a prior of the form $p[h(\theta)] = constant$ (a flat prior on $h[\theta]$) is suggested. Using a change of variables to transform $p[h(\theta)]$ back onto the θ scale suggests a prior on θ of the form

$$p(\theta) \propto \left|\frac{\delta h(\theta)}{d\theta}\right|$$

Example 4.

Suppose the likelihood function assumes data follow an exponential distribution,

$$\boldsymbol{\ell}(\theta|x) = \left({}^{x}/_{\theta}\right) exp\left(-{}^{x}/_{\theta}\right)$$

Noting that

$$\frac{x}{\theta} = exp\left[ln\left(\frac{x}{\theta}\right)\right] = exp[lnx - ln\theta]$$

We can express the likelihood as

$$l(\theta|x) = exp[(lnx - ln\theta) - \exp(lnx - ln\theta)]$$

Hence, in a data-translated format the likelihood function becomes

$g(y) = exp[y - \exp(y)]$, $t(x) = lnx,\ g(\theta) = \ln(\theta)$

The "natural scale" for θ in this likelihood function is thus $ln\theta$, and a natural prior is $p(ln\theta) = constant$, giving the prior as

$$p(\theta) \propto \left|\frac{\delta ln\theta}{\delta\theta}\right| = \frac{1}{\theta}$$

The Jeffrey's' Prior

Suppose we cannot easily find the natural scale on which the likelihood is in data-translated format, or that such a decomposition does not exist. Jeffreys (1961) proposed a general prior in such cases, based on the Fisher information I of the likelihood. Recall that

$$I(\theta|x) = -E_x\left(\frac{\delta^2 ln\, \ell(\theta|x)}{\delta\theta^2}\right) \qquad (3u)$$

Jeffreys' rule (giving the Jeffreys' prior) is to take as the prior

$$p(\theta) \propto \sqrt{I(\theta|x)}$$

More information on this can be found in Lee(1997, section 3.3).

Example.

Consider the likelihood for n independent draws from a binomial,

$$\boldsymbol{\ell(\theta|x) = C\theta^{x}(1-\theta)^{n-x}}$$

Where the constant $\boldsymbol{C}$ does not involve $\boldsymbol{\theta}$. Taking logs give,

$$L(\theta|x) = ln[\boldsymbol{\ell(\theta|x)}] = lnC + xln\theta + (n-x)\ln(1-\theta)$$

Thus

$$\frac{\delta L(\theta|x)}{\delta\theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

And likewise

$$\frac{\delta^2 L(\theta|x)}{\delta\theta^2} = -\frac{x}{\theta^2} - (-1)\times(-1)\frac{n-x}{(1-\theta)^2} = -\left(\frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2}\right)$$

Since $E[x] = n\theta$, we have

$$-E_x\left(\frac{\delta^2 \ln \boldsymbol{\ell(\theta|x)}}{\delta\theta^2}\right) = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = n\theta^{-1}(1-\theta)^{-1}$$

Hence the Jeffreys' prior becomes

$$p(\theta) \propto \sqrt{\theta^{-1}(1-\theta)^{-1}} \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

This is a Beta distribution (which we discuss later in this section)

When there are multiple parameters, I is the Fisher Information matrix, the matrix of the expected second partials,

$$I(\Theta|x)_{ij} = -E_x\left(\frac{\partial^2 \ln \ell(\Theta|x)}{\partial\theta_i \partial\theta_j}\right)$$

In this case, the Jeffrey's prior becomes

$$p(\Theta) \propto \sqrt{det[I(\theta|x)]}$$

Posterior Distributions under Normality Assumptions

To introduce the basic ideas of Bayesian analysis, consider a case when data is drawn from a normal distribution, so that the likelihood function for the ith observation, x_i is

$$\ell(\mu, \sigma^2|x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

the resulting full likelihood for all n data points is

$$\ell(\mu|x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2\sigma^2}\left(-\sum_{i=1}^{n} x_i{}^2 - 2\mu n\bar{x} + n\mu^2\right)\right]$$

Known Variance and Unknown mean

Assume the variance σ^2 is known, while the mean μ is unknown. For a Bayesian analysis, it remains to specify the prior for, $p(\mu)$. Suppose we assume a Gaussian prior, $\mu \sim N(\mu_0, {\sigma^2}_0)$, so that

$$p(\mu) = \frac{1}{\sqrt{2\pi{\sigma^2}_0}} exp\left(-\frac{(\mu-\mu_0)^2}{2{\sigma^2}_0}\right)$$

The mean and variance of the prior μ_0, $and\ {\sigma^2}_0$ are reffered to as hyperparameters.

One important trick we will use when calculating the posterior distribution is to ignore terms that are constants with respect to the unknown parameters. Suppose x denotes the data and Θ_1 is a vector of known model parameters, while Θ_2 is a vector of unknown parameters. If we can write the posterior as

$$p(\Theta_2|x,\Theta_1) = f(x,\Theta_1)\cdot g(x,\Theta_1,\Theta_2)$$

With the prior given by the above equation;

$$p(\mu) = \frac{1}{\sqrt{2\pi{\sigma^2}_0}} exp\left(-\frac{(\mu-\mu_0)^2}{2{\sigma^2}_0}\right)$$

Then we can express the resulting posterior distribution as

$$p(\mu|x) \propto \ell(\mu|x)\cdot p(\mu)$$

$$\propto exp\left(\frac{(\mu-\mu_0)^2}{2{\sigma^2}_0} - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} {x_i}^2 - 2\mu n\bar{x} + n\mu^2\right]\right)$$

We can factor out additional terms not involving μ to give

$$p(\mu|x) \propto exp\left(-\frac{\mu^2}{2{\sigma^2}_0} + \frac{\mu\mu_0}{{\sigma^2}_0} + \frac{\mu n\bar{x}}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right)$$

Factoring in terms of μ, the term in the exponential becomes

$$-\frac{\mu^2}{2}\left(\frac{1}{{\sigma^2}_0} + \frac{n}{\sigma^2}\right) + \mu\left(\frac{\mu_0}{{\sigma^2}_0} + \frac{n\bar{x}}{\sigma^2}\right) = -\frac{\mu^2}{{\sigma^2}_*} + \frac{2\mu\mu_*}{2{\sigma^2}_*}$$

Where

$\sigma^2{}_* = \left(\frac{1}{\sigma^2{}_0} + \frac{n}{\sigma^2}\right)^{-1}$ and $\mu_* = \sigma^2{}_* \left(\frac{\mu_0}{\sigma^2{}_0} + \frac{n\bar{x}}{\sigma^2}\right)$

Finally, by completing the square, we have

$$p(\mu|x) \propto \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma^2{}_0} + f(x, \mu_0, \sigma^2, \sigma^2{}_0)\right)$$

The posterior density function for μ thus become

$$p(\mu|x) \propto exp\left(-\frac{(\mu-\mu_*)^2}{2\sigma^2{}_*}\right)$$

Recalling that the density function for $z \sim N(\alpha, \beta)$ is

$$p(z) \propto exp\left(-\frac{(z-\alpha)^2}{2\beta}\right)$$

Shows that the posterior density function for μ is a normal distribution with mean μ_* and variance $\sigma^2{}_*$, e.g.,

$$\mu|(x, \sigma^2) \sim N(\mu_*,\ \sigma^2{}_*\ )$$

Notice that the posterior density is in the same form as the prior. This occurred because the prior conjugated with the likelihood function- the product of the prior and likelihood returned a distribution in the same family as the prior. The use of such **conjugate priors** (for a given likelihood) is a key concept in Bayesian analysis and we explore it more fully below.

We are now in a position to inquire about the relative importance of the prior verses the data. Under assumed prior, the mean (and mode) of the posterior distribution is given by

$$\mu_* = \mu_0 \frac{\sigma^2{}_*}{\sigma^2{}_0} + \bar{x} \frac{\sigma^2{}_*}{\sigma^2/_n}$$

Note with a very diffuse prior on μ (i.e., $\sigma^2{}_0 \gg \sigma^2$), that $\sigma^2{}_* \to \sigma^2/n$ and $\mu_* \to \bar{x}$. Also note that as we collect enough data, $\sigma^2{}_* \to \sigma^2/n$ and $\mu_* \to \bar{x}$.

Gamma, Inverse-Gamma, Chi-square ($\mathcal{X}^2$), and $\mathcal{X}^{-2}$), Distributions

Before we examine a Gaussian likelihood with unknown variance, a brief side is needed to develop $\mathcal{X}^{-2}$, the inverse chi-square distribution. We do this via the gamma and inverse gamma distributions.

The $\mathcal{X}^2$ is a special case of the gamma distribution , a two parameter distribution. A gamma-distributed variable is denoted by $x{\sim}Gamma(\alpha, \beta)$, with density function

$$p(x|\alpha,\beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \; x > 0, \;\; \alpha, \beta > 0 \\ 0, \qquad\qquad\qquad\qquad Elsewhere \end{cases} \qquad (3v)$$

As a function of x note that

$$p(x) \propto x^{\alpha-1} e^{-\beta x}$$

We can parameterize a gamma in terms of its mean and variance by noting that

$$\mu_x = \frac{\alpha}{\beta}, \; \sigma^2{}_x = \frac{\alpha}{\beta^2}$$

$\Gamma(\alpha)$, gamma function evaluated at α (which normalized the gamma distribution) is defined as

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$$

The gamma function is the generalization of the factorial function ($n!$) to all positive numbers (and as integration by parts will show) satisfies the following identities

$$\Gamma(\alpha) = (1-\alpha)\,\Gamma(\alpha-1), \; \Gamma(1) = 1, \; \Gamma_{\frac{1}{2}} = \sqrt{\pi}$$

The $\mathcal{X}^2$ distribution is a special case of the gamma distribution, with a $\mathcal{X}^2$ with n digrees of freedom being a gamma random variable with $\alpha = \frac{n}{2}$ and $\beta = \frac{1}{2}$, i.e.,

$\mathcal{X}^2{}_n \sim Gamma(\frac{n}{2}, \frac{1}{2})$,

$$p(x \mid n) = \frac{2^{-n/2}}{\Gamma(n/2)} x^{-(n/2-1)} e^{-1/(2x)}$$

With mean and variance given by

$$\mu_x = \frac{1}{n-2}, \qquad \sigma_x^2 = \frac{2}{(n-2)^2(n-4)}$$

The scaled inverse chi-square distribution is more typically used, where

$$p(x \mid n) \propto x^{-(n/2-1)} e^{-\sigma_0^2/(2x)}$$

So that the $\frac{1}{2x}$ term in the exponential is replaced by an $\frac{\sigma^2{}_0}{2x}$ term. If x follows this distribution, then $\sigma^2{}_0 \cdot x$ follows a standard $\boldsymbol{\mathcal{X}^2}$ **distribution. The scaled-inverse** $\boldsymbol{\mathcal{X}^2}$ **distribution thus involves two parameters,** $\sigma^2{}_0$ and n and it is denoted by $SI - \boldsymbol{\mathcal{X}^2}(\boldsymbol{n},$ $\sigma^2{}_0)$. Note that if

$$x \sim \chi^{-2}_{(n,\sigma_0^2)}$$

Then

$$\sigma_0^2\, x \sim \chi_n^{-2}$$

Unknown Variance: Inverse-$\mathcal{X}^2$ Priors

Now suppose the data are drawn from a normal with mean μ, but unknown variance σ^2. The resulting likelihood function becomes

$$\ell(\sigma^2 \,|\, \mathrm{x}, \mu) \propto (\sigma^2)^{-n/2} \cdot \exp\left(-\frac{nS^2}{2\sigma^2}\right)$$

Where

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Notice that we condition on x and μ (i.e., the values are known), the S^2 is a constant. Further observe that, as a function of the unknown variance σ^2, the likelihood is proportional to a scaled inverse-$\mathcal{X}^2$ distribution. Thus, taking the prior for the unknown variance also as a scaled invers-$\mathcal{X}^2$ with hyperparameters v_0 and ${\sigma^2}_0$, the posterior becomes

$$\begin{aligned} p(\sigma^2 \,|\, \mathrm{x}, \mu) &\propto (\sigma^2)^{-n/2} \exp\left(-\frac{nS^2}{2\sigma^2}\right)(\sigma^2)^{-\nu_0/2-1} \cdot \exp\left(-\frac{\sigma_0^2}{2\sigma^2}\right) \\ &= (\sigma^2)^{-(n+\nu_0)/2-1} \exp\left(-\frac{nS^2 + \sigma_0^2}{2\sigma^2}\right) \end{aligned}$$

Unknown Mean and Variance

Putting all the pieces together, the posterior density for draws from a normal with unknown mean and variance is obtained as follows. First, write the joint prior by conditioning on the variance,

$$p(\mu, \sigma^2) = p(\mu \,|\, \sigma^2) \cdot p(\sigma^2)$$

As above, assume a scaled inverse chi-square distribution for the variance and, conditioned on the variance, normal prior for the mean with hyper parameters μ_0 and ${\sigma^2}/{k_0}$. We write the variance for the conditional mean prior this way because σ^2 is known (as we condition on it) and we scale this by the hyper parameter k_0. Hence, we assume

$$\sigma^2 \sim \chi^{-2}(\nu_0, \sigma_0^2), \qquad (\mu \,|\, \sigma^2) \sim \mathrm{N}\left(\mu_0, \frac{\sigma^2}{\kappa_O}\right)$$

The resulting posterior marginal become

$$\sigma^2 \,|\, \mathrm{x} \sim \chi^{-2}(\nu_n, \sigma_n^2), \quad \text{and} \quad \mu \,|\, \mathrm{x} \sim \mathrm{t}_{\nu_n}\left(\mu_n, \frac{\sigma_n^2}{\kappa_n}\right)$$

Where $t_n(\mu_n, \sigma^2{}_n)$ denotes a t-distribution with v_n degrees of freedom, mean μ_n and variance$\sigma^2{}_n$. Here

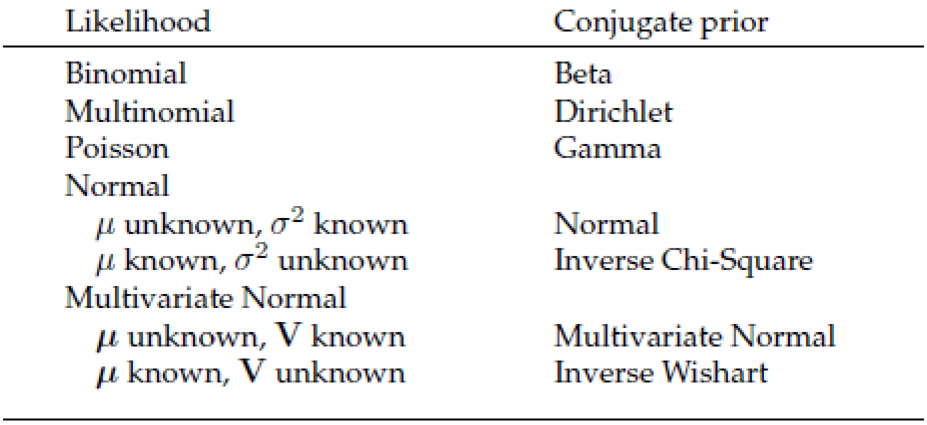
$$\nu_n = \nu_0 + n, \qquad \kappa_n = \kappa_0 + n$$

$$\mu_n = \mu_0 \frac{\kappa_0}{\kappa_n} + \overline{x} \frac{n}{\kappa_n} = \mu_0 \frac{\kappa_0}{\kappa_0 + n} + \overline{x} \frac{n}{\kappa_0 + n}$$

$$\sigma_n^2 = \frac{1}{\nu_n} \left(\nu_0 \sigma_0^2 + \sum_{i=1}^{n} (x_i - \overline{x})^2 + \frac{\kappa_0 n}{\kappa_n} (\overline{x} - \mu_0)^2 \right)$$

Conjugate priors

The use of a prior density that conjugates the likelihood allows for analytic expressions of the posterior density. The table below gives the conjugate priors for several common likelihood functions.



Likelihood	Conjugate prior
Binomial	Beta
Multinomial	Dirichlet
Poisson	Gamma
Normal	
μ unknown, σ^2 known	Normal
μ known, σ^2 unknown	Inverse Chi-Square
Multivariate Normal	
μ unknown, $\mathbf{V}$ known	Multivariate Normal
μ known, $\mathbf{V}$ unknown	Inverse Wishart

We first review some of the additional distributions introduced in the table above and conclude by discussion conjugate priors for members of the exponential family of distributions.

The Beta and Dirichlet Distributions

Where we have frequency data, such as for data drawn from a binomial or multinomial likelihood, the Dirichlet distribution an appropriate prior. Here,

$x \sim Dirichlet(\alpha_1, \alpha_2, \ldots \alpha_k)$, with

$$p(x_1, \cdots x_k) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} x_1^{\alpha_1 - 1} \cdots x_k^{\alpha_k - 1} \qquad (3w)$$

Where

$$\alpha_0 = \sum_{i=1}^{k} \alpha_i, \quad 0 \le x_i < 1, \quad \sum_{i=1}^{k} x_i = 1, \quad \alpha_i > 0$$

Where

$$\mu_{x_i} = \frac{\alpha_i}{\alpha_0}, \quad \sigma^2(x_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}, \quad \sigma^2(x_i, x_j) = -\frac{\alpha_i \, \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$$

An important case of the Dirichlet is the **Beta distribution**,

$$p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha - 1}(1 - x)^{\beta - 1} \quad \text{for} \quad 0 < x < 1, \quad \alpha, \beta > 0$$

Given that Uniform distribution with parameters (0,1) is a special case of Beta distribution with parameters (1,1), therefore, **uniform distribution is a special case of Dirichlet process**.

Wishart and Inverse Wishart Distribution

The Wishart distribution can be thought of as the multivariate extension of the

$\mathcal{X}^2$ distribution. In particular, if $x_1, x_2, \ldots x_n$ are independent and identically distributed with $x_i \sim MVN_k(0, \mathrm{V})$ $-$ that is , each is drawn from a $k - dimensional$ multivariate normal with mean vector zero and variance-covariance matrix V, then the random $(k \times k\ symetric, positive\ definate)$ matrix

$$\mathbf{W} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \sim \mathrm{W}_n(\mathbf{V}) \qquad (3x)$$

Hence the sum follows a Wishart with n degrees of freedom and parameter V. For the special case of $k = 1$ with $V = (1)$, this is just a $\mathcal{X}^2(n)$ distribution. The Wishart distribution is the sampling distribution for covariance matrices (Just like the $\mathcal{X}^2$ is associated with the distribution of a sample variance). The probability density function for a Wishart is given by

$$p(\mathbf{W}) = 2^{-nk/2}\pi^{-k(k-1)/k} \mid \mathbf{V} \mid^{-n/2} \mid \mathbf{W} \mid^{(n+k+1)/2} \frac{\exp\left(-\frac{1}{2}\mathrm{tr}\left[\mathbf{V}^{-1}\mathbf{W}\right]\right)}{\prod_{i=1}^{k} \Gamma\left(\frac{n+1-i}{2}\right)}$$

If $Z \sim W_n(V)$, then $Z^{-1} \sim W_n{}^{-1}(V^{-1})$, where W^{-1} denotes the Inverse-Wishart distribution. The density function for an Inverse-Wishart distributed random matrix W is

$$p(\mathbf{W}) = 2^{-nk/2}\pi^{-k(k-1)/k} \mid \mathbf{V} \mid^{n/2} \mid \mathbf{W} \mid^{-(n+k+1)/2} \frac{\exp\left(-\frac{1}{2}\mathrm{tr}\left[\mathbf{V}\mathbf{W}^{-1}\right]\right)}{\prod_{i=1}^{k} \Gamma\left(\frac{n+1-i}{2}\right)}$$

Thus, the Inverse-Wishart distribution is the distribution of the Inverse of the sample covariance matrix.

Conjugate priors for the Exponential Family of Distributions

Many common distributions (normal, gamma, Poisson, binomial, etc.) are members of the exponential family, whose general form is given by the equation below. Note that this should not be confused with the simple exponential distribution, which is just one particular member from this family. When the likelihood is in the form of an exponential family of an exponential family, a conjugate prior (also a member of exponential family of distributions) can be found.

Suppose the likelihood for a single observation (out of n) is in the form of an exponential family,

$$\ell(y_i \mid \theta) = g(\theta) h(y) \exp\left(\sum_{j=1}^{m} \phi_j(\theta)\, t_j(y_i) \right) \qquad (3y)$$

Using the prior

$$p(\theta) \propto [\,g(\theta)\,]^{b} \exp\left(\sum_{j=1}^{m} \phi_j(\theta)\, a_j\right)$$

Yields the posterior density

$$\begin{aligned} p(\theta\,|\,y) &\propto \left[\prod_{i=1}^{n} \ell(y_i\,|\,\theta)\right] p(\theta) \\ &=\propto [\,g(\theta)\,]^{b+n} \exp\left(\sum_{j=1}^{m} \phi_j(\theta)\, d_j(y)\right) \end{aligned}$$

Where

$$d_j = a_j + \sum_{i=1}^{n} t_j(y_i)$$

From the above, $p(\theta)$ is the conjugate prior for the likelihood given by

$\ell(y_i\,|\,\theta)$

With the posterior having the same form as the prior, with $n + b$ (in the posterior) replacing b and a_j.

CHAPTER FOUR

4.1 INTEGRATED NESTED LAPLACE APPROXIMATION (INLA)

In a (very small!) nutshell, Bayesian inference boils down to the computation of posterior/predictive distributions.

Note: $p(\theta|y) = p(\theta)p(y)$. The naïve version of probability that we learn of events that are independent.

$$p\ (\theta/y) = \frac{P\ (y|\theta)P(\theta)}{\int p(y|\theta)P(\theta)d\theta}$$

Where; $p(\theta|y)$ = posterior distribution; $p(\theta)$ = prior distribution and $p(y|\theta)$=likelihood distribution

Let $p(y|\theta)p(\theta) = g(y)$

N:B: The $p(\theta|y) \propto p(y|\theta)$ when we use flat priors.

Fundamental of INLA (Rue et al, 2009);

$p(\theta\,|\,y)$= $\frac{p(y,\theta)}{p(y)}$

The second "ingredient" is the Laplace approximation

$$\int g(\mathrm{y})d\,y^{*}$$

Main idea: approximate $log\ g(y)$ using Taylor series expansion around the mode $\hat{y}$ and use Laplace approximation to obtain prior and likelihood equivalence to Gaussian distribution

$log\ g(y) \approx \log g(\hat{y}) + \frac{\delta \log g(\hat{y})}{\delta y}(y-\hat{y})$+$\frac{1}{2}\frac{\delta^2}{\delta y^2}(y-\hat{y})^2$+R

$$= \log g(\hat{y})+\frac{1}{2}\frac{\delta^2 \log g(y)}{\delta y^2}(y-\hat{y})^2$$

(Since $\frac{\delta \log g(\hat{y})}{\delta y}$=0)

Setting $\hat{\sigma}^2$=$-1\Big/\frac{\delta^2 \log g(y)}{\delta y^2}$, we can re-write

$\log g\ (x) \approx \log g(\hat{y})$-$\frac{1}{2\hat{\sigma}^2}(y-\hat{y})^2$

Or equivalently;

$$\int g(y) = \int e^{\log g(y)}dy \approx const \int exp\left|-\frac{(y-\hat{y})^2}{2\hat{\sigma}^2}\right|dx$$

Thus under INLA $g(y) = normalizing\ constant \approx Normal(\hat{y}, \hat{\sigma}^2)$

N: B. As will be illustrated in the following chapter, INLA thinks in terms of Latent Gaussian models (LGMs) and Gaussian Markov Random Fields (GMRFs). Thus a common mean for all the observations is specified in terms of regression.

4.2 Laplace Approximation Example

Consider a χ^2 distribution: $p(x) = \frac{g(x)}{c} = \frac{x^{\frac{k}{2}-1}e^{\frac{-x}{2}}}{c}$

❶ $l(x) = \log g(x) = \left(\frac{k}{2} - 1\right) \log x - \frac{x}{2}$

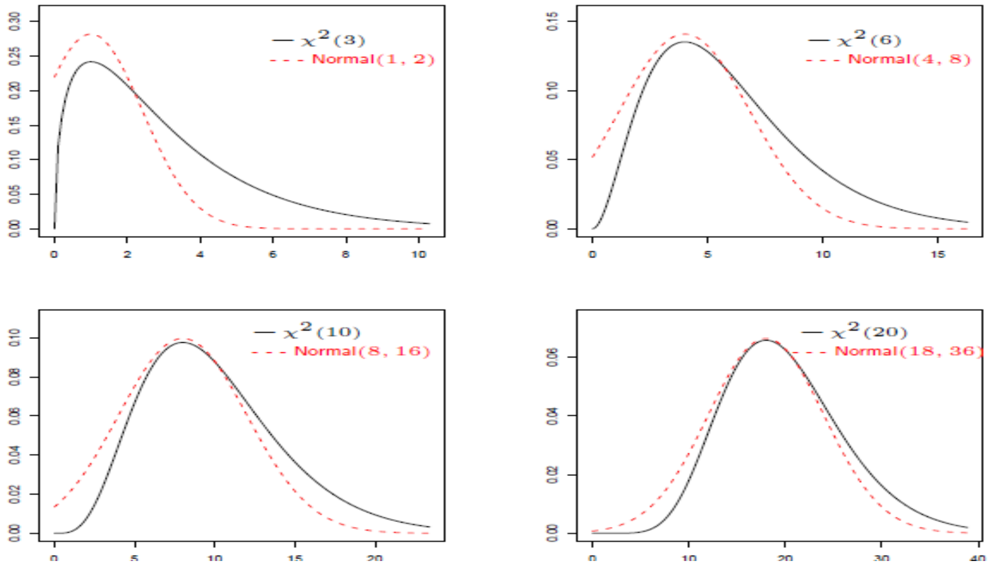
❷ $l'(x) = \frac{\partial \log g(x)}{\partial x} = \left(\frac{k}{2} - 1\right) x^{-1} - \frac{1}{2}$

❸ $l''(x) = \frac{\partial^2 \log g(x)}{\partial x^2} = -\left(\frac{k}{2} - 1\right) x^{-2}$

- Then
  - Solving $l'(x) = 0$ we find the mode: $\hat{x} = k - 2$
  - Evaluating $-\frac{1}{l''(x)}$ at the mode gives $\hat{\sigma}^2 = 2(k-2)$

- Consequently, we can approximate $p(x)$ as

$$p(x) \approx \tilde{p}(x) = \text{Normal}(k-2, 2(k-2))$$



- The general idea is that using the fundamental probability equations, we can approximate a generic conditional (posterior) distribution as

$$\tilde{p}(z \mid w) = \frac{p(x, z \mid w)}{\tilde{p}(x \mid z, w)},$$

where $\tilde{p}(x \mid z, w)$ is the Laplace approximation to the conditional distribution of x given z, w

- This idea can be used to approximate any generic required posterior distribution

Objective of Bayesian estimation

- In a Bayesian LGM, the required distributions are

$$\begin{aligned} p(\theta_j \mid \boldsymbol{y}) &= \int p(\theta_j, \boldsymbol{\psi} \mid \boldsymbol{y}) d\boldsymbol{\psi} = \int p(\boldsymbol{\psi} \mid \boldsymbol{y}) p(\theta_j \mid \boldsymbol{\psi}, \boldsymbol{y}) d\boldsymbol{\psi} \\ p(\psi_k \mid \boldsymbol{y}) &= \int p(\boldsymbol{\psi} \mid \boldsymbol{y}) d\boldsymbol{\psi}_{-k} \end{aligned}$$

- Thus we need to estimate:
  (1.) $p(\boldsymbol{\psi} \mid \boldsymbol{y})$, from which also all the relevant marginals $p(\psi_k \mid \boldsymbol{y})$ can be obtained;

  (2.) $p(\theta_j \mid \boldsymbol{\psi}, \boldsymbol{y})$, which is needed to compute the marginal posterior for the parameters

(1.) can be easily estimated as

$$\begin{aligned}
p(\psi \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{\theta}, \psi \mid \boldsymbol{y})}{p(\boldsymbol{\theta} \mid \psi, \boldsymbol{y})} \\
&= \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}, \psi) p(\boldsymbol{\theta}, \psi)}{p(\boldsymbol{y})} \frac{1}{p(\boldsymbol{\theta} \mid \psi, \boldsymbol{y})} \\
&= \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \psi) p(\psi)}{p(\boldsymbol{y})} \frac{1}{p(\boldsymbol{\theta} \mid \psi, \boldsymbol{y})} \\
&\propto \frac{p(\psi) p(\boldsymbol{\theta} \mid \psi) p(\boldsymbol{y} \mid \boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \psi, \boldsymbol{y})} \\
&\approx \left. \frac{p(\psi) p(\boldsymbol{\theta} \mid \psi) p(\boldsymbol{y} \mid \boldsymbol{\theta})}{\tilde{p}(\boldsymbol{\theta} \mid \psi, \boldsymbol{y})} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\psi)} =: \tilde{p}(\psi \mid \boldsymbol{y})
\end{aligned}$$

where

- $\tilde{p}(\theta \mid \psi, \boldsymbol{y})$ is the Laplace approximation of $p(\boldsymbol{\theta} \mid \psi, \boldsymbol{y})$
- $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\psi)$ is its mode

(2.) is slightly more complex, because in general there will be more elements in $\boldsymbol{\theta}$ than there are in ψ and thus this computation is more expensive

- One easy possibility is to approximate $p(\theta_j \mid \psi, \boldsymbol{y})$ directly using a Normal distribution, where the precision matrix is based on the Cholesky decomposition of the precision matrix $\boldsymbol{Q}$. While this is very fast, the approximation is generally not very good

- Alternatively, we can write $\boldsymbol{\theta} = \{\theta_j, \boldsymbol{\theta}_{-j}\}$, use the definition of conditional probability and again Laplace approximation to obtain

$$\begin{aligned}
p(\theta_j \mid \psi, \boldsymbol{y}) &= \frac{p\left(\{\theta_j, \boldsymbol{\theta}_{-j}\} \mid \psi, \boldsymbol{y}\right)}{p(\boldsymbol{\theta}_{-j} \mid \theta_j, \psi, \boldsymbol{y})} = \frac{p\left(\{\theta_j, \boldsymbol{\theta}_{-j}\}, \psi \mid \boldsymbol{y}\right)}{p(\psi \mid \boldsymbol{y})} \frac{1}{p(\boldsymbol{\theta}_{-j} \mid \theta_j, \psi, \boldsymbol{y})} \\
&\propto \frac{p\left(\boldsymbol{\theta}, \psi \mid \boldsymbol{y}\right)}{p(\boldsymbol{\theta}_{-j} \mid \theta_j, \psi, \boldsymbol{y})} \quad \propto \frac{p(\psi) p(\boldsymbol{\theta} \mid \psi) p(\boldsymbol{y} \mid \boldsymbol{\theta})}{p(\boldsymbol{\theta}_{-j} \mid \theta_j, \psi, \boldsymbol{y})} \\
&\approx \left. \frac{p(\psi) p(\boldsymbol{\theta} \mid \psi) p(\boldsymbol{y} \mid \boldsymbol{\theta})}{\tilde{p}(\boldsymbol{\theta}_{-j} \mid \theta_j, \psi, \boldsymbol{y})} \right|_{\boldsymbol{\theta}_{-j}=\hat{\boldsymbol{\theta}}_{-j}(\theta_j, \psi)} =: \tilde{p}(\theta_j \mid \psi, \boldsymbol{y})
\end{aligned}$$

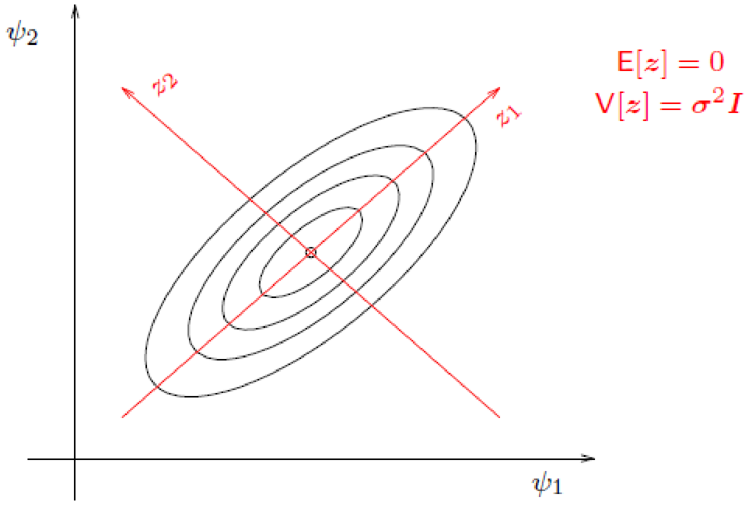
- Because $(\boldsymbol{\theta}_{-j} \mid \theta_j, \boldsymbol{\psi}, \boldsymbol{y})$ are reasonably Normal, the approximation works generally well
- However, this strategy can be computationally expensive

- The most efficient algorithm is the "**Simplified Laplace Approximation**"
  - Based on a Taylor's series expansion up to the third order of both numerator and denominator for $\tilde{p}(\theta_j \mid \boldsymbol{\psi}, \boldsymbol{y})$
  - This effectively "corrects" the Gaussian approximation for location and skewness to increase the fit to the required distribution

- This is the algorithm implemented by default by `R-INLA`, but this choice can be modified
  - If extra precision is required, it is possible to run the full Laplace approximation — of course at the expense of running time!

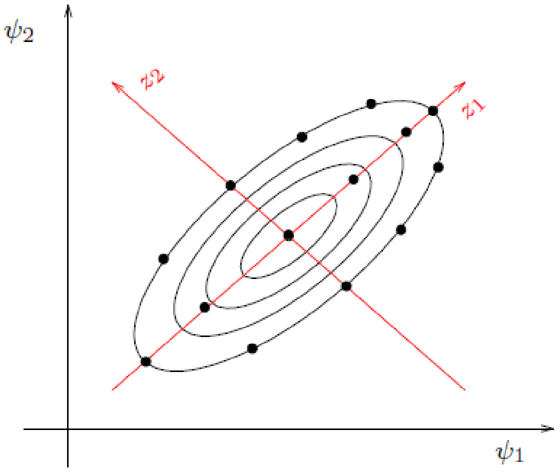
Operationally, the INLA algorithm proceeds with the following steps:

i. Explore the marginal joint posterior for the hyper-parameters $\tilde{p}(\boldsymbol{\psi} \mid \boldsymbol{y})$
   - Locate the mode $\hat{\boldsymbol{\psi}}$ by optimising $\log \tilde{p}(\boldsymbol{\psi} \mid \boldsymbol{y})$, eg using Newton-like algorithms
   - Compute the Hessian at $\hat{\boldsymbol{\psi}}$ and change co-ordinates to standardise the variables; this corrects for scale and rotation and simplifies integration


ψ_2
z_2
z_1
$\mathsf{E}[\boldsymbol{z}] = 0$
$\mathsf{V}[\boldsymbol{z}] = \sigma^2 \boldsymbol{I}$
ψ_1

Operationally, the INLA algorithm proceeds with the following steps:

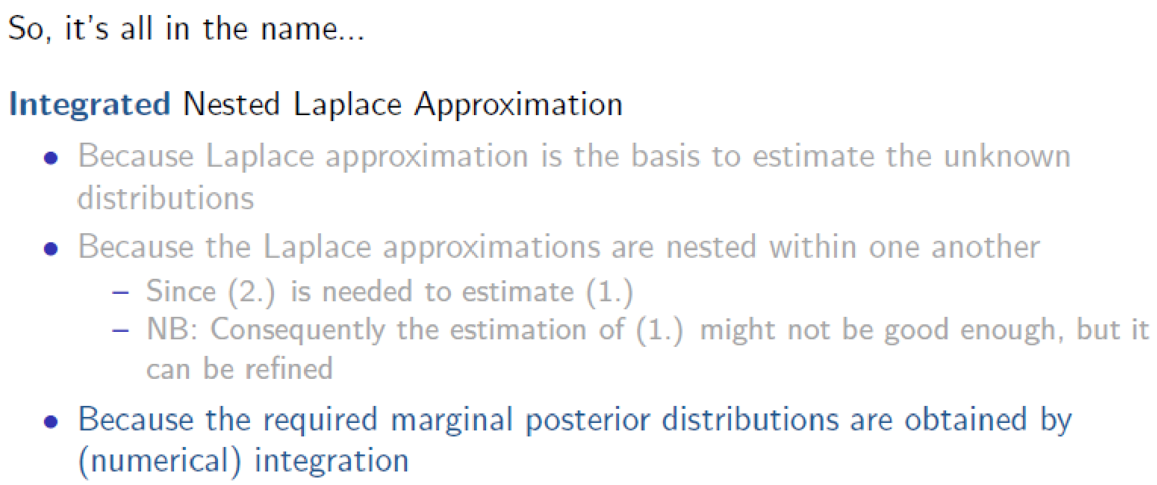
i. Explore the marginal joint posterior for the hyper-parameters $\tilde{p}(\psi \mid \boldsymbol{y})$
   - Locate the mode $\hat{\psi}$ by optimising $\log \tilde{p}(\psi \mid \boldsymbol{y})$, eg using Newton-like algorithms
   - Compute the Hessian at $\hat{\psi}$ and change co-ordinates to standardise the variables; this corrects for scale and rotation and simplifies integration
   - Explore $\log \tilde{p}(\psi \mid \boldsymbol{y})$ and produce a grid of H points $\{\psi_h^*\}$ associated with the bulk of the mass, together with a corresponding set of area weights $\{\Delta_h\}$



ii. For each element ψ_h^* in the grid,
   - Obtain the marginal posterior $\tilde{p}(\psi_h^* \mid \boldsymbol{y})$, using interpolation and possibly correcting for (probable) skewness by using log-splines;
   - Evaluate the conditional posteriors $\tilde{p}(\theta_j \mid \psi_h^*, \boldsymbol{y})$ on a grid of selected values for θ_j;

iii. Marginalise ψ_h^* to obtain the marginal posteriors $\tilde{p}(\theta_j \mid \boldsymbol{y})$ using **numerical integration**

$$\tilde{p}(\theta_j \mid \boldsymbol{y}) \approx \sum_{h=1}^{H} \tilde{p}(\theta_j \mid \psi_h^*, \boldsymbol{y}) \tilde{p}(\psi_h^* \mid \boldsymbol{y}) \Delta_h$$

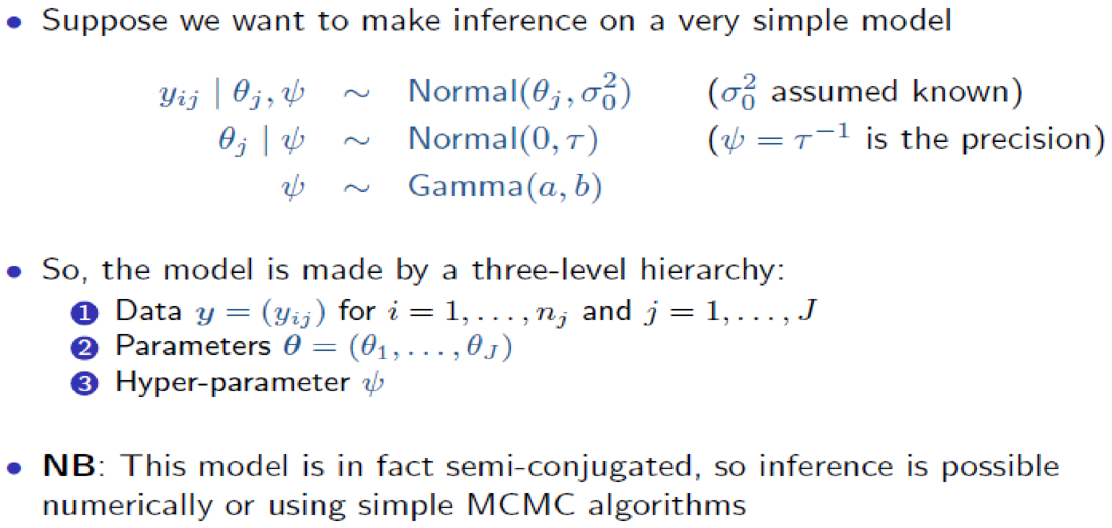


So, it's all in the name...

Integrated Nested Laplace Approximation

- Because Laplace approximation is the basis to estimate the unknown distributions
- Because the Laplace approximations are nested within one another
  - Since (2.) is needed to estimate (1.)
  - NB: Consequently the estimation of (1.) might not be good enough, but it can be refined
- Because the required marginal posterior distributions are obtained by (numerical) integration

4.3 INLA Example



- Suppose we want to make inference on a very simple model

$$\begin{aligned} y_{ij} \mid \theta_j, \psi &\sim \text{Normal}(\theta_j, \sigma_0^2) && (\sigma_0^2 \text{ assumed known}) \\ \theta_j \mid \psi &\sim \text{Normal}(0, \tau) && (\psi = \tau^{-1} \text{ is the precision}) \\ \psi &\sim \text{Gamma}(a, b) \end{aligned}$$

- So, the model is made by a three-level hierarchy:
  1. Data $y = (y_{ij})$ for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$
  2. Parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$
  3. Hyper-parameter ψ

- **NB**: This model is in fact semi-conjugated, so inference is possible numerically or using simple MCMC algorithms

- Because of semi-conjugacy, we know that

$$\boldsymbol{\theta}, \boldsymbol{y} \mid \psi \sim \text{Normal}(\cdot, \cdot)$$

and thus we can compute (numerically) all the marginals

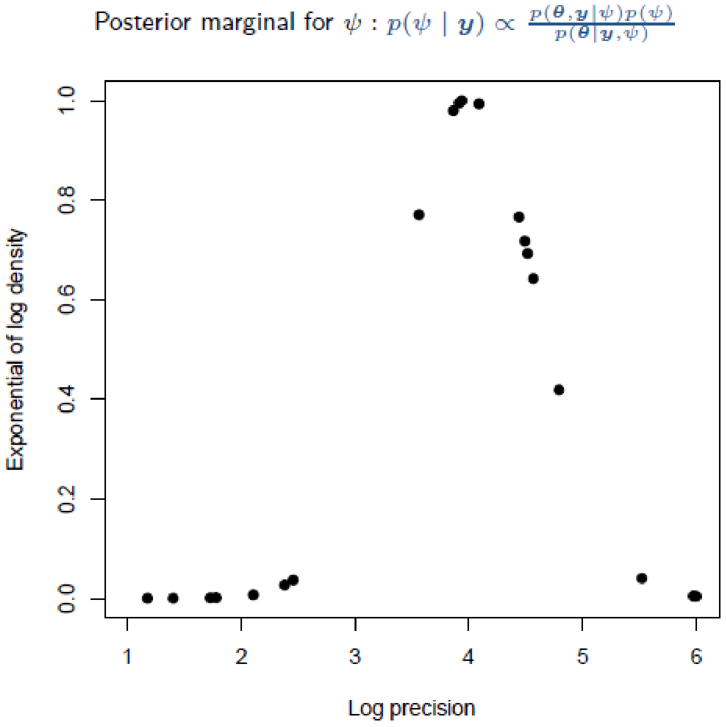
- In particular

$$\begin{aligned} p(\psi \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \psi) p(\psi) \\ &\propto \frac{\overbrace{p(\boldsymbol{\theta}, \boldsymbol{y} \mid \psi)}^{\text{Gaussian}} p(\psi)}{\underbrace{p(\boldsymbol{\theta} \mid \boldsymbol{y}, \psi)}_{\text{Gaussian}}} \end{aligned}$$

- Moreover, because $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \sim \text{Normal}(\cdot, \cdot)$ and so are all the resulting marginals (ie for every element j), it is easy to compute

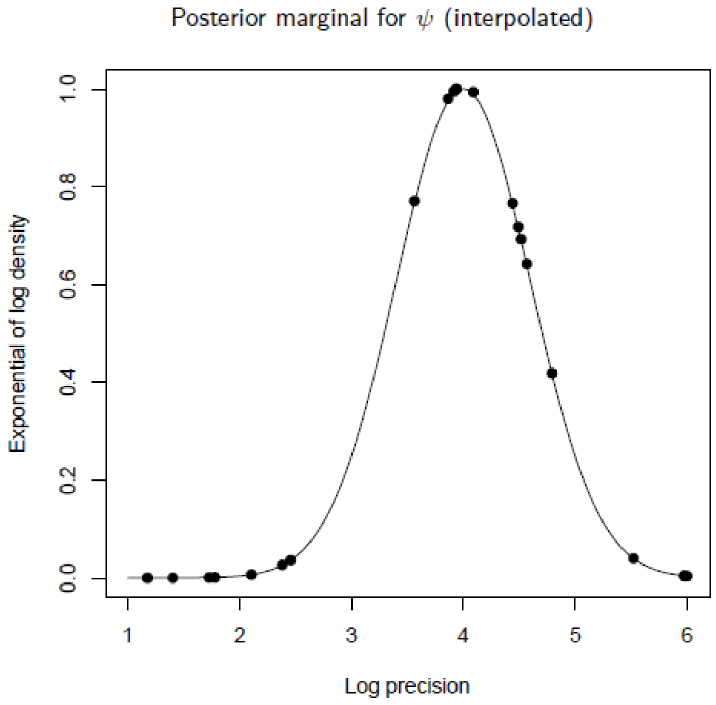
$$p(\theta_j \mid \boldsymbol{y}) = \int \underbrace{p(\theta_j \mid \boldsymbol{y}, \psi)}_{\text{Gaussian}} \; \underbrace{p(\psi \mid \boldsymbol{y})}_{\text{Approximated}} \; d\psi$$

1. Select a grid of H points for ψ ($\{\psi_h^*\}$) and the associated area weights ($\{\Delta_h\}$)

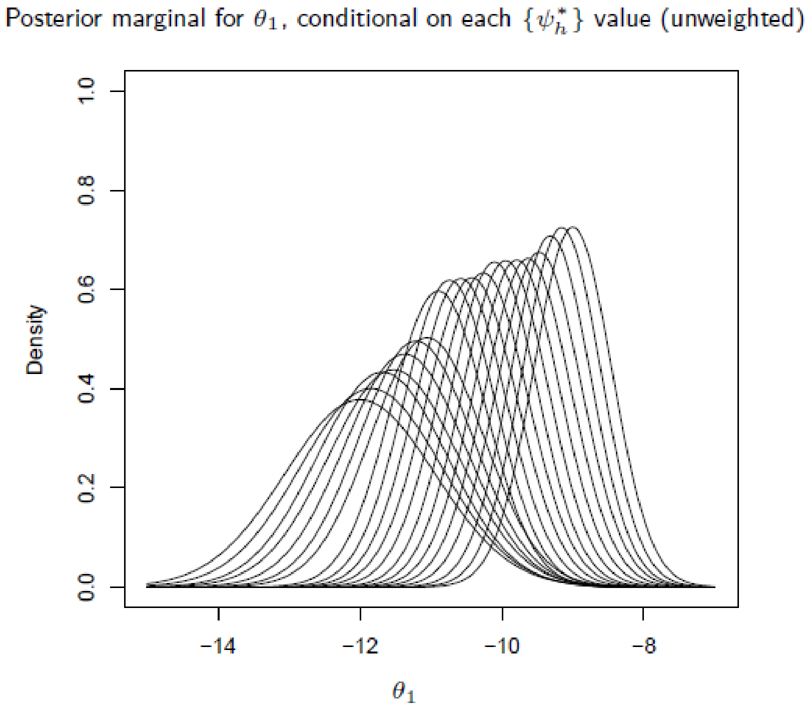


Posterior marginal for ψ : $p(\psi \mid \boldsymbol{y}) \propto \frac{p(\boldsymbol{\theta}, \boldsymbol{y} \mid \psi) p(\psi)}{p(\boldsymbol{\theta} \mid \boldsymbol{y}, \psi)}$

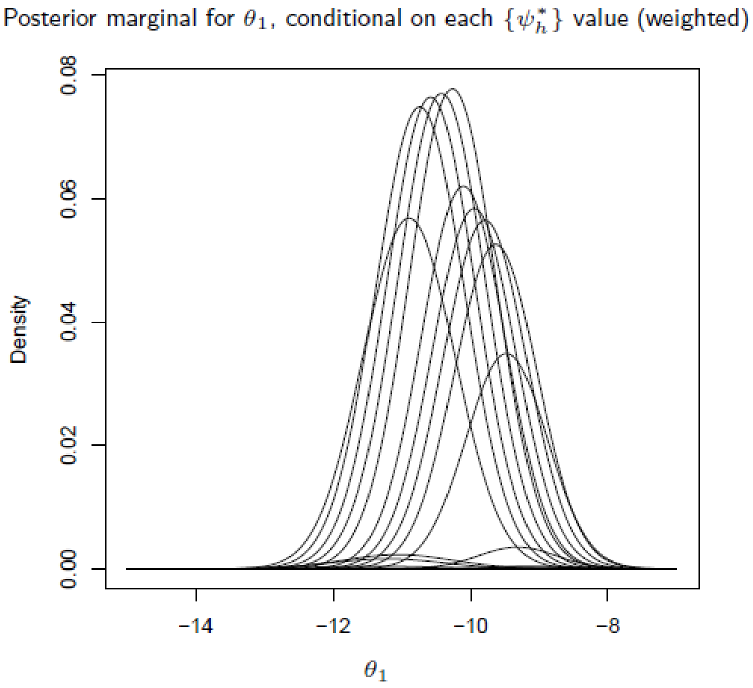
2. Interpolate the posterior density to compute the approximation to the posterior



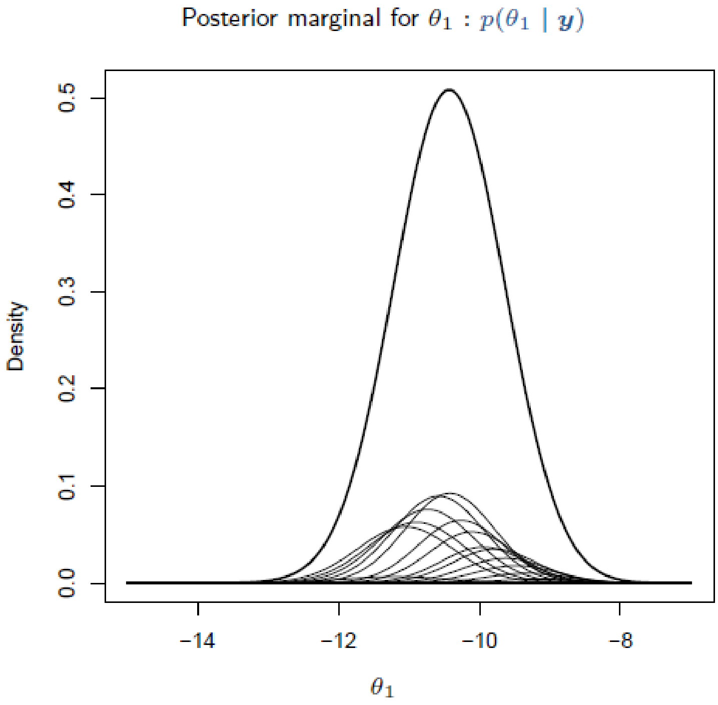
3. Compute the posterior marginal for each θ_j given each ψ on the $H-$dimensional grid



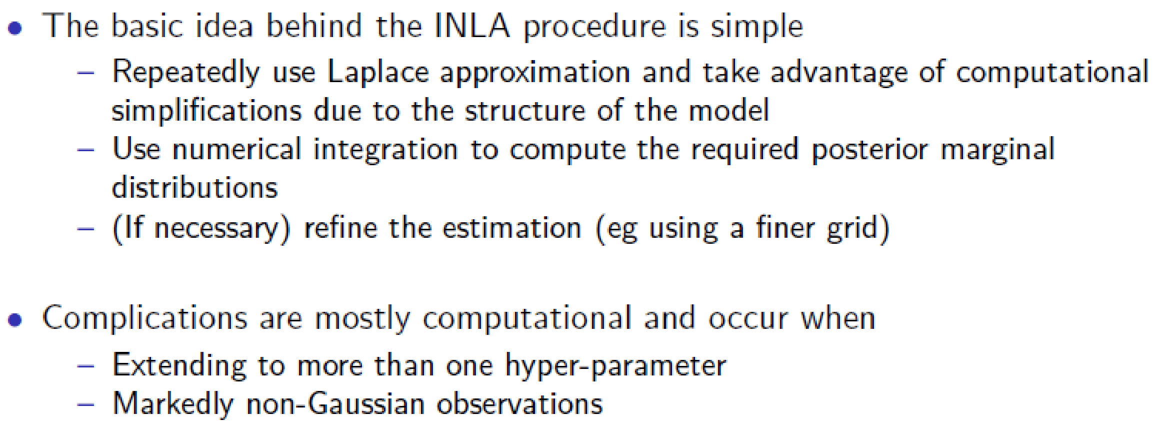
4. Weight the resulting (conditional) marginal posteriors by the density associated with each ψ on the grid



5. (Numerically) sum over all the conditional densities to obtain the marginal posterior for each of the elements θ_j



INLA Summary



- The basic idea behind the INLA procedure is simple
  - Repeatedly use Laplace approximation and take advantage of computational simplifications due to the structure of the model
  - Use numerical integration to compute the required posterior marginal distributions
  - (If necessary) refine the estimation (eg using a finer grid)
- Complications are mostly computational and occur when
  - Extending to more than one hyper-parameter
  - Markedly non-Gaussian observations

4.4 The Gaussian field

To fix the notion, let s any location on the study area and X(s) is the random effect at this location. We have X(s) a stochastic process, with s∈ D, where D is the domain area of the locations and D∈ $\boldsymbol{\mathfrak{R}^d}$**.** Suppose, for example, that we have D and country and we have any data measured on geographical locations, d=2, within this country (Lindgren & Rue, 2011).

Suppose that we assume that we have a realization of (s_i) , $i = 1,2,\dots,n$, a realization of $x(s)$ in n locations. It is commonly assumed that $x(s)$ has a multivariate Gaussian distribution. Also, if we assume that $x(s)$ is continuous over space, we have continuously indexed **Gaussian field (GF).** It is because we suppose it is possible that we get data in any location within the study region. To complete the specification of the distribution of $x(s)$, it is necessary to define its mean and covariance.

A very simple option is the definition of a correlation function based only on Euclidean distance between locations $s_i\ and\ s_j$. This assumes that if we have two pairs of points separated same distance h, both pairs have same correlation. Also is intuitive to choose any function decreasing with h. There is some work about the GF and correlation functions in (Abrahamsen, 1997), (Krainski, 2013)

A very popular correlation function is the Matérn correlation function, that depends on a scale parameter $k > 0$ and a smoothness parameter $v > 0$. Considering two locations s_i ands_j, the stationery and isotropic Matérn correlation function is;

$$cor_M\big(x(s_i), x(s_j)\big) = \frac{2^{1-v}}{\Gamma_v\, 2^{v-1}}\ \left(k\|s_i - s_j\|\right)^v K_v\ \left(k\|s_i - s_j\|\right) \qquad (4m)$$

Where ||·|| denotes the Euclidean distance and K_v is the modified Bessel function of second kind order and order. Also, we define the Matérn covariance function by $\sigma_x cor_M\left(x(s_i), x(s_j)\right), where\ \sigma_x\ is\ the\ marginal\ variance\ of\ the\ process.$

If we have a realization $x(s)$ on n locations, we write the joint correlation, or joint covariance, matrix $\boldsymbol{\Sigma}$ making each entry $\Sigma_{ij} = \sigma_x cor_M\left(x(s_i), x(s_j)\right)$. It is common to assume that $x(s)$ has a zero mean. So, we have completely defined a multivariate distribution to $x(s)$.

Now, suppose now that we have a data y_i observed at locations $s_i,\ i = 1, . . . , n$. If we suppose that we have an underlie GF that generates these data, we are going to fit the parameters of this process, making the identity$y(s_i) = x(s_i), and\ y(s_i)\ is\ just\ a\ realization\ of\ the\ GF$. In this case, the likelihood function is the multivariate distribution with mean $\mu_x\ and\ covariance\ \Sigma$. If we assume that $\boldsymbol{\mu}_x = \beta_0$, then we have four parameters to estimate.

In many situations we assume that we have an underlie GF but we are unable to observe it , therefore, observe a data with measurement error, i.e. $y(s_i) = x(s_i) + \varepsilon_i$. Additionally, it is common to assume that ε_i and ε_j are independent for all $i \neq j$ and $\varepsilon_i \sim N(0, \sigma^2{}_\varepsilon)$. This additional parameter, σ_ε, measures the noise effect, called the nugget effect. In this case, the covariance of marginal distribution of $y(s)$ is$\sigma^2{}_\varepsilon I + \Sigma$. This model is a short of extension of the basic GF model, and in this case we have one additional parameter to estimate as in (Diggle and Ribeiro, 2007).

It is possible to describe this model with larger class of models, the hierarchical models. Suppose that we have observations y_i on locations $s_i,\ i = 1, \dots, n.$ we start with

$$y_i | \theta, \beta, x_i, F_i \sim N(y_i | \mu_i, \phi)$$

$$x \sim GF(0, \Sigma)$$

Where $\mu_i = h(F^T{}_i\beta + x_i)$, F is a matrix of covariates, x is random effects, $\boldsymbol{\theta}$ are parameters of random effects, $\boldsymbol{\beta}$ are covariate coefficients, $h(\cdot)$ is a function of mapping the linear predictor $F^T{}_i\beta + x_i$ to $E(y_i) = \mu_i$ and ϕ is a dispersion parameter of the distribution, in the exponential family, assumed to y_i , with variance ${\sigma^2}_{\varepsilon}$ and x as GF.

We have many extensions of this basic hierarchical model. But if we know the properties of the GF, we are able to study all the practical models that contain or are based on, this random effect.

It is mentioned that the data, or the random effect, on a finite number of n points that we have observed is considered a realization of a multivariate Gaussian distribution. But to evaluate the likelihood function or the random effect distributions of the multivariate Gaussian density. So, we have,

$$\log \pi(x) = -\frac{n}{2}\log(2\pi) - |\Sigma| - \frac{1}{2}(x(s) - \mu_x)^T \Sigma^{-1}(x(s) - \mu_x),$$

Where $\boldsymbol{\Sigma}$ is a dense nxn matrix. To compute this, we need a factorization of this matrix. Because this matrix is dense, this amount to an operation of order $O(n^3)$, so is one "big n problem".

An alternative used in some software that allow the practice of classical geostatistical analysis, is the use of empirical variogram to fit the parameters of the correlation function. This option doesn't use any likelihood for the data and the multivariate Gaussian distribution to the random effects. A good technique of this technique is made on (Cressie, 1991).

However, it is adequate to assume any likelihood for the data and a GF for the spatial dependence, the model based approach on geostatistics, (Diggle and Ribeiro, 2007). So, in some times we need the use of the multivariate Gaussian distribution to the random effects. But, if the dimension of the GF is big, it is impractical to make model based inference.

In another area of the spatial statistics like for the analysis of areal data, there is a model specified by conditional distributions that implies a joint distribution with a sparse precision matrix. These models are called the Gaussian Markov random fields (GMRF), (Rue and Held, 2005). So, inference when we use GMRF is more easy to do than when we use GF, because to work with two dimensional GMRF models, we have cost of $\mathrm{O}(n^{3/2})$ on the computations with its precision matrix. So, it is easier to make analysis with "big n".

Additionally, it is common to assume

The parameter v, which is usually kept fixed, measures the degree of smoothness of the process and its integer value determines the mean squared differentiability of the process. Instead, $k > 0$ is a scaling parameter related to the range(ρ), i.e the distance at which the spatial correlation becomes small enough for the observation at site s_i and s_j to be declared independent.

4.5 The Gaussian Markov Random Fields (GMRFs)

Gaussian Markov random fields are frequently used as computationally efficient models in spatial statistics, Simpson et al, (2011). Unfortunately, it has traditionally been difficult to link GMRFs with the more traditional Gaussian random field models as the Markov property is difficult to deploy in continuous space.

From a practical perspective, the primary difficulty with spatial Gaussian models in applied statistics is dimension. The bad situation becomes worse with increasing dimension. Computationally speaking, this is a disaster Simpson et al, (2011). Time series models, for example, can suffer from the same problems. In the temporal case, such problems have been controlled by adding a conditional independence (Markovian), structure to the model, Simpson et al (2011). The key advantage of Markov property for time series models is that the computational burden then grows only linearly (rather than area wise or cubically) in the dimension, which makes inference on these models feasible for long time series, Simpson et al, (2011) and (Cameletti, Lindgren, Simpson, & Rue, 2012).

However, Markov property has had a less exalted role in spatial statistics. Almost, all instances where the Markov property has been used in spatial modeling has been in the form of Markov random fields, Simpson et al (2011) and (Blangiardo, Cameletti, Baio, & Rue, 2013), defined over a set of discrete Markov random fields, in which the value of the random fields at the nodes is jointly Gaussian Rue and Held (2005).

GMRFs are typically written as;

$x \sim N(\mu, Q^{-1})$,

Where Q is a precision matrix and the Markov property is equivalent to requiring that Q is sparse, that is

$$Q_{ij} = 0 \; iff \; x_i \; and \; x_j \;\; are \; conditionally \; independent \; (Rue \; and \; Held, 2005).$$

As problems in spatial statistics are usually concerned with inferring a spatially continuous effect over a domain of interest, Simpson et al (2011), it is difficult to directly apply the fundamentally discrete GMRFs. For this reason, it is commonly stated that there are two essential fields in spatial statistics (Simpson et al, 2011) and (Simpson & Lindgren, 2010). The one that uses GMRFs and the one that uses continuously indexed Gaussian random fields. In a recent paper, (Lindgren et al, 2011) showed that these two approaches are not distinct.

By carefully utilizing the continuous space of Markov property, it is possible to construct a Gaussian random fields for which all quantities of interest can be computed using GMRFs (Simpson et all, 2011) .

The most exciting aspect of the Markovian models of (Lindgren et al, 2011), is there flexibility (Simpson et al, 2011). There is no barrier conceptually or computationally and the methods can be extended to non-Gaussian models, semi continuous models, joint modeling a covariate with misalignment and modeling a non-stationary models all of which I will want to look into more details during my PhD program. This type of flexibility is not found in any other method for constructing Gaussian random field models (Simpson et al, 2011).

4.6 Computations with Gaussian Markov random fields

As in temporal setting, the Markovian property allows for first computation of samples, likelihoods and other quantities of interest (Rue and Held, 2005). This allows the investigation of much larger models than would be available using general multivariate Gaussian models, (Simpson et al, 2011).

The situation is not, however, as good as it is in the one dimensional case, where all of these quantities can be computed using O(n) operations, (Simpson et al, 2011) where n is the dimension of the GMRF. Instead, for the two dimensional spatial models, samples and likelihoods can be computed in O$(n)^{3/2}$ operations , where is still a significant saving on the O$(n)^3$ operations required for a generall Gaussian models , (Simpson et al, 2011).

The key object when computing with GMRFs is the Cholesky decomposition$Q = LL^T$, where $L\ \ is\ a\ lower\ triangular\ matrix. When\ Q\ is\ sparse, its\ Cholesky\ decomposition\ can\ be\ computed$ $very\ efficiently\ (Davis, 2006)$. One the Cholesky triangle has been computed, it is easy to show that $x = \mu + \ L^{-T}Z$ is a sample from the GMRF $x{\sim}N(\mu, Q^{-1})$ where$z{\sim}N(0, I)$. Similarly, the log density for a GMRF can be computed as;

$$\log \pi(x) = -\frac{n}{2}\log(2\pi) + \sum_{i=1}^{n} logL_{ii} - \frac{1}{2}(x-\mu)'Q(x-\mu),$$

Where L_{ii} is the i^{th} diagonal element of L and the inner product

$(x-\mu)'Q(x - \mu)\ the\ quadratic\ form\ can\ be\ computed\ in\ \mathrm{O(n)}$calculations using the sparsity of Q. It is also possible to use the Cholesky triangle L to compute $diag(Q^{-1})$, which are the marginal variances of the GMRF (Rue and Martino, 2007).

Furthermore, it is possible to sample from x conditioned on a small number of linear constraints$x|\beta x = b$, where $\beta \epsilon \Re^{kxn}$ is usually a dense matrix and the number of constraints,k is very small.

According to (Simpson et al, 2011), this occurs when the GMRF is constrained to sum to zero. However, if one wishes to sample conditional on the data, which usually corresponds to a large number of linear constraints, the alternative methods are usually more efficient.

While direct calculation of the conditional dense is possible, when $\boldsymbol{\beta}$ is a dense matrix, conditioning destroys the Markov structure of the problem (Simpson et al, 2011). It is still possible to sample efficiently using a technique known as conditioning by kriging (Rue and Held, 2005), whereby an unconditional sample x is drawn from $N(\mu, Q^{-1})$ and then corrected using the equation;

$$x^* = x - Q^{-1}\beta^T(\beta Q^{-1}\beta^T)^{-1}(\beta x - b).$$

Where k is small (Simpson et al, 2011).

The conditioning by kriging update can be computed efficiently from the Cholesky factorization. Simpson et al (2011), however, reiterate that when there are a large number of constraints, the conditional by kriging method will be inefficient, and if $\boldsymbol{\beta}$ is sparse (as is the case when conditioning on data), it is usually better to use alternative methods which includes but not limited to Bayesian methodologies.

The conditional kriging sampling can be done by kriging update $x^* = x - \boldsymbol{\delta x}\ and\ can\ be\ computed\ by\ solving\ the\ augmented\ system\ of\ equation;$

$\begin{pmatrix} Q & \beta^T \\ \beta & 0 \end{pmatrix} x \begin{pmatrix} \delta x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ \beta x - b \end{pmatrix}$, where y is an auxiliary variable.

Fast Bayesian Inference

The second and the most appealing property of GMRFs is that they behave well under conditioning. Consider the simple Bayesian hierarchical model;

$y|x \sim N(Ax, {Q_y}^{-1})$ and $x \sim N(\mu\,, {Q^{-1}}_x)$ where $A, Q_x, and\ Q_y$ are sparse matrices. A simple manipulation shows that $(x^T, y^T)^T$ is jointly a Gaussian markov random field with precision matrix;

$$Q_{xy} = \begin{pmatrix} Q_x + A^T Q_y A & -A^T Q_y \\ -Q_y A & Q_y \end{pmatrix}$$

And the mean defined implicitly through the equation;

$$Q_{xy}\mu_{xy} = \begin{pmatrix} Q_x\mu \\ 0 \end{pmatrix}$$

As $(x^T, y^T)^T$ is jointly a GMRF, According to (Simpson et al, 2011), it is easy to see as proved in (Rue and Held, 2005) that;

$x|y \sim N(\mu + \left(Q_x + A^T Q_y A\right)^{-1} A^T Q_y\ (y - A\mu)(Q_x + A^T Q_y A)^{-1})$, Simpson et al (2011).

It is important to note that the precision matrices for the joint distribution in the above matrix ofQ_{xy} and the conditional distribution $x|y\ above\ are\ only\ sparse\ and\ the\ corresponding\ fields\ are$ only GMRFs - if A is sparse. This observation directly links the structure of the matrix A to the availability of efficient inference methods, (Simpson et al, 2011).

If we group the unknown parameters into a vector $\boldsymbol{\theta}$, we obtain the following hierarchical model;

$$y|x,\theta \sim N(Ax, Q_y\theta^{-1})$$

$$x|\theta \sim N(\mu, Q_x\theta^{-1})$$

$$\theta \sim \pi(\theta)$$

In order to perform inference on the above three models, it is common to use Markov Chain Monte-Carlo (MCMC) methods for sampling from the posterior distribution;

$\pi(x,\theta|y)$, however, this is not necessary (Simpson et al, 2011). It is an easy exercise in Gaussian density manipulation to show that the marginal posterior for the parameters denoted by $\pi(\theta|y)$ can be computed without integration and is given by;

$$\pi(\theta|y)\ \alpha\ \frac{\pi(x,y|\theta)\ \pi(\theta)}{\pi(x|y,\theta)}\ for\ x = x^*$$

Where x^*can be any point, but is typically taken to be the conditional model $E(x|y,\theta)$ and the corresponding marginal's $\pi(\theta_j|y)$ can be computed using numerical integration (Simpson et al, 2011), and the usual observation that for everyθ, $\pi(x,y|\theta)$ is a GMRF (Rue and Martino, 2007) and (Rue et al, 2009) holds.

It follows that for models with Gaussian observations, it is possible to perform deterministic inference that is exact up to the error in the numerical integration (Simpson et al, 2011). In particular, if there are only a moderate number of parameters, this will be extremely fast, (Simpson et al, 2011). For non-Gaussian observation process, exact deterministic inference is no longer possible, (Simpson et al, 2011). However, (Rue et al, 2009), showed that it is possible to construct extremely accurate approximate inference schemes by cleverly deploying series of Laplace approximations. The integrated nested

Laplace approximations (INLA) has been used successfully on a large number of spatial problems, (Simpson et al, 2011); (Fong et al, 2010); (Akerkar et al, 2010); (Schrodle and Held, 2011); (Riebler et al, 2011); (Cameletti et al, 2011); (Cameletti et al, 2012); (Illian et al, 2011); (Lindgren et al, 2011); (Lindgren et al, 2012) among many others.

4.7 Continuously specified, Markovian Gaussian random fields

One of the primary aims of spatial statistics is to infer a spatially continuous surface $x(s)$ over the region of interest (Simpson et al, 2011). It is therefore necessary, to build probability distributions over the space of functions, and the standard way of doing this is to construct Gaussian random fields, which are the generalization of functions of multivariate Gaussian distributions in the sense that for any collection of points; $(s_1,\ s_2\ .\ .\ .\ s_p)^T$, the field evaluated at those points is jointly Gaussian, (Simpson et al, 2011).

In particular, $x \equiv (x(s_1), x(s_2), \ , \ , x(s_p))^{\,T} \sim N(\mu, \Sigma)$, where the covariance matrix is given by;

$\Sigma_{ij} = c(s_i, s_j)$ for some positive definite covariance function c(. . .). In most commonly used cases, the covariance function is non-zero everywhere and as a result $\boldsymbol{\Sigma}$ is a dense matrix, (Simpson et al, 2011).

It is clear that we would like to transfer some of the pleasant computational properties of GMRFs which are outlined above to the Gaussian random field setting.

The obvious barrier to this is that classical GMRF models are strongly tied to discrete sets of points. Throughout this work, we would borrow heavily from the recent development lead by (Lindgren et al, 2011).

4.8 The spatial Markov property

For temporal processes, defining Markov property is greatly simplified by the structure of time (Simpson et al, 2011). Its directional nature and the clear distinction between past, present and the future allow for every natural discursion of neighborhoods. Unfortunately, space is far less structured and as such the Markov property is harder to define exactly

(Simpson et al, 2011). Informally, a Gaussian random field $x(s)$ has the spatial markov property if for every appropriate set S separating $A \; and \; B$, the values of $x(s) \; in \; A$ are conditionally independent of the values in B given the values in S.

A formal definition of the spatial Markov property can be found in (Rozanov, 1977). In our view, this follows shortly after the first geography law (Tobler, 1970) in which observations that are separated by large distances are less related than those separated by short distances. Even only in line with this fact, a Markov property can be informally deduced.

As far as (Cameletti et al, 2011) is concerned, it is not immediately clear how the spatial Markov property can be used for computational inference. However, in an almost ignored paper, (Rozanov, 1877) provided the vital characterization of Markovian Gaussian random fields in terms of their power spectra. The power spectrum of a stationery Gaussian random field is defined as the Fourier transform of its covariance function$c(h)$, that is;

$$R(k) \equiv \frac{1}{(2\pi)^d} \int_{R^d}^{w} \exp(-ik^T h) \, c(h)$$

Rozanov showed that a stationery field is Markovian if and only if;

$R(k) = \frac{1}{p(k)}$, where $p(k)$ is a positive, symmetric polynomial. More on this could be found on (Rozanov et al, 1977) and (Cameletti et al, 2011) where it is discussed into more detail. For the purpose of this work, we will consider another alternative that links GMRF with the continuous processes, that is Stochastic Partial Differential Equations (SPDE) settings.

Preliminary and Main result

We will further discuss the Matérn covariance model and explain its relationship with SPDE methodology.

4.8 Matérn covariance model and its stochastic Partial Differential Equations (SPDE)

Let ||·|| denote the Euclidean distance inR^d. The Matérn covariance function between two locations $u, v \in R^d$ is defined as;

$r(u,v) = \frac{\sigma^2}{2^{v-1}\,\Gamma_v}(k\|v-u\|)^v K_v(k\|v-u\|)$, $(4n)$

Here, K_v is the modified Bessel function of second kind and order $v > 0$, $k > 0$ is a scaling parameter and σ^2 is the marginal variance.

The integer value of v determines the mean squared differentiability of the underlying process, which matters for predictions that are made by using such a model. However, v is usually fixed since it is poorly identified by the users in typical applications. A more natural interpretations of the scaling parameter k is as a range parameter ρ (rho); the Euclidean distance where $x(u) and\ x(v)$ are almost independent. Lacking a simple relationship, we shall use the empirically derived definition $\rho = \sqrt{8v}/_k$ which correspond to correlation

near 0.1 at the distance ρ, for all v. This relationship is explained into much more detail in (Lindgren et al, 2011).

Covariance function appears naturally in various scientific fields (Guttorp and Gneiting, 2006), but the important relationship that we shall make use of is that a GF $x(u)$ with the Mat$\acute{e}$rn covariance function is a solution to the linear fractional SPDE.

$(k^2 - \Delta)^{\alpha/2}\ \ x(u) = w(u),\ u\epsilon\Re^d,\ \alpha = v + {}^{d}/_{2},\ k > 0,\ v > 0.$ $(4o)$

Where $(k^2 - \Delta)^{\alpha/2}$ is a pseudo differential operator that is defined later in equation four. The innovation process $w\ is\ spatial\ Gaussian\ white\ noise\ \ with\ unit\ variance$, Δ is the Laplacian .

$$\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial {x^2}_i}$$

And the marginal variance is;

$$\Delta = \frac{\Gamma v}{\Gamma\left(v+\frac{d}{2}\right)(4\pi)^{d/2}\,k^{2v}}$$

We shall name any solution to equation two a Mat$\acute{e}$rn field in what follows. However, the limiting solutions to the SPDE in equation two as $k \to 0\ or\ v \to 0$ do not have mater covariance functions but the SPDE still has solutions when $k = 0\ or\ v = 0$ which are well defined random measures according to (Lindgren et al, 2011). Further, there is implicit assumption of appropriate boundary conditions for the SPDE, as **α**≥ 2 the null space of the differential operator is non-trivial, containing for example, the functions $\exp(ke^T u)$, for all $\|e\| = 1$.

The Matérn fields are the only stationary solutions to SPDE. The proof that was given by Whittle (1954, 1963) is to show that the number of spectrum of a stationary solution is;

$R(k) = (2\pi)^{-d}(k^2 + \|k\|^2)^{-\alpha}$,

Using the Foriour transform definition of the fractional Laplacian in $\Re^d$,

$$\{F(k^2 - \Delta)^{\alpha/2}(k) = (k^2 + \|k\|^2)^{-\alpha}(F\phi)(k)\}, \qquad (4p)$$

Where ϕ is a function on $\Re^d$ for which the right hand side of the definition has a well-defined inverse Fourier function.

To construct a GMRF representation of the mater field on the triangulated lattice, we start with a stochastic weak formulation of SPDE in equation two. Define the inner product;

$(f, g) = \int f(u)g(u)\, du$, eq5

Where the integral is over the region of interest. The stochastic weak solution of the SPDE is found by requiring that

$$\{(\phi_j,\ (k^2 - \Delta)^{\alpha/2}, x), j = 1, \dots\ , m\} \triangleq \{(\phi_j, w),\ j = 1, \dots., m\}$$

For every appropriate finite set of test functions $\{\phi_j(u),\ j = 1, \dots, m\}, where\ " \triangleq "$ denotes equality in distribution.

The next step is to construct a finite element representation of the solution to the SPDE (Brenner and Scott, 2007);

$$x(u) = \sum_{k=1}^{n} \psi_k\,(u) w_k \qquad (4q)$$

For some chosen basis functions $\{\psi_k\}$ and Gaussian distribution weights$\{w_k\}$ and n is the number of vertices in the triangulation. If the functions ψ_k are piecewise linear in each triangle, ψ_k is 1 at vertex k and 0 at all other vertices.

The second result is obtained using the $n\ x\ n\ matrices\ C, G, and\ K$ with entries

$C_{i.j} = \langle \psi_i, \psi_j \rangle,\ G_{i.j} = \langle \Delta\psi_i, \Delta\psi_j \rangle,\ (K_{k^2})_{i.j} = k^2 C_{i.j} + G_{i.j}$ to get the precision matrix $Q_{\alpha,k}$ as a function of k^2 and $\boldsymbol{\alpha}$.

$$Q_{1,k^2} = K_{k^2},$$

$$Q_{2,k^2} = K_{k^2}\ C^{-1} K_{k^2},$$

$$Q_{\alpha,k^2} = K_{k^2}\ C^{-1}\ Q_{\alpha-2.k^2}\ C^{-1}\ K_{k^2}, for\ \alpha = 3, 4, \dots$$

Here too we have the notation that if v increases; we need a more dense precision matrix.

The Q precision matrix is generalized for a fractional values of $\boldsymbol{\alpha}$ (or$v$) using a Taylor approximation, this is thoroughly discussed in (Lindgren et al, 2011 [authors discussion response]). From this approximation, we have the polynomial of order $p = [\alpha]$ for the precision matrix

$$Q = \sum_{i=0}^{p} b_i\, C\, (C^{-1}G)^i \qquad (4r)$$

For $\boldsymbol{\alpha}$=1 and $\boldsymbol{\alpha}$=2 we have $cor_M\big(x(s_i), x(s_j)\big) = \frac{2^{1-v}}{\Gamma v\ 2^{v-1}}\ (k\|s_i - s_j\|)^v\ K_v\ (k\|s_i - s_j\|)$ as indicated above. This is because for $\boldsymbol{\alpha}$=1 we have $b_0 = k^2$ and $b_1 = 1$, and for $\boldsymbol{\alpha}$=2 we have $b_0 = k^4$ and $b_1 = \alpha k^4$ and $b_2 = 1$. For fractional $\boldsymbol{\alpha}$=$^1/_2$, we have $b_0 = \frac{3k}{4}$ and $b_1 = k^{-1}\frac{3}{8}$. And for $\alpha = {^3/_2}$, $(v = 0.5, the\ exponential\ case)$, $b_0 = \frac{15k^3}{16}$, $b_1 = \frac{15k}{8}$, $b_2 = \frac{15k^{-1}}{128}$. Using these results combined with recursive construction for $\alpha > 2$, we have GMRF approximations for all positive integers and half integers. Therefore Mat$\acute{e}$rn correlation function connects MGRF to the continuous setting.

Although the approach does give a GMRF representation of the Mat$\acute{e}$rn field on the triangulated region, it is truly an approximation to the stochastic weak solution as we use only a subset of the possible test functions. Please, refer to (Lindgren et al, 2011) for more detailed information on how SPDE and other approaches have linked MGRF to the continuous settings.

4.9 Summary of Stochastic Partial Differential Equations (SPDE) in INLA

The simplest model for $x(s)$ currently implemented in R-INLA is the SPDE/GMRF version of the stationery Matérn family, obtained as the stationery solutions to $(k^2 - \mathit{\Delta})^{\alpha/2}\ (\tau x(u)) = w(s), s \in \mathit{\Omega},$

Where $\mathit{\Delta}$ is the laplacian, k is the spatial scale parameter, α controls the smoothness of the realizations, τ controls the variance, and Ω is the spatial domain. The right hand side of the equation $w(s)$ is Gaussian spatial white noise process. As noted by Whittle (1954, 1963), the stationery solutions on $\mathcal{R}^d$ have Matérn covariances,

$$Cov(x(u), x(v)) = \frac{\sigma^2}{2^{v-1}\,\Gamma_v}(k\|v-u\|)^v K_v(k\|v-u\|)$$

As illustrated above, the parameters in the two formulations are coupled so that the Matérn smoothness is $v = \alpha - {}^{d}\!/_{2}$ and the marginal variance is

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\nu}\tau^2}.$$

From this we can identify the exponential covariance with $v = {}^{1}\!/_{2}$ and $\alpha = {}^{3}\!/_{2}$ and note that fields with $\alpha \leq 1$ give $v \leq 0$ and that such fields have no point-wise interpretation (but do have well defined integration properties). From spectral theory one can show that integer values for α gives continuous domain Markov fields (Rozanov 1982), and these are the easiest for which to provide discrete basis representations as introduced in the authors discussion response in Lindgren et al, 2011.

$$x(s) = \sum_{k=1}^{n} \psi_k(s) x_k,$$

Where the joint distribution of $x = \{x_1, \ldots, x_n\}$ is chosen so that the distribution of the functions $x(s)$ approximates the distribution of solutions to the SPDE on the domain. To obtain a Markov structure, and to preserve it when conditioning on local observations, we use basis function with compact support. The construction is done by projecting the SPDE onto the basis representation in what is essentially a Finite Element Method (Lindgren & Rue, 2013).

To allow easy and explicit evaluation, for two dimensional domains we use piece-wise linear basis functions defined by a triangulation of the domain of interest. This yield sparse matrices $C, G_1,\ and\ G_2$ such that the appropriate precision matrix for the weights is given by

$$Q = \tau^2(\kappa^4 C + 2\kappa^2 G_1 + G_2)$$

For the default case $= 2$, so that the elements of Q have explicit expression as functions of $k\ and\ \tau$ which have been estimated above. Assuming the Gaussian distribution $x \sim N(0, Q^{-1})$ now generates continuously defined functions $x(s)$ that are approximate solutions to the SPDE (in a stochastically weak sense)

The simplest internal representation of the parameters in the model interface is $\log(\tau) = \theta_1$ and $\log(k) = \theta_2$ where v and θ_2 are assigned a joint normal prior distribution. Since $\tau\ and\ k$ has a joint influence on the marginal variances of the resulting field, it is often more natural to construct the parameter model using the standard deviation σ and range ρ where $\rho = \frac{(8v)^{1/2}}{k}$ is the distance for which correlation functions has fallen to approximately 0.13 for all $v > 1/2$.

Translating this into $\tau\ and\ k$ yields

$$\log \tau = \frac{1}{2} \log \left(\frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}} \right) - \log \sigma - \nu \log \kappa,$$
$$\log \kappa = \frac{\log(8\nu)}{2} - \log \rho.$$

The internal parameterization is then obtained by setting

$$\begin{aligned}\log\sigma &= \log\sigma_0 + \theta_1,\\ \log\rho &= \log\rho_0 + \theta_2.\end{aligned}$$

CHAPTER FIVE

COMBINATION OF BAYESIAN AND SPATIAL STATISTICS

5.1 Latent Gaussian models

We propose to use the Integrated Nested Laplace Approximation (INLA). This is an approach proposed by Rue et al. (2009) to perform approximate fully Bayesian inference on the class of Latent Gaussian models (LGMs). INLA makes use of deterministic Nested Laplace approximations and, as an algorithm tailored to the class of LGMs, it provides a faster and more accurate alternative to simulation-based MCMC schemes (Rue, Simpson & Lindgren, 2013).

The INLA framework was designed to deal with latent Gaussian models, where the observation (or response) variable y_i is assumed to belong to a distribution family (not necessarily part of exponential family) where some parameter of the family $\emptyset_i$ is linked to a constructed additive predictor η_i through a link function $g(\cdot)$ so that $g(\emptyset_i) = \eta_i$. The structured additive predictive η_i accounts for the effects of various covariates in an additive way:

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_b} \beta_k \, z_{ki} + \epsilon_i \qquad (4a)$$

Where $\{f^{(j)}()\}$'s are unknown functions of the covariates u, used for example to relax the linear relationship of covariates and to model temporal and/or spatial dependence, the $\{\beta_k\}$'s represent the linear effect of covariate z and the $\{\epsilon_i\}$'s are unstructured terms. Then a Gaussian prior is assigned to α, $\{f^{(j)}()\}$,$\{\beta_k\}$ and $\{\epsilon_i\}$.

We can also write the model described above using hierarchical structure, where the first stage is formed by the likelihood function with conditional independence properties given the latent field $x = (\eta, \ \alpha, \ f, \ \beta)$ and possible hyperparameter θ_1, where each data point $\{y_i, i = 1, \ \dots \ n_d\}$ is connected to one element in the latent field x_i (Martins, Simpson, Lindgren, & Rue, 2013)

Bayes rule only provides mathematically correct re-allocation of credibility across the candidate parameter values (Kruschke, 2012). The result reveals how strongly we should believe in each candidate parameter value given the data. We assume that the prior parameter values for each of the fixed effects ($\alpha_z, \alpha_y, altitude \ and \ seaDist$) have equal credibility across possibilities (non-informative Gaussian distribution with mean 0 and variance 1000).

The random effects, that is spatial random field and temporal will be Stochastic Partial Differential Equation approach (SPDE) (Lindgren & Rue, 2011) and autoregressive process of order 1 (ar1) respectively and the two are linked using group and control-group object in the model. In other words, each time point the spatial locations are linked by the SPDE model object, while across time, the process evolves according to an ar1 process.

With INLA, missing rainfall data (outcome) on either occurrence or amount component will contribute nothing to the likelihood though too many missing data may compromise richness of information from the data. In fact, information from the non-informative prior tends to prevail as the number of missing rainfall data increases. Data on covariates are internally modified to be equal to zero hence is not used in prediction (Simpson & Lindgren, 2011).

5.2 The Model

Many climatic scenarios, if defined continuously over space and time, can be monitored and measured only at a limited number of spatial locations and time points (Krainski, 2013). This is the case, for example of, amount of rainfall in a region and other meteorological fields including temperature, precipitation among others as well as geo-hydrological approaches.

Data coming from such fields are assumed to be realization of a continuously indexed spatial process (random field) changing in time denoted by;

$$Y(s,t) \equiv \{y(s,t) \; : \; (s,t)\, \epsilon\, D \subseteq \Re^2\, x\, \Re\,\} \qquad (4f)$$

These realizations are used to make inference about the process that generated the data and subsequently predict it at unvisited locations. Usually, we deal with a Gaussian field (GF) that is completely specified by its mean and Spatio-temporal covariance function as illustrated by (Simpson & Lindgren, 2012)

$$cov\big(y(s,t), y(s',t')\big) = \boldsymbol{\sigma}^{\mathbf{2}}\boldsymbol{C}((s,t),(s',t'))$$

Defined for each (s,t) and (s',t') for all $s \neq s'$ and $t \neq t'$ in $\Re^2\, x\, \Re$. In addition, the process is said to be second order stationery if its mean is not a function of time and the spatio-temporal covariance depends only on the locations and time points through the spatial distance vector

$h = (s - s'\,)\, \epsilon \Re^2$ and the spatio-temporal lag $l = (t - t'\,)\, \epsilon\, \Re$

As illustrated in Ranergee et al (2004), Lindgren et al (2011), Cameletti et al., (2012) and Lindgren,(2012) even if a GF is easily defined directly through its first and second moments, its implementation suffers from the so called “bin n” problem that arises especially in case of large datasets in space and operations required for model fitting, spatial interpolation and prediction.

In this paradigm, it has been suggested in the literature that possible ways to deal with the problem of “big n” includes covariance tapering, predictive process models and low rank kriging (this is documented in Furrer et al (2006), Banergee et al (2008), Cressie and Johannesson (2008) among others).

These approaches share one thing in common, they try to reduce the dimension or simplify the structure of the dense covariance matrix of the Gaussian field (GF).

In this project, we consider approaches that include representing a continuously indexed GF with Mat*é*rn covariance function as a discretely indexed random process. This is commonly referred to as a Gaussian Markov Random Field (GMRF) as illustrated in Rue and Held (2005), Rue et al (2009) and (Lindgren & Rue, 2011). This proposal is based on the work is illustrated further in Cameletti et al, (2012) and a more detailed step by step in Lindgren et al (2011), where an explicit link between GFs and GMRFs was reformulated as a basis function representation. It is provided as a Stochastic Partial Differential Equations (SPDE) approach. The most important point here is that the Spatio-temporal covariance and the dense covariance matrix are substituted respectively by a neighborhood structure and by a sparse precision matrix that together form a GMRF.

The idea behind moving from GF to GMRF stems from good computational properties that GMRF enjoys. According to Cameletti et al, (2012) which is the key paper in this project, GMRFs are defined by a precision matrix with sparse structure for which it is possible to use computationally effective numerical methods especially Integrated Nested Laplace Approximations (INLA) algorithm which was only recently proposed in Rue et al (2009) a more effective and efficient alternative to the traditional MCMC methods. The most outstanding advantage of INLA as indicated by Rue et al (2009), Lindgren et al (2011), Lindgren (2012) and Cameletti et al (2012) among others is computational because it produces almost immediately accurate approximate approximations to posterior distributions including in cases of complex models. Therefore, the joint use of SPDE approach together with INLA algorithm is a powerful solution in overcoming the computational challenges related to GF modeling. In addition, the common inability of MCMC algorithm to converge leading to a false positive posterior distribution is another reason why INLA should be preferred.

The main objective of this project is to illustrate the implantation of SPDE method using Integrated Nested Laplace Approximations focusing on the amount of rainfall in Kenya taken from the months of January to March 2013 from 24 weather stations each day. This would be a spatio-temporal model considering the effect of time component in the model.

The model of analysis would be a modified version of what is presented by Cameletti et al (2012).

$$Y(s,t) \equiv \{y(s,t) \;\;:\;\; (s,t)\,\epsilon\, D \;\subseteq\; \Re^2\; x\; \Re\,\} \tag{4g}$$

Where;

$$y(s,t) = z(s,t)\boldsymbol{\beta} \;+\; \boldsymbol{\xi}(\boldsymbol{s},\boldsymbol{t}) +\; \boldsymbol{\epsilon}(\boldsymbol{s},\boldsymbol{t})$$

$$\boldsymbol{\xi}(\boldsymbol{s},\boldsymbol{t}) = \;\; \boldsymbol{a\xi}(\boldsymbol{s},\boldsymbol{t}-\mathbf{1}) +\; \boldsymbol{\omega}(\boldsymbol{s},\boldsymbol{t})$$

AS shown by Camelleti et al (2012), the equations define a hierarchical model characterized by a GF $y(s,t)$ built from covariate information$z(s,t)$, measurement error $\epsilon(s,t)$ and a first order autoregressive dynamic model for the latent process $\xi(s,t)$ with spatially correlated innovations$\omega(s,t)$.

5.3 The spatio-temporal model

Let $y(s_i,t)$ denote the realizations of the spatio-temporal process $Y(.,.)$ that represents the amount of rainfall at station $i = 1, \dots d$ located at s_i and day $t = 1, \dots, T$.

Further, let r_i the rain on the locations_i. We use the augmented data by definition of two new variables. We define the occurrence variable

$$z = \begin{cases} 1, & if\; r_i > 0 \\ 0, & \quad Otherwise \end{cases} \tag{4h}$$

And the amount variable

$$y = \begin{cases} NA, & if\ r_i = 0 \\ r_i, & Otherwise \end{cases} \qquad (4i)$$

With above modification, we use a model with two likelihoods. We use the Bernoulli for y_1

And the gamma likelihood for y_2 such that

$z \sim Bernoulli(p_i)$ and $y \sim Gamma(a_i, b_i)$

So we need a two column response matrix one for each response and $2xn\ rows$. we put the first response on first column at first n rows and the second response on the last n rows of second column.

$$y(s_i, t) = z(s_i, t)\boldsymbol{\beta} + \boldsymbol{\xi}(s_i, t) + \boldsymbol{\epsilon}(s_i, t)$$

$$\underbrace{y(s_i, t)}_{Outcome} = \underbrace{z(s_i, t)\boldsymbol{\beta}}_{model\ fixed\ effects} + \underbrace{\boldsymbol{\xi}(s_i, t)}_{\substack{\boldsymbol{Temporal\ effects} \\ \boldsymbol{Autoregressive\ process\ of\ order\ one}}} + \underbrace{\boldsymbol{\epsilon}(s_i, t)}_{\boldsymbol{White\ noise}\ (\boldsymbol{error})}$$ Where

$$z(s_i, t) = z_1(s_i, t) + , \ . \ . \ . \ , + z_p(s_i, t)$$

Denotes the vector of p variates for site s_i at time t, and $\beta = (\beta_1, . . . , \beta_p)$ is the coefficient vector. Moreover, $\boldsymbol{\epsilon}(s_i, t) \sim N(0, \sigma^2{}_{\varepsilon})$ is the measurement error defined by a Gaussian white noise process, both serially and spatially uncorrelated. In classical geostatistical literature, the term $z(s_i, t)\boldsymbol{\beta}$ is commonly referred to as the **trend surface** while the error variance $\sigma^2{}_{\varepsilon}$ is commonly referred to as the nugget effect in variogram analysis. Finally, $\boldsymbol{\xi}(s_i, t)$ is the realization of the so called state process, i.e. the true unobserved level of rainfall. According to Cameletti et al(2012), such phenomenon are assumed to be spatio-temporal Gaussian field that changes in time with first order autoregressive dynamics with coefficient a and equation given by;

$$\xi(s_i, t) = a\xi(s_i, t-1) + w(s_i, t) \qquad (\boldsymbol{4j})$$

For $t = 2, \ . \ . \ ., \ T$ where $|a| < 1$ normally the weights of autoregressive process and

$\xi(s_i, 1)$ derives from the stationery distribution which is

$N(0, \, {}^{\sigma^2{}_w}/_{1-a^2})$

Furthermore, $w(s_i, t)$ has a zero mean Gaussian distribution, is assumed to be temporally independent and is characterized by the spatio-temporal covariance function (Tijms, 2003) and (Schrödle & Held, 2011).

$$cov\left(w(s_i, t), w\left(s_j, t'\right)\right) = \begin{cases} 0, & if\ t \neq t' \\ \sigma^2{}_w c(h), & if\ \ t = t' \end{cases}$$

For $i \ \neq j$. The purely spatial correlation function $c(h)$ depends on the location of $s_i \ \ and \ \ s_j$ only through the Euclidean spatial distance$h = \left|s_i \ - \ s_j\right| \epsilon \ \Re$; thus, the process is assumed to be second order stationery and isotropic (Cressie, 1993).

According to Cameletti et al (2012), such a process follows immediately that;

$var\big(w(s_i, t)\big)$ = $\sigma^2{}_w$, for each s_i and t.

The spatial correlation function $c(h)$ is defined by the Mat$\acute{e}$rn function and is given by;

$$c(h) = \frac{1}{\Gamma v \ 2^{v-1}} \ (kh)^v \, K_v \, (kh) \qquad (4k)$$

With K_v denoting the modified Bessel function of second kind and order v>0. The parameter v, which is usually kept fixed, measures the degree of smoothness of the process and its integer value determines the mean squared differentiability of the process. Instead, $k > 0$ is a scaling parameter related to the range(ρ), i.e the distance at which the spatial correlation becomes small enough for the observation at site s_i and s_j to be declared independent.

In classical geostatistics, we use the empirically derived definition;

$\rho = \frac{\sqrt{8v}}{k}$, with

ρ Corresponding to the distance where the spatial correlation is close to 0.1, for each$v$.

$Collecting\ all\ the\ observations\ measured$ at time t in a vector denoted by $y_t = (y_1(s_1, t) + , \ . \ . \ . \ , + \ y_d(s_d, t))'$

It follows that equation one and two can be written as;

$y_t = \ z_t\beta + \ \xi_t + \ \boldsymbol{\varepsilon_t},$ where $\boldsymbol{\varepsilon_t} \sim N(0, \sigma^2{}_\varepsilon \, I_d)$

And $\xi(t) = a\xi_t + \ w_t$, where $w_t \ \sim \ N(0, \Sigma = \sigma^2{}_w \, \Sigma)$

Where I_d is the identity matrix of dimension$d$, and

$z_t = (z(s_1, t)' \, , . \ . \ . \ , \ z(s_d, t)')'$ and

$z_t = (\xi(s_1, t),. \ . \ . \ , \ \xi(s_d, t))'$ with ξ_1 coming from the stationery distribution of the $AR(1)$ process $N(0, \frac{\Sigma}{1-a^2})$.

$Moreover, \hat{\Sigma}$ is the dense correlation matrix of dimension d with elements $c\|s_i - \ s_j\|$, where $c(\,.\,)$ is the Mat$\acute{e}$rn function given by equation four and is parameterized by $k \ and \ v$ as shown above.

$Let\ \boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{\sigma^2}_w, \boldsymbol{k}\}$ denote the parameter vector to be estimated. The joint posterior distribution is given by; $\pi(\theta, \xi|y) \ \alpha \ \ \pi(y|\theta, \xi) \ \boldsymbol{x} \ \ \pi(\xi|\theta) \ \boldsymbol{x} \, \pi(\theta)$

Where the notation $\pi(\,.\,)$ is used for the probability density function,

$y = \{y_t\}$ and $\xi = \ \xi_t$ with $t = 1,. \ . \ . \ , \ T$.

Usually. Independent prior distributions are chosen for the parameters, so that;

$$\pi(\theta) = \prod_{i=1}^{\dim(\theta)} \pi(\theta_i)$$

$considering\ \ that\ the\ observations\ y_t\ are\ serially\ independent\ conditional\ on\ \boldsymbol{\xi}$ and that the state process follows a Markovian time dynamic process . With this additional information, equation seven can be written as follows; $\pi(\theta, \xi|y) \ \ \alpha \ \ \prod_{t=1}^{T} \pi(y_t|\xi_t, \theta) \ \ x \ \ \pi(\xi_1|\theta) \prod_{t=2}^{T} \pi(\xi_t| \ \xi_{t-1}, \ \theta) \ \ x \ \ \pi(\theta)$

From the Gaussian distributions defined in equations five and six, it follows immediately that the joint posterior distribution in equation eight is given by;

$$\pi(\theta,\xi|y)\ \ \alpha\ (\sigma^{2}{}_{\varepsilon})^{\frac{-dT}{2}}\exp\left(\frac{1}{2\sigma^{2}{}_{\varepsilon}}\sum_{t=1}^{T}(y_t - z_t\beta - \xi_t)'\ x\ \ (y_t - z_t\beta\right.$$

$$\left.- \xi_t)\right)\ x\ \ \left(\frac{\sigma_w{}^2}{1-a^2}\right)^{\frac{-d}{2}}\ \ |\hat{\Sigma}|^{\frac{1}{2}}\ \exp\left(-\frac{1-a^2}{2\sigma_w{}^2}\xi'{}_1\,\hat{\Sigma}^{-1}\ \ \xi_1\right)\ x$$

$$\left((\sigma_w{}^2)^{-\frac{d(T-1)}{2}}\ \ |\widehat{\Sigma}|^{-\frac{(T-1)}{2}}\ \exp\left(-\frac{1}{2\sigma_w{}^2}\textstyle\sum_2^T(\xi_t - a\xi_{t-1})'\right)\ \ x\ \ \hat{\Sigma}^{-1}(\xi_t -\right.$$

$$\left.a\xi_{t-1})\right)\ x\ \ \prod_{i=1}^{\dim(\theta)}\pi\,(\theta_i) \qquad\qquad (4l)$$

where $|\widehat{\Sigma}|$ *is the determinant of the dense* $d -$ *dimensional covariance matrix* $\hat{\Sigma}$.

In Bayesian framework, the common approach to make inference for this model (i.e parameter estimation and spatial prediction)is the traditional MCMC sampling method. This traditional method work has been thoroughly discredited by Rue et al, (2009)*and Lindgren et al,* (2011)*besides many other key papers* . However, we will heavily rely on a summary of this paradigm as provide by Cameletti et al, (2012), Lindgren et al, (2012), Simpson et al, (2012) and unpublished reports including the Elias’s book chapters in progress on geostatistics using stochastic partial differential equations (SPDE).

CHAPTER SIX

6.1 BAYESIAN SPATIAL APPLICATION TO RAINFALL DATA

The daily rainfall data just like other climate related data like precipitation is one type of data that is not easy to fit using simple models. This is because there are days when it rains and days when it does not rain moreover, whether it rains at station i depends on the status of rainfall at station j for $i\ and\ j\ are\ neighbours$. This as earlier illustrated brings to the model the effect due to space with decreases with increasing distance from the ith station. In addition, the complexity increases when the data is taken on the same stations over a time period say one year. There exist the effect due to time and just like in time series observations between adjacent days are more similar than those separated by many days. As indicated in the theory of SPDE, Markov chain which traditionally is associated with discrete time points observations; it could also begin to think continuous in paradigm.

The distribution of rainfall (Assymmetric)

Distributions that can be used in this set up could be

$$f(x) = \begin{cases} \frac{\beta^{\alpha}}{\Gamma\alpha} \exp(-\beta x) x^{\alpha-1}, & x > 0,\ \ \beta > 0\ ,\alpha > 0 \\ \qquad 0, & Elsewhere \end{cases}$$

Which is gamma distributed. As can be seen above x takes values that are greater than zero but not zero.

We therefore, let r_i be the rain on location s_i. We use the augmented data by definition of two new variables. That is the occurrence and the amount variable i.e;

$$z = \begin{cases} 1, & if\ r_i > 0 \\ 0, & \quad Otherwise \end{cases}$$

And the amount variable

$$y = \begin{cases} NA, & if\ r_i = 0 \\ \ \ r_i, & \quad Otherwise \end{cases}$$

We will therefore use two likelihoods in the model building. We use the Bernoulli likelihood for z_i and the gamma for y_i.

$z_i \sim Bernoulli(p_i)$ and $y_i \sim Gamma(\alpha_i, \beta_i)$

So we need a two column matrix, one for the response and $2xn\ row$. We put the first response on the first column at first n rows and the second response on last n rows of second column.

Later we associate on the model formulae, one likelihood for each column. We define the predictor to the first component by

$$logit(p_i) = \alpha_z + x_i$$

Where α_z is an intercept and x_i is a random effect modeled by a Gaussian field through the SPDE approach.

To second component we consider that $E(y_i) = \mu_i = \frac{a_i}{b_i}\ \ and\ var(y_i) = \frac{a_i}{b^2{}_i} = \frac{\mu_i}{\emptyset}, where\ \emptyset\ is\ a\ precision\ parameter\ and\ we\ define\ a\ linear\ predictor\ to\ \log(\mu_i).$ we have;

$$\log(\mu_i) = \alpha_y + \beta x_i$$

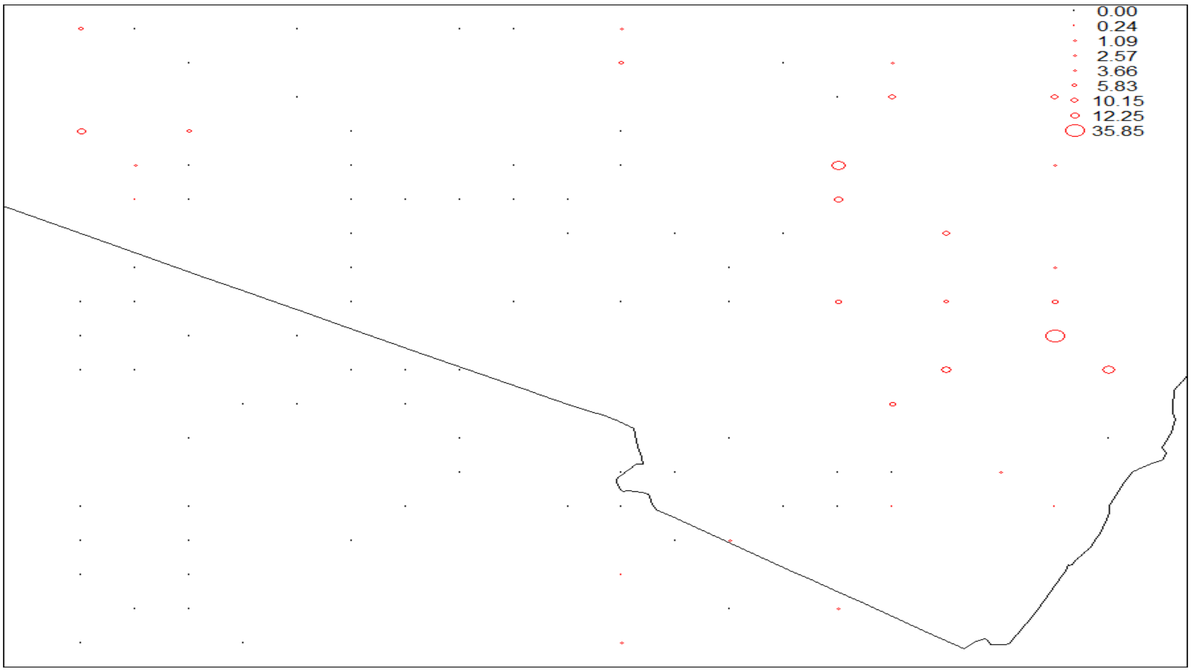
Where α_y is an intercept and β is a scaling parameter to $x_i\ that\ is\ a\ shared\ \ random\ effect\ with\ first\ componemt\ of\ the\ model.$ A time random effect which is exchangeable must be added to the model in addition to the covariate which for the purpose of the project include elevation and distance from the sea.

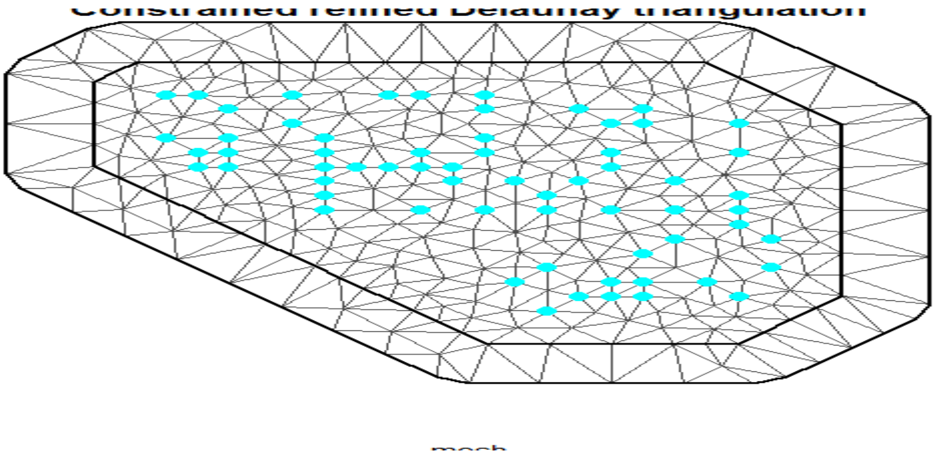
With this modification, we use a Bayesian method and develop a model with two likelihoods. We use the Bernoulli likelihood for z_i and a gamma for y_i. We define a linear predictor for first component $logit(p_i) = \alpha_z + x_i$ where α_z is an intercept and x_i is a random effect modeled by a Gaussian field through the stochastic partial differential equations approach (SPDE). To second component, we consider $\log(\mu_i) = \alpha_y + \beta x_i$ where α_y an intercept is and β is a scaling parameter to x_i that is a shared random effect with the first component of the model.

We propose to use the Integrated Nested Laplace Approximation (INLA). This is an approach proposed by Rue et al. (2009) to perform approximate fully Bayesian inference on the class of Latent Gaussian models (LGMs). INLA makes use of deterministic Nested Laplace approximations and, as an algorithm tailored to the class of LGMs, it provides a faster and more accurate alternative to simulation-based MCMC schemes.

6.2 Exploratory data analysis

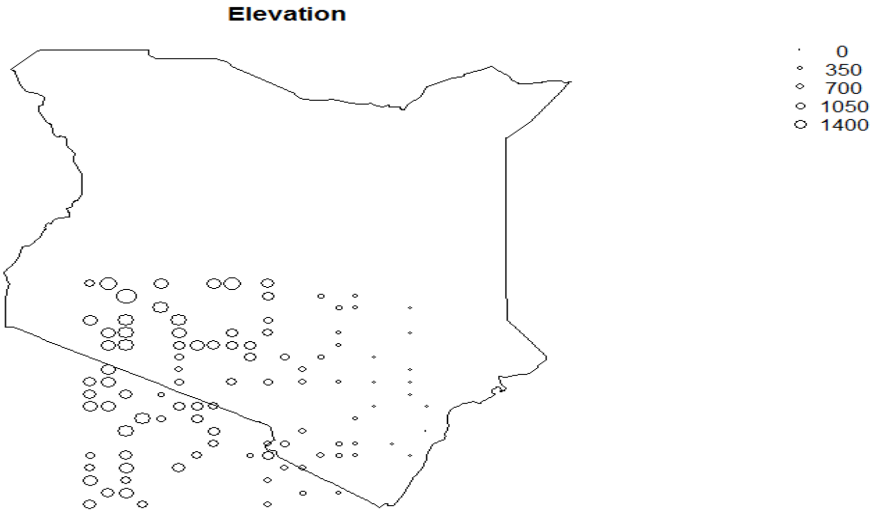
This is a typical example of a daily rainfall recorded for only a single day for 100 locations for 365 Julian days of 1998.



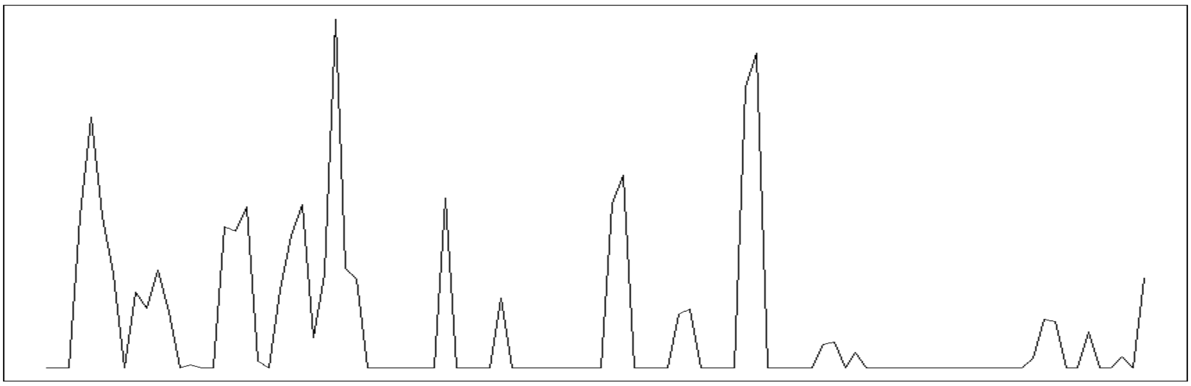


Taking a point at which no rain was recorded, and then it is likely that there is no rain at neighboring sites. In addition, choosing a site where larger rainfall was recorded then we are likely to find rainfall at neighboring sites. If daily rainfall is recorded say, for a year or so, then we have an outcome with both spatial and temporal effects. Besides this, rainfall is a mixture of both occurrence and amount variables and as such no form of data transformation would be appropriate for such scenarios.

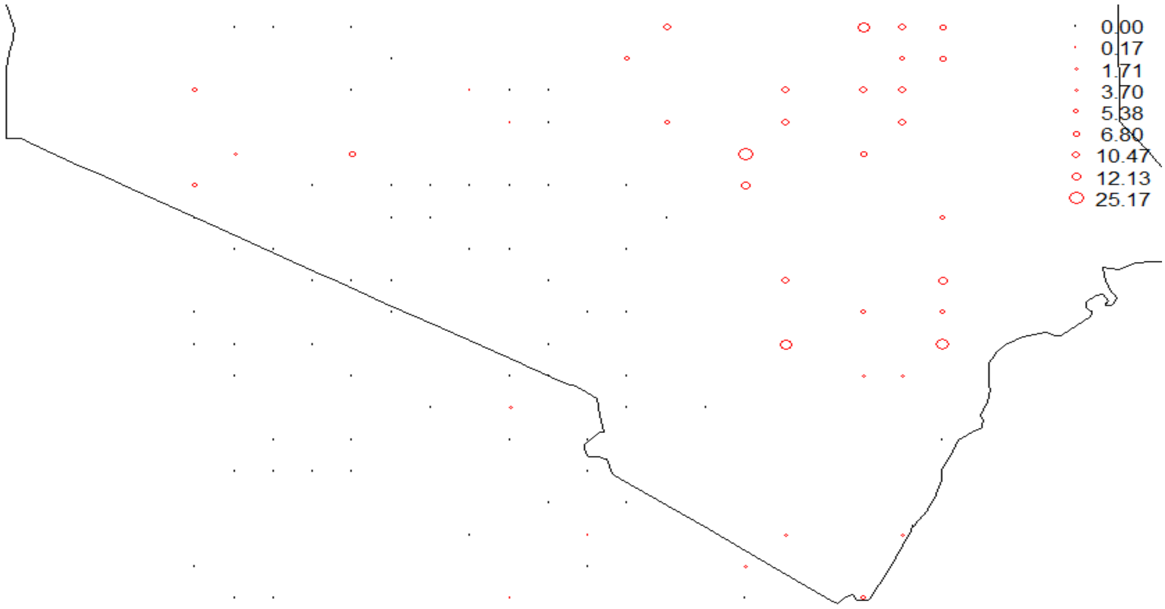
Here is a map of area elevation



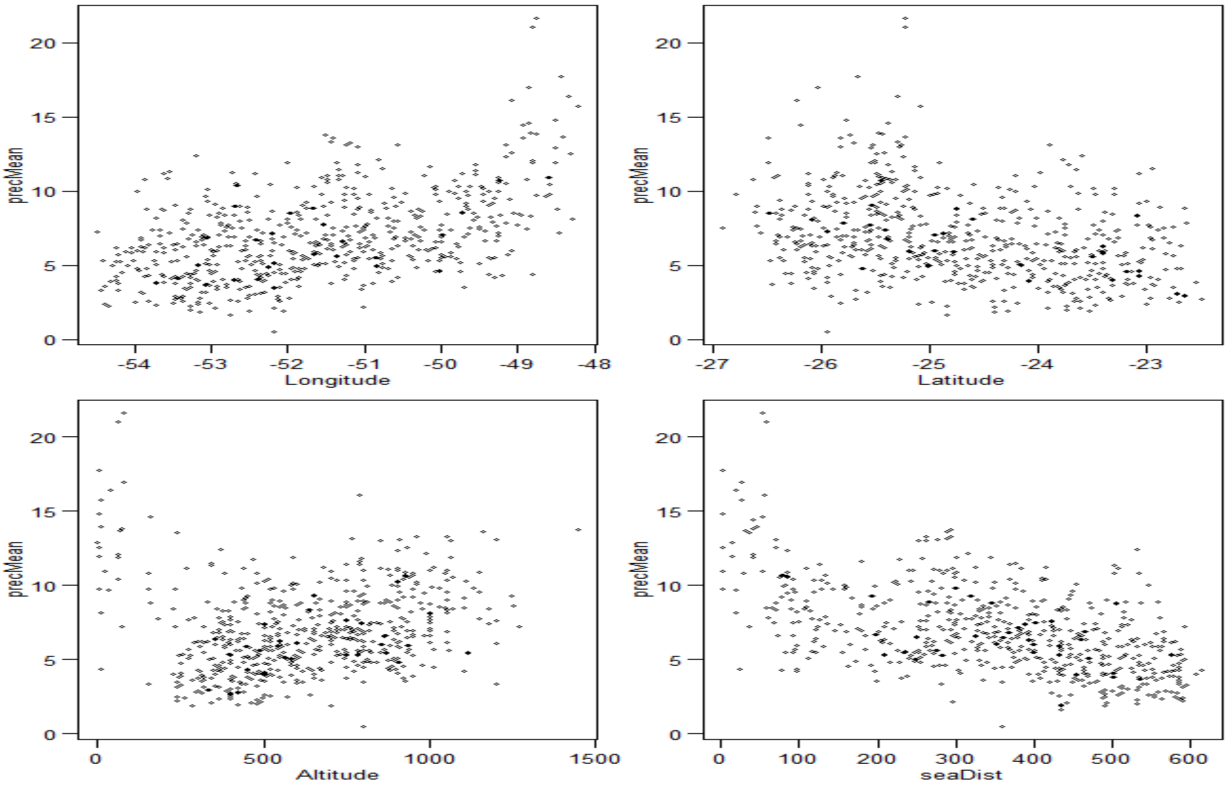
Time series plot of amount of rainfall over time



The mean annual rainfall over the 15 years of study



To make an initial exploration of relationship between the precipitation and the covariates, we visualize some dispersion graphs. After preliminary tests, we see that it is more adequate to construct a new covariate: distance from each station to the Indian Ocean. We found the coordinates of Kenya border that share a frontier with the sea and compute the distance from each station to the neighbor coordinate of this line. We see the dispersion plots in the figure below



From the above graphics, we observe that we do not have a well-defined non-linear relationship with altitude and there is a similar, but inverse, relation with sea distance. Therefore, we build two models, one with elevation as a covariate call it model one and another with distance from the sea as a covariate call it model two.

In addition, compute the Deviance Information Criterion (DIC) to decide what model is more parsimonious.

DIC is a measure of complexity and fit, introduced in Spiegelhalter et al, (2002) and used to compare complex hierarchical models. It is defined as:

$$DIC = \overline{D} + P_D$$

Where $\overline{D}$ the posterior is mean of the deviance of the model and P_D is the effective number of parameters. Smaller DIC values indicate a better trade-off between complexity and fit.

Notice that by specifying both the spatial and temporal effect, we indicate that each time point the spatial locations are linked by the SPDE model object, while across the time, the process evolves according to an ar(1) process.

Models selection

Model	DIC
1	2514.754
2	3011.499

It is true that if we want a model for prediction, smaller models tend to be more generalizable than bigger models. We really must be parsimonious to avoid quark and idiosyncrasies relationships in the model. However, if we were interested in only controlling for confounding, bigger models are always better. In this case, we tend to be more liberal. Therefore, since our one objective is prediction, we chose a model with elevation as the covariate given that it has a better tradeoff between model complexity and the fit (DIC). We retrieve the posterior summary statistics of the fixed effects from the selected model. The posterior marginal of the precision $\tau_c = {}^{1}/_{\sigma^2{}_c}$ is included in the summary statistics. If we are interested in $\sigma^2{}_c$, we employ the function inla.emerginal for computing the expected

6.3 Posterior Summary statistics

Now, we look at the summary of the posterior distribution to $\alpha_z, \alpha_y,\ \beta_{elevation}, \qquad \beta_y$ (Scaling parameter), $Prec_{gamma}, \theta_1 = \log(\tau)\,,\ \theta_2 = \log(k), rho_{ar1}, prec_{ar1}, k$ (spatial scale parameter), $\sigma^2{}_x(variance\ of\ the\ process)\ and$

$practical\ range\ of\ the\ process$

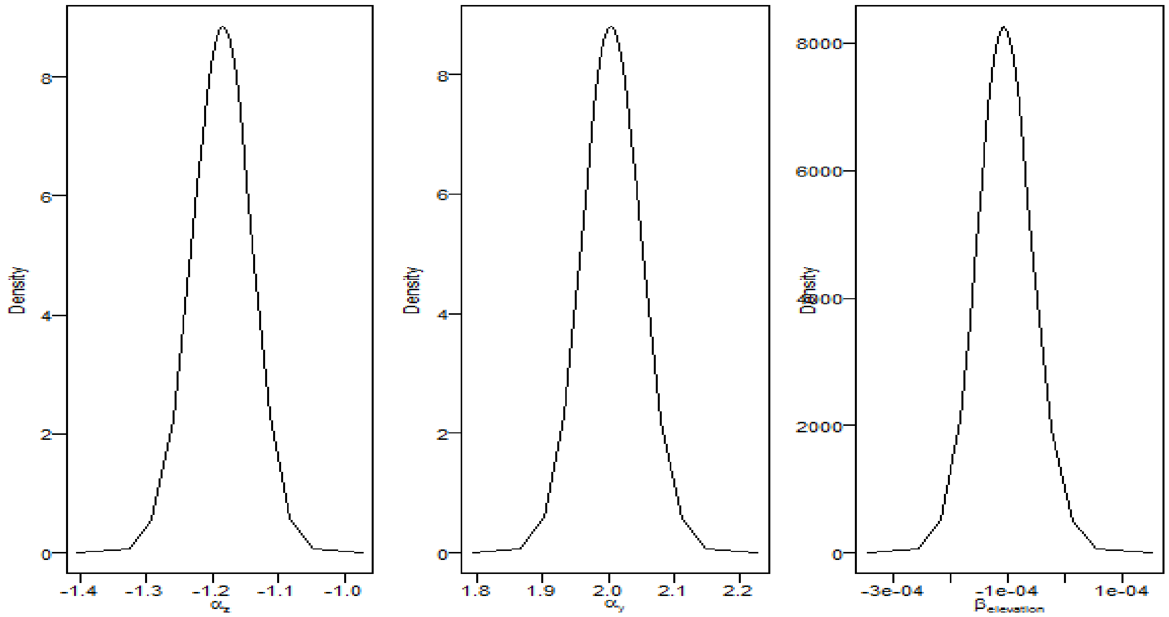
Fixed effects:

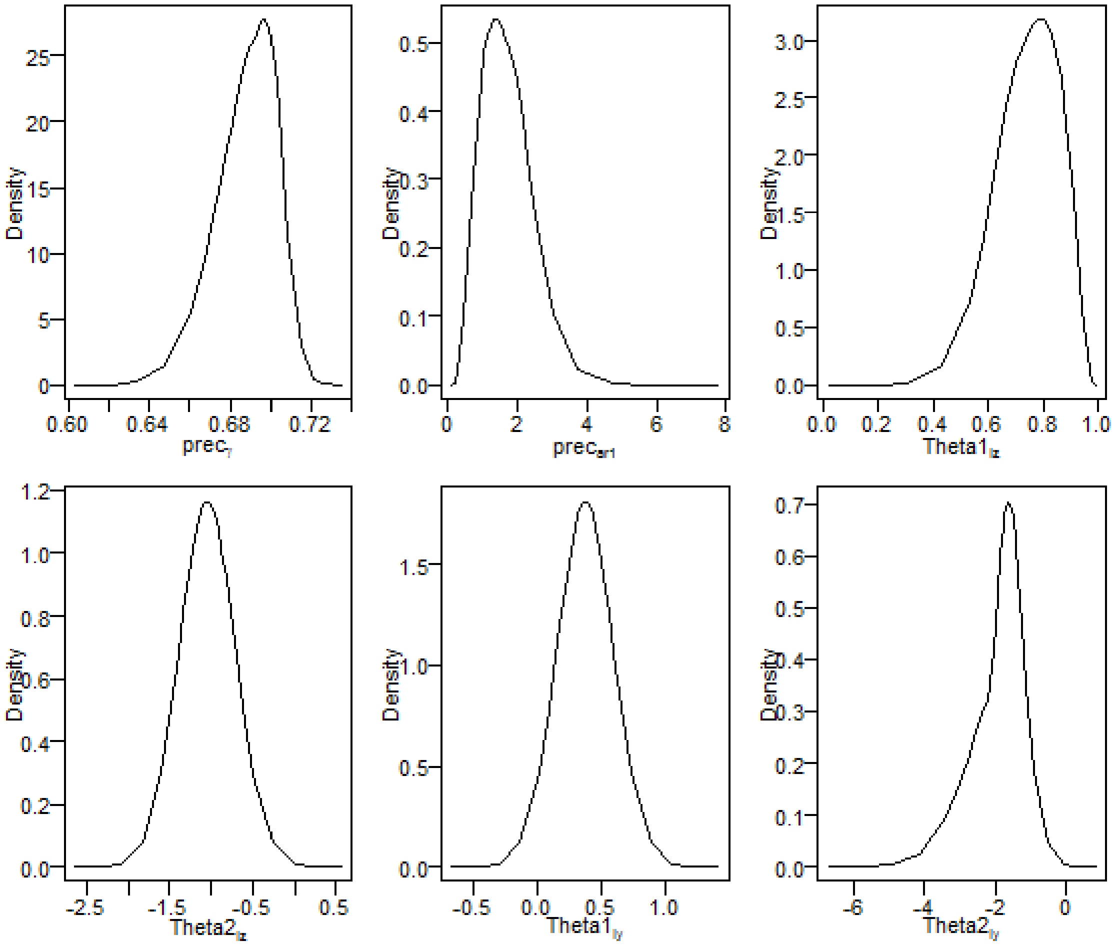
```
              mean     sd 0.025quant 0.5quant 0.975quant   kld
zIntercept -1.1185 0.4165    -1.8812  -1.1323    -0.2511 1e-04
yIntercept  2.0672 0.4165     1.3044   2.0536     2.9347 1e-04
Elevation  -0.0001 0.0000    -0.0002  -0.0001     0.0000 0e+00
```

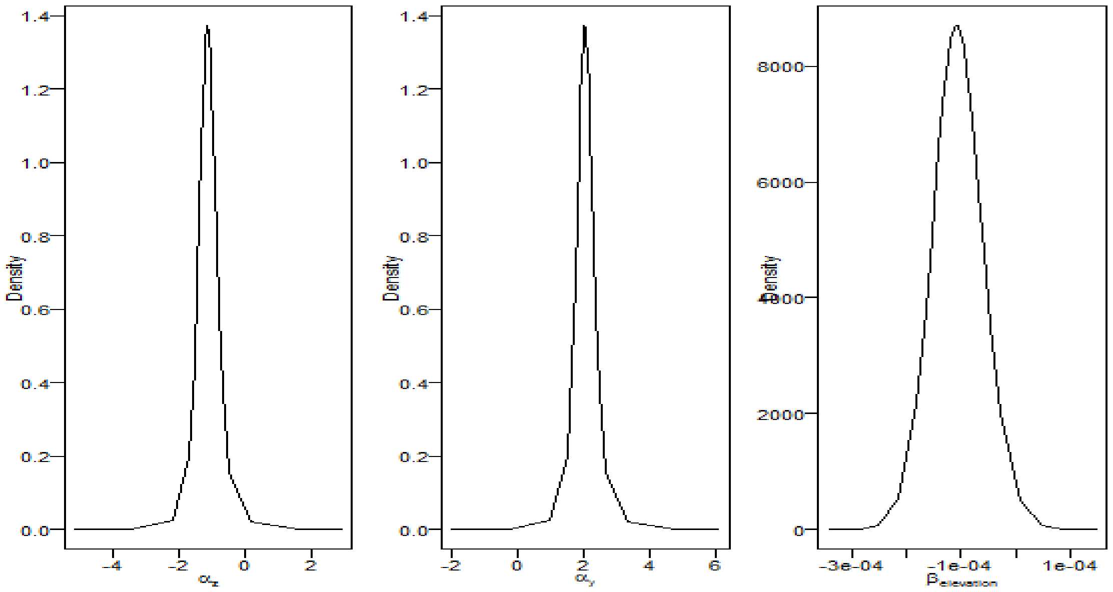
Model hyperparameters:

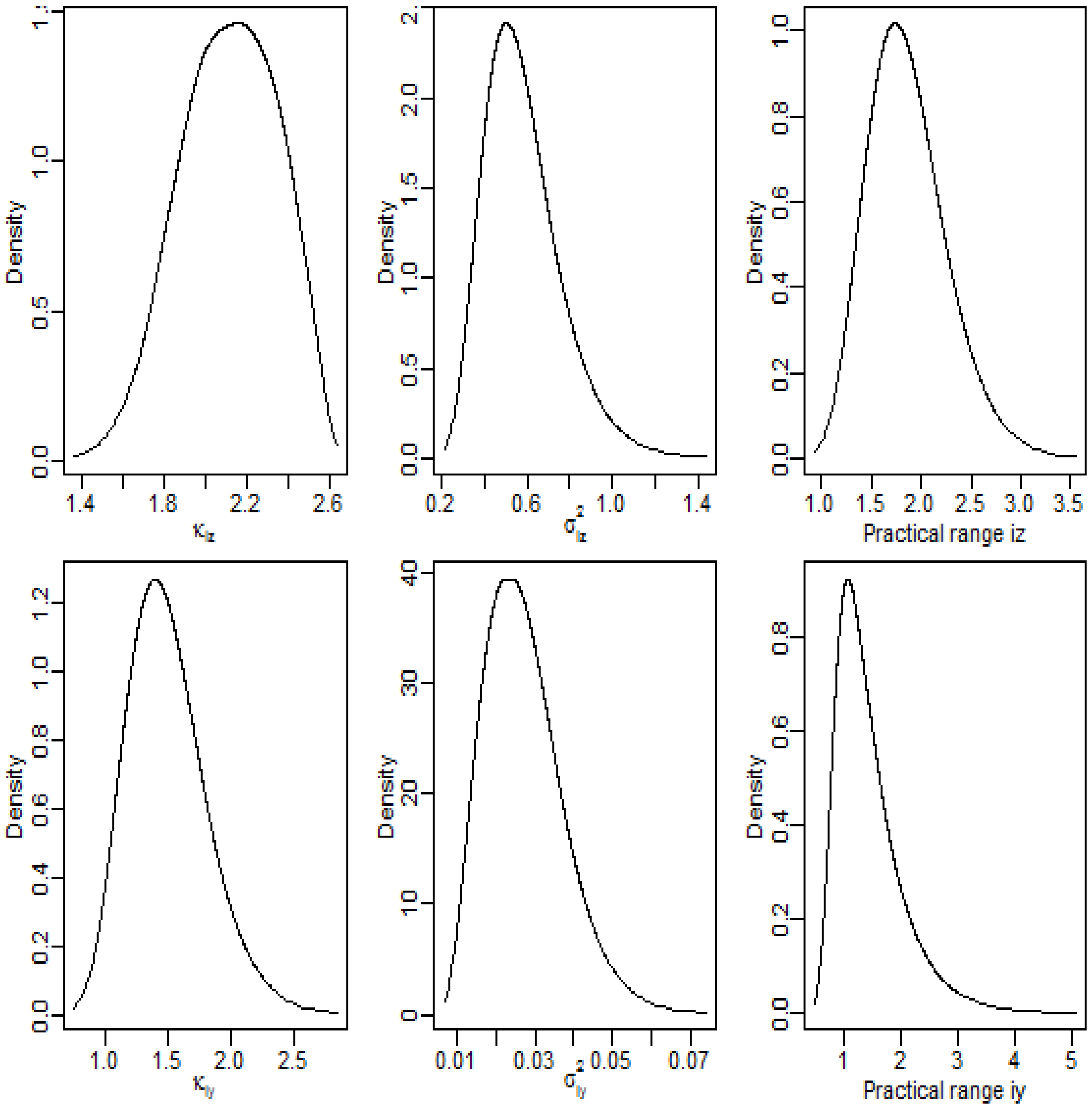
```
                          mean    sd      0.025quant 0.5quant 0.975quant
Precision(gamma)          0.6877  0.0148  0.6547     0.6896     0.7114
Precision for time        1.6952  0.7283  0.5548     1.6132     3.3352
Rho for time              0.7407  0.1163  0.4882     0.7515     0.9270
Theta1 for iz            -1.0484  0.3374 -1.7132    -1.0490    -0.3851
Theta2 for iz             0.3769  0.2166 -0.0513     0.3772     0.8015
Theta1 for iy            -1.9413  0.7558 -3.7339    -1.7918    -0.7535
Theta2 for iy             0.0566  0.3899 -0.6393     0.0293     0.8818
```

6.4 Posterior densities



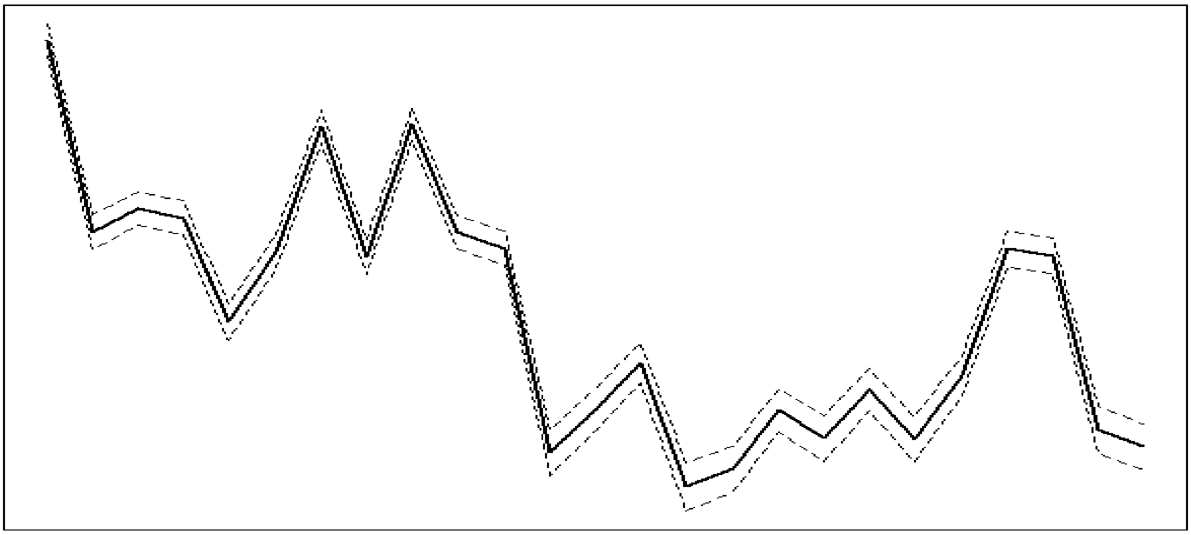

Density
25
20
15
10
5
0
0.60
0.64
0.68
0.72
$prec_y$
Density
0.5
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
$prec_{ar1}$
Density
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
$Theta1_{lz}$
Density
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-2.5
-1.5
-0.5
0.5
$Theta2_{lz}$
Density
1.5
1.0
0.5
0.0
-0.5
0.0
0.5
1.0
$Theta1_{ly}$
Density
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-6
-4
-2
0
$Theta2_{ly}$


Density
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-4
-2
0
2
α_z
Density
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-2
0
2
4
6
α_y
Density
8000
6000
4000
2000
0
-3e-04
-1e-04
1e-04
$\beta_{elevation}$



The posterior mean distribution with 2.5% and 97.5% credibility interval of the amount of rainfall in the country that includes both visited and unvisited stations including non-weather stations for the 365 days in this study is given by;

time



From the above posterior means over time, it is clear that the amount of rainfall is relatively uniform for the first three months of the year. However, around 20th of January tend to be relatively higher compared to the other days under investigation. Nevertheless, from the graph, there exist seasonal effects after every 20 days the amount of rainfall tends to be higher. This might not be generalizable due to the small number of days considered in this study.

As illustrated above, this is an autoregressive process of order one (ar1) with the coefficient given in the posterior means summary result for the hyper parameter for rho as 0.7407 and the posterior mean for precision for rho is 0.1163 whose prior is inverse gamma distributed. In theory, it means that the value of the immediate past is lower than the present value by 0.7407.

So that the ar1 process for the time effect will be given by;

$$x_t = 0.7407\ x_{t-1} + \in_t$$

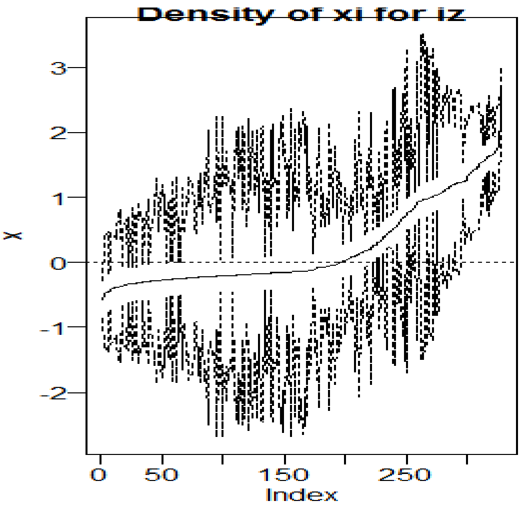
To decide if the occurrence of rain and the amount of rain at a particular point have spatial dependency, we test for the significance of x_i.

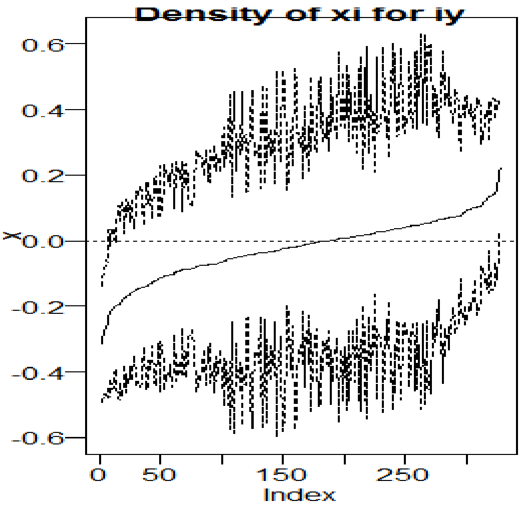
One approach is looking at the 2.5% and 97.5% quartiles of each element of the random effect. If for some, they have quantiles with same signal, then they are significantly none null which can be plotted.

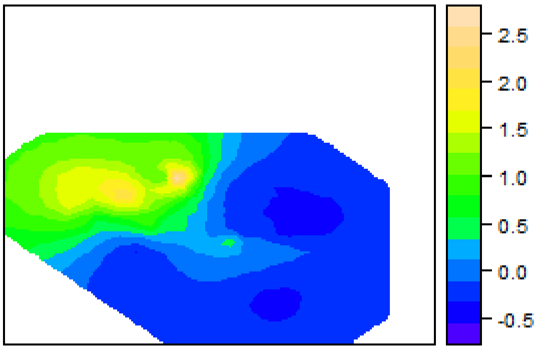
The other approach, we prefer is to compare the DIC of the model with x with the DIC of the model without it. We can also extend this to an additional model with x and other effects excluding the SeaDist just to assess whether excluding distance from the sea might have led to even a better fit. The model with x plus all the other effects will be our full model for the purpose of this work as indicated above in the model fitting procedures at the beginning of this chapter. As shown, the full model excluding distance from the sea was the most parsimonious model. However, the reduced model (reduced model four) without temporal effect was the worst candidate model therefore our model cannot be reliable without temporal effect. We remember, the practical range gives the extent of spatial dependence beyond which the spatial relationship between stations s_i and s_j (where $i\ and\ j\ may\ or\ may\ not\ be\ neighbours$) is negligible.

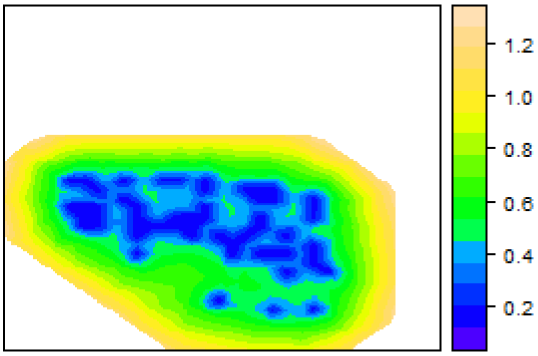
6.5 Posterior densities for the predictive distributions

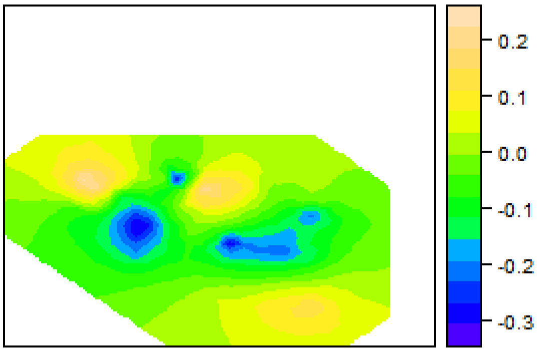
The posterior estimated means are interpolated for both the visited and unvisited sites for the occurrence, its standard deviations and posterior distributions of amount variable as well as its standard deviation respectively.

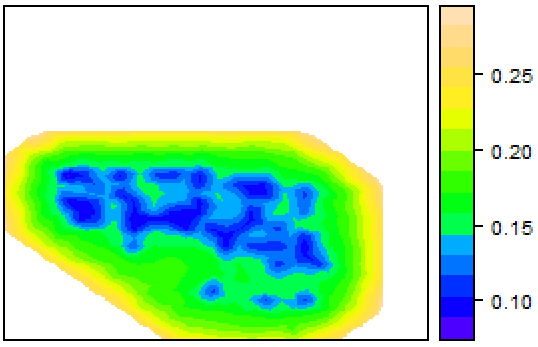












6.6 DISCURSION

On average, the spatial practical range of the occurrence of rainfall is 1.8653km with standard deviation of 0.5818. Similarly, the average spatial practical range of the amount of rainfall is 1.4293km with standard deviation of 0.0272.

The above imply that beyond 1.8653km from a particular point (location) the correlation coefficient between rainfall occurrences is less than 0.13. Similarly, beyond 1.4293km from a particular point (location) the correlation coefficient between their corresponding rainfall amounts is less than 0.13.

On the other hand, the average amount of rainfall at any given time is higher than the amount on the previous day by 0.7407 if indeed it rained on the previous day.

The posterior mean of the amount of rainfall is generally higher towards the northern region which according to the grid is the region slightly below the equator.

There was a dismal occurrence of rainfall towards the Indian Ocean. However, this result is limited to rainfall as experienced on the 200th day, 1998 in Kenya region below the equator.

Generally, amount of rainfall experienced across the spatial domain had a continuous trend though occurrence had a decline on around the 170th index.

CHAPTER SEVEN
CONCLUSIONS AND RECOMMENDATIONS

7.1 CONCLUSIONS

In this project we have provided Bayesian model based approach to geostatistics using Integrated Laplace Approximation (INLA) for spatial and Spatio-temporal models, their general complexity remains, potentially, a fundamental issue within the Bayesian approach.

The INLA approach is in general able to provide reliable estimations in lower computational time than their corresponding traditional MCMC-based estimations. One of the fundamental differences between MCMC and INLA methods is that the former provide (asymptotically) exact inference, while the latter give, by definition, an approximation to the relevant posterior distributions. This particularly is significant especially with large datasets dealing with geostatistical inferences. The use of SPDE algorithms produce massive savings in computational times and allow the user to work with relatively complex models as we have seen above with semi-continuous model in an efficient way.

Because of its recent inception, INLA is less established than MCMC-methods. Consequently, its development is still on-going particularly in regard to more advanced features that frequent users require.

INLA framework has become a daily tool for many applied researchers from different areas of application ranging from (generalized) linear mixed to spatial and spatio-temporal models. Combined with the Stochastic Partial Differential Equation approach (SPDE, Lindgren et al, 2011), one can easily accommodate all kinds of geographically referenced data, including areal and geostatistical ones, as well as spatial point process data and epidemiological related studies.

Because of its inception, INLA is less established than MCMC methods (although we acknowledge a flurry of activity in the development of new MCMC algorithms, e.g Girolami and Calderhead, 2011; and Hoffman and Gelman, 2011). Consequently, its development is still ongoing, particularly with respect to some more advanced features (e.g. the SPDE approaches) (Lindgren & Havard, 2013).

Moreover, Bayesian inference is simple in principle and provides a single recipe for coherent inference, all based on the posterior distribution. Inference is conditional on the observed, and not on data that were possible but not observed, obeying the likelihood principle. In addition, it tells how to update prior beliefs and how to add additional information. Uncertainty is naturally framed as probability statements based on the posterior in a way that non-statisticians easily relate to (what else could 'statistical inference about θ' mean?) With Bayesian framework, complicated hierarchical models can be naturally constructed. In addition, Bayesian results often have good frequentist properties and frequentist inference is sometimes a special case of Bayesian results under a particular prior.

7.3 RECOMMENDATIONS FOR FURTHER RESEARCH

7.3.1 Joint modeling a covariate with misalignment

Here we wish to focus on a situation when we have a response y and a covariate c. But, we have misalignment, i.e.; we have y observed at n_y locations and c observed at n_c locations. We design a solution that not depends if we have or not some common observed locations for $c\ and\ y$.

A restriction is the assumption that c have spatial dependency. This restriction is made because we want a good model to predict c at locations of y. So, the goodness of prediction is proportional to the spatial dependency.

Taking into account that c have spatial dependency, a simple approach is to define a model for c, predict it on locations that we have y and build a model for y. But, in this two stage model, we don't take into account the prediction error of c on the second model. The measurement error model is an approach to solve similar problems, (Muff et al, 2013). But, here we are able to consider the spatial dependency on c. So, we build a spatial model for c and another spatial model for y, and do the estimation process jointly.

Taking an example of effect of a covariate c (amount of rainfall) on maize production. We know that the amount of rainfall has spatial dependency and since there are days when it rains and others when it doesn't rain at certain points we develop approach of model building for covariate first that has been described in this project. So we must build a spatial model for rainfall first and predict it onto points (farms) so as to enable us build another model for y (maize yield) taking into account both occurrence and amount of rainfall as described.

We consider semi-continuous likelihood (binomial for occurrence and gamma for amount) and build the model. Having predicted amount of rainfall at required points (available farms), we build the second model;

Let the following model for y; $y_i \sim N(\alpha_y + \beta c_i + x_i,\ \sigma^2{}_y)$

Where α_y is an intercept, β is the regression coefficient on c, c_i is the predicted covariate at locations y_i, x_i is an zero mean random field and $\sigma^2{}_y$ measures the error that remain unexplained on Y.

A particular case is when we don't have the x term in the model for y. Another case, is when $\sigma^2{}_c = 0$ and we don't have white noise in the covariate i.e. covariate is considered just a realization of a random field.

REFERENCE

Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology*, *4*, 33–49.

Cameletti, M., Lindgren, F., Simpson, D., & Rue, H. (2012). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, *97*(2), 109–131.

Congdon, P. (2006). *Book. Bayesian Statistical Modelling*.

Cressie, N. (1993). *Book.* Statistics for Spatial Data.

Diggle, P. J., Menezes, R., & Su, T. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *59*(2), 191–232.

Diggle, P., & Ribeiro. (2007). *Book. Model Based Geostatistics*.

Hengl, T. (2009). *Book. A Practical Guide to Geostatistical Mapping*.

Ioannidis, J. P. A. (2005). Why most published research findings are false. (W. Jantsch & F. Schaffler, Eds.) *PLoS Medicine*, *2*(8), e124.

Kruschke, J. K. (2010). *Book. Doing Bayesian Data Analysis : A Tutorial with R and BUGS*.

Lindgren, F., & Havard, R. (2013). Journal of Statistical Software with R-INLA. *Bayesian Spatial and Spatio-temporal Modelling with R-INLA*, *VV*(Ii).

Lindgren, F., & Rue, H. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields : the stochastic, 423–498.

Lindgren, F., & Rue, H. (2013). Bayesian Spatial and Spatio-temporal Modelling with R-INLA, (Section 1).

Martins, T. G., Simpson, D., Lindgren, F., & Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, *67*, 68–83.

Ntzoufras, I. (2009). *Book. Bayesian Modeling using WinBugs*.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(2), 319–392.

Rue, Simpson, D., & Lindgren, F. (2013). Bayesian computing with INLA : new features, (2009), 1–29.

Schrödle, B., & Held, L. (2011). Spatio-temporal disease mapping using INLA. *Environmetrics*, *22*(6), 725–734.

Sharon Bertsch. (2010). *Book. The Theory That will Not Die*.

Simpson, D., & Lindgren, F. (2011). Think continuous : Markovian Gaussian models in spatial statistics by PREPRINT NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY.

Simpson, D., & Lindgren, F. (2012). Bayesian computing with INLA : new features, (2009), 1–27. PREPRINT NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY.

Tijms, H. C. (2003). *Book. A First Course in Stochastic Models*. Chichester, UK: John Wiley & Sons, Ltd.

Yu, H., Xu, J., Okuto, E., & Luedeling, E. (2012). Seasonal Response of Grasslands to Climate Change on the Tibetan Plateau, *7*(11). doi:10.1371/journal.pone.0049230