# Bayesian Notes for building the geostatistical MANOVA-KNN pipeline

Roxana Tesileanu

roxana.te@web.de
INCDS, Romania

November 2017

## Contents

# 1 Introduction

From Gelman et al. 2014

BAYESIAN INFERENCE is the process of fitting a probability model to a set of data and SUMMARIZING THE RESULT BY A PROBABILITY DISTRIBUTION ON:

1. THE PARAMETERS OF THE MODEL and on
2. THE UNOBSERVED QUANTITIES SUCH AS PREDICTIONS FOR NEW OBSERVATIONS.

=> make inferences from data using probability models for quantities we observe and for quantities we wish to learn. **THE ESSENTIAL CHARACTERISTIC OF BAYESIAN MODELS IS THEIR EXPLICIT USE OF PROBABILITY FOR QUATIFYING UNCERTAINTY IN INFERENCES BASED ON STATISTICAL DATA ANALYSIS**. This is the main idea of the MANOVA-KNN pipeline, to analyse errors and reduce them. I've already sketched a geometric approach. It needs to be better backed up by probability theory. Which is what I expect to find in these Bayesian texts.

Steps of Bayesian Data Analysis:

1. setting up A FULL PROBABILITY MODEL - a JOINT PROBABILITY DISTRIBUTION FOR ALL OBSERVABLE AND UNOBSERVABLE QUANTITIES IN A PROBLEM.
2. CONDITIONING ON OBSERVED DATA - calculating and interpreting the appropriate POSTERIOR DISTRIBUTION (which is the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data).
3. EVALUATING THE FIT OF THE MODEL AND THE IMPLICATIONS OF THE RESULTING POSTERIOR DISTRIBUTION: how well does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modeling assumptions in step 1? **IN RESPONSE, ONE CAN ALTER OR EXPAND THE MODEL AND REPEAT THE THREE STEPS**.

Advances in carrying the third step alleviate to some degree the need to assume correct model specification at the first attempt. In particular, the much-feared dependence of conclusions on "subjective" prior distributions can be examined and explored. A primary motivation of Bayesian thinking is that it facilitates a COMMON-SENSE INTERPRETATION OF STATISTICAL CONCLUSIONS. Also, the Bayesian paradigm provides a conceptually simple method for **coping with multiple paramerers**.

## 1.1 Notation for statistical inference

**Parameters, data and predictions**

- let $\theta$ denote unobservable vector quantities or POPULATION PARAMETERS OF INTEREST.
- $y$ denotes the observed data
- $\tilde{y}$ denotes unknown, but potentially observable, quantities

In general these symbols represent MULTIVARIATE QUANTITIES. When using matrix notation, we consider vectors as column vectors throughout. Data are gathered on each of a set of n objects or units (like in classical multivariante analyses). And we can write the data as a vector: $y = (y_1, ..., y_n)$. If several variables are measured on each unit, then each $y_i$ is a vector. **The entire dataset is a matrix with n rows**. The y variables are called the **"outcomes" and are considered "random"** in the sense that, when making inferences, we wish to allow for the possibility that the observed values of the variables could have turned out otherwise, due to the sampling process and the natural variation of the population. We commonly model data from an exchangeable distribution as independently and identically distributed (iid) given some unknown parameter vector $\theta$ with distribution $p(\theta)$. The explanatory variables (covariates) are denoted with $x$. We use $X$ to denote the entire set of explanatory variables for all $n$ units. If there are $k$ explanatory variables, then $X$ is a matrix with $n$ rows and $k$ columns (similar to conventional notation from classical multivariate statistics). $X$ is treated as random.

**Bayesian inference** Bayesian STATISTICAL CONCLUSIONS about a parameter $\theta$, or unobserved data $\tilde{y}$, are made in terms of PROBABILITY STATEMENTS. These probability statements are conditional on the observed value of $y$, and in our notation they are simply written as: $p(\theta|y)$ or $p(\tilde{y}|y)$. This is also written as $p(\cdot|\cdot)$. For a marginal distribution we use $p(\cdot)$.

We may use the notation $Pr(\cdot)$ for the probability of an event (for example in $Pr(\theta > 2) = \int_{\theta>2} p(\theta)\mathrm{d}\theta$). Also the probability density functions are denoted like for example: $\theta \sim N(\mu, \sigma^2)$ meaning that $\theta$ has a Normal distribution with mean $\mu$ and variance $\sigma^2$. Equiv. we can write $p(\theta|\mu, \sigma^2) = N(\theta|\mu, \sigma^2)$.

## 1.2 Bayes' Rule

In order to make probability statements about $\theta$ given $y$, we must begin with a model providing a JOINT PROBABILITY DISTRIBUTION FOR $\theta$ and $y$. The joint probability mass or density function can be written as a PRODUCT OF TWO DENSITIES that are often referred to as the PRIOR DISTRIBUTION $p(\theta)$ and the SAMPLING DISTRIBUTION (or DATA DISTRIBUTION) $p(y|\theta)$:

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

Simply CONDITIONING ON THE KNOWN DATA, using the Bayes' Rule yields the POSTERIOR DENSITY:

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

where $p(y) = \int p(\theta)p(y|\theta)\mathrm{d}\theta$ (for $\theta$ continuous), and $1/p(y)$ represents the constant of proportionality ensuring that the posterior density integrates to 1 as a proper probability density must [2]. The constant of proportionality can be formally written as [2]:

$$[\int p(y|\theta)p(\theta)\mathrm{d}\theta]^{-1}$$

Leaving the $1/p(y)$ off, we get the UNNORMALIZED POSTERIOR DENSITY which is the right side of:

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

We can thus state that the **POSTERIOR IS PROPORTIONAL TO THE PRIOR TIMES THE LIKELIHOOD** [2].

**Prediction** Make inferences about an unknown observable (predictive inferences). Before the data are considered, the distribution of the unknown but observable $y$ is:

$$p(y) = \int p(y, \theta)\mathrm{d}\theta = \int p(\theta)p(y|\theta)\mathrm{d}\theta$$

This is called the **marginal distribution of y** and it is also called the **prior predictive distribution**. After the data $y$ have been observed, we can predict an unknown observable, $\tilde{y}$, from the same process. The distribution of $\tilde{y}$ is called the **posterior predictive distribution**. It is CONDITIONAL ON THE OBSERVED $y$:

$$p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y)\mathrm{d}\theta = \int p(\tilde{y}|\theta, y)p(\theta|y)\mathrm{d}\theta = \int p(\tilde{y}|\theta)p(\theta|y)\mathrm{d}\theta$$

**Likelihood** Using Bayes' Rule WITH A CHOSEN PROBABILITY MODEL means that the data $y$ affect the posterior inference only through $p(y|\theta)$, which is called the **likelihood function**.

**Example for a single-parameter problem** From Jackman 2009.

Suppose we have the success probability $\theta \in [0, 1]$ underlying a binomial process. There might be more combinations of a prior, a likelihood and a posterior distribution. If we use a uniform prior, $\theta \sim Unif(0, 1)$, the information of the prior is "absorbed" into the constant of proportionality, i.e. is an uninformative prior, resulting in a posterior density over $\theta$ that is proportional to the likelihood. **The Uniform prior is used when we have no prior information about**

**the value of $\theta$ and hence no way to a priori prefer one set of values for $\theta$ over any other**. Also note that likelihood based analyses of data (so with a uniform prior) assume prior ignorance (which must be plausible). **Usually, Bayesian inference works with more or less informative priors for $\theta$.** In these cases, the mean of the posterior distribution is a **precision-weighted average** of the prior and the likelihood, i.e. it reduces the variance observed in the prior and likelihood. This is one of the consequences of working with so-called **conjugate priors** in the exponential family.

**Conjugate priors** Priors represent convenient ways of mathematically expressing prior beliefs over parameters. Many statistical models are in the exponential family (but not all), for which conjugate priors are convenient ways of mathematically and computationally quite simple.

*Definition: Suppose a prior density $p(\theta)$ belongs to a class of parametric densities F. Then the prior density is said to be conjugate with respect to a likelihood $p(y|\theta)$ if the posterior density $p(\theta|y)$ is also in F.* Where the definition of a "class of parametric densities" will be made clear along the way.

An example is the use of a Beta prior with Binomial data (in the beta-binomial model). For the likelihood function formed with binomial data, any Beta density over $\theta$ is a conjugate prior. The idea of the beta-binomial model is that, if prior beliefs about $\theta$ can be represented as a Beta density, then after those beliefs have been updated (via Bayes Rule) in light of binomial data, posterior beliefs about $\theta$ are also characterized by a Beta density.

When **CONJUGACY HOLDS**, Bayesian analysis with conjugate priors is equivalent to COMBINING INFORMATION FROM THE PRIOR WITH THE INFORMATION IN THE DATA **with the relative contributions of prior and data to the posterior being proportional to their repective precisions**. Again, Bayesian analysis with conjugate priors over a parameter $\theta$ is equivalent to taking a precision-weighted average of prior information about $\theta$ and the information in the data about $\theta$. Thus, when prior beliefs about $\theta$ are "vague", "diffuse" or in the extreme case "uninformative", the posterior will be dominated by the likelihood. In the extreme case of an uninformative prior, the only information about the parameter is the information in the data, and the posterior has the same shape as the likelihood function. When prior information is available, the posterior incorporates it rationally, in the sense of being consistent with the laws of probability via Bayes Theorem. In fact, WHEN PRIOR BELIEFS ARE QUITE PRECISE RELATIVE TO THE DATA, IT IS POSSIBLE THAT THE LIKELIHOOD IS LARGELY IGNORED, AND THE POSTERIOR DISTRIBUTION WILL LOOK ALMOST EXACTLY LIKE THE PRIOR. Very dangerous when you have false prior information. In the limiting case of a degenerate, infinitely-precise, "spike prior" (all probability concentrated on one point), the data are completely ignored, and the posterior is also a degenerate "spike" distribution. Should you hold such a dogmatic prior, no amount of data will ever result in you changing your mind about the issue.

**Bayesian Updating as Information Accumulation** Bayesian procedures accumulate information in the sense that the posterior distribution will through repeated applications of the data generation process eventually modify its precision, eventually overwhelming any non-degenerate prior. This means that different beliefs will eventually coincide provided you see enough data and you update your beliefs using Bayes' Theorem.

**Probability as a measure of uncertainty** From Gelman et al. 2014

Since the uses of probability within a Bayesian framework are much broader than within non-Bayesian statistics, the features of probabilities are briefly presented within this paragraph. The mathematical definition of probabilities is that they are NUMERICAL QUANTITIES, DEFINED ON A SET OF OUTCOMES, ARE NONNEGATIVE, ARE ADDITIVE OVER MUTUALLY EXCLUSIVE OUTCOMES, AND SUM TO 1 OVER ALL POSSIBLE MUTUALLY EXCLUSIVE OUTCOMES.
! You don't have negative probabilities !
! A probability can be between 0 and 1, not greater than 1 !
! All probabilities defined on a set of outcomes sum to 1 !
In Bayesian statistics, probability is used as the fundamental measure of uncertainty. Bayesian methods enable statements about how sure we are concerning some situation. The guiding principle is that the state of knowledge (i.e. how sure we are) about anything unknown is described by a probability distribution. Why is probability a reasonable way of quantifying uncertainty? Uncertainty is related to RANDOMNESS, and randomness relates to random events characterized as "probably" or "unlikely".

# 2 Single-parameter models

## 2.1 Estimating a probability from binomial data

In the simple beta-binomial model, the aim is to estimate an unknown population proportion from the results of a sequence of Bernoulli trials, that is, data **y** is either 0 (failure) or 1 (success). Because of the exchangeability, the data can be summarized by the total number of successes denoted by $y$ in the $n$ trials. The binomial sampling model is:

$$p(y|\theta) = Binomial(y|n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

Example: estimation of the sex ratio within a population of human births. Let $y$ be the number of girls in $n$ recorded births. To perform Bayesian inference in the binomial model, we must specify a prior distribution for $\theta$. For the beginning, assume that the prior distribution for $\theta$ is uniform in the interval [0,1]. Elementary application of Bayes' Rule gives the posterior density for $\theta$ as:

$$p(\theta|y) \propto \theta^y (1-\theta)^{n-y}$$

The binomial coefficient, does not depend on the unknown parameter $\theta$ and so it can be treated as a constant when calculating the posterior density of $\theta$. We can recognize the kernel of a Beta density. The unnormalized posterior density for binomial parameter $\theta$ (the Beta kernel) becomes more precise with increasing $n$ for the same proportion of successes. We conventionally write the result (i.e. the posterior density) as:

$$p(\theta|y) \sim Beta(y+1, n-y+1)$$

In order to describe the posterior density, commonly used summaries of location are the mean, median, and mode(s) of the distribution; variation is commonly summarized by the standard deviation, the interquartile range and other quantiles.

**Informative prior distributions** The prior distribution can be seen as representing a population of possible parameter values, from which the $\theta$ of current interest has been drawn. Sometimes we don't have such a population, and in this case, we must express our knowledge (and uncertainty i.e. probability) about $\theta$ and should include all plausible values of $\theta$. Beta densities are conjugate priors with respect to binomial likelihood. The Beta density takes two shape parameters as its arguments ($\alpha$, $\beta$). It is confined to the unit interval. (The Uniform density is a special case of Beta density arising when $\alpha = \beta = 1$.) The Beta kernel is $\theta^{\alpha-1}(1-\theta)^{\beta-1}$. So the prior is: $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

From Jackman 2009:
**Conjugacy of Beta prior, Binomial data: Given a Binomial likelihood over r successes in n Bernoulli trials, each independent conditional on an unknown success parameter $\theta \in [0,1]$ then the prior density $p(\theta) = Beta(\alpha, \beta)$ is conjugate with respect to the binomial likelihood, generating the posterior density $p(\theta|r,n) = Beta(\alpha + r, \beta + n - r)$.**

To obtain the posterior Beta according to Bayes' Rule (posterior $\propto$ likelihood x prior):

$$p(\theta|y) \propto \theta^r (1-\theta)^{n-r} \theta^{\alpha-1} (1-\theta)^{\beta-1} = \theta^{r+\alpha-1} (1-\theta)^{n-1+\beta-1} = Beta(\alpha+r, \beta+n-r)$$

Once again regarding the conjugate Beta prior, $p(\theta; \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$, the interpretation of its parameters in data equivalent terms makes it possible to specify conjugate priors for PROBABILITY, RATES AND PROPORTIONS. The interpretations is:

- we have $\alpha - 1$ successes

- we have $\beta - 1$ failures
- we have $\alpha + \beta - 2$ total trials (n)

For example, suppose you believe in 50 trials, you get 12 successes and 38 failures. Then the conjugate Beta prior has the $\alpha = 13$ and $\beta = 39$.

## Note

This document is "under construction". It contains older notes of mine on Bayesian data analysis. Some were used in technical reports of mine (see https://www.researchgate.net/publication/317549069_poisson_model) and also new sections aiming at creating the background necessary for the implementation of the MANOVA-KNN pipeline in geostatistics using the idea of **posterior predictive checks** (Introduction and Deduction in Bayesian Data Analysis, Andrew Gelman, 2011) [1]. For this purpose, I will have to work through books building up my skills, fortunately I was given a hint (and a copy) by a friend on "Bayesian Data Analysis for Social Sciences" by Simon Jackman (Wiley, 2009) [2] and "Bayesian Data Analysis" by Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari and Donald B. Rubin (CRC, 2014) [3]. Please download the current version from my GitHub profile under the multivariate_analyses project repository: https://github.com/RoxanaTesileanu/multivariate_analyses/blob/master/literature_analysis/geospatial_scala/bayesian_notes_geosp.pdf.

The statistical plots in this document were generated in Scala using the JavaPlot package developed by Panayotis Katsaloulis [4]. You can find the scala source files used for generating them under the link: https://github.com/RoxanaTesileanu/multivariate_analyses/tree/master/DeepLearning/src/main/scala/com/mai/scalaPlot.

The present document was edited using Latex [5] (https://www.latex-project.org/). The source .tex file of the present document is also available in the multivariate_analyses repository on my GitHub profile. Special thanks to Gustavo Mezzetti for the Latex halloweenmath package: http://mirrors.concertpass.com/tex-archive/macros/latex/contrib/halloweenmath/halloweenmath-man.pdf!

## References

[1] A. Gelman, “Introduction and Deduction in Bayesian Data Analysis,” *Rationality, Markets and Morals Journal*, vol. 2, pp. 67–78, 2011.

[2] S. Jackman, *Bayesian analyses for social sciences*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, 2009.

[3] A. Gelman, J. B. Carlin, H. S. Stern, Dunson, David B., A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. CRC Press, 2014.

[4] P. Katsaloulis, “JavaPlot,” 2017. [Online]. Available: http://javaplot.panayotis.com/

[5] C. Vellage, “LaTeX-Tutorial.com.” [Online]. Available: https://www.latex-tutorial.com/