Introduction to Bayesian (geo)-statistical modelling

D G Rossiter

Cornell University, Soil & Crop Sciences Section

March 17, 2020

Background

- Bayes' 1763 paper [2]: theory of inverse probability in order to make probabilistic statements about the future
  - A simple use of conditional probability: "Bayes' Rule"
  - Later extended to statistical distributions: "Bayesian" = "Bayes-like"
- Focus is on **decision-making under uncertainty**
- A useful way of thinking about probability.
- An increasingly common way of making inferences, because of its flexibility
  - Can handle arbitrarily complex models, e.g., hierarchical
  - Modern computing methods make this accessible

History

- **Frequentist**
  - R A Fisher at Rothamstead Experimental Station (England), 1920's and 1930's
  - developed by well-known workers (Yates, Snedecor, Cochran . . .)
  - Common statistical computing packages follow this
- **Bayesian**
  - named for Thomas Bayes (1701–1761)
  - developed since the 1960's (Jeffreys, de Finetti, Wald, Savage, Lindley . . .)
  - requires sophisticated computing and complex mathematics

Principal differences

- Interpretation of the meaning of **probability**
- Hypothesis testing
- Prediction
- Presentation of probabilistic results
  - e.g. confidence intervals vs. credible intervals
- Computational methods

What is probability?

Frequentist the probability of an outcome is the proportion of experiments in which the outcome occurs, in some hypothetical repetitions of the experiment under the same conditions and with the same population

Bayesian subjective belief in the probability of an outcome, consistent with some axioms

In both cases, experiments/observations of a sample are used for inference.

Bayesian concept of probability

- the *degree of rational belief* that something is true;
  - so certain rules of *consistency* must be followed
- All probability is *conditional* on evidence;
- Any statement has a *probability distribution*;
- any value of a parameter has a defined probability;
- Probability is continuously *updated* in view of new evidence.
- So, there is a degree of *subjectivity*; but this is reduced as more *evidence* is accumulated.

Types of probability

- *Prior* probability: before observations are made, with previous knowledge;
- *Posterior* probability: after observations are made, using this new information;
- *Unconditional* probability: not taking into account other events, other than general knowledge and agreed-on facts;
- *Joint* probability: of two or more event(s);
- *Conditional* probability: in light of other information, i.e., some other event(s) that may affect it.

Bayesian thinking about statistical distributions

- Parameters of statistical distributions are *random variables*, i.e., they also have their own statistical distributions, which in turn have parameters, often called *hyperparameters*
- Statistical inferences are based on a *posterior* ("after the fact") distribution of parameters of statistical distributions
- These are updated versions of *prior* ("before the fact") beliefs based on data from experiments or observations.
- The updating depends on the *likelihood* of each possible value of the parameters, given the data actually observed.

Subjectivity in Bayesian thinking

- It is required to have *prior* probability distributions, set by the analyst
- "Solution": *non-informative* (actually, "minimum prior information") priors
- But do we want these? In most situations we have prior evidence to incorporate in the decision-making.
- The selection of **model form** in both Bayesian and classical approaches is subjective
  - although the fit of the model form to the data can be compared (internal evaluation).

Bayes' Rule (1)

- One aspect of Bayesian computation is not controversial: Bayes' Rule derived from the definition of *conditional probability*.
- $P(A), P(B)$ *unconditional* probability of two events
- *Joint probability* $P(A \cap B)$ of two events A and B, i.e., that both occur.
- Reformulated in terms of *conditional probability*, i.e., that one event occurs conditional on the other having occurred:

$$P(A \cap B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A) \tag{1}$$

where | indicates that the event on the left is conditional on the event on the right.

Bayes' Rule (2)

- Equating the two right-hand sides and rearranging gives Bayes' Rule:

$$P(A \mid B) = P(A) \cdot \frac{P(B \mid A)}{P(B)} \tag{2}$$

or

$$P(B \mid A) = P(B) \cdot \frac{P(A \mid B)}{P(A)} \tag{3}$$

- $P(B \mid A)/P(B)$, $P(A \mid B)/P(A)$ are *likelihood ratios* – the additional strength of evidence

Reformulation

The denominator $P(B)$ can also be written as the sum of the two mutually-exclusive intersection probabilities, one if event A occurs $P(A)$ and one where it does not occur $P(\neg A)$:

$$P(B) = P(B \mid A) \cdot P(A) + P(B \mid \neg A) \cdot P(\neg A) \tag{4}$$

We rename the probabilities to correspond to the concept of an observed "event" E and an unobserved or unknowable event for which we want to estimate the probability H ("hypothesis").

Bayes' Rule for the binary case then can be written:

$$P(H \mid E) = P(H) \cdot \frac{P(E \mid H)}{P(E \mid H) \cdot P(H) + P(E \mid \neg H) \cdot P(\neg H)} \tag{5}$$

Example – land cover classification (1)

- $P(H)$ the probability that a pixel in the image covers an area of water
- $P(E)$ pixel NDVI is below a certain threshold, say 0.1
- $P(H|E)$ the probability that, given that a pixel's NDVI is below the threshold, it covers water
  - this is what we want to know
- $P(H \cap E)$: the probability of a pixel in the image covers water *and* its NDVI is below the threshold
- $P(H \cap \neg E)$: the probability of a pixel in the image covers water, but its NDVI is *not* below the threshold
  - water body contains many aquatic plants, specular reflection ...
- $P(E|H)$ the probability that, given that a pixel covers water, its NDVI is below the threshold
- $P(E|\neg H)$ the probability that, given that a pixel covers water, its NDVI is *not* below the threshold

Example (2)

- We want to classify the image into water/non-water: *hypothesis H* is that the area represented by a pixel is in fact mostly covered by water
- We have a training sample with some pixels in each class
- For each of these, we compute the NDVI of the pixel, from the imagery: *event E* that we can observe is that a pixel's NDVI < 0.1.
- $P(H)$ is the prior probability that a random pixel area mostly covers water
  - proportion from training sample or prior estimate
- $P(E \mid H)$ if a pixel really does cover water, what is the conditional probability it will have a low NDVI: *sensitivity*
- $P(E \mid \neg H)$: false positives, inverse of *specificity*

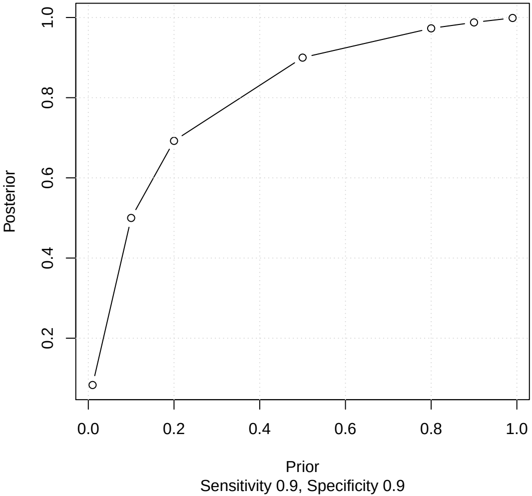
Example – computation

```
# prior estimate 20% of the image covered by water
p.h <- 0.2
# sensitivity: 90% of water pixels have low NDVI
#    (from training sample)
p.e.h <- 0.9
# false positive rate: 10% of non-water pixels have low NDVI
#    (from training sample)
p.e.nh <- 0.1
# denominator of likelihood ratio:
# predicted overall proportion of low-NDVI pixels in the iamge
#       (p.e <- (p.e.h * p.h) + (p.e.nh * (1 - p.h)))
## [1] 0.26
# likelihood ratio: increase in probability of hypothesis
#        given the evidence
(lr.h <- p.e.h/p.e)
## [1] 3.461538
# posterior probability
(p.h.e <- p.h * lr.h)
## [1] 0.6923077
```

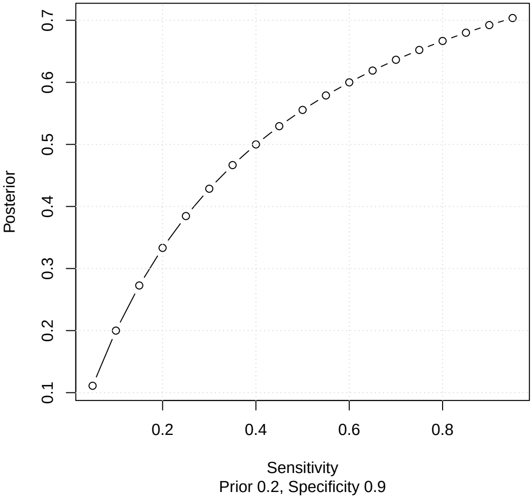
What affects the posterior probability?

1. $P(H)$, the prior probability of the hypothesis.
   - The higher the prior, the higher the posterior, other factors being equal. In the absence of any information in a two-class problem, we could set this to 0.5.
2. $P(E \mid H)$, the sensitivity of the hypothesis to the evidence.
   - The higher this is, the more diagnostic is the NDVI; it is in the numerator of the likelihood ratio.
3. $P(E \mid \neg H)$, the false positive rate (complement of the specificity).
   - The higher this is, the less diagnostic is the NDVI, since it is in the denominator.

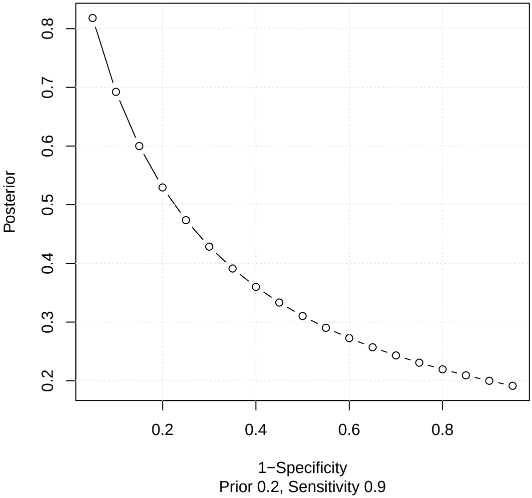
Effect of prior



Effect of sensitivity



Effect of specificity



Bayes' Rule for multivariate outcomes

This can be generalized to a sequence of n mutually-exclusive hypotheses H_n, given some evidence E.
The posterior probability of one of the hypotheses H_j is:

$$P(H_j \mid E) = P(H_j) \cdot \frac{P(E \mid H_j)}{P(E)} \tag{6}$$

$P(E) = \sum_{j=1}^{n} P(E \mid H_j) \cdot P(H_j)$ is the overall probability of the event.
This normalizes the conditional probability $P(H_j \mid E)$.

Bayesian statistical inference

The term "Bayesian" has been extended to a form of inference for *statistical models* where we:

- update a *prior probability distribution* ("**before** observations or experiments") of *model parameters* . . .
- with some evidence to obtain a *posterior probability distribution* ("**after** observations or experiments") of model parameters . . .
- based on the *likelihood* of the results of observations or experiments considering possible values of the parameters.
- This step is called *estimation* of the model parameters . . .
- We can then use these estimates for *prediction* of the target variable(s).

Bayesian view of statistical models

A *statistical model* has the following general form, using the notation $[\cdot]$ to indicate a probability distribution:

$$[Y, S \mid \theta] \tag{7}$$

- Y is the joint distribution of some variable(s) for given values of model parameter(s) θ
- the values of the variables are determined by some unobservable process S: the *signal*
- we can not account the *noise*, i.e., random variations not accounted for by the process.
- decompose as:

$$[Y, S \mid \theta] = [S \mid \theta]\,[Y \mid S, \theta] \tag{8}$$

Bayesian inference from data

1. Assume some *model form*, with unknown parameters θ, which is supposed to produce signal S
2. Observe some of the Y produced by the signal S
3. use these to *estimate* a probability distribution for θ
4. then use the statistical model to *predict* other values produced by the process.

$$[S \mid Y] = \int_\theta [S \mid Y, \theta] \, [\theta \mid Y] \, d\theta \tag{9}$$

Note that the prediction depends on the entire *posterior distribution* of the parameters θ

Model parameters are random variables

- In Bayesian inference we assume that the true values of model parameters θ are *random variables*, and therefore have a *joint* probability distribution with the observations:

$$[Y, \theta] = [Y \mid \theta] \, [\theta] \tag{10}$$

- The term $[\theta]$ is the *marginal* distribution for θ, i.e., before any data is known; therefore it is called the *prior* distribution of θ.
- Inference is then based on sampling from the posterior distributions of the different model parameters.
- Can find the most likely value, but also use the full distribution for simulating possible scenarios.
- Example: linear regression: a joint probability distribution of the parameters of the regression model (coefficients, their errors, their inter-correlation).

Frequentist view

- parameters of statistical models are considered to be *fixed*, but unknowable by finite experiment.
- Conduct more experiments, collect more evidence → come closer to the "true" value as a point estimate
- Assume an error distribution → confidence intervals around the "true" value
- Assumes that there *is* a "true" population value.

Bayesian inference for the Binomial distribution

- The Binomial distribution: a *continuous* probability distribution, with one parameter $\theta \in [0 \ldots 1]$

$$p(k, n) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{11}$$

  - k is the number of "successes" in n independent, exchangeable Bernoulli trials
  - i.e., with two mutually-exclusive possible outcomes conventionally referred to as "successes" and "failures", 0/1, True/False
- It models any situation where a number of independent observations n is made, each with one of two mutually-exclusive outcomes.
- The process S is thus some process that only gives one of these outcomes for each observation.

Example

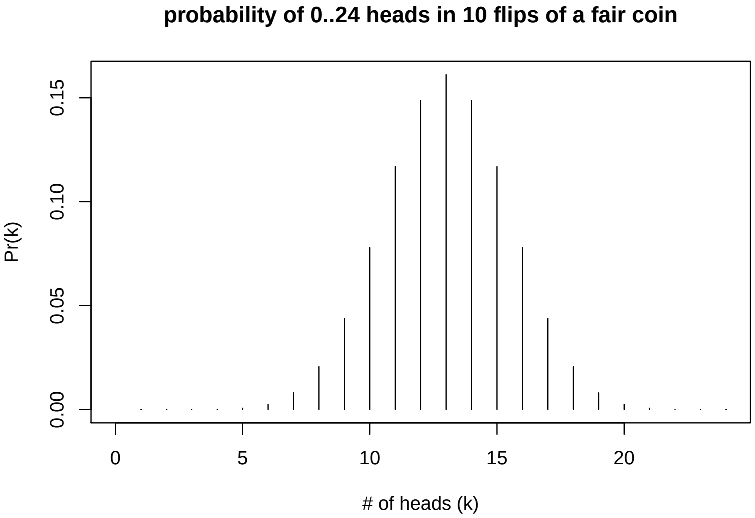
(1) Plot a histogram of the probability of 0 . . . 24 heads in 24 flips of a fair coin with the dbinom "binomial density" function.

(2) Compute the probability of exactly 10 heads in 24 flips.

$$\binom{24}{10} 0.5^{10} (1 - 0.5)^{24-10} = 0.1169$$

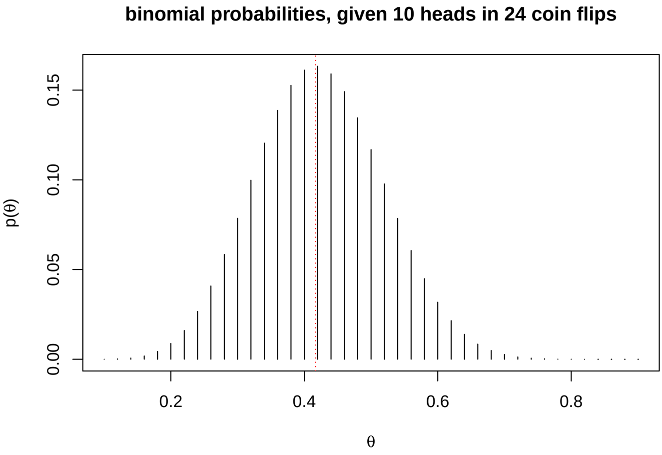
```
> plot(dbinom(0:24, size=24, prob=0.5), type="h",
   xlim=c(0,24),
   xlab="# of heads (k)", ylab="Pr(k)",
   main="probability of 0..24 heads in 10 flips of a fair coin")
> dbinom(10, size=24, prob=0.5)
[1] 0.1169
```

Binomial probabilities



The inverse view

Looking at this distribution from the opposite perspective, we see that if we observe *any* number $0 \ldots 24$ heads in 24 trials, this is *evidence* of different strength for **all** values of θ.



Probability distribution of a model parameter

- The aim of Bayesian inference is to have a *full probability distribution* for a parameter, here θ of the Binomial distribution.
- That is, we do *not* want to determine a single most probable value for θ;
- Instead we want to determine the probability of *any* value, or that the value is within a certain range, or that the value exceeds a certain number.
- For this we need a distribution for θ, parameterized by one or more *hyperparameters*.

Likelihood ratio

We extend Bayes' Rule to full distributions of a parameter, given the evidence of k successes in n trials:

$$p(\theta \mid k, n) = p(\theta) \cdot \frac{p(k, n \mid \theta)}{p(k, n)} \tag{12}$$

- The *posterior* probability of any proportion of successes θ, given that we observe k successes in n trials:
  - the *prior* probability distribution of $\theta \in [0 \ldots 1]$ from previous evidence or knowledge ...
  - ... multiplied by the *likelihood ratio*

$$\frac{p(k, n \mid \theta)}{p(k, n)} \tag{13}$$

LR: probability of finding a given number k success in n trials for a known value of θ...
... divided by the probability of finding k successes in n trials, no matter what value of θ.

Denominator of the likelihood ratio

For the binomial distribution, the denominator is an integral over all possible values of θ, which reduces to a very simple form:

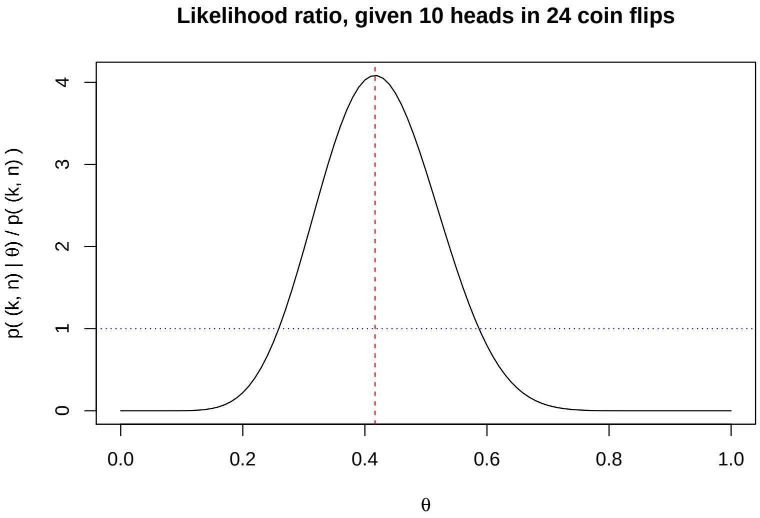
$$\begin{aligned} p(k,n) &= \int_{\theta=0}^{1} p(k,n|\theta)\,d\theta \\ &= \binom{n}{k} \cdot \mathrm{Beta}(k+1,(n-k)+1) \\ &= \binom{n}{k} \cdot \frac{\Gamma(k+1)\Gamma((n-k)+1)}{\Gamma(n+2)} \\ &= \binom{n}{k} \cdot \frac{k!(n-k)!}{(n+1)!} \\ &= \frac{n!}{k!(n-k)!} \cdot \frac{k!(n-k)!}{(n+1)!} \\ &= \frac{1}{n+1} \end{aligned} \tag{14}$$

Most distributions do *not* integrate so easily! In those cases numerical integration must be used.

Likelihood function

Plot the *continuous* distribution of the likelihood:

```
> (p.k.n <- 1/(n+1))  # normalizing constant
[1] 0.04
> curve(dbinom(k, size=n, prob=x)/p.k.n,
     xlab=expression(theta),
     ylab=expression(paste(plain("p( (k, n) | "),
                           theta, plain(") / p( (k, n) )"))),
     main="Likelihood ratio, given 10 heads in 24 coin flips")
> abline(v=k/n, col="red", lty=2)
> abline(h=1, col="blue", lty=3)
```


Likelihood ratio, given 10 heads in 24 coin flips
p((k, n) | θ) / p((k, n))
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
θ

Likelihood function (2)

The likelihood ratio can also be written with the reverse functional relation, i.e., θ as a function of k, n:

$$\ell(\theta \mid k, n) = p(k, n \mid \theta) \tag{15}$$

where the ℓ function is read as "the likelihood of".

This is another way of thinking about the relation between the observations and the parameter: the likelihood that the parameter has a certain value, knowing the observations, i.e., considering the data as fixed.

Computing the unnormalized posterior distribution

- The likelihood function is also called the *sampling density* because it depends on having taken a sample, i.e., having made a trial.
- Once we have the *prior probability distribution* and the *likelihood function*, we compute the *(un-normalized) posterior probability distribution* by a modification of Bayes' Rule, applying to distributions:

$$p(\theta \mid x) \propto p(\theta) \cdot \ell(\theta \mid x) \tag{16}$$

  Note $\propto$ "proportional to", not $=$ "equals".
- **This is the fundamental equation of Bayesian inference.**

Probability distribution for the binomial parameter

- θ can take any value from $[0 \ldots 1]$
- we need to find a *probability distribution* for it
    - function $f(\theta)$: domain $\mathbb{R} \in [0 \ldots 1]$ (possible values of θ) and range $[0 \cdots 1]$ (their probability)
    - $\int_0^1 f(\theta) = 1$
- this distribution will be parametrized by one or more *hyperparameters*

Conjugate prior distribution

- preferable to find a function that has the same form *prior* and *posterior*, i.e., after being multiplied by the likelihood
- this is called a *conjugate* prior
- It is desirable because we may want to later use the posterior distribution as a prior in further analysis

Conjugate prior for the binomial distribution

- Beta distribution with two hyperparameters α and β

$$\mathrm{Beta}(\theta; \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{17}$$

- The first term is a normalizing constant to ensure that the total probability integrates to 1, using the Beta function:

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{18}$$

- $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$, the generalization to the real numbers of the factorial. For integer x, $\Gamma(x+1) = x!$.
- So, the normalizing constant is:

$$1/\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{19}$$

Why is it conjugate? (1)

$$\begin{aligned} p(\theta) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ \ell(k,n \mid \theta) &\propto \theta^{k}(1-\theta)^{n-k} \\ \\ p(\theta \mid k,n) &\propto p(\theta)\cdot\ell(k,n \mid \theta) \\ p(\theta \mid k,n) &\propto \theta^{\alpha+k-1}(1-\theta)^{\beta+(n-k)-1} \end{aligned} \quad (20)$$

So the posterior also has the form of a Beta distribution

Why is it conjugate? (2)

- If prior $p(\theta) \sim \mathrm{Beta}(\alpha, \beta)$, and the number of successes k in n trials follows the binomial distribution with parameter θ, then the posterior becomes

$$p(\theta \mid k, n) \sim \mathrm{Beta}(\alpha + k, \beta + (n - k))$$

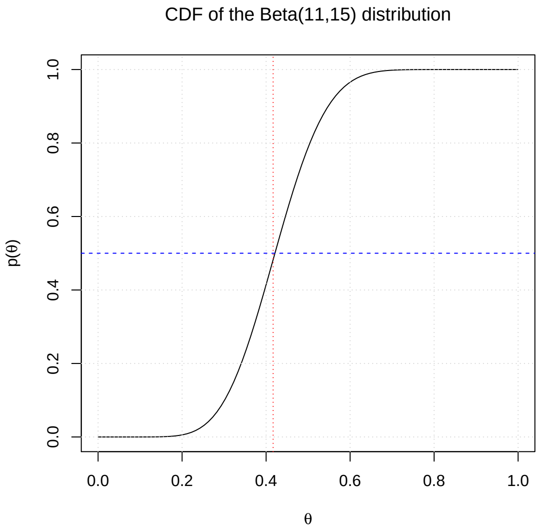
- This simple updating formula allows us to modify a prior Beta distribution to posterior Beta distribution that takes into account the data.
- Note that the larger the n, the less important are the prior values of the hyperparameters.

Hyperparameters of the Beta distribution

α $\alpha + 1$ number of "successes"

β $\beta + 1$ number of "failures"

$(\alpha + \beta - 2)$ total number of trials


CDF of the Beta(11,15) distribution
p(θ)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
θ

Parameterizing the Beta distribution

- Expected value of a Beta-distributed θ is:

$$\begin{aligned} \mathbf{E}\theta &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta d\theta \\ &= \alpha/(\alpha+\beta) \end{aligned} \tag{21}$$

- Because of the -1 in the exponents of the Beta distribution, this number is better given as $(\alpha+\beta+2)$, and the numerator as $(\alpha+1)$
- Then expected proportion is $(\alpha-1)/(\alpha+\beta-2)$

Selecting prior hyperparameters

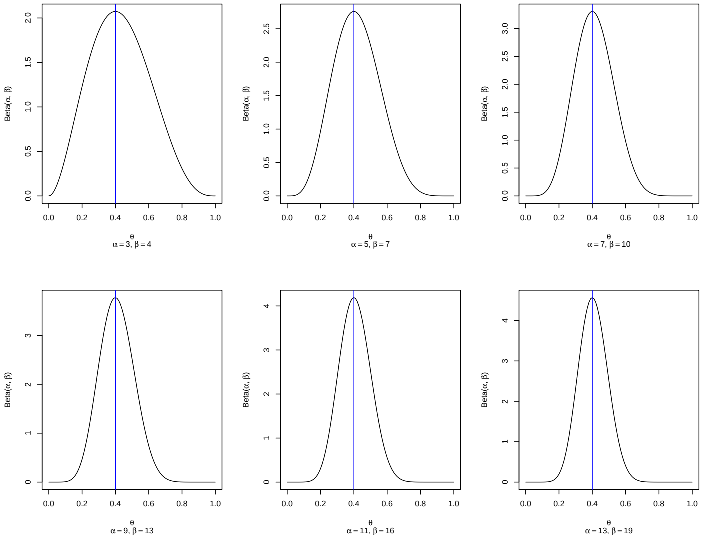
1. α as the modal number of "successes" in $(\alpha + \beta)$ trials.
   - more trials → more prior evidence
2. Use the expected mean and the variance to solve two equations in two unknowns to obtain the two hyperparameters:

$$\mathbf{E}\theta = \frac{\alpha}{(\alpha + \beta)} \tag{22}$$

$$\mathbf{Var}\theta = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{23}$$

   - As the number of trial increases, the variance decreases
   - This requires an expert judgement of a variance, which is not as intuitive as a mean and sample size.

Plot of informative priors



Non-informative prior

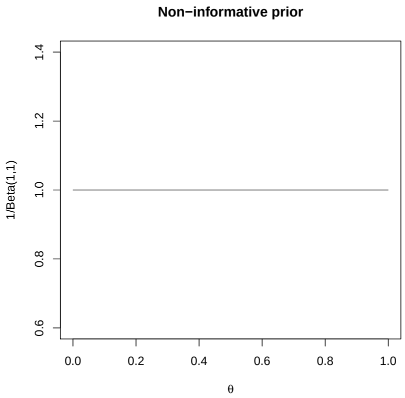
- Parameterize the Beta distribution such that all values of θ are *a priori* equally likely, and all inferences about the distribution of θ come from the data.
  - "Non-informative" is not really good terminology, as even absence of information is information. The idea is to represent in some sense the least amount of information, i.e., maximum *a priori* ignorance, consistent with the form of the prior distribution.
- One choice[1] is $\alpha = \beta = 1$:

$$\mathrm{Beta}(x; 1, 1) = \frac{1}{\mathrm{B}(1,1)} x^{1-1} (1-x)^{1-1} = \frac{1}{\mathrm{B}(1,1)} = 1 \quad (24)$$

Uniform on $[0 \ldots 1]$, does not depend on the Binomial parameter

[1] used by Bayes in his Essay

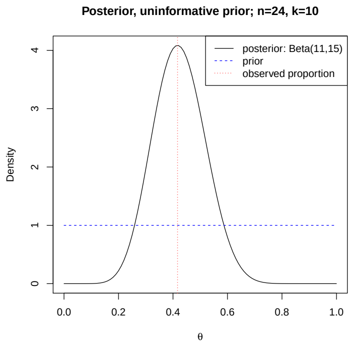
Plot of non-informative prior

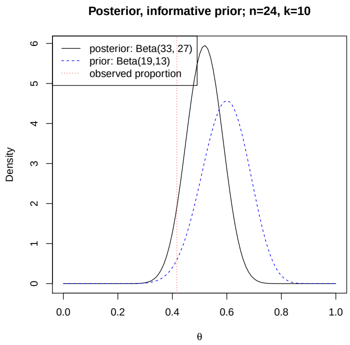


Updating based on evidence

- Suppose we observe 10 "successes" in 24 Bernoulli trials
- What is the distribution of the parameter of the Binomial distribution θ ...
  - starting from the non-informative prior $\alpha = \beta = 1$
    - posterior $\alpha = 11, \beta = 15$
  - starting from an informative prior somewhat far from this, $\alpha = 19, \beta = 13$; total "prior evidence" 30 trials.
    - posterior $\alpha = 33, \beta = 27$

Posterior distributions for θ





Credible intervals

- compute a *credible interval* within which we believe, with some probability, the parameter lies
- We obtain credible intervals from the quantiles of the distribution, prior or posterior.
- To do this, we find the upper limit c of the definite integral of the distribution, such that it equals the desired quantiles q, for example $q = 0.05$ and $q = 0.95$ for the 90% credible interval.

$$\int_0^c p(\theta|k,n)\,d\theta = q \tag{25}$$

```
> ## informative prior
> (cred.inf.pre <- qbeta(c(0.05, 0.95),
                         shape1=prior.a, shape2=prior.b))
[1] 2
> (cred.inf.post <- qbeta(c(0.05, 0.95),
                         shape1=prior.a+k, shape2=prior.b+(n-k)))
[1] 0.4085025 0.6264798

> ## non-informative prior
> (cred.non.inf.pre <- qbeta(c(0.05, 0.95),
                         shape1=1, shape2=1))
[1] 0.05 0.95
> (cred.non.inf.post <- qbeta(c(0.05, 0.95),
                         shape1=1+k, shape2=1+(n-k)))
[1] 0.2698531 0.5831620
```

Note the narrower credible interval 0.218 from the informative prior vs. the non-informative prior 0.281.

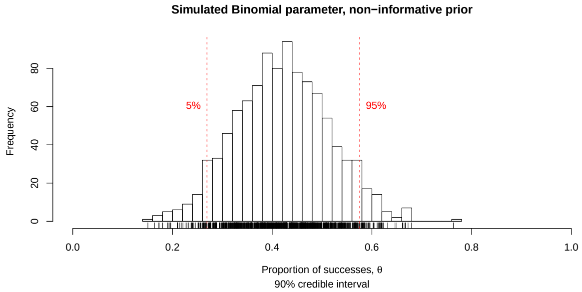
Simulation of posterior intervals

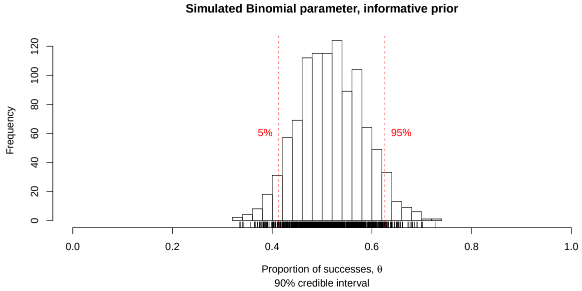
Can also compute intervals by *simulation*:

1. draw samples with the `rbeta` "random value from the beta distribution" function
2. find the quantiles of the simulated draw with the `quantile` function
3. compute any summary ($>$, $<$ some quantile, within some range ...)

Note: not necessary in this case because the posterior is expressible analytically, but this method works for any posterior distribution

Simulated credible intervals for θ





Prediction

- Recall the general form of the predictive distribution:

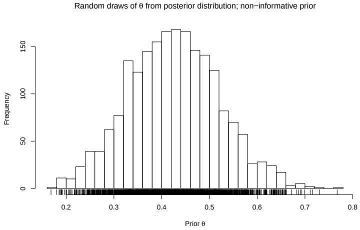
$$[S \mid Y] = \int_\theta [S \mid Y, \theta]\,[\theta \mid Y]\,d\theta \qquad (9)$$

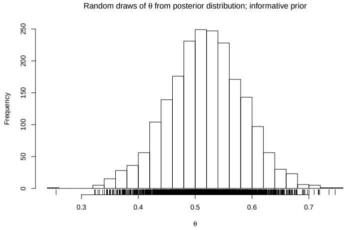
- Here the process S is the set of Bernoulli trials
- We want to predict results of a *future* set, based on the set we've seen (Y) and the posterior distribution of the parameter of the Binomial process (θ).
- Integrate the predictions from *each* value of the parameter based on its posterior probability:

$$p(\tilde{y} \mid y) = \int p(\tilde{y} \mid \theta)\,p(\theta|y)\,d\theta \qquad (26)$$

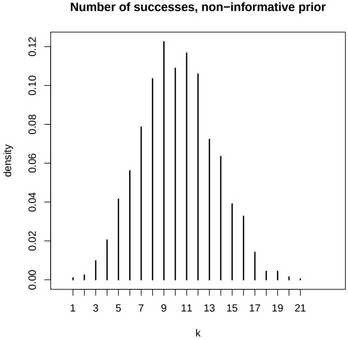
- Can evaluate this by simulation of θ, and from that $p(k, n)$.

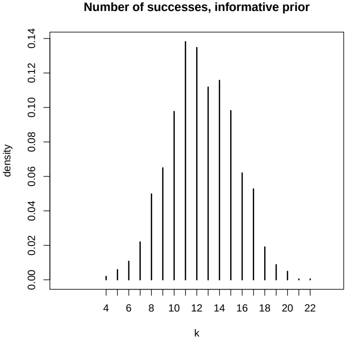
Frequency of θ for 2048 draws from its posterior distribution



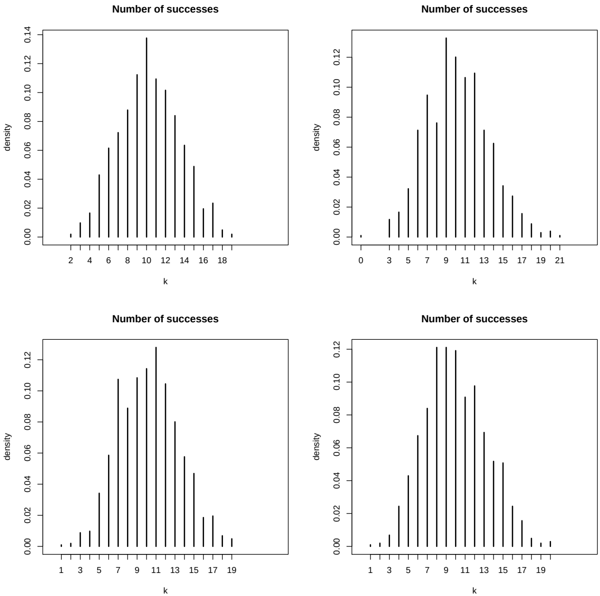


Density of $p(k, 24)$ for 2048 draws from the posterior distribution



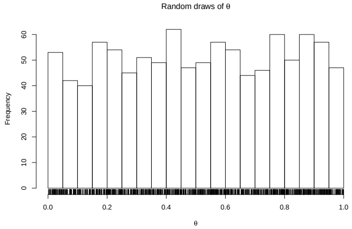


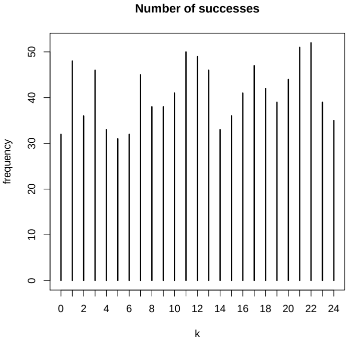
Compare 4 simulations



Prediction from non-informative prior

Density of $p(k, 24)$ for 2048 draws from the posterior distribution – in theory all values are equally likely





Hierarchical models

- A *hierarchical* model, also called a *multilevel* model, is one where several posterior distributions must be estimated, with some depending on others.
- Example: a *multinomial mixture* of *binomial distributions* The population is divided into m groups, each with its own separate binomial distribution:

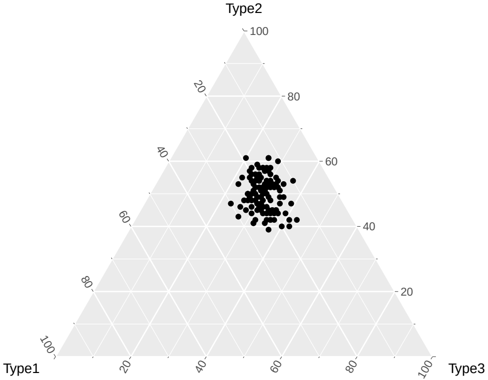
$$p(k_j) = \binom{n_j}{k_j} \theta_j^{k_j} (1 - \theta_j)^{n_j - k_j} \tag{27}$$

- The division of the population into groups is also probabilistic and represented by a *multinomial* distribution:

$$\begin{aligned} f(n_1, n_2, \ldots n_m; n; \psi_1, \psi_2, \ldots \psi_m) &= \Pr(X_1 = n_1, X_2 = n_2, \ldots X_m = n_m \\ &= \frac{n!}{n_1! n_2! \ldots n_m!} \psi_1^{n_1} \psi_2^{n_2} \ldots \psi_m^{n_m} \end{aligned} \tag{28}$$

Example

128 draws of 100 items each from $\psi_i = 0.2, \psi_2 = 0.5, \psi_3 = 0.3$



Application

- A soil sampling campaign where we will make a fixed number n of *spatially-random observations*, constrained by the budget, to determine the proportion of soils that require some intervention based on a critical limit.
- Several soil types: a *multinomial* distribution
- *Within each soil type*, a proportion of soils θ_j that exceed the limit: k_j of the n_j samples *of that soil type* will require intervention: a set of *binomial* distributions
- Q: Why not just use the maximum likelihood binomial mean/standard deviation from the completely random sample?
- A: The hierarchical approach allows the use of *prior* probability distributions. This is especially important with small sample size.

Multilevel model

Level 1 $k_j \mid \theta_j, n_j \sim \mathrm{Binomial}(\theta_j, n_j)$, the number of observations of the total n_j in soil type j requiring intervention;

Level 2 $\theta_j \mid \alpha_j, \beta_j \sim \mathrm{Beta}(\alpha_j, \beta_j)$, the distribution for the binomial parameter θ_j in soil type j;

Level 3 $n_j \mid \psi_1, \psi_2, \ldots \psi_m, n \sim \mathrm{Multinomial}(\psi_1, \psi_2, \ldots \psi_m, n)$, the number of observations of soil type j, out of the total number of observation n, for each of the m possible soil types;

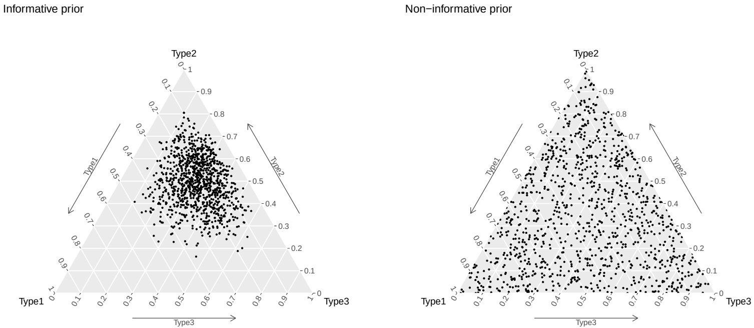
Level 4 $\psi_j \mid \alpha_1, \alpha_2 \ldots \alpha_m \sim \mathrm{Dirichlet}(\alpha_1, \alpha_2 \ldots \alpha_m)$, the distribution of the m multinomial parameters.

The Dirichlet distribution is the multivariate analogue of the Beta distribution:

$$\mathrm{D}(\alpha) = \frac{1}{\mathbf{B}(\alpha)} \prod_{j=1}^{m} x_j^{\alpha_j - 1} \tag{29}$$

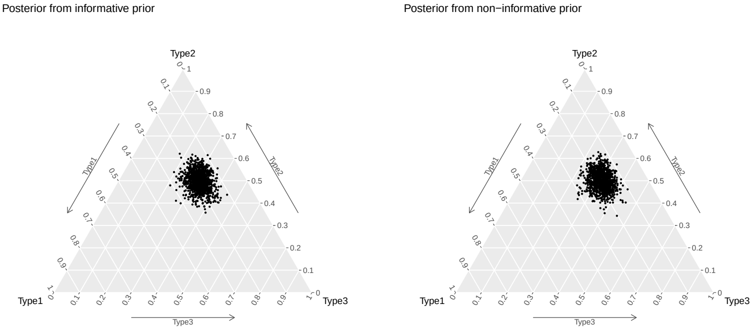
Draws from Dirichlet distribution

Informative: estimate $(0.2, 0.5, 0.3)$; non-informative all $0.\overline{3}$



Posterior proportions

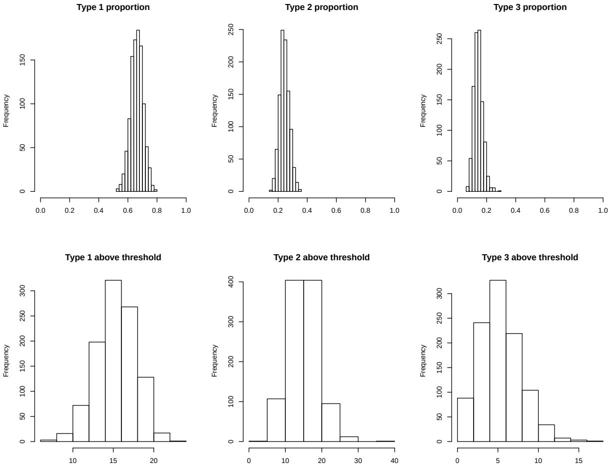
Suppose 128 observations in classes $(24, 64, 40)$:



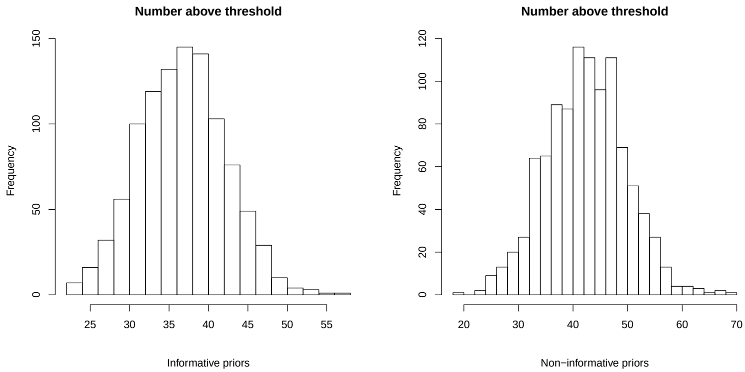
Note how information concentrates the posterior distributions of $\mathrm{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$

Posterior counts – per soil type

Suppose the soils requiring intervention are $(12/24, 20/64, 10/40)$; all with non-informative prior



Posterior counts – for all soil types

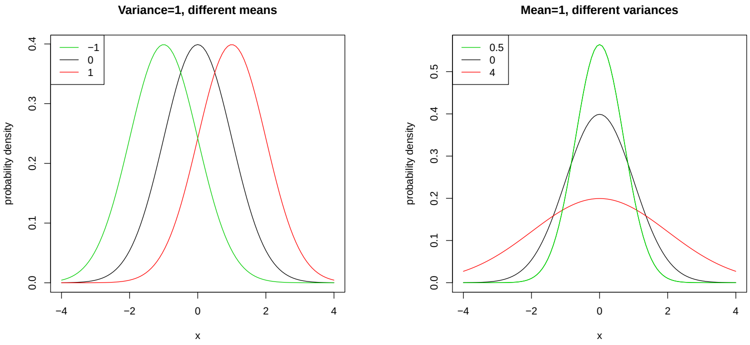


Multi-parameter models

- Models have > 1 parameter; in general not independent; their *joint* as well as *marginal* distributions must be estimated
- Example: *normal* ("Gaussian") distribution; two parameters:
  1. the *location* μ, also called the *mean*;
  2. the *dispersion* σ^2, also called the *variance*.
     Can be convenient to work with the inverse $1/\sigma^2$, called the *precision*, written as τ.

  The density function is:

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} \qquad (30)$$


Variance=1, different means
−1
0
1
probability density
x
Mean=1, different variances
0.5
0
4
probability density
x

Likelihood

$$\ell(\mu, \sigma^2 \mid \mathbf{x}) = \tag{31}$$

$$p(\mathbf{x} \mid \mu, \sigma^2) = \prod_{i=1}^{n} p(x_i \mid \mu, \sigma^2) \tag{32}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right) \tag{33}$$

As the parameters μ and σ^2 change, so does the likelihood of having observed the values $\mathbf{x}$.

Distributions for the Normal distribution parameters

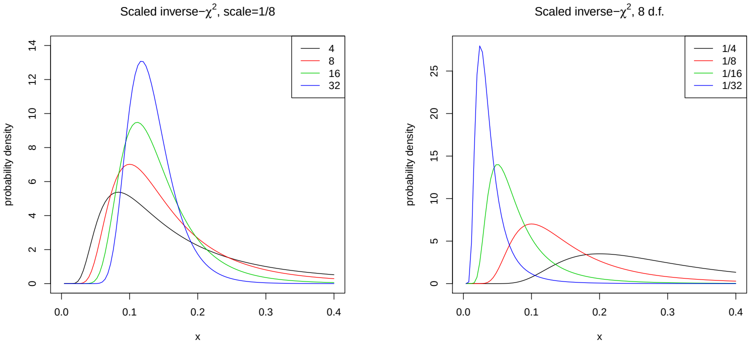
Most common:

- For μ, another Normal distribution
  - *hyperparameters* (μ_0, σ_0^2);
- For σ^2, an inverse χ^2 distribution
  - *hyperparameter* ν, the degrees of freedom:

$$\chi_\nu^{-2}(x) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} x^{-(\nu/2)-1} e^{-1/(2x)} \tag{34}$$

    More degrees of freedom → more probable that the variance σ^2 is small.
  - Usually *scaled*: additional parameter $\tau^2 = 1/\sigma^2$, the inverse of the variance of the process.

Inverse χ^2 distribution



Multivariate normal distribution

- Several variables; all *marginal* distributions are normal, each with their own parameters
- The variables may be *correlated*, i.e., instead of a variance, there is a *variance-covariance* matrix
- Parameters:

  μ vector of means
  Σ variance-covariance matrix

- PDF: a generalization of the univariate normal distribution:

$$\det(2\pi\Sigma)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma(\mathbf{x}-\mu)\right\}$$

Numerical methods

- Most models can *not* be reduced to analytical forms.
- Their posterior distributions can *not* be computed as a closed form
- This is often because the denominator (proportionality constant) in the fundamental Bayesian inference formula has no closed form.

$$\int p(\theta) \cdot p(Y \mid \theta) d\theta$$

- The required integration over the parameter space must be done by *numerical* simulations of the posterior distribution
- This requires substantial computer power and some mathematical tricks.

MCMC

- The most common method to simulate posterior distributions is the **Markov chain Monte Carlo**[2] (MCMC) method.
- This is an algorithm for sampling from a (multivariate) probability distribution that can not be expressed as a closed form, based on constructing a *Markov chain* that has the desired distribution, e.g., posterior or predictive, as its equilibrium distribution.
- Markov chain: sequence of values of parameter(s) where value at θ_{t+1} depends only on previous value θ_t, not on the entire history of the chain
  - so, conditional on the *present* value, future and past values are independent.

[2] Just a fancy name for "random"

The Gibbs sampler

Repeatedly sample from the full *conditional* distribution of each of the k parameters in the posterior distribution, one parameter i at a time: $p(\theta_i \mid \theta_{j \neq i}, i = 1, 2, \ldots k)$

1. Pick arbitrary starting values $x^0 = (x_1^0, \ldots x_k^0)$. This does not depend (yet) on the observations Y.
2. Make a *random* drawing from the full conditional distribution $\pi(x_i \mid x_{-i}, i = 1, \ldots k)$, as follows:

   x_1^1 from $\pi(x_1 \mid x_{-1}^0 \mid Y)$
   x_2^1 from $\pi(x_2 \mid x_1^1, x_3^0, \ldots x_k^0 \mid Y)$
   x_3^1 from $\pi(x_3 \mid x_1^1, x_2^1, x_4^0, \ldots x_k^0 \mid Y)$
   ...
   x_k^1 from $\pi(x_k \mid x_{-k}^1 \mid Y)$

   This results in an *updated* full conditional distribution $x^1 = (x_1^1, \ldots x_k^1 \mid Y)$.

Under certain conditions this converges to a steady-state distribution.

Multivariate regression

- This model has the well-known form: $y_i = (X_i)^T \beta + \varepsilon_i$, with i.i.d. Gaussian errors: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
- Can be directly solved by OLS, but that assumes independence of the β.
- β is a *vector* of regression coefficients; each of these has its own standard error *and* these may be correlated with each other
- The priors are *semi-conjugate* and *a priori* independent:

$$\beta \sim \mathbf{MVN}(b_0, B_0^{-1}) \qquad (35)$$

$$1/\sigma^2 = \tau \sim \Gamma(c_0/2, d_0/2) \qquad (36)$$

- Assume *a priori* (without evidence) that the distribution of the β vector is independent of the distribution of the $1/\sigma^2 = \tau$

Conditional posterior probability for β

$$p(\beta \mid \sigma^2, \mathbf{y}, \mathbf{X}) \sim \mathrm{MVN}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}', \sigma^2(\mathbf{X}'\mathbf{X})^{-1}) \tag{37}$$

which is the OLS formulation.

Note how the variance-covariance matrix of the regression parameters depends on the residual variance of the regression.

Marginal posterior distributions for β_m

This requires integrating out the variance:

$$p(\beta_m \mid \mathbf{y}, \mathbf{X}) = \int_0^{+\infty} p(\beta_m \mid \sigma^2, \mathbf{y}, \mathbf{X}) d\sigma^2 \quad (38)$$

Similarly, for the marginal posterior distribution of the regression variance σ^2, we need to integrate out the regression coefficients.

Computation

- The `MCMCregress` function of the `MCMCpack` package generates a sample from the posterior distribution of a (multiple) linear regression model with Gaussian errors, using using Gibbs sampling.
- The prior distribution for the β vector (regressors) must be multivariate Gaussian, and that for the error variance an inverse-Γ prior.
- The returned sample from the posterior distribution can be analyzed with functions provided in the coda "Convergence Diagnosis and Output Analysis and Diagnostics for MCMC" package

The MCMCregress function

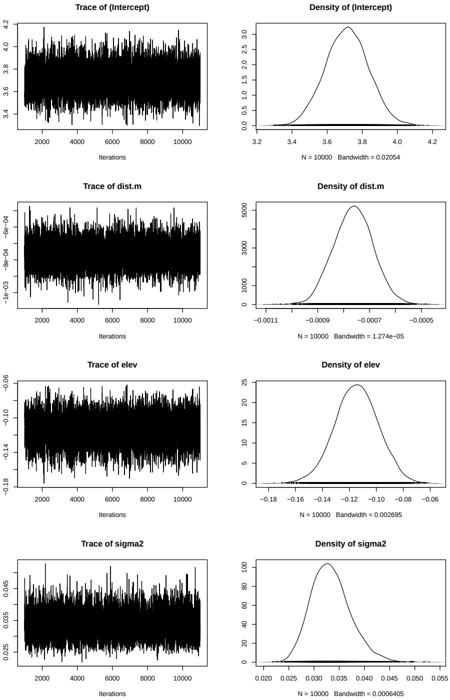
- generates a sample from the posterior distribution of a (multiple) linear regression model with Gaussian errors, using the Gibbs sampler.
- Hyperparameters:
  - `b0` a vector of the mean prior values of β;
  - `B0` a matrix of the prior precisions of each β; this can be a full matrix (precisions of different predictors are correlated).
  - `c0` $c_0/2$ is the *shape* parameter of the inverse-Γ prior for σ^2; the amount of information represents c_0 pseudo-observations;
  - `d0` $d_0/2$ is the *scale* parameter of the inverse-Γ prior for σ^2; it represents the sum of squared errors of the c_0 pseudo-observations;
- Control arguments:
  - `burnin` the number of *burn-in* iterations, i.e., before statistics are collected for the posterior distribution; default 1000;
  - `mcmc` The number of MCMC iterations after burn-in; default 10000.

Example: Meuse River soil pollution

```
> m <- MCMCregress(log10(zinc) ~ dist.m + elev, data=meuse)
> summary(m)
Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:
                  Mean        SD
(Intercept)  3.7131559 1.223e-01
dist.m      -0.0007622 7.581e-05
elev        -0.1145986 1.604e-02
sigma2       0.0333455 3.902e-03

2. Quantiles for each variable:
                  2.5%        25%        50%       75%      97.5%
(Intercept)  3.4770168  3.6302124  3.7135632  3.794888  3.9520484
dist.m      -0.0009079 -0.0008138 -0.0007618 -0.000711 -0.0006137
elev        -0.1466163 -0.1253493 -0.1146035 -0.103790 -0.0837966
sigma2       0.0265521  0.0305927  0.0330553  0.035702  0.0419692
```


Trace of (Intercept)
Iterations
Density of (Intercept)
N = 10000 Bandwidth = 0.02054
Trace of dist.m
Iterations
Density of dist.m
N = 10000 Bandwidth = 1.274e−05
Trace of elev
Iterations
Density of elev
N = 10000 Bandwidth = 0.002695
Trace of sigma2
Iterations
Density of sigma2
N = 10000 Bandwidth = 0.0006405

Compare to OLS fit

```
> summary(m <- lm(log10(zinc) ~ dist.m + elev, data=meuse))
Coefficients:
              Estimate Std. Error t value
(Intercept)  3.713e+00  1.223e-01  30.366
dist.m      -7.607e-04  7.489e-05 -10.158
elev        -1.146e-01  1.604e-02  -7.144

Residual standard error: 0.1815 on 152 degrees of freedom

> (summary(m)$sigma)^2    # sigma^2 of residuials
[1] 0.03294231

> coefficients(m)[3] +  # 97.5 quantile of elevation coef
    (summary(m)$coefficients[3,"Std. Error"]*qnorm(0.975))
       elev
-0.08317467
```

Mean values of coefficients, σ^2, 97.5% confidence limit/credible limit not too different.

Meuse River soil pollution – informative priors

Large negative coefficients for elevation, slope; precise; but large s.e.

```
> m.i <- MCMCregress(log10(zinc) ~ dist.m + elev, data=meuse,
                     b0=c(0, -0.3 , -0.3),
                     B0=c(1e-6, .0001, .0001),
                     c0=10, d0=10)
> summary(m.i)
                  Mean        SD
(Intercept)  3.7139349 0.2049347
dist.m      -0.0007631 0.0001272
elev        -0.1146691 0.0268936
sigma2       0.0936901 0.0106301

                 2.5%        25%        50%        75%      97.5%
(Intercept)  3.317704  3.5752025  3.7143326  3.8510817  4.1149562
dist.m      -0.001007 -0.0008501 -0.0007626 -0.0006769 -0.0005147
elev        -0.168530 -0.1326239 -0.1146034 -0.0965921 -0.0630443
sigma2       0.075109  0.0861624  0.0929222  0.1001137  0.1170061
```

Effect of informative priors

```
> m <- MCMCregress(log10(zinc) ~ dist.m + elev, data=meuse)
> m.i <- MCMCregress(log10(zinc) ~ dist.m + elev, data=meuse,
+                      b0=c(0, -0.3 , -0.3),
+                      B0=c(1e-6, .0001, .0001),
+                      c0=10, d0=10)

> summary(m)$statistics[2:3,"Mean"]
       dist.m          elev
-0.0007622489 -0.1145985676
> summary(m.i)$statistics[2:3,"Mean"]
       dist.m          elev
-0.0007631305 -0.1146691391

> summary(m)$statistics["sigma2","Mean"]
[1] 0.0333455
> summary(m.i)$statistics["sigma2","Mean"]
[1] 0.09369014
```

Comparing models with the Bayes factor

- Bayes Factor: the ratio of posterior likelihoods of the data, given the fitted models:

$$\mathrm{BF} = \frac{p(y \mid X, m_a)}{p(y \mid X, m_b)} \tag{39}$$

m_a, m_b two models to compare, X design matrix, y observed data.

- The Bayes factor quantifies the support from the data for one model compared to another.
- Jeffreys [7] subjective scale:

factor	ln(factor)	strength of evidence for m_a
$< 10^0$	< 0	negative, supports m_b
$10^0 \ldots 10^{0.5}$	$0 \ldots \approx 1.5$	barely worth mentioning
$10^{0.5} \ldots 10^1$	$\approx 1.5 \ldots \approx 2.3$	substantial
$10^1 \ldots 10^{3/2}$	$\approx 2.3 \ldots \approx 3.5$	strong
$10^{3/2} \ldots 10^2$	$\approx 3.5 \ldots \approx 4.6$	very strong
$> 10^2$	$>\approx 4.6$	decisive

Bayes Factor example

```
> m <- lm(log10(zinc) ~ x + y + dist.m + elev, data=meuse)
> lm.1.posterior <- MCMCregress(formula(m),
      data=meuse,
      B0=c(1e-6, .01, .01, .01, .01), marginal.likelihood="Chib95"
> lm.2.posterior <- MCMCregress(update(formula(m), . ~ . -x -y),
      data=meuse,
      B0=c(1e-6, .01, .01), marginal.likelihood="Chib95")
> round(summary(lm.1.posterior)$statistics[2:5,"Mean"],6)
        x         y    dist.m      elev
-0.000061  0.000062 -0.000680 -0.117053
> round(summary(lm.2.posterior)$statistics[2:3,"Mean"],6)
   dist.m      elev
-0.000762 -0.114598
> (bf.1.2 <- BayesFactor(lm.1.posterior, lm.2.posterior))
The matrix of the natural log Bayes Factors is:
               lm.1.posterior lm.2.posterior
lm.1.posterior            0.0          -23.6
lm.2.posterior           23.6            0.0
lm.1.posterior : log marginal likelihood =  -15.92415
lm.2.posterior : log marginal likelihood =  7.685869
```

The more complex model is preferred.

Frequentist model comparison

```
> lm.1 <- lm(formula(m), data=meuse)
> lm.2 <- update(lm.1, ~ . - x - y)

> summary(lm.1)$adj.r.squared
[1] 0.6697626
> summary(lm.2)$adj.r.squared
[1] 0.6648385

> anova(lm.1,lm.2)

Model 1: log10(zinc) ~ x + y + dist.m + elev
Model 2: log10(zinc) ~ dist.m + elev
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    150 4.8687
2    152 5.0072 -2  -0.13848 2.1332  0.122

> AIC(lm.1); AIC(lm.2)
[1] -84.52019
[1] -84.17308
```

The more complex model (include coördinates) is preferred.

Spatial Bayesian analysis

The same kind of reasoning for *non-spatial* models applies to *spatial* models:

- We have a *model form*, which usually includes a *model of spatial dependence*.
- We consider the parameters of the model to be *random variables* each with a *distribution*.
- These have *prior* distributions, updated by the evidence to *posterior* distributions.
- Predictions are made by sampling from the posterior distributions.

R packages: `spBayes` [4], `geoR`, [14], `geoGLM`

`geoR`: Bayesian methods for point geostatistics, analogous to the `gstat`, `spatial` and `fields` packages that take a frequentist approach to geostatistical inference

Gaussian Model

General linear model with a linear regression for the spatial trend and residual spatial correlation:

$$[Y] \sim \mathcal{N}(X\beta, \sigma^2 R(\phi) + \tau^2 I) \tag{40}$$

- X $n \times p$ matrix of covariates
- β vector of regression parameters (coefficients)
- R spatial correlation function depending on a decay ("range") parameter ϕ
  - spherical, exponential ...
  - generalized exponential/Gaussian: *Matérn*, extra parameter κ (see next slide)
- σ^2 overall variance of the residual spatial process ("sill")
- τ^2 nugget effect, pure noise of the process

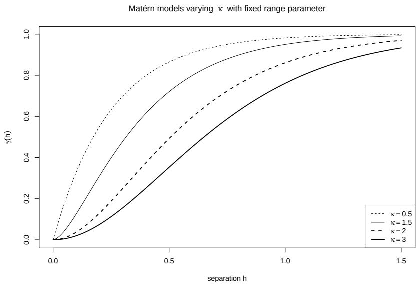
Matérn model of spatial covariance

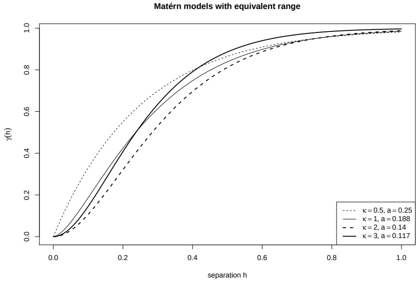
A general model with variable shape, adds a shape parameter κ to the scale parameter needed by all spatial covariance functions; Reference: [12].

$$p(h) = \left\{2^{\kappa-1}\Gamma(\kappa)\right\}^{-1} (h/\phi)^{\kappa} K_{\kappa}(h/\phi) \tag{41}$$

- $K_{\kappa}(\cdot)$ a modified Bessel function of order κ
- $\phi > 0$ scale parameter with the dimensions of distance
- $\kappa > 0$ the *order*: a shape parameter which determines the analytic smoothness of the spatial process
  - $\kappa = 0.5$ exponential $\exp(-h/\phi)$
  - $\kappa \to \infty$ Gaussian $\exp\left\{(-(h/\phi)^2\right\}$
  - generally try a few values of κ, not fit by likelihood over the whole range

Matérn model



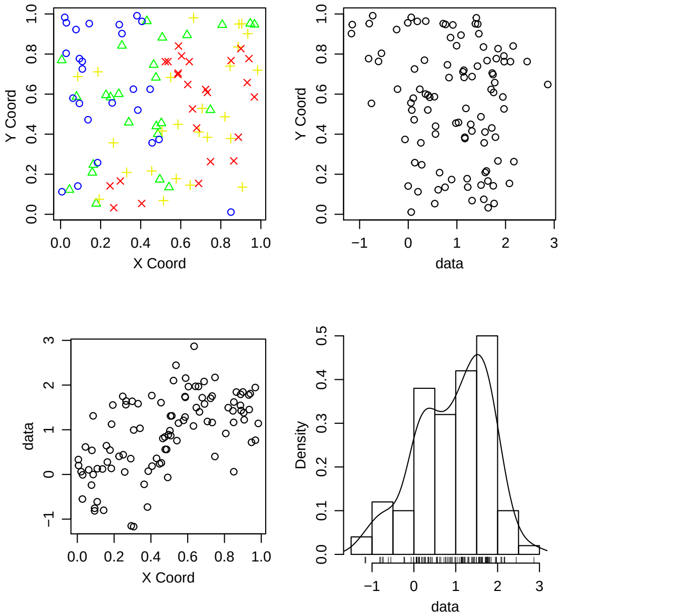


Distributions

- For fixed ϕ (range), priors for β, σ^2 as for the Normal distribution: Normal – scaled inverse χ^2
- For variable ϕ:

  $p(\phi \mid y) \propto \pi(\phi) \left| V_{\tilde{\beta}} \right|^{1/2} |R|^{-1/2} (S^2)^{-(n+n_\sigma)/2}$

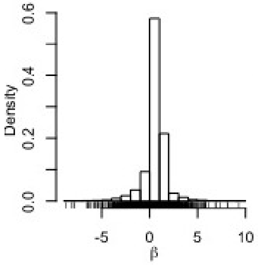
Example dataset – elevation points



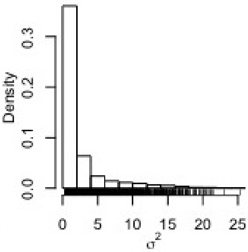
Model fitting

```
> bsp4 <- krige.bayes(s100, loc = loci,
                   prior = prior.control(phi.discrete =
                                           seq(0,5,l=101),
                                         phi.prior="rec"),
                   output=output.control(n.post=5000))
> summary(bsp4)
             Length Class                Mode
posterior      6    posterior.krige.bayes list
predictive     7    -none-               list
prior          4    prior.geoR           list
model          6    model.geoR           list
.Random.seed 626    -none-               numeric
max.dist       1    -none-               numeric
call           5    -none-               call
```

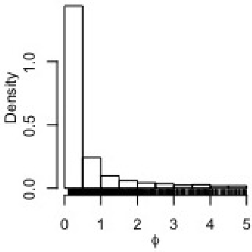
Posterior distribution of parameters



spatial trend



covariance sill



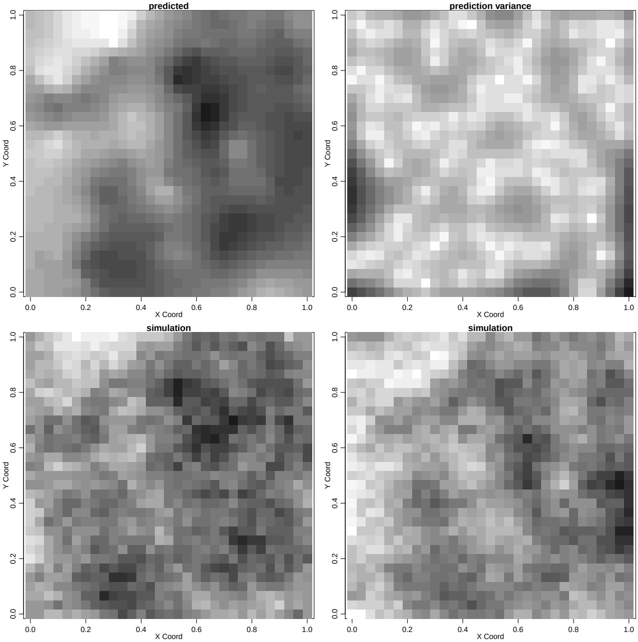
covariance range

Prediction

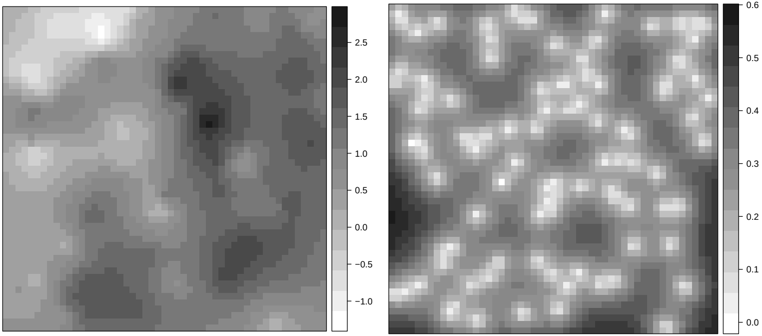
```
> ## prediction grid
> pred.grid <-  expand.grid(seq(0,1, l=31), seq(0,1, l=31))

> ## best prediction
> bsp4 <- krige.bayes(s100, loc = loci,
                  prior = prior.control(phi.discrete =
                                        seq(0,5,l=101),
                                        phi.prior="rec"),
                  output=output.control(n.post=5000))

> ## simulation
> bsp <- krige.bayes(s100, loc = pred.grid,
                  prior = prior.control(phi.discrete =
                                        seq(0,5,l=51)),
                  output=output.control(n.predictive=2))
```


predicted
prediction variance
simulation
simulation
X Coord
Y Coord
0.0
0.2
0.4
0.6
0.8
1.0

Compare to conventional kriging



References

- texts: [5, 6, 8, 9]
- computation in R: [1, 10, 11, 13]
- historical: [2]
- MCMC: [3, 15]
- spatial: [4, 14]

Bibliography I

[1] Jim Albert. *Bayesian Computation with R*. Springer-Verlag New York, 2nd edition, 2009. ISBN 978-0-387-92298-0. URL http://site.ebrary.com/id/10294526.

[2] G. A. Barnard and Thomas Bayes. Studies in the history of probability and statistics: IX. Thomas Bayes's 'Essay Towards Solving a Problem in the Doctrine of Chances'. *Biometrika*, 45(3/4):293–315, 1958. doi: 10.2307/2333180.

[3] George Casella and Edward I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[4] Andrew O. Finley, Sudipto Banerjee, and Bradley P. Carlin. spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19(4), 2007. doi: 10.18637/jss.v019.i04. URL http://www.jstatsoft.org/v19/i04/.

[5] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC, 3 edition edition, Nov 2013. ISBN 978-1-4398-4095-5.

[6] Peter D Hoff. *A first course in Bayesian statistical methods*. Springer Verlag, 2009. ISBN 978-0-387-92407-6. URL https://link.springer.com/openurl?genre=book&isbn=978-0-387-92299-7.

Bibliography II

[7] Harold Jeffreys. *Theory of probability*. Clarendon Press, 3d ed. edition, 1961.

[8] Peter M Lee. *Bayesian statistics: an introduction*. Arnold, 2004. ISBN 0 340 81405 5. URL http://www-users.york.ac.uk/~pml1/bayes/book.htm.

[9] Jean-Michel Marin and Christian P. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer Texts in Statistics. Springer New York, 2007. ISBN 978-0-387-38979-0. doi: 10.1007/978-0-387-38983-7.

[10] Andrew D. Martin and Kevin M. Quinn. Applied Bayesian inference in R using MCMCpack. *R News*, 6(1):2–7, 2006.

[11] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, Dec 2015. ISBN 978-1-4822-5344-3.

[12] Budiman Minasny and Alex B. McBratney. The Matérn function as a general model for soil variograms. *Geoderma*, 128(3–4):192–207, 2005. doi: 10.1016/j.geoderma.2005.04.003.

[13] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6(1): 7–11, 2006.

Bibliography III

[14] Jr. Ribeiro and Peter J. Diggle. geoR: A package for geostatistical analysis. *R News*, 1(2):14–18, 2001.

[15] A. F. M. Smith and G. O. Roberts. Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1):3–23, Jan 1993. doi: 10.2307/2346063.

End