UBC Spatial Stats Course II

Jürgen Pilz

Institut für Statistik
Universität Klagenfurt
Universitätsstr. 65-67, 9020 Klagenfurt, Austria
juergen.pilz@uni-klu.ac.at

November 2, 2010 / UBC Vancouver

Bayesian approach

- Classical statistics: model parameters are *fixed* and unknown.
- Bayesians think of parameters as *random*, and thus having distributions (just like the data). We can thus think about unknowns for which no reliable frequentist experiment exists, e.g. θ = proportion of men in some region with untreated prostate cancer.
- Bayesians write down a prior guess for parameter(s) θ, say $p(\theta)$, then combine this with the information provided by the observed data $\mathbf{y}$ to obtain the posterior distribution of θ, which we denote by$p(\theta|\mathbf{y})$.
- Statistical inference (point and interval estimation, hypothesis testing) then follow from posterior summaries. For example, the posterior means/medians/modes offer point estimates of θ, while the quantiles yield credible intervals.

- The key to Bayesian inference is “learning” or “updating” of prior beliefs via **Bayes’s Theorem**
- The classical approach is not wrong, but limited in scope. The Bayesian approach expands the class of models and easily handles settings that are precluded (or much more complicated) in classical settings.
- Moreover, we have **complete class theorems**
- (Maximum) likelihood results are often obtained as limiting Bayesian results w.r.t. vague or **non-informative priors**

Bayes Theorem

- We start with a model (likelihood) $p(\mathbf{y}|\theta)$ for the observed data $\mathbf{y} = (y_1, \ldots, y_n)^T$ given a vector of unknown parameters θ
- We add a prior (probability) density $p(\theta)$
- The posterior density of θ is then given by

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{\int_{\Theta} p(\mathbf{y}|\theta)p(\theta)d\theta} = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}$$

  where Θ denotes the parameter space
- This is often written as

$$\textbf{posterior} \propto \textbf{likelihood} * \textbf{prior}$$

  since $p(\mathbf{y})$ is just a *normalizing constant*

- Note: The normalizing constant (proportionality factor) is (usually) difficult to evaluate: multi-dimensional integration
- Computational tools such as Markov Chain Monte Carlo and software packages such as WinBUGS come to the rescue!
- All statistical inferences (point and interval estimates, hypothesis tests) then follow from posterior summaries. For example, the posterior means/medians/modes offer *point estimates*, while the quantiles yield *credible intervals*.

- Sometimes a further (hierarchical) step is invoked: If the prior is not completely known, but only up to some *hyperparameter*, say η, then we assign a prior, $p(\eta)$, and obtain a so-called **three stage hierarchical Bayes model**

$$\mathbf{y} \sim p(\mathbf{y}|\theta) \quad \theta \sim p(\theta|\eta) \quad \eta \sim p(\eta)$$

- We then seek:

$$p(\theta, \eta|y) \propto p(\eta) * p(\theta|\eta) * p(y|\theta)$$

- This represents the joint posterior from the hierarchical model. The marginal posterior distribution for θ is:

$$p(\theta|\mathbf{y}) = \int p(\eta) * p(\theta|\eta) * p(\mathbf{y}|\theta) d\eta$$

Books and Software

- Some good introductory books, e.g.
  Bolstad, W.M.: Introduction to Bayesian Statistics. 2nd ed., Wiley-Interscience 2007
  Gill, J.: Bayesian methods. 2nd ed. Chapman and Hall/CRC 2008
  Hoff, P.D.: A first course in Bayesian statistical methods. Springer 2009
- Books on Bayesian data analysis:
  Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B.: Bayesian Data Analysis, 2nd ed., Chapman and Hall/CRC 2003
  Carlin, B.P. and Louis, T.A.: Bayesian Methods for Data Analysis, 3rd ed., Chapman and Hall/CRC 2008
  Efron, B.: Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge Univ. Press 2010
  Congdon, P.: Applied Bayesian Hierarchical Methods. Chapman and Hall/CRC 2010

Books cont'd.

- Bayesian computation in R:
  Albert, J.: Bayesian computation with R. 2nd ed., Springer 2009
  Marin, J.-M. and Robert,Ch.: Bayesian Core: A Practical Approach to Computational Bayesian Statistics. Springer 2008
  Ntzoufras, I.: Bayesian Modeling Using WinBUGS. Wiley 2009
- Advanced books:
  O'Hagan, A. and Forster, J. J.: Bayesian Inference, 2nd edition, volume 2B of "Kendall's Advanced Theory of Statistics". Arnold, London 2004
  Bernardo, J.M. and Smith, A.F.M.: Bayesian Theory. Wiley 1994
  Robert, Ch.: The Bayesian Choice. 2nd ed. , Springer 2007

Software for Bayesian Analysis

- Generic packages with R-links:
  - JAGS (R2jags)
  - OpenBUGS
  - WinBUGS
  - R2WinBUGS calls WinBUGS on Windows and Linux systems

  WinBUGS software downloadable from:
  www.mrc-bsu.cam.ac.uk/bugs/
- Bayes linear and generalized linear (mixed) models:
  - MCMCpack
  - MCMCglmm

- Bayesian spatial analysis for Gaussian RF
  - spBayes (incl. multivariate settings) and
  - ramps (reparameterized and marginalized posterior sampling)
- Bayesian geostatistics including non-Gaussian RFs:
  - geoR and geoRglm
  - intamap (and psgp)
- R packages for learning and first steps:
  - Bolstad (companion to the book) and
  - LearnBayes (1- and 2-param. prpblems)

Conjugate priors

Specific parametric family with analytical properties

Definition : A family $\mathcal{P}$ of probability distributions on Θ is conjugate for a likelihood function $p(\mathbf{y}|\theta)$ if, for every $p(\theta) \in \mathcal{P}$, the posterior $p(\theta|\mathbf{y})$ also belongs to $\mathcal{P}$.

Only of interest when $\mathcal{P}$ is parameterized : switching from prior to posterior distribution is reduced to an updating of the corresponding parameters.

This is convenient for *Exponential families*!

Conjugacy in Exponential families

Normal (*known variance*) : *Normal*

Full Normal: *Normal* − *Gamma* ($\theta = (mean, precision)$)

Poisson : *Gamma*

Gamma (*known shape*) : *Gamma* ($\theta = rate$)

Binomial : *Beta*

Negative Binomial : *Beta*

Multinomial : *Dirichlet*

Ex. 1: Normal observations, known variance

$y \sim \mathcal{N}(\theta, \sigma^2)$; assume σ^2 is known

$\theta \sim \mathcal{N}(\mu, \tau^2)$ = conjugate prior for θ

With i.i.d. observations $\mathbf{y} = (y_1, \ldots, y_n)^T$ we get

$$p(\mathbf{y}|\theta) \propto exp(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta)^2) \propto exp(-\frac{n}{2\sigma^2}(\theta - \bar{y})^2)$$

where $\bar{y}$ is the sample mean, and thus

$$p(\theta|\mathbf{y}) \propto exp(-\frac{n}{2\sigma^2}(\theta - \bar{y})^2) * exp(-\frac{1}{2\tau^2}(\theta - \mu)^2)$$

$$\propto exp(-\frac{a}{2}(\theta - \bar{y})^2) * exp(-\frac{b}{2}(\theta - \mu)^2)$$

Here $a = n/\sigma^2$ is the *sample precision*

and $b = 1/\tau^2$ is the *prior precision*

This finally leads to

$p(\theta|\mathbf{y}) \propto exp(-\frac{1}{2}\hat{\tau}^{-2}(\theta - \hat{\mu})^2)$
i.e. the posterior distribution of θ is normal with mean

$$\hat{\mu} = \frac{a * \bar{y} + b * \mu}{a + b} = \frac{\frac{n}{\sigma^2} * \bar{y} + \frac{1}{\tau^2} * \mu}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

and variance

$$\hat{\tau}^2 = (a + b)^{-1} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

Interpretation

- posterior mean $\hat{\mu}$ is **weighted mean** of sample mean $\bar{y}$ and prior mean μ, with weights equal to the corresponding precisions a and b, resp.

- posterior mean is a **convex combination** of sample mean and prior mean,

$$\hat{\mu} = \alpha * \bar{y} + (1 - \alpha) * \mu$$

  with $\alpha = a/(a+b) \in (0, 1)$.

- posterior variance is the **harmonic mean** of sample variance a^{-1} and prior variance b^{-1},

$$\hat{\tau}^2 = \frac{1}{1/a^{-1} + 1/b^{-1}}$$

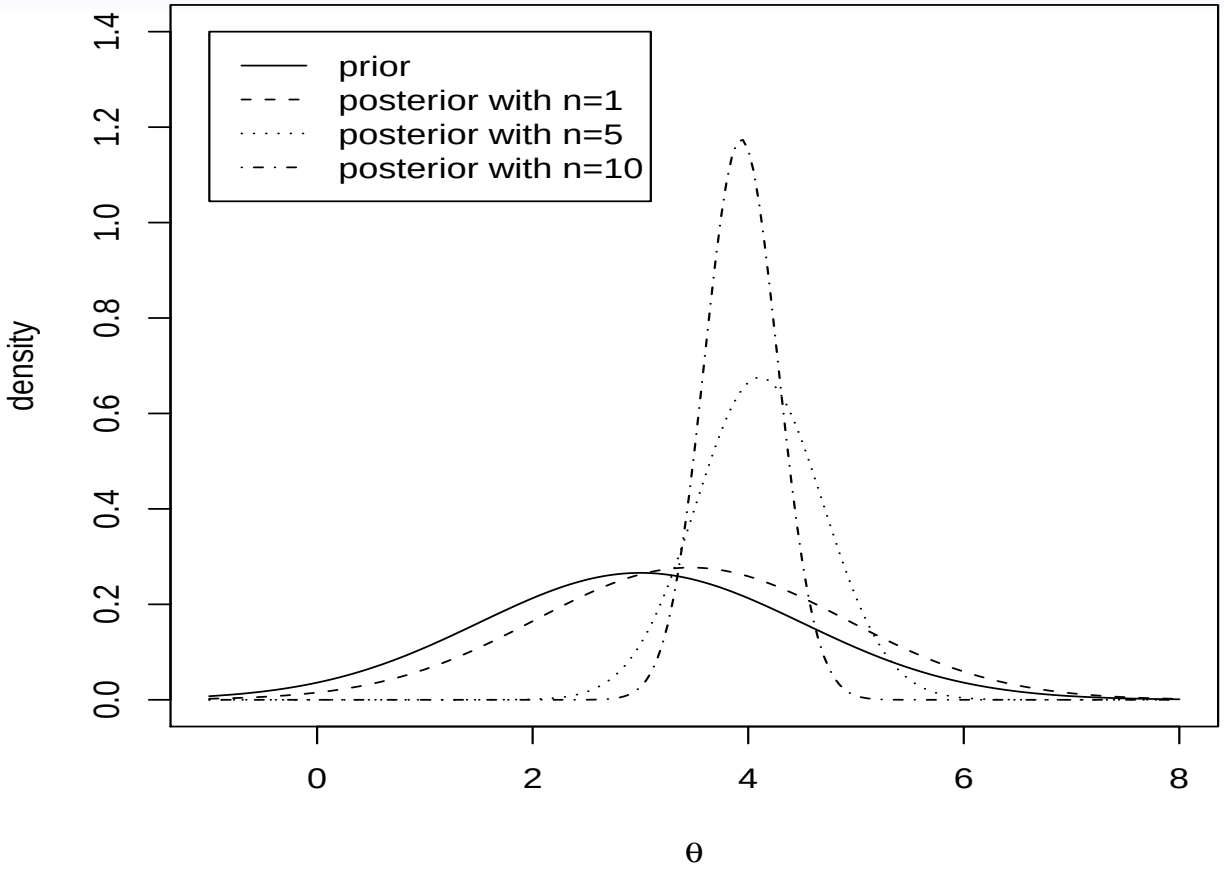

Figure: Normal observations and normal prior

Ex. 2: Poisson distribution: $y \sim \mathcal{P}(\theta = \lambda)$

With i.i.d. observations $\mathbf{y} = (y_1, \ldots, y_n)^T$ we get

$$p(\mathbf{y}|\theta) = \theta^{\sum y_i} * exp(-n\theta) / \prod y_i!$$

Considered as a function of θ, this is the density of a Gamma

distribution $\mathcal{G}(1 + \sum y_i, n)$, therefore we assume a conjugate **Gamma prior** :

$$\theta \sim \mathcal{G}(shape = a, rate = b) : p(\theta) \propto \theta^{a-1} exp(-b\theta)$$

This implies a gamma posterior:

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta) * p(\theta) \propto \theta^{\sum y_i + a - 1} * e^{-(b+n)\theta}$$

We thus have

$$\theta|\mathbf{y} \sim \mathcal{G}(a + \sum y_i, b + n)$$

Homework I:

Derive the mean, mode and variance of this posterior distribution and discuss the results for various choices of a, b and n, using simulated Poisson data.

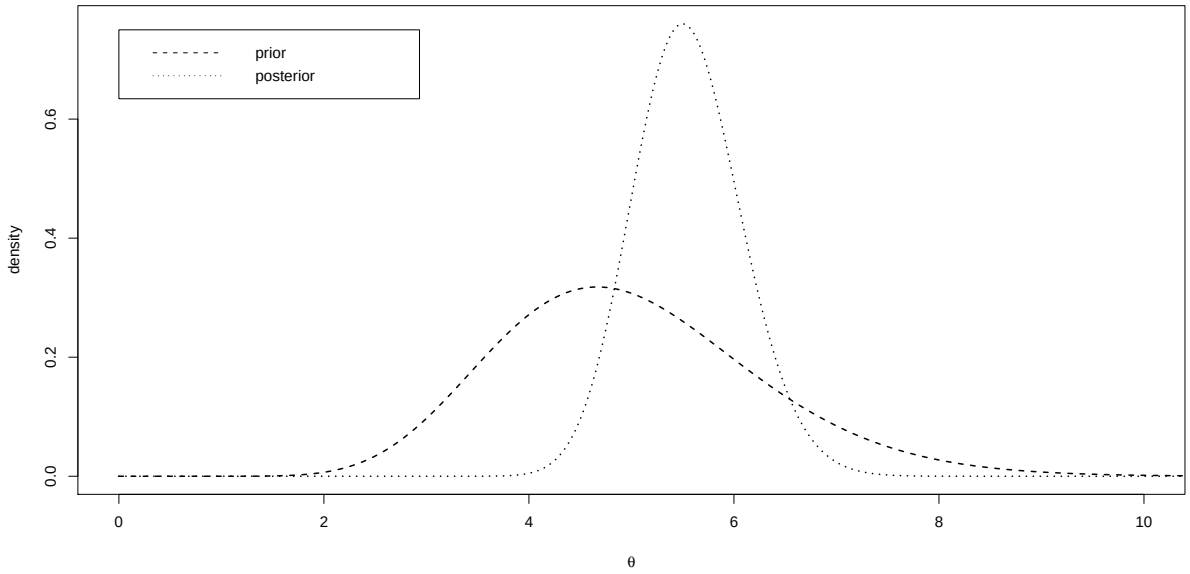

Figure: Poisson d. observations and Gamma prior

Improper prior distribution

Extension from a prior distribution to a prior σ -finite measure on Θ such that

$$\int_{\Theta} p(\theta) d\theta = +\infty$$

Justifications

- Often only way to derive a prior in noninformative/automatic settings
- Performances of associated estimators usually good
- Often occur as limits of proper distributions
- More robust answer against possible misspecifications of the prior
- Improper priors (infinitely!) preferable to vague proper priors such as a $\mathcal{N}(0, 100^2)$ distribution [e.g. in BUGS]

Jeffreys' prior

Based on Fisher information the Jeffreys prior distribution is defined by

$$p(\theta) \propto [det(IF(\theta))]^{1/2}$$

Pros and Cons

- Relates to information theory
- Agrees with most invariant priors
- Parameterization invariant
- Suffers from dimensionality curse

We require, however, for improper and/or Jeffreys' prior:

$$\int_{\Theta} p(\mathbf{y}|\theta) * p(\theta) d\theta < \infty$$

Evaluating estimators

- Purpose of most inferential studies: to provide the statistician/client with a **decision** d
- Requires an evaluation criterion, called **loss function** for decisions and estimators: $L(\theta, d)$
- There exists an axiomatic derivation of the existence of a loss function, which goes back to *DeGroot*, 1970.

Bayesian estimation principle:

Integrate over Θ to get the **posterior expected loss**

$$E_{p(\theta|\mathbf{y}}[L(\theta, d)|\mathbf{y}] = \int_{\Theta} L(\theta, d) * p(\theta|\mathbf{y}) d\theta$$

and minimize w.r.t. d

Bayes estimator for **quadratic loss**
$L(\theta, d) = (\theta - d)^2$

Historically, this loss function had already been used by Legendre, Gauss, and Laplace
Least squares principle!

The minimizer is just the **posterior mean**

$$\hat{\theta}(\mathbf{y}) = \int_{\Theta} \theta * p(\theta|\mathbf{y}) d\theta = \frac{\int_{\Theta} \theta * p(\mathbf{y}|\theta) * p(\theta) d\theta}{\int_{\Theta} p(\mathbf{y}|\theta) * p(\theta) d\theta}$$

MAP estimator and Credible region

- The maximum a posteriori (**MAP**) estimator is often used as an alternative:

$$\hat{\theta}(\mathbf{y}) = \arg\max_{\theta} p(\mathbf{y}|\theta) * p(\theta)$$

  It can be considered as a Penalized likelihood estimator.

- A *credible region* for θ is any region $C \subset R^k$ such that

$$P(\theta \in C|\mathbf{y}) \geq 1 - \alpha$$

  where $1 - \alpha$ is a given level of credibility, $\alpha \in (0, 1)$.

HPD- and Equal tail regions

- Highest posterior density (**HPD**) region: For given α, find the largest k_α such that

$$C_\alpha = \{\theta : p(\theta|\mathbf{y}) > k_\alpha\}; P(\theta \in C_\alpha|\mathbf{y}) \geq 1 - \alpha$$

  The HPD regions give the highest probabilities of containing θ for a given volume.

- Simpler alternative: Equal-tailed credible interval

$$C_\alpha = (q_{\alpha/2}, q_{1-\alpha/2})$$

  where $q_{\alpha/2}$ and $q_{1-\alpha/2}$ are the $\alpha/2$ and $1 - \alpha/2$- quantiles of the posterior distribution of θ, resp.

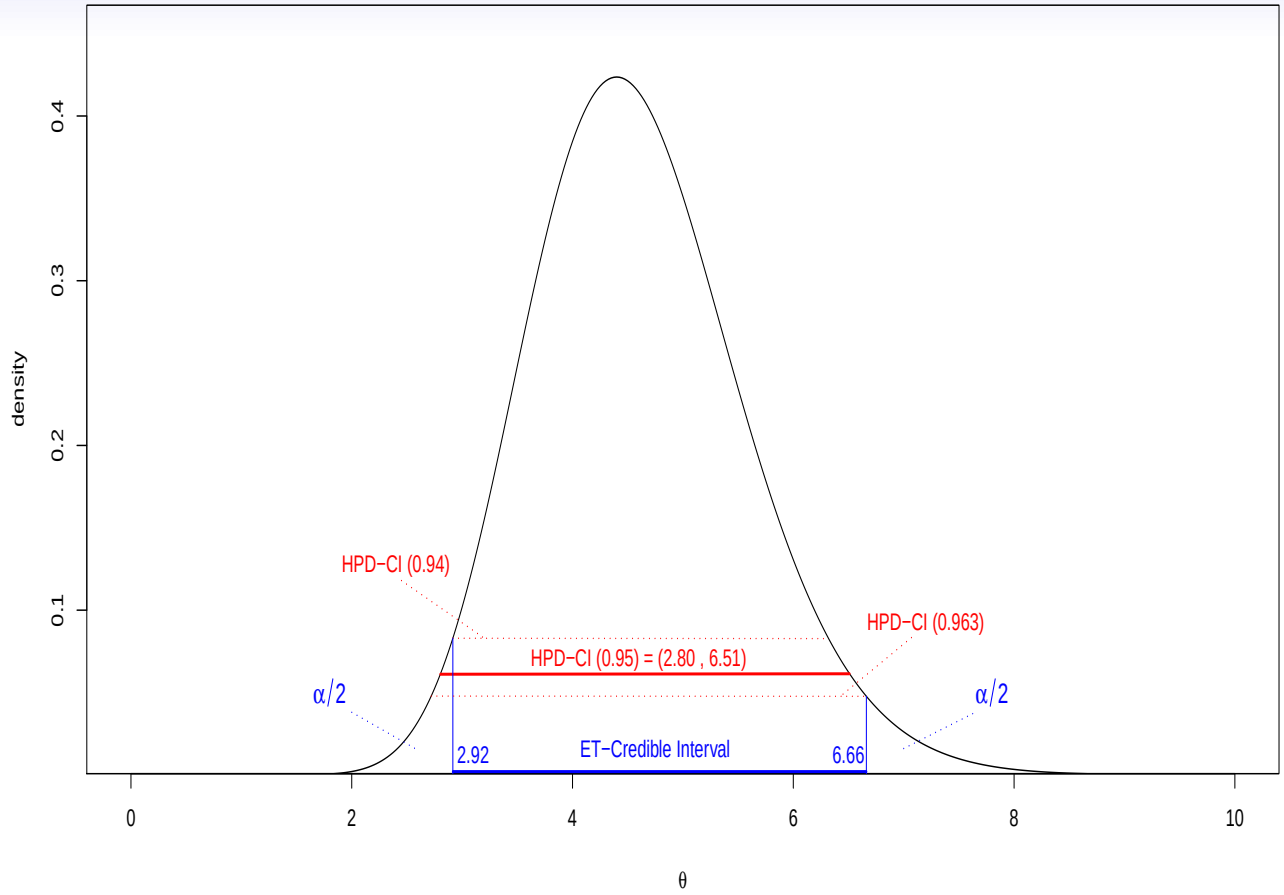

Figure: Equal-tailed and HPD- Credible Intervals

Monte Carlo integration

- Generic problem of evaluating an integral

$$I = E_f[h(\theta)] = \int_{\Theta} h(\theta) * f(\theta) d\theta$$

  where Θ is uni- or multidimensional, f is a closed form, partly closed form, or implicit (prior or posterior) density, and h is a given function

- Monte Carlo Principle: Use a sample $\theta_1, \ldots, \theta_N$ from the density f to approximate the integral I by the empirical average $h_N = (1/N) \sum_{i=1}^{N} h(\theta_i)$
- Convergence of the average $h_N \rightarrow E_f[h(\theta)]$ follows from the "Strong Law of Large Numbers"

Importance sampling

- Simulation from f (the true density) is not necessarily optimal
  Alternative to direct sampling from f is importance sampling, based on the alternative representation

$$E_f[h(\theta)] = \int_\Theta \frac{h(\theta)f(\theta)}{g(\theta)} * g(\theta) = E_g[hf/g]$$

  which allows us to use other distributions than f

- IS-Algorithm : Generate a sample $\theta_1, \ldots, \theta_N$ from a distribution g
  Use the approximation $\frac{1}{N}\sum_{i=1}^{N} \frac{f(\theta_i)}{g(\theta_i)} * h(\theta_i)$
- Approx. converges for any choice of the distribution g as long as

$$supp(g) \supset supp(f)$$

Ex. 3: Bayes linear regression

Assume

$$\mathbf{y} = X\theta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

with known variance σ^2. The likelihood function is then given by

$$p(\mathbf{y}|\theta) \propto exp(-\frac{1}{2\sigma^2}(\mathbf{y} - X\theta)^T(\mathbf{y} - X\theta))$$

Recalling the decomposition

$$(\mathbf{y} - X\theta)^T(\mathbf{y} - X\theta) = (\theta - \hat{\theta})^T X^T X(\theta - \hat{\theta}) + \mathbf{y}^T A\mathbf{y}$$

where $A = I_n - X(X^T X)^{-1} X^T$ is the projector matrix, it becomes clear that the conjugate prior for θ is normal, too.

Bayes linear regression cont'd

Therefore, assume $\theta \sim \mathcal{N}(\mu, \Phi)$ and obtain

$$p(\theta|\mathbf{y}) \propto exp(-\frac{1}{2\sigma^2}(\theta-\hat{\theta})^T X^T X(\theta-\hat{\theta})) * exp(-\frac{1}{2}(\theta-\mu)^T \Phi^{-1}(\theta-\mu))$$

$$\propto exp(-\frac{1}{2}(\theta-\hat{\theta}_B)^T(\sigma^{-2}X^T X + \Phi^{-1})(\theta-\hat{\theta}_B))$$

where

$$\hat{\theta}_B = (\sigma^{-2}X^T X + \Phi^{-1})^{-1}(\sigma^{-2}X^T \mathbf{y} + \Phi^{-1}\mu)$$

Obviously, the posterior of θ is normal, $\theta|\mathbf{y} \sim \mathcal{N}(\hat{\theta}_B, V_B)$, with posterior mean $\hat{\theta}_B$ and posterior covariance matrix

$$V_B = (\sigma^{-2}X^T X + \Phi^{-1})^{-1}$$

Bayes kriging predictor

Homework II: Compute the Bayes kriging predictor for the "meuse" data, using

- geoR
- spBayes

Do this for the case of known estimated variance (sill) and known range and smoothness, using both a normal and a non-informative prior for the regression parameter $\theta = \beta$.

Note : The Bayes kriging predictor is identical to the Universal kriging predictor with the GLS $\hat{\beta}$ replaced by the Bayes linear regression estimator $\hat{\theta}_B$.