Model uncertainty and model choice: Bayesian tools

Christian P. Robert

Université Paris Dauphine and CREST-INSEE
http://www.ceremade.dauphine.fr/~xian

Journées Suisses de Statistique/Schweizer Statistiktage, Zurich
November 11, 2005

Outline

1. Bayesian Model Choice
2. Compatible priors for variable selection
3. k-nearest-neighbour classification

1 Bayesian Model Choice

1. Bayesian Model Choice
   - Introduction
   - Bayesian resolution
   - Problems
   - Bayes factors
2. Compatible priors for variable selection
3. k-nearest-neighbour classification

[Joint book with J.M. Marin]

Setup

Choice of models

Several models available for the same observation

$$\mathfrak{M}_i : x \sim f_i(x|\theta_i), \qquad i \in \mathfrak{I}$$

where $\mathfrak{I}$ can be finite or infinite

Bayesian resolution

Bayesian Framework

Probabilises the entire model/parameter space

Bayesian resolution

Bayesian Framework

Probabilises the entire model/parameter space

This means:

- allocating probabilities p_i to all models $\mathfrak{M}_i$
- defining priors $\pi_i(\theta_i)$ for each parameter space Θ_i

Formal solution

Resolution

1. Compute

$$p(\mathfrak{M}_i|x) = \frac{p_i \displaystyle\int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\displaystyle\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j}$$

Formal solution

Resolution

1. Compute

$$p(\mathfrak{M}_i|x) = \frac{p_i \displaystyle\int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\displaystyle\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j}$$

2. Take largest $p(\mathfrak{M}_i|x)$ to determine ‘‘best’’ model, or use averaged predictive

$$\sum_j p(\mathfrak{M}_j|x) \int_{\Theta_j} f_j(x'|\theta_j)\pi_j(\theta_j|x)\mathrm{d}\theta_j$$

Several types of problems

- Concentrate on selection perspective:
  - averaging = estimation = non-parsimonious = no-decision
  - how to integrate loss function/decision/consequences

Several types of problems

- Concentrate on selection perspective:
  - averaging = estimation = non-parsimonious = no-decision
  - how to integrate loss function/decision/consequences
  - representation of parsimony/sparcity (Occam's rule)
  - how to fight overfitting for nested models

Several types of problems

- Concentrate on selection perspective:
  - averaging = estimation = non-parsimonious = no-decision
  - how to integrate loss function/decision/consequences
  - representation of parsimony/sparcity (Occam's rule)
  - how to fight overfitting for nested models

Which loss function?

Several types of problems (2)

- Choice of prior structures
  - adequate weights p_i:
    if $\mathfrak{M}_1 = \mathfrak{M}_2 \cup \mathfrak{M}_3$,

Several types of problems (2)

- Choice of prior structures
  - adequate weights p_i:
    if $\mathfrak{M}_1 = \mathfrak{M}_2 \cup \mathfrak{M}_3$, $p(\mathfrak{M}_1) = p(\mathfrak{M}_2) + p(\mathfrak{M}_3)$?
  - priors distributions
    - $\pi_i(\theta_i)$ defined for every $i \in \mathfrak{I}$

Several types of problems (2)

- Choice of prior structures
    - adequate weights p_i:
      if $\mathfrak{M}_1 = \mathfrak{M}_2 \cup \mathfrak{M}_3$, $p(\mathfrak{M}_1) = p(\mathfrak{M}_2) + p(\mathfrak{M}_3)$?
    - priors distributions
        - $\pi_i(\theta_i)$ defined for every $i \in \mathfrak{I}$
        - $\pi_i(\theta_i)$ *proper* (Jeffreys)

Several types of problems (2)

- Choice of prior structures
  - adequate weights p_i:
    if $\mathfrak{M}_1 = \mathfrak{M}_2 \cup \mathfrak{M}_3$, $p(\mathfrak{M}_1) = p(\mathfrak{M}_2) + p(\mathfrak{M}_3)$?
  - priors distributions
    - $\pi_i(\theta_i)$ defined for every $i \in \mathfrak{I}$
    - $\pi_i(\theta_i)$ *proper* (Jeffreys)
    - $\pi_i(\theta_i)$ coherent (?) for nested models

Several types of problems (2)

- Choice of prior structures
  - adequate weights p_i:
    if $\mathfrak{M}_1 = \mathfrak{M}_2 \cup \mathfrak{M}_3$, $p(\mathfrak{M}_1) = p(\mathfrak{M}_2) + p(\mathfrak{M}_3)$?
  - priors distributions
    - $\pi_i(\theta_i)$ defined for every $i \in \mathfrak{I}$
    - $\pi_i(\theta_i)$ *proper* (Jeffreys)
    - $\pi_i(\theta_i)$ coherent (?) for nested models

Warning

Parameters common to several models must be treated as separate entities!

Several types of problems (3)

- Computation of predictives and marginals
  - infinite dimensional spaces

Several types of problems (3)

- Computation of predictives and marginals
  - infinite dimensional spaces
  - integration over parameter spaces

Several types of problems (3)

- Computation of predictives and marginals
  - infinite dimensional spaces
  - integration over parameter spaces
  - integration over different spaces

Several types of problems (3)

- Computation of predictives and marginals
  - infinite dimensional spaces
  - integration over parameter spaces
  - integration over different spaces
  - summation over (too) many models (2^k)

[MCMC resolution = another talk]

A function of posterior probabilities

Definition (Bayes factors)

Models $\mathfrak{M}_1$ vs. $\mathfrak{M}_2$

$$B_{12} = \frac{\Pr(\mathcal{M}_1|x)}{\Pr(\mathcal{M}_2|x)} \bigg/ \frac{\Pr(\mathcal{M}_1)}{\Pr(\mathcal{M}_2)}$$

$$= \frac{\displaystyle\int f_1(x|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}{\displaystyle\int f_2(x|\theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2}$$

[Good, 1958 & Jeffreys, 1961]

▸ Goto Poisson example

Self-contained concept

- eliminates choice of $\Pr(\mathfrak{M}_i)$

Self-contained concept

- eliminates choice of $\Pr(\mathfrak{M}_i)$
- but depends on the choice of $\pi_i(\theta_i)$

Self-contained concept

- eliminates choice of $\Pr(\mathfrak{M}_i)$
- but depends on the choice of $\pi_i(\theta_i)$
- Bayesian/marginal likelihood ratio

Self-contained concept

- eliminates choice of $\Pr(\mathfrak{M}_i)$
- but depends on the choice of $\pi_i(\theta_i)$
- Bayesian/marginal likelihood ratio
- Jeffreys' scale of evidence

A difficulty

Improper priors not allowed here

If

$$\int_{\Theta_1} \pi_1(\mathrm{d}\theta_1) = \infty \quad \text{or} \quad \int_{\Theta_2} \pi_2(\mathrm{d}\theta_2) = \infty$$

then either π_1 or π_2 cannot be normalised uniquely

A difficulty

Improper priors not allowed here

If

$$\int_{\Theta_1} \pi_1(d\theta_1) = \infty \quad \text{or} \quad \int_{\Theta_2} \pi_2(d\theta_2) = \infty$$

then either π_1 or π_2 cannot be normalised uniquely but the normalisation matters in the Bayes factor

◂ Recall Bayes factor

Constants matter

Example (Poisson versus Negative binomial)

If $\mathfrak{M}_1$ is a $\mathscr{P}(\lambda)$ distribution and $\mathfrak{M}_2$ is a $\mathscr{N}\mathscr{B}(m,p)$ distribution, we can take

$$\begin{aligned}
\pi_1(\lambda) &= 1/\lambda \\
\pi_2(m,p) &= \frac{1}{M}\,\mathbb{I}_{\{1,\cdots,M\}}(m)\,\mathbb{I}_{[0,1]}(p)
\end{aligned}$$

Constants matter (cont'd)

Example (Poisson versus Negative binomial (2))

then

$$
\begin{aligned}
B_{12} &= \frac{\displaystyle\int_0^\infty \frac{\lambda^{x-1}}{x!} e^{-\lambda} \mathrm{d}\lambda}{\displaystyle\frac{1}{M}\sum_{m=1}^{M}\int_0^\infty \binom{m}{x-1} p^x (1-p)^{m-x} dp} \\
&= 1 \Bigg/ \frac{1}{M}\sum_{m=x}^{M}\binom{m}{x-1}\frac{x!(m-x)!}{m!} \\
&= 1 \Bigg/ \frac{1}{M}\sum_{m=x}^{M} x/(m-x+1)
\end{aligned}
$$

Constants matter (cont'd)

Example (Poisson versus Negative binomial (3))

- does not make sense because $\pi_1(\lambda) = 10/\lambda$ leads to a different answer, **ten times larger!**

Constants matter (cont'd)

Example (Poisson versus Negative binomial (3))

- does not make sense because $\pi_1(\lambda) = 10/\lambda$ leads to a different answer, **ten times larger!**
- same thing when both priors are improper

Constants matter (cont'd)

Example (Poisson versus Negative binomial (3))

- does not make sense because $\pi_1(\lambda) = 10/\lambda$ leads to a different answer, **ten times larger!**
- same thing when both priors are improper

Note

Improper priors on common (nuisance) parameters do not matter (so much)

Vague proper priors are not the solution

▸ To compatible priors

Taking a proper prior and take a “very large” variance (e.g., BUGS)

Vague proper priors are not the solution

▸ To compatible priors

Taking a proper prior and take a “very large” variance (e.g., BUGS) will most often result in an undefined or ill-defined limit

Vague proper priors are not the solution

▸ To compatible priors

Taking a proper prior and take a "very large" variance (e.g., BUGS) will most often result in an undefined or ill-defined limit

Example (Lindley's paradox)

If testing $H_0 : \theta = 0$ when observing $x \sim \mathcal{N}(\theta, 1)$, under a normal $\mathcal{N}(0, \alpha)$ prior $\pi_1(\theta)$,

$$B_{01}(x) \xrightarrow{\alpha \longrightarrow \infty} 0$$

Vague proper priors are not the solution (cont'd)

Example (Poisson versus Negative binomial (4))

$$
\begin{aligned}
B_{12} &= \frac{\displaystyle\int_0^1 \frac{\lambda^{\alpha+x-1}}{x!} e^{-\lambda\beta} \mathrm{d}\lambda}{\displaystyle\frac{1}{M}\sum_m \frac{x}{m-x+1}\,\frac{\beta^\alpha}{\Gamma(\alpha)}} \quad \text{if } \lambda \sim \mathcal{G}a(\alpha,\beta) \\
&= \frac{\Gamma(\alpha+x)}{x!\,\Gamma(\alpha)}\beta^{-x} \Big/ \frac{1}{M}\sum_m \frac{x}{m-x+1} \\
&= \frac{(x+\alpha-1)\cdots\alpha}{x(x-1)\cdots 1}\beta^{-x} \Big/ \frac{1}{M}\sum_m \frac{x}{m-x+1}
\end{aligned}
$$

Vague proper priors are not the solution (cont'd)

Example (Poisson versus Negative binomial (4))

$$
\begin{aligned}
B_{12} &= \frac{\displaystyle\int_0^1 \frac{\lambda^{\alpha+x-1}}{x!} e^{-\lambda\beta} \mathrm{d}\lambda}{\displaystyle\frac{1}{M}\sum_m \frac{x}{m-x+1}\,\frac{\beta^\alpha}{\Gamma(\alpha)}} \quad \text{if } \lambda \sim \mathcal{G}a(\alpha,\beta) \\
&= \frac{\Gamma(\alpha+x)}{x!\,\Gamma(\alpha)}\beta^{-x} \bigg/ \frac{1}{M}\sum_m \frac{x}{m-x+1} \\
&= \frac{(x+\alpha-1)\cdots\alpha}{x(x-1)\cdots 1}\beta^{-x} \bigg/ \frac{1}{M}\sum_m \frac{x}{m-x+1}
\end{aligned}
$$

depends on choice of $\alpha(\beta)$ or $\beta(\alpha) \longrightarrow 0$

2 Compatible priors

1. Bayesian Model Choice
2. Compatible priors for variable selection
   - Principle
   - Linear regression
   - Variable selection
   - Application
3. k-nearest-neighbour classification

[Joint work with C. Celeux, G. Consonni and J.M. Marin]

Principle

Difficult to simultaneously find priors on a collection of models $\mathfrak{M}_i$ $(i \in \mathfrak{I})$

Principle

Difficult to simultaneously find priors on a collection of models $\mathfrak{M}_i$ $(i \in \mathfrak{I})$

Easier to start from a single prior on a "big" model and to derive the other priors from a coherence principle

[Dawid & Lauritzen, 2000]

Projection approach

For $\mathfrak{M}_2$ submodel of $\mathfrak{M}_1$, π_2 can be derived as the distribution of $\theta_2^{\perp}(\theta_1)$ when $\theta_1 \sim \pi_1(\theta_1)$ and $\theta_2^{\perp}(\theta_1)$ is a projection of θ_1 on $\mathfrak{M}_2$, e.g.

$$d(f(\cdot\,|\theta_1), f(\cdot\,|{\theta_1}^{\perp})) = \inf_{\theta_2 \in \Theta_2} d(f(\cdot\,|\theta_1)\,, f(\cdot\,|\theta_2))\,.$$

where d is a divergence measure

[McCulloch & Rossi, 1992]

Projection approach

For $\mathfrak{M}_2$ submodel of $\mathfrak{M}_1$, π_2 can be derived as the distribution of $\theta_2^{\perp}(\theta_1)$ when $\theta_1 \sim \pi_1(\theta_1)$ and $\theta_2^{\perp}(\theta_1)$ is a projection of θ_1 on $\mathfrak{M}_2$, e.g.

$$d(f(\cdot\,|\theta_1), f(\cdot\,|{\theta_1}^{\perp})) = \inf_{\theta_2 \in \Theta_2} d(f(\cdot\,|\theta_1)\,, f(\cdot\,|\theta_2))\,.$$

where d is a divergence measure

[McCulloch & Rossi, 1992]

Or we can look instead at the posterior distribution of

$$d(f(\cdot\,|\theta_1), f(\cdot\,|{\theta_1}^{\perp}))$$

[Goutis & Robert, 1998]

Kullback proximity

Alternative solution

Definition (Compatible prior)

Given a prior π_1 on a model $\mathfrak{M}_1$ and a submodel $\mathfrak{M}_2$, a prior π_2 on $\mathfrak{M}_2$ is *compatible* with π_1

Kullback proximity

Alternative solution

Definition (Compatible prior)

Given a prior π_1 on a model $\mathfrak{M}_1$ and a submodel $\mathfrak{M}_2$, a prior π_2 on $\mathfrak{M}_2$ is *compatible* with π_1 when it achieves the minimum Kullback divergence between the corresponding marginals:
$m_1(x;\pi_1) = \int_{\Theta_1} f_1(x|\theta)\pi_1(\theta)\mathrm{d}\theta$ and
$m_2(x);\pi_2 = \int_{\Theta_2} f_2(x|\theta)\pi_2(\theta)\mathrm{d}\theta$,

Kullback proximity

Alternative solution

Definition (Compatible prior)

Given a prior π_1 on a model $\mathfrak{M}_1$ and a submodel $\mathfrak{M}_2$, a prior π_2 on $\mathfrak{M}_2$ is *compatible* with π_1 when it achieves the minimum Kullback divergence between the corresponding marginals:
$m_1(x;\pi_1) = \int_{\Theta_1} f_1(x|\theta)\pi_1(\theta)\mathrm{d}\theta$ and
$m_2(x);\pi_2 = \int_{\Theta_2} f_2(x|\theta)\pi_2(\theta)\mathrm{d}\theta$,

$$\pi_2 = \arg\min_{\pi_2} \int \log\left(\frac{m_1(x;\pi_1)}{m_2(x;\pi_2)}\right) m_1(x;\pi_1)\,\mathrm{d}x$$

Difficulties

- Does not give a working principle when $\mathfrak{M}_2$ is not a submodel $\mathfrak{M}_1$

Difficulties

- Does not give a working principle when $\mathfrak{M}_2$ is not a submodel $\mathfrak{M}_1$
- Depends on the choice of π_1

Difficulties

- Does not give a working principle when $\mathfrak{M}_2$ is not a submodel $\mathfrak{M}_1$
- Depends on the choice of π_1
- Prohibits the use of improper priors

Difficulties

- Does not give a working principle when $\mathfrak{M}_2$ is not a submodel $\mathfrak{M}_1$
- Depends on the choice of π_1
- Prohibits the use of improper priors
- Worse: useless in unconstrained settings...

Linear regression

$\mathfrak{M}_1$ and $\mathfrak{M}_2$ are two nested Gaussian linear regression models with Zellner's g-priors and the same variance $\sigma^2 \sim \pi(\sigma^2)$:

1. $\mathfrak{M}_1$:

   $$y|\beta_1, \sigma^2 \sim \mathcal{N}(X_1\beta_1, \sigma^2), \quad \beta_1|\sigma^2 \sim \mathcal{N}\left(s_1, \sigma^2 n_1 (X_1^\mathsf{T} X_1)^{-1}\right)$$

   where X_1 is a $(n \times k_1)$ matrix of rank $k_1 \leq n$
2. $\mathfrak{M}_2$:

   $$y|\beta_2, \sigma^2 \sim \mathcal{N}(X_2\beta_2, \sigma^2), \quad \beta_2|\sigma^2 \sim \mathcal{N}\left(s_2, \sigma^2 n_2 (X_2^\mathsf{T} X_2)^{-1}\right),$$

   where X_2 is a $(n \times k_2)$ matrix with span$(X_2) \subseteq$ span(X_1)

Compatible g-priors

Since σ^2 is a nuisance parameter, we can minimize the Kullback-Leibler divergence between the two marginal distributions conditional on σ^2: $m_1(y|\sigma^2; s_1, n_1)$ and $m_2(y|\sigma^2; s_2, n_2)$

Compatible g-priors

Since σ^2 is a nuisance parameter, we can minimize the Kullback-Leibler divergence between the two marginal distributions conditional on σ^2: $m_1(y|\sigma^2; s_1, n_1)$ and $m_2(y|\sigma^2; s_2, n_2)$

Theorem

Conditional on σ^2, the conjugate compatible prior of $\mathfrak{M}_2$ wrt $\mathfrak{M}_1$ is

$$\beta_2|X_2, \sigma^2 \sim \mathcal{N}\left(s_2^*, \sigma^2 n_2^* (X_2^T X_2)^{-1}\right)$$

with

$$\begin{aligned} s_2^* &= (X_2^T X_2)^{-1} X_2^T X_1 s_1 \\ n_2^* &= n_1 \end{aligned}$$

Variable selection

Regression setup where y regressed on a set $\{x_1, \ldots, x_p\}$ of p **potential explanatory** regressors (plus intercept)

Variable selection

Regression setup where y regressed on a set $\{x_1, \dots, x_p\}$ of p **potential explanatory** regressors (plus intercept)

Corresponding 2^p submodels $\mathfrak{M}_\gamma$, where $\gamma \in \Gamma = \{0,1\}^p$ indicates inclusion/exclusion of variables by a binary representation,

Variable selection

Regression setup where y regressed on a set $\{x_1, \ldots, x_p\}$ of p **potential explanatory** regressors (plus intercept)

Corresponding 2^p submodels $\mathfrak{M}_\gamma$, where $\gamma \in \Gamma = \{0, 1\}^p$ indicates inclusion/exclusion of variables by a binary representation, e.g. $\gamma = 101001011$

Notations

For model $\mathfrak{M}_\gamma$,

- q_γ variables included
- $t_1(\gamma) = \{t_{1,1}(\gamma), \ldots, t_{1,q_\gamma}(\gamma)\}$ indices of those variables and $t_0(\gamma)$ indices of the variables *not* included
- For $\beta \in \mathbb{R}^{p+1}$,

$$\begin{aligned} \beta_{t_1(\gamma)} &= \left[\beta_0, \beta_{t_{1,1}(\gamma)}, \ldots, \beta_{t_{1,q_\gamma}(\gamma)}\right] \\ X_{t_1(\gamma)} &= \left[1_n \middle| x_{t_{1,1}(\gamma)} \middle| \ldots \middle| x_{t_{1,q_\gamma}(\gamma)}\right] . \end{aligned}$$

Notations

For model $\mathfrak{M}_\gamma$,

- q_γ variables included
- $t_1(\gamma) = \{t_{1,1}(\gamma), \ldots, t_{1,q_\gamma}(\gamma)\}$ indices of those variables and $t_0(\gamma)$ indices of the variables *not* included
- For $\beta \in \mathbb{R}^{p+1}$,

$$\begin{aligned}
\beta_{t_1(\gamma)} &= \left[\beta_0, \beta_{t_{1,1}(\gamma)}, \ldots, \beta_{t_{1,q_\gamma}(\gamma)}\right] \\
X_{t_1(\gamma)} &= \left[1_n | x_{t_{1,1}(\gamma)} | \ldots | x_{t_{1,q_\gamma}(\gamma)}\right] .
\end{aligned}$$

Submodel $\mathfrak{M}_\gamma$ is thus

$$y|\beta, \gamma, \sigma^2 \sim \mathcal{N}\left(X_{t_1(\gamma)}\beta_{t_1(\gamma)}, \sigma^2 I_n\right)$$

Global and compatible priors

Use Zellner's g-prior, i.e. a normal prior for β conditional on σ^2,

$$\beta|\sigma^2 \sim \mathcal{N}(\tilde{\beta}, c\sigma^2(X^\mathsf{T}X)^{-1})$$

and a Jeffreys prior for σ^2,

$$\pi(\sigma^2) \propto \sigma^{-2}$$

▸ Noninformative g

Global and compatible priors

Use Zellner's g-prior, i.e. a normal prior for β conditional on σ^2,

$$\beta|\sigma^2 \sim \mathcal{N}(\tilde{\beta}, c\sigma^2(X^\mathsf{T}X)^{-1})$$

and a Jeffreys prior for σ^2,

$$\pi(\sigma^2) \propto \sigma^{-2}$$

▸ Noninformative g

Resulting compatible prior

$$\mathcal{N}\left(\left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)}\right)^{-1} X_{t_1(\gamma)}^\mathsf{T} X\tilde{\beta}, c\sigma^2 \left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)}\right)^{-1}\right)$$

Global and compatible priors

Use Zellner's g-prior, i.e. a normal prior for β conditional on σ^2,

$$\beta|\sigma^2 \sim \mathcal{N}(\tilde{\beta}, c\sigma^2(X^\mathsf{T}X)^{-1})$$

and a Jeffreys prior for σ^2,

$$\pi(\sigma^2) \propto \sigma^{-2}$$

▸ Noninformative g

Resulting compatible prior

$$\mathcal{N}\left(\left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)}\right)^{-1} X_{t_1(\gamma)}^\mathsf{T} X\tilde{\beta}, c\sigma^2 \left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)}\right)^{-1}\right)$$

[Surprise!]

Model index

For the hierarchical parameter γ, we use

$$\pi(\gamma) = \prod_{i=1}^{p} \tau_i^{\gamma_i} (1 - \tau_i)^{1-\gamma_i},$$

where τ_i corresponds to the prior probability that variable i is present in the model.

Model index

For the hierarchical parameter γ, we use

$$\pi(\gamma) = \prod_{i=1}^{p} \tau_i^{\gamma_i}(1-\tau_i)^{1-\gamma_i},$$

where τ_i corresponds to the prior probability that variable i is present in the model.
Typically, when no prior information is available, $\tau_1 = \ldots = \tau_p = 1/2$, ie a uniform prior

$$\pi(\gamma) = 2^{-p}$$

Posterior model probability

Can be obtained in closed form:

$$\pi(\gamma|y) \propto (c+1)^{-(q_\gamma+1)/2} \left[y^\mathsf{T} y - \frac{c y^\mathsf{T} P_1 y}{c+1} + \frac{\tilde\beta^\mathsf{T} X^\mathsf{T} P_1 X \tilde\beta}{c+1} - \frac{2 y^\mathsf{T} P_1 X \tilde\beta}{c+1} \right]^{-n/2} .$$

Posterior model probability

Can be obtained in closed form:

$$\pi(\gamma|y) \propto (c+1)^{-(q_\gamma+1)/2} \left[y^\mathsf{T} y - \frac{c y^\mathsf{T} P_1 y}{c+1} + \frac{\tilde\beta^\mathsf{T} X^\mathsf{T} P_1 X \tilde\beta}{c+1} - \frac{2 y^\mathsf{T} P_1 X \tilde\beta}{c+1} \right]^{-n/2} .$$

Conditionally on γ, posterior distributions of β and σ^2:

$$\begin{aligned}
\beta_{t_1(\gamma)}|\sigma^2, y, \gamma &\sim \mathcal{N}\left[\frac{c}{c+1}(U_1 y + U_1 X \tilde\beta / c), \frac{\sigma^2 c}{c+1} \left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)} \right)^{-1} \right], \\
\sigma^2|y, \gamma &\sim \mathcal{IG}\left[\frac{n}{2}, \frac{y^\mathsf{T} y}{2} - \frac{c y^\mathsf{T} P_1 y}{2(c+1)} + \frac{\tilde\beta^\mathsf{T} X^\mathsf{T} P_1 X \tilde\beta}{2(c+1)} - \frac{y^\mathsf{T} P_1 X \tilde\beta}{c+1} \right].
\end{aligned}$$

Noninformative case

Use the same compatible informative g-prior distribution with $\tilde{\beta} = 0_{p+1}$ and a hierarchical diffuse prior distribution on c,

$$\pi(c) \propto c^{-1} \mathbb{I}_{\mathbb{N}^*}(c)$$

▸ Recall g-prior

Noninformative case

Use the same compatible informative g-prior distribution with $\tilde{\beta} = 0_{p+1}$ and a hierarchical diffuse prior distribution on c,

$$\pi(c) \propto c^{-1} \mathbb{I}_{\mathbb{N}^*}(c)$$

▸ Recall g-prior

The choice of this hierarchical diffuse prior distribution on c is due to the model posterior sensitivity to large values of c:

Noninformative case

Use the same compatible informative g-prior distribution with $\tilde{\beta} = 0_{p+1}$ and a hierarchical diffuse prior distribution on c,

$$\pi(c) \propto c^{-1} \mathbb{I}_{\mathbb{N}^*}(c)$$

▸ Recall g-prior

The choice of this hierarchical diffuse prior distribution on c is due to the model posterior sensitivity to large values of c:

$$\boxed{\textbf{Taking} \quad \tilde{\beta} = 0_{p+1} \quad \textbf{and } c \textbf{ large does not work}}$$

Influence of c

▸ Erase influence

Consider the 10-predictor full model

$$y|\beta,\sigma^2 \sim \mathcal{N}\left(\beta_0+\sum_{i=1}^{3}\beta_i x_i+\sum_{i=1}^{3}\beta_{i+3}x_i^2+\beta_7 x_1 x_2+\beta_8 x_1 x3+\beta_9 x_2 x_3+\beta_{10} x_1 x_2 x_3,\sigma^2 I_n\right)$$

where the x_is are iid $\mathscr{U}(0,10)$

[Casella & Moreno, 2004]

Influence of c

Erase influence

Consider the 10-predictor full model

$$y|\beta,\sigma^2 \sim \mathcal{N}\left(\beta_0 + \sum_{i=1}^{3}\beta_i x_i + \sum_{i=1}^{3}\beta_{i+3}x_i^2 + \beta_7 x_1 x_2 + \beta_8 x_1 x3 + \beta_9 x_2 x_3 + \beta_{10} x_1 x_2 x_3, \sigma^2 I_n\right)$$

where the x_is are iid $\mathscr{U}(0,10)$

[Casella & Moreno, 2004]

True model: two predictors x_1 and x_2, i.e. $\gamma^* = 110\ldots0$, $(\beta_0,\beta_1,\beta_2) = (5,1,3)$, and $\sigma^2 = 4$.

Influence of c^2

$t_1(\gamma)$	$c = 10$	$c = 100$	$c = 10^3$	$c = 10^4$	$c = 10^6$
0,1,2	0.04062	0.35368	0.65858	0.85895	0.98222
0,1,2,7	0.01326	0.06142	0.08395	0.04434	0.00524
0,1,2,4	0.01299	0.05310	0.05805	0.02868	0.00336
0,2,4	0.02927	0.03962	0.00409	0.00246	0.00254
0,1,2,8	0.01240	0.03833	0.01100	0.00126	0.00126

Noninformative case (cont'd)

In the noninformative setting,

$$\pi(\gamma|y) \propto \sum_{c=1}^{\infty} c^{-1}(c+1)^{-(q_\gamma+1)/2}\left[y^{\mathsf{T}}y - \frac{c}{c+1}y^{\mathsf{T}}P_1 y\right]^{-n/2}$$

converges for all y's

Casella & Moreno's example

$t_1(\gamma)$	$\sum_{i=1}^{10^6} \pi(\gamma\|y,c)\pi(c)$
0,1,2	0.78071
0,1,2,7	0.06201
0,1,2,4	0.04119
0,1,2,8	0.01676
0,1,2,5	0.01604

Gibbs approximation

When p large, impossible to compute the posterior probabilities of the 2^p models.

Gibbs approximation

When p large, impossible to compute the posterior probabilities of the 2^p models.
Use of a Monte Carlo approximation of $\pi(\gamma|y)$

Gibbs approximation

When p large, impossible to compute the posterior probabilities of the 2^p models.
Use of a Monte Carlo approximation of $\pi(\gamma|y)$

Gibbs sampling

- At $t = 0$, draw γ^0 from the uniform distribution on Γ
- At t, for $i = 1, \ldots, p$, draw $\gamma_i^t \sim \pi(\gamma_i|y, \gamma_1^t, \ldots, \gamma_{i-1}^t, \ldots, \gamma_{i+1}^{t-1}, \ldots, \gamma_p^{t-1})$

Gibbs approximation (cont'd)

Example (Simulated data)

Severe multicolinearities among predictors for a 20-predictor full model

$$y|\beta,\sigma^2 \sim \mathcal{N}\left(\beta_0 + \sum_{i=1}^{20} \beta_i x_i, \sigma^2 I_n\right)$$

where $x_i = z_i + 3z$, the z_i's and z are iid $\mathcal{N}_n(0_n, I_n)$.

Gibbs approximation (cont'd)

Example (Simulated data)

Severe multicolinearities among predictors for a 20-predictor full model

$$y|\beta,\sigma^2 \sim \mathcal{N}\left(\beta_0 + \sum_{i=1}^{20} \beta_i x_i, \sigma^2 I_n\right)$$

where $x_i = z_i + 3z$, the z_i's and z are iid $\mathcal{N}_n(0_n, I_n)$.
True model with $n = 180$, $\sigma^2 = 4$ and seven predictor variables

$$x_1, x_3, x_5, x_6, x_{12}, x_{18}, x_{20},$$
$$(\beta_0, \beta_1, \beta_3, \beta_5, \beta_6, \beta_{12}, \beta_{18}, \beta_{20}) = (3, 4, 1, -3, 12, -1, 5, -6)$$

Gibbs approximation (cont'd)

Example (Simulated data (2))

γ	$\pi(\gamma\|y)$	$\widehat{\pi(\gamma\|y)}^{GIBBS}$
0,1,3,5,6,12,18,20	0.1893	0.1822
0,1,3,5,6,18,20	0.0588	0.0598
0,1,3,5,6,9,12,18,20	0.0223	0.0236
0,1,3,5,6,12,14,18,20	0.0220	0.0193
0,1,2,3,5,6,12,18,20	0.0216	0.0222
0,1,3,5,6,7,12,18,20	0.0212	0.0233
0,1,3,5,6,10,12,18,20	0.0199	0.0222
0,1,3,4,5,6,12,18,20	0.0197	0.0182
0,1,3,5,6,12,15,18,20	0.0196	0.0196

Gibbs ($T = 100,000$) results for $\tilde{\beta} = 0_{21}$ and $c = 100$

Processionary caterpillar

Influence of some forest settlement characteristics on the development of caterpillar colonies

Processionary caterpillar

Influence of some forest settlement characteristics on the development of caterpillar colonies

Processionary caterpillar

Influence of some forest settlement characteristics on the development of caterpillar colonies

Response y log-transform of the average number of nests of caterpillars per tree on an area of 500 square meters ($n = 33$ areas)

Processionary caterpillar (cont'd)

Potential explanatory variables

x_1 altitude (in meters), x_2 slope (in degrees),

x_3 number of pines in the square,

x_4 height (in meters) of the tree at the center of the square,

x_5 diameter of the tree at the center of the square,

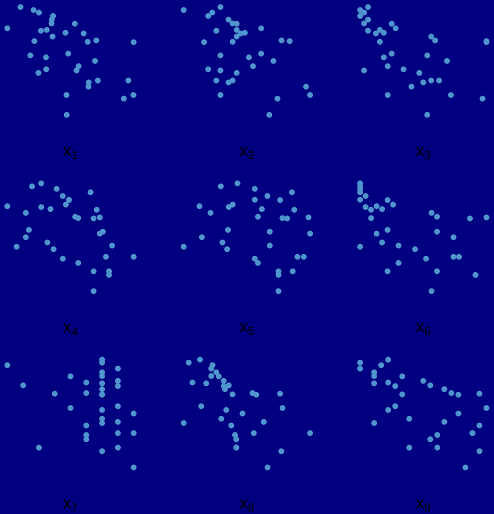
x_6 index of the settlement density,

x_7 orientation of the square (from 1 if southb'd to 2 ow),

x_8 height (in meters) of the dominant tree,

x_9 number of vegetation strata,

x_{10} mix settlement index (from 1 if not mixed to 2 if mixed).


X_1
X_2
X_3
X_4
X_5
X_6
X_7
X_8
X_9

Bayesian regression output

	Estimate	BF	log10(BF)
(Intercept)	9.2714	26.334	1.4205 (***)
X1	-0.0037	7.0839	0.8502 (**)
X2	-0.0454	3.6850	0.5664 (**)
X3	0.0573	0.4356	-0.3609
X4	-1.0905	2.8314	0.4520 (*)
X5	0.1953	2.5157	0.4007 (*)
X6	-0.3008	0.3621	-0.4412
X7	-0.2002	0.3627	-0.4404
X8	0.1526	0.4589	-0.3383
X9	-1.0835	0.9069	-0.0424
X10	-0.3651	0.4132	-0.3838

evidence against H0: (****) decisive, (***) strong, (**) subtantial, (*) poor

Bayesian variable selection

$t_1(\gamma)$	$\pi(\gamma\|y,X)$	$\widehat{\pi}(\gamma\|y,X)$
0,1,2,4,5	0.0929	0.0929
0,1,2,4,5,9	0.0325	0.0326
0,1,2,4,5,10	0.0295	0.0272
0,1,2,4,5,7	0.0231	0.0231
0,1,2,4,5,8	0.0228	0.0229
0,1,2,4,5,6	0.0228	0.0226
0,1,2,3,4,5	0.0224	0.0220
0,1,2,3,4,5,9	0.0167	0.0182
0,1,2,4,5,6,9	0.0167	0.0171
0,1,2,4,5,8,9	0.0137	0.0130

Noninformative G-prior model choice and Gibbs estimations

3 Classification via k-nearest-neighbour

1. Bayesian Model Choice
2. Compatible priors for variable selection
3. k-nearest-neighbour classification
   - Principle
   - Statistical reformulation
   - Bayesian inference in k mean models
   - Ripley's benchmark
   - Global classification

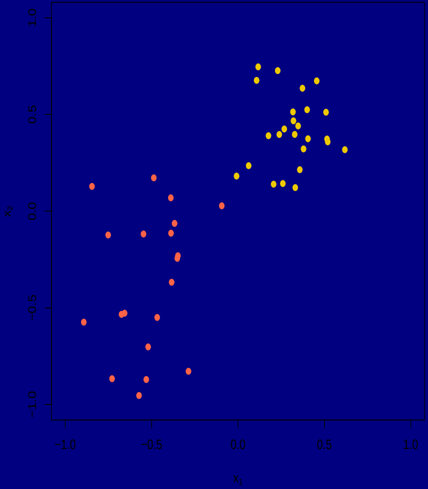
[Joint work with C. Celeux, J.M. Marin and D.M. Titterington]

Idea

Use for classification purposes of a training dataset

$$\left((y_i^{\text{tr}}, x_i^{\text{tr}})\right)_{i=1,\dots,n}$$

with class label $1 \le y_i^{\text{tr}} \le Q$ and predictor variables x_i^{tr}



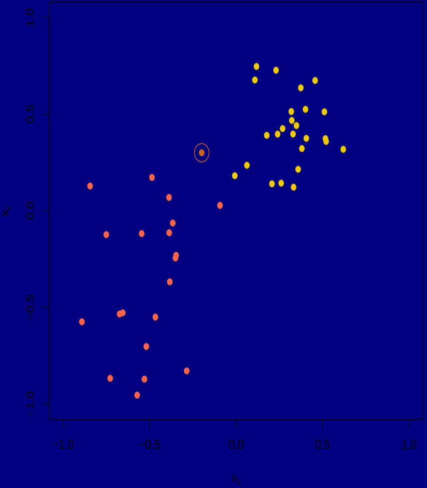
Classification

▸ Skip animation

Principle

Prediction for a new point $(y_j^{\text{te}}, x_j^{\text{te}})$ $(j = 1, \ldots, m)$: the most common class amongst the k nearest neighbours of x_j^{te} in the training set

Neighbourhood based on a distance metric



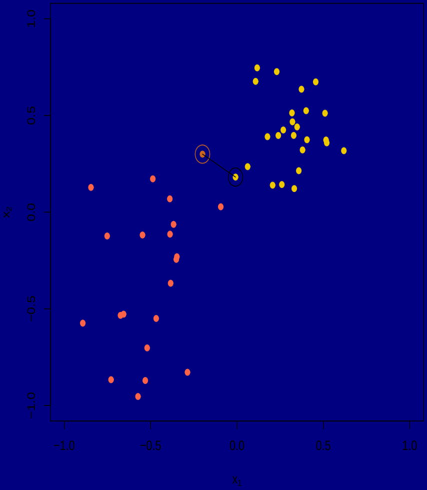
Classification

▸ Skip animation

Principle

Prediction for a new point $(y_j^{\text{te}}, x_j^{\text{te}})$ $(j = 1, \ldots, m)$: the most common class amongst the k nearest neighbours of x_j^{te} in the training set

Neighbourhood based on a distance metric



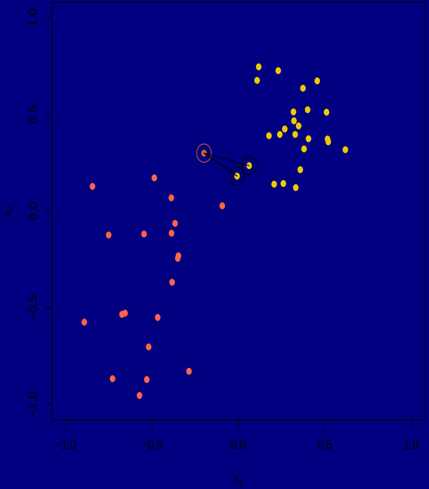
Classification

▸ Skip animation

Principle

Prediction for a new point $(y_j^{\text{te}}, x_j^{\text{te}})$ $(j = 1, \ldots, m)$: the most common class amongst the k nearest neighbours of x_j^{te} in the training set

Neighbourhood based on a distance metric



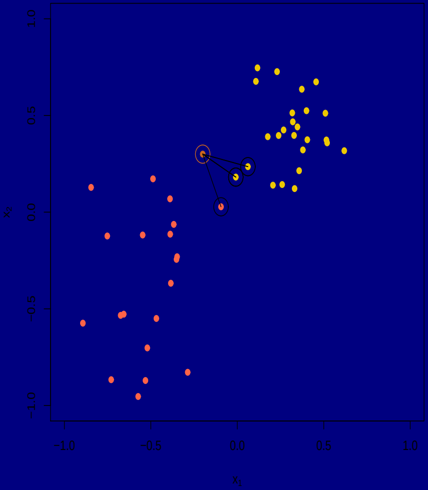
Classification

▸ Skip animation

Principle

Prediction for a new point $(y_j^{\text{te}}, x_j^{\text{te}})$ $(j = 1, \ldots, m)$: the most common class amongst the k nearest neighbours of x_j^{te} in the training set

Neighbourhood based on a distance metric



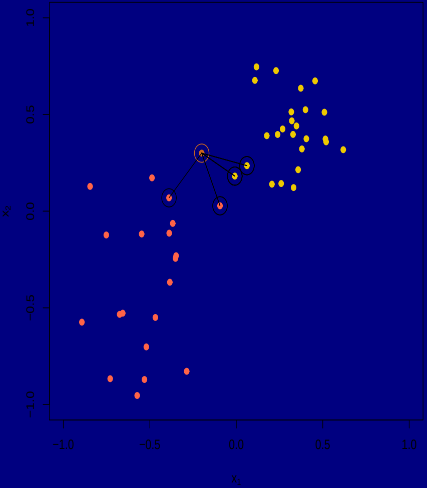
Classification

▸ Skip animation

Principle

Prediction for a new point $(y_j^{\text{te}}, x_j^{\text{te}})$ $(j = 1, \ldots, m)$: the most common class amongst the k nearest neighbours of x_j^{te} in the training set

Neighbourhood based on a distance metric



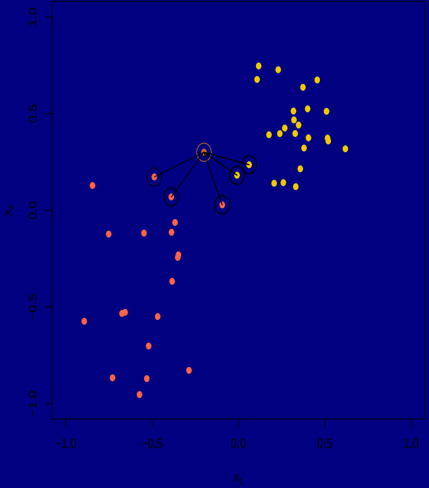
Classification

▸ Skip animation

Principle

Prediction for a new point $(y_j^{\text{te}}, x_j^{\text{te}})$ $(j = 1, \ldots, m)$: the most common class amongst the k nearest neighbours of x_j^{te} in the training set

Neighbourhood based on a distance metric



Model choice perspective

◂ Back to idea

Choice of k?

Usually chosen by minimizing cross-validated misclassification rate (non-parametric or even non-probabilist!)

Formalisation thru a probabilty model

k nearest neighbour model

Based on full conditional distributions ($\omega \in \{C_1, \ldots, C_Q\}$)

$$\mathbb{P}(y_i^{\text{tr}} = \omega | y_{-i}^{\text{tr}}, x^{\text{tr}}, \beta, k) \propto \exp\left(\beta \sum_{l \overset{k}{\sim} i} \delta_\omega(y_l^{\text{tr}}) \Big/ k\right) \quad \beta > 0$$

where $l \overset{k}{\sim} i$ is the k nearest neighbour relation

[Holmes & Adams, 2002]

Drawback

Because the neighbourhood structure is not symmetric (x_i may be one of the k nearest neighbours of x_j and x_j not one of the k nearest neighbours of x_i),

Drawback

Because the neighbourhood structure is not symmetric (x_i may be one of the k nearest neighbours of x_j and x_j not one of the k nearest neighbours of x_i), **there usually is no joint probability distribution corresponding to these "full conditionals"!**

Resolution

Symmetrize the neighbourhood relation:

Resolution

Symmetrize the neighbourhood relation:

if x_i^{tr} belongs to the k-nearest-neighbour set for x_j^{tr} and x_j^{tr} does not belong to the k-nearest-neighbour set for x_i^{tr}, x_j^{tr} is added to the set of neighbours of x_i^{tr}

Consequence

Given the full conditionals

$$\mathbb{P}(y_i^{\text{tr}} = \omega | y_{-i}^{\text{tr}}, x^{\text{tr}}, \beta, k) \propto \text{exp}\left(\beta \sum_{l \overset{k}{\sim} i} \delta_\omega(y_l^{\text{tr}}) \Big/ N(i) \right)$$

where $l \overset{k}{\sim} i$ is the symmetrized k nearest neighbour relation, and $N(i)$ denotes the size of the symmetrized k-nearest neighbourhood of x_i^{tr}

Consequence

Given the full conditionals

$$\mathbb{P}(y_i^{\text{tr}} = \omega | y_{-i}^{\text{tr}}, x^{\text{tr}}, \beta, k) \propto \exp\left(\beta \sum_{l \overset{k}{\sim} i} \delta_\omega(y_l^{\text{tr}}) \Big/ N(i) \right)$$

where $l \overset{k}{\sim} i$ is the symmetrized k nearest neighbour relation, and $N(i)$ denotes the size of the symmetrized k-nearest neighbourhood of x_i^{tr} **there exists a corresponding joint distribution**

Extension to the unclassified points

Use for the predictive distribution of y_j^{te} $(j = 1, \ldots, m)$

$$\mathbb{P}(y_j^{\text{te}} = \omega | x_j^{\text{te}}, y^{\text{tr}}, x^{\text{tr}}, \beta, k) \propto \exp\left(\beta \sum_{l \overset{k}{\#} j} \delta_\omega(y_l^{\text{tr}}) \Big/ k \right)$$

where $l \overset{k}{\#} j$ denotes the symmetrized k-nearest-neighbour relation wrt the set $\{x_1^{\text{tr}}, \ldots, x_n^{\text{tr}}\}$

Bayesian global inference

Within the Bayesian paradigm, assign a prior $\pi(\beta, k)$ and use the marginal predictive distribution of y_j^{te} given x_j^{te} $(j = 1, \ldots, m)$

Bayesian global inference

Within the Bayesian paradigm, assign a prior $\pi(\beta, k)$ and use the marginal predictive distribution of y_j^{te} given x_j^{te} $(j = 1, \ldots, m)$

$$\int \mathbb{P}(y_j^{\text{te}} = \omega | x_j^{\text{te}}, y^{\text{tr}}, x^{\text{tr}}, \beta, k) \pi(\beta, k | y^{\text{tr}}, x^{\text{tr}}) \text{d}\beta \, \text{d}k$$

where $\pi(\beta, k | y^{\text{tr}}, x^{\text{tr}}) \propto f(y^{\text{tr}} | x^{\text{tr}}, \beta, k) \pi(\beta, k)$ posterior distribution of (β, k) given the training dataset y^{tr}

[$\widehat{y_j^{\text{te}}} =$ MAP estimate]

Bayesian global inference

Within the Bayesian paradigm, assign a prior $\pi(\beta, k)$ and use the marginal predictive distribution of y_j^{te} given x_j^{te} $(j = 1, \ldots, m)$

$$\int \mathbb{P}(y_j^{\text{te}} = \omega | x_j^{\text{te}}, y^{\text{tr}}, x^{\text{tr}}, \beta, k) \pi(\beta, k | y^{\text{tr}}, x^{\text{tr}}) \text{d}\beta \, \text{d}k$$

where $\pi(\beta, k | y^{\text{tr}}, x^{\text{tr}}) \propto f(y^{\text{tr}} | x^{\text{tr}}, \beta, k) \pi(\beta, k)$ posterior distribution of (β, k) given the training dataset y^{tr}

[$\widehat{y}_j^{\text{te}}$ = MAP estimate]

Note

Model choice *without* varying dimension because β is the same on all models

Difficulty

To compute $f(y^{\text{tr}}|x^{\text{tr}}, \beta, k)$ requires a normalisation constant that is not readily available

Difficulty

To compute $f(y^{\text{tr}}|x^{\text{tr}},\beta,k)$ requires a normalisation constant that is not readily available

Approximation

Use instead a pseudo-likelihood $\widehat{f}(y^{\text{tr}}|x^{\text{tr}},\beta,k)$ equal to

$$\prod_{i=1}^{n}\left[\mathbb{P}(y_i^{\text{tr}}=0|y_{-i}^{\text{tr}},x^{\text{tr}},\beta,k)\right]^{1-y_i^{\text{tr}}}\left[1-\mathbb{P}(y_i^{\text{tr}}=0|y_{-i}^{\text{tr}},x^{\text{tr}},\beta,k)\right]^{y_i^{\text{tr}}}$$

Further difficulty

Even with this approximation, the computation of $\mathbb{P}(y_j^{\text{te}} = \omega | x_j^{\text{te}}, y^{\text{tr}}, x^{\text{tr}})$ is not feasible.

Further difficulty

Even with this approximation, the computation of $\mathbb{P}(y_j^{\text{te}} = \omega | x_j^{\text{te}}, y^{\text{tr}}, x^{\text{tr}})$ is not feasible.

Use instead a Monte Carlo approximation of $\pi(\beta, k | y^{\text{tr}}, x^{\text{tr}})$,

$$M^{-1} \sum_{i=1}^{M} \mathbb{P}\left(y_j^{\text{te}} = 0 \,\middle|\, x_j^{\text{te}}, y^{\text{tr}}, x^{\text{tr}}, (\beta, k)^{(i)} \right)$$

where $(\beta, k)^{(i)}$ simulated by MCMC with r-neighbour random-walk proposal on k: $\mathcal{U}\left(\{k-r, k-r+1, \ldots, k+r-1, k+r\}\right)$

[Gibbs too costly]

MCMC for k-nearest-neighbours

Random walk k-nearest-neighbours

At time 0, generate $\beta^{(0)} \sim \mathcal{N}\left(0, \tau^2\right)$ and $k^{(0)} \sim \mathcal{U}_{\{1,\dots,K\}}$
At time $1 \leq t \leq T$,

1. Generate $\log \tilde{\beta} \sim \mathcal{N}\left(\log \beta^{(t-1)}, \tau^2\right)$ and $\tilde{k} \sim \mathcal{U}\left(\{k-r, k-r+1, \dots, k+r-1, k+r\}\right)$

MCMC for k-nearest-neighbours

Random walk k-nearest-neighbours

At time 0, generate $\beta^{(0)} \sim \mathcal{N}\left(0, \tau^2\right)$ and $k^{(0)} \sim \mathcal{U}_{\{1,\ldots,K\}}$
At time $1 \leq t \leq T$,

1. Generate $\log \tilde{\beta} \sim \mathcal{N}\left(\log \beta^{(t-1)}, \tau^2\right)$ and $\tilde{k} \sim \mathcal{U}\left(\{k-r, k-r+1, \ldots, k+r-1, k+r\}\right)$
2. Calculate Metropolis-Hastings acceptance probability $\rho(\tilde{\beta}, \tilde{k}, \beta^{(t-1)}, k^{(t-1)})$

MCMC for k-nearest-neighbours

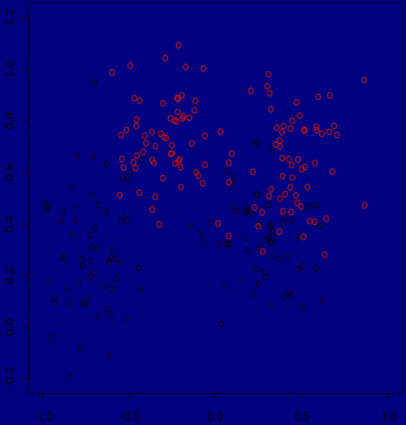
Random walk k-nearest-neighbours

At time 0, generate $\beta^{(0)} \sim \mathcal{N}\left(0, \tau^2\right)$ and $k^{(0)} \sim \mathcal{U}_{\{1,\dots,K\}}$
At time $1 \leq t \leq T$,

1. Generate $\log \tilde{\beta} \sim \mathcal{N}\left(\log \beta^{(t-1)}, \tau^2\right)$ and $\tilde{k} \sim \mathcal{U}\left(\{k-r, k-r+1, \dots, k+r-1, k+r\}\right)$
2. Calculate Metropolis-Hastings acceptance probability $\rho(\tilde{\beta}, \tilde{k}, \beta^{(t-1)}, k^{(t-1)})$
3. Move to $\left(\beta^{(t)}, k^{(t)}\right)$ by Metropolis-Hastings step

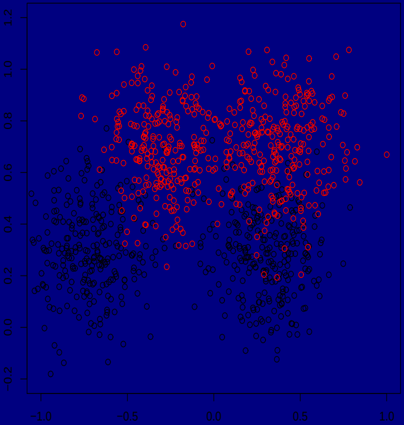
Benchmark

Dataset from Ripley (1994), with two classes where each population of x_i's from a mixture of two bivariate normal distributions.
Training set of $n = 250$ points and testing set on a set of $m = 1,000$ points



Benchmark

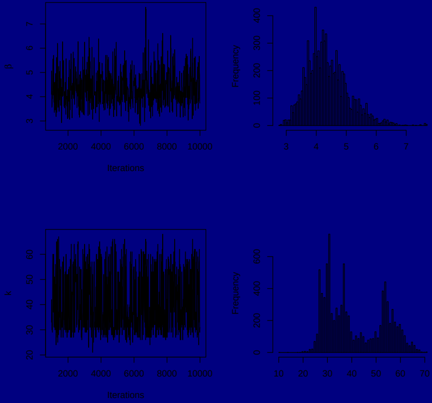
Dataset from Ripley (1994), with two classes where each population of x_i's from a mixture of two bivariate normal distributions.
Training set of $n = 250$ points and testing set on a set of $m = 1,000$ points



Gibbs output

Use of the prior

$$\pi(\beta, k) \propto \mathbb{I}_{(0,15)}(\beta)\, \mathbb{I}_{\{1,\ldots,\lfloor n/2 \rfloor\}}(k)$$

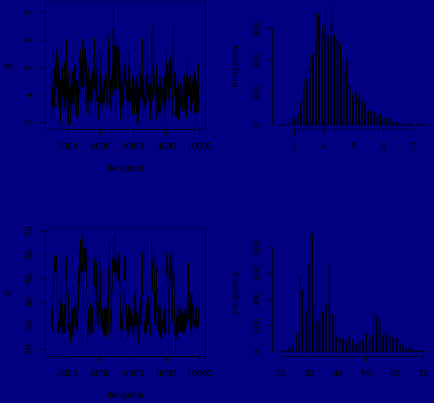


Hybrid Gibbs output

Gibbs output

Use of the prior

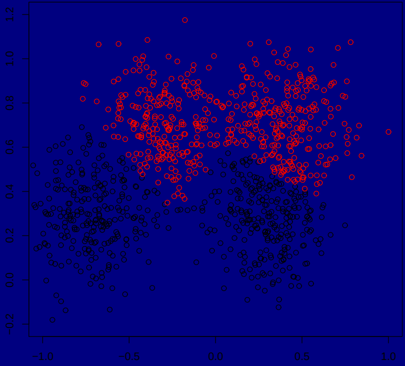
$$\pi(\beta, k) \propto \mathbb{I}_{(0,15)}(\beta)\,\mathbb{I}_{\{1,\ldots,\lfloor n/2\rfloor\}}(k)$$



Metropolis–Hastings output

Prediction performances

Same label allocation and same misclassification rate (8.4%) for both algorithms



Alternative perspective

Lack of coherence of previous predictive:

- Each testing point processed marginaly

Alternative perspective

Lack of coherence of previous predictive:

- Each testing point processed marginaly
- Different distribution for training and testing points

Alternative perspective

Lack of coherence of previous predictive:

- Each testing point processed marginaly
- Different distribution for training and testing points
- No global assessment of uncertainty

Alternative perspective

Lack of coherence of previous predictive:

- Each testing point processed marginaly
- Different distribution for training and testing points
- No global assessment of uncertainty
- Unless notified otherwise, testing sample = missing at random

Joint k-nearest-neighbour distribution

Full exchangeability of training and testing samples
$y = (y^{\text{tr}}, y^{\text{te}}) = (y_1, \ldots, y_{n+m})$ and
$x = (x^{\text{tr}}, x^{\text{te}}) = (x_1, \ldots, x_{n+m})$

Joint k-nearest-neighbour distribution

Full exchangeability of training and testing samples
$y = (y^{\text{tr}}, y^{\text{te}}) = (y_1, \ldots, y_{n+m})$ and
$x = (x^{\text{tr}}, x^{\text{te}}) = (x_1, \ldots, x_{n+m})$

$$\mathbb{P}(y_i = \omega | y_{-i}, x, \beta, k) \propto \exp\left(\beta \sum_{l \overset{k}{\#} i} \delta_0(y_l) \Big/ N(i)\right)$$

where $l \overset{k}{\#} i$ is the symmetrized k-nearest-neighbour relation in the set $\{x_1, \ldots, x_{n+m}\}$ and $N(i)$ the number of symmetrized k-nearest-neighbours of x_i $(1 \leq i \leq n + m)$

Pseudo-likelihood

Same difficulty with joint distribution (normalizing constant)

Pseudo-likelihood

Same difficulty with joint distribution (normalizing constant)
Use instead pseudo-likelihood

$$\prod_{i=1}^{m+n} \left[\mathbb{P}(y_i = 0|y_{-i}, x, \beta, k)\right]^{1-y_i} \left[1 - \mathbb{P}(y_i = 0|y_{-i}, x, \beta, k)\right]^{y_i}$$

Gibbs implementation

Process the y_j^{te}'s as missing data

Hybrid Gibbs k-nearest-neighbour classification

At time $1 \le t \le T$,

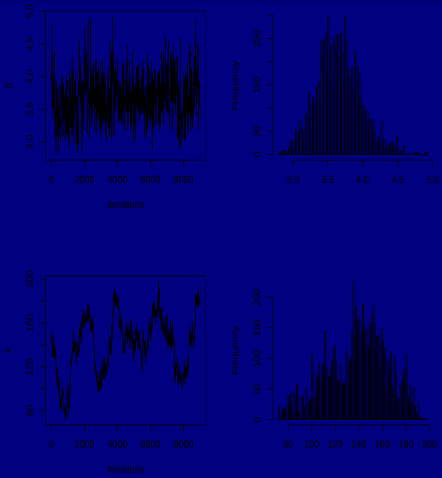
1. For $n+1 \le i \le n+m$, compute $q_i = \mathbb{P}\left(y_i = 1 \middle| y_{-i}^{(t)}, x, \beta^{(t-1)}, k^{(t-1)}\right)$ and generate $y_i^{(t)} \sim \mathcal{B}(1, q_i)$
2. Generate $\log \tilde\beta \sim \mathcal{N}\left(\log \beta^{(t-1)}, \tau^2\right)$ and $\tilde k \sim \mathcal{U}\left(\{k^{(t-1)} - r, \ldots, k^{(t-1)} + r\}\right)$
3. Accept $(\tilde\beta, \tilde k)$ with M-H probability $\rho(\tilde\beta, \beta^{(t-1)}, k^{(t-1)})$ otherwise replicate $(\beta^{(t-1)}, k^{(t-1)})$

Benchmark illustration

For Ripley's benchmark and testing sample of $1,000$ points, use of prior

$$\pi(\beta, k) \propto \mathbb{I}_{0\leq\beta\leq 15)}\, \mathbb{I}_{\{1,\ldots,\lfloor\frac{m+n}{2}\rfloor\}}(k)$$

and misclassification rate 8.3%



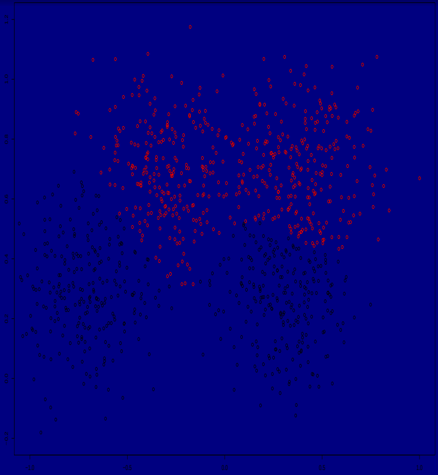
Hybrid Gibbs output

Benchmark illustration

For Ripley's benchmark and testing sample of 1,000 points, use of prior

$$\pi(\beta, k) \propto \mathbb{I}_{0\leq\beta\leq 15)}\, \mathbb{I}_{\{1,\ldots,\lfloor\frac{m+n}{2}\rfloor\}}(k)$$

and misclassification rate 8.3%



Testing allocation

Extensions

- Assessment and representation of uncertainty on buffer points
- k dependent β's
- Behaviour of marginal/local versus global/exchangeable when m goes to ∞
- Selection of the significant components of x ($=$ imbedded principal components)