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

\[ M_i : x \sim f_i(x|\theta_i), \quad i \in \mathcal{I} \]

where \( \mathcal{I} \) 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 $M_i$
- defining priors $\pi_i(\theta_i)$ for each parameter space $\Theta_i$
Formal solution

Resolution

1. Compute

\[
p(M_i | x) = \frac{p_i \int_{\Theta_i} f_i(x | \theta_i) \pi_i(\theta_i) d\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x | \theta_j) \pi_j(\theta_j) d\theta_j}
\]
Formal solution

**Resolution**

1. Compute

\[
p(M_i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j}
\]

2. Take largest \(p(M_i|x)\) to determine ‘‘best’’ model, or use averaged predictive

\[
\sum_j p(M_j|x) \int_{\Theta_j} f_j(x'|\theta_j)\pi_j(\theta_j|x)d\theta_j
\]
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 $M_1 = M_2 \cup M_3$, 

Several types of problems (2)

Choice of prior structures
- adequate weights $p_i$:
  if $M_1 = M_2 \cup M_3$, $p(M_1) = p(M_2) + p(M_3)$?
- priors distributions
  - $\pi_i(\theta_i)$ defined for every $i \in I$
Several types of problems (2)

Choice of prior structures
- adequate weights $p_i$: 
  if $M_1 = M_2 \cup M_3$, $p(M_1) = p(M_2) + p(M_3)$?
- priors distributions
  - $\pi_i(\theta_i)$ defined for every $i \in I$
  - $\pi_i(\theta_i)$ proper (Jeffreys)
Several types of problems (2)

- **Choice of prior structures**
  - adequate weights $p_i$:
    - if $M_1 = M_2 \cup M_3$, $p(M_1) = p(M_2) + p(M_3)$?
  - priors distributions
    - $\pi_i(\theta_i)$ defined for every $i \in I$
    - $\pi_i(\theta_i)$ *proper* (Jeffreys)
    - $\pi_i(\theta_i)$ *coherent* (?) for nested models
Several types of problems (2)

- **Choice of prior structures**
  - adequate weights $p_i$:
    - if $M_1 = M_2 \cup M_3$, $p(M_1) = p(M_2) + p(M_3)$?
  - priors distributions
    - $\pi_i(\theta_i)$ defined for every $i \in I$
    - $\pi_i(\theta_i)$ *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 $\mathcal{M}_1$ vs. $\mathcal{M}_2$

\[
B_{12} = \frac{\Pr(\mathcal{M}_1|x)}{\Pr(\mathcal{M}_2|x)} \frac{\Pr(\mathcal{M}_1)}{\Pr(\mathcal{M}_2)}
\]

\[
= \int \frac{f_1(x|\theta_1)\pi_1(\theta_1)}{f_2(x|\theta_2)\pi_2(\theta_2)} d\theta_1
\]

[Good, 1958 & Jeffreys, 1961]
Self-contained concept

- eliminates choice of $\Pr(M_i)$
Self-contained concept

- eliminates choice of $\Pr(M_i)$
- but depends on the choice of $\pi_i(\theta_i)$
Self-contained concept

- eliminates choice of $\Pr(M_i)$
- but depends on the choice of $\pi_i(\theta_i)$
- Bayesian/marginal likelihood ratio
Self-contained concept

- eliminates choice of $\Pr(M_i)$
- 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

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

then either $\pi_1$ or $\pi_2$ 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 \( \pi_1 \) or \( \pi_2 \) cannot be normalised uniquely but the normalisation matters in the Bayes factor
Constants matter

Example (Poisson versus Negative binomial)

If $M_1$ is a $P(\lambda)$ distribution and $M_2$ is a $NB(m, p)$ distribution, we can take

$$\pi_1(\lambda) = \frac{1}{\lambda}$$
$$\pi_2(m, p) = \frac{1}{M} \mathbb{I}_{\{1, \ldots, M\}}(m) \mathbb{I}_{[0,1]}(p)$$
Example (Poisson versus Negative binomial (2))

then

\[ B_{12} = \frac{\int_0^\infty \frac{\lambda^{x-1}}{x!} e^{-\lambda} d\lambda}{\frac{1}{M} \sum_{m=1}^M \int_0^\infty \binom{m}{x-1} p^x (1-p)^{m-x} dp} \]

\[ = \frac{1}{\frac{1}{M} \sum_{m=x}^M \binom{m}{x-1} \frac{x!(m-x)!}{m!}} \]

\[ = \frac{1}{\frac{1}{M} \sum_{m=x}^M \frac{x}{(m-x+1)}} \]
Example (Poisson versus Negative binomial (3))

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

Taking a proper prior and take a “very large” variance (e.g., BUGS)
Vague proper priors are not the solution

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

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 $\pi_1(\theta)$,

$$B_{01}(x) \xrightarrow{\alpha \to \infty} 0$$
Vague proper priors are not the solution (cont’d)

Example (Poisson versus Negative binomial (4))

\[ B_{12} = \int_0^1 \frac{\lambda^{\alpha+x-1} e^{-\lambda \beta}}{x!} \, d\lambda \]

if \( \lambda \sim \text{Ga}(\alpha, \beta) \)

\[ = \frac{\Gamma(\alpha + x)}{x! \, \Gamma(\alpha)} \beta^{-x} \left/ \frac{1}{M} \sum_m \frac{x}{m - x + 1} \right. \]

\[ = \frac{(x + \alpha - 1) \cdots \alpha}{x(x - 1) \cdots 1} \beta^{-x} \left/ \frac{1}{M} \sum_m \frac{x}{m - x + 1} \right. \]
Vague proper priors are not the solution (cont’d)

Example (Poisson versus Negative binomial (4))

\[ B_{12} = \frac{\int_0^1 \frac{\lambda^{\alpha+x-1}}{x!} e^{-\lambda} \beta^\alpha \lambda d\lambda}{\frac{1}{M} \sum_{m} \frac{x}{m - x + 1} \frac{\beta^\alpha}{\Gamma(\alpha)}} \quad \text{if } \lambda \sim \text{Ga}(\alpha, \beta) \]

\[ = \frac{\Gamma(\alpha + x)}{x! \Gamma(\alpha)} \frac{\beta^{-x}}{\beta^{-x}} \frac{1}{M} \sum_{m} \frac{x}{m - x + 1} \]

\[ = \frac{(x + \alpha - 1) \cdots \alpha}{x(x - 1) \cdots 1} \frac{\beta^{-x}}{\beta^{-x}} \frac{1}{M} \sum_{m} \frac{x}{m - x + 1} \]

depends on choice of \( \alpha(\beta) \) or \( \beta(\alpha) \) \( \to 0 \)
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]
Difficult to simultaneously find priors on a collection of models $M_i$ ($i \in I$)
Difficult to simultaneously find priors on a collection of models $M_i$ ($i \in I$) 
Easier to start from a single prior on a “big” model and to derive the other priors from a coherence principle 

[Dawid & Lauritzen, 2000]
Model uncertainty and model choice: Bayesian tools
Compatible priors for variable selection
Principle

**Projection approach**

For $M_2$ submodel of $M_1$, $\pi_2$ 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 $\theta_1$ on $M_2$, 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 $\mathcal{M}_2$ submodel of $\mathcal{M}_1$, $\pi_2$ 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 $\theta_1$ on $\mathcal{M}_2$, 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 $\pi_1$ on a model $M_1$ and a submodel $M_2$, a prior $\pi_2$ on $M_2$ is \textit{compatible} with $\pi_1$. 
Compatible priors for variable selection

Principle

**Kullback proximity**

**Alternative solution**

**Definition (Compatible prior)**

Given a prior $\pi_1$ on a model $\mathcal{M}_1$ and a submodel $\mathcal{M}_2$, a prior $\pi_2$ on $\mathcal{M}_2$ is *compatible* with $\pi_1$ when it achieves the minimum Kullback divergence between the corresponding marginals:

$m_1(x ; \pi_1) = \int_{\Theta_1} f_1(x | \theta) \pi_1(\theta) d\theta$ and $m_2(x ; \pi_2) = \int_{\Theta_2} f_2(x | \theta) \pi_2(\theta) d\theta$, \[ \text{min} \]
Kullback proximity

Alternative solution

Definition (Compatible prior)

Given a prior $\pi_1$ on a model $M_1$ and a submodel $M_2$, a prior $\pi_2$ on $M_2$ is \textit{compatible} with $\pi_1$ when it achieves the minimum Kullback divergence between the corresponding marginals:

$m_1(x; \pi_1) = \int_{\Theta_1} f_1(x|\theta)\pi_1(\theta)d\theta$ and $m_2(x; \pi_2) = \int_{\Theta_2} f_2(x|\theta)\pi_2(\theta)d\theta$,

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

- Does not give a working principle when $M_2$ is not a submodel $M_1$
Difficulties

- Does not give a working principle when \( M_2 \) is not a submodel \( M_1 \)
- Depends on the choice of \( \pi_1 \)
Difficulties

- Does not give a working principle when $M_2$ is not a submodel $M_1$
- Depends on the choice of $\pi_1$
- Prohibits the use of improper priors
Difficulties

- Does not give a working principle when $M_2$ is not a submodel $M_1$
- Depends on the choice of $\pi_1$
- Prohibits the use of improper priors
- Worse: useless in unconstrained settings...
Linear regression

$M_1$ and $M_2$ are two nested Gaussian linear regression models with Zellner’s $g$-priors and the same variance $\sigma^2 \sim \pi(\sigma^2)$:

1. **$M_1$**:
   
   $y | \beta_1, \sigma^2 \sim \mathcal{N}(X_1 \beta_1, \sigma^2), \quad \beta_1 | \sigma^2 \sim \mathcal{N} \left( s_1, \sigma^2 n_1 (X_1^T X_1)^{-1} \right)$
   
   where $X_1$ is a $(n \times k_1)$ matrix of rank $k_1 \leq n$

2. **$M_2$**:
   
   $y | \beta_2, \sigma^2 \sim \mathcal{N}(X_2 \beta_2, \sigma^2), \quad \beta_2 | \sigma^2 \sim \mathcal{N} \left( s_2, \sigma^2 n_2 (X_2^T X_2)^{-1} \right)$
   
   where $X_2$ is a $(n \times k_2)$ matrix with span($X_2$) $\subseteq$ span($X_1$)
Since $\sigma^2$ is a nuisance parameter, we can minimize the Kullback-Leibler divergence between the two marginal distributions conditional on $\sigma^2$: $m_1(y|\sigma^2; s_1, n_1)$ and $m_2(y|\sigma^2; s_2, n_2)$
Compatible $g$-priors

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

**Theorem**

*Conditional on $\sigma^2$, the conjugate compatible prior of $M_2$ wrt $M_1$ 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*

$$s_2^* = (X_2^T X_2)^{-1} X_2^T X_1 s_1$$

$$n_2^* = n_1$$
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 $\{x_1, \ldots, x_p\}$ of $p$ potential explanatory regressors (plus intercept)

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

For model $\mathcal{M}_\gamma$,
- $q_\gamma$ 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 $\beta \in \mathbb{R}^{p+1}$,
  \[
  \beta_{t_1(\gamma)} = \left[ \beta_0, \beta_{t_{1,1}(\gamma)}, \ldots, \beta_{t_{1,q(\gamma)}(\gamma)} \right],
  \]
  \[
  X_{t_1(\gamma)} = \left[ 1_n, x_{t_{1,1}(\gamma)}, \ldots, x_{t_{1,q(\gamma)}(\gamma)} \right].
  \]
Notations

For model $M_\gamma$,

- $q_\gamma$ 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 $\beta \in \mathbb{R}^{p+1}$,

\[
\begin{align*}
\beta_{t_1(\gamma)} &= \begin{bmatrix} \beta_0, \beta_{t_{1,1}(\gamma)}, \ldots, \beta_{t_{1,q_\gamma}(\gamma)} \end{bmatrix} \\
X_{t_1(\gamma)} &= \begin{bmatrix} 1_n \mid x_{t_{1,1}(\gamma)} \mid \ldots \mid x_{t_{1,q_\gamma}(\gamma)} \end{bmatrix}.
\end{align*}
\]

Submodel $M_\gamma$ is thus

\[
y \mid \beta, \gamma, \sigma^2 \sim \mathcal{N} \left( X_{t_1(\gamma)} \beta_{t_1(\gamma)}, \sigma^2 I_n \right)
\]
Global and compatible priors

Use Zellner’s $g$-prior, i.e. a normal prior for $\beta$ conditional on $\sigma^2$,

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

and a Jeffreys prior for $\sigma^2$,

$$\pi(\sigma^2) \propto \sigma^{-2}$$
Global and compatible priors

Use Zellner’s $g$-prior, i.e. a normal prior for $\beta$ conditional on $\sigma^2$,

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

and a Jeffreys prior for $\sigma^2$,

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

Resulting compatible prior

$$
N\left( \left( X_{t_1(\gamma)}^T X_{t_1(\gamma)} \right)^{-1} X_{t_1(\gamma)}^T \tilde{\beta}, c\sigma^2 \left( X_{t_1(\gamma)}^T X_{t_1(\gamma)} \right)^{-1} \right)
$$
Global and compatible priors

Use Zellner's $g$-prior, i.e. a normal prior for $\beta$ conditional on $\sigma^2$,

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

and a Jeffreys prior for $\sigma^2$,

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

Resulting compatible prior

$$\mathcal{N}\left(\left((X_{t1(\gamma)}^T X_{t1(\gamma)})^{-1} X_{t1(\gamma)}^T X \tilde{\beta}, c\sigma^2 \left(X_{t1(\gamma)}^T X_{t1(\gamma)}\right)^{-1}\right)\right)$$
For the hierarchical parameter $\gamma$, we use

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

where $\tau_i$ corresponds to the prior probability that variable $i$ is present in the model.
For the hierarchical parameter $\gamma$, we use

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

where $\tau_i$ 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

$$
\pi(\gamma) = 2^{-p}
$$
Model uncertainty and model choice: Bayesian tools
Compatible priors for variable selection
Variable selection

Posterior model probability

Can be obtained in closed form:

$$\pi(\gamma|y) \propto (c+1)^{-(q\gamma+1)/2} \left[ y^T y - \frac{cy^T P_1 y}{c + 1} + \frac{\beta^T X^T P_1 X \beta}{c + 1} - \frac{2y^T P_1 X \beta}{c + 1} \right]^{-n/2}.$$
Can be obtained in closed form:

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

Conditionally on \(\gamma\), posterior distributions of \(\beta\) and \(\sigma^2\):

\[
\beta_{t_1(\gamma)}|\sigma^2, y, \gamma \sim \mathcal{N}\left[\frac{c}{c+1}(U_1 y + U_1 X \tilde{\beta}/c), \frac{\sigma^2 c}{c+1} \left(X_{t_1(\gamma)}X_{t_1(\gamma)}\right)^{-1}\right],
\]

\[
\sigma^2 | y, \gamma \sim \mathcal{IG}\left[\frac{n}{2}, \frac{y^T y}{2} - \frac{cy^T P_1 y}{2(c+1)} + \frac{\tilde{\beta}^T X^T P_1 X \tilde{\beta}}{2(c+1)} - \frac{y^T P_1 X \tilde{\beta}}{c+1}\right].
\]
Noninformative case

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

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

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

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

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 \( c \),

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

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

Taking \( \tilde{\beta} = 0_{p+1} \) and \( c \) large does not work
Consider the 10-predictor full model

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

where the \( x_i \)s are iid \( \mathcal{U}(0,10) \)

[Casella & Moreno, 2004]
Consider the 10-predictor full model

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

where the \( x_i \)s are iid \( \mathcal{U}(0, 10) \)

[Casella & Moreno, 2004]

True model: two predictors \( x_1 \) and \( x_2 \), i.e.  \( \gamma^* = 110 \ldots 0 \), \( (\beta_0, \beta_1, \beta_2) = (5, 1, 3) \), and \( \sigma^2 = 4 \).
### Influence of $c^2$

<table>
<thead>
<tr>
<th>$t_1(\gamma)$</th>
<th>$c = 10$</th>
<th>$c = 100$</th>
<th>$c = 10^3$</th>
<th>$c = 10^4$</th>
<th>$c = 10^6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,1,2</td>
<td>0.04062</td>
<td>0.35368</td>
<td>0.65858</td>
<td>0.85895</td>
<td>0.98222</td>
</tr>
<tr>
<td>0,1,2,7</td>
<td>0.01326</td>
<td>0.06142</td>
<td>0.08395</td>
<td>0.04434</td>
<td>0.00524</td>
</tr>
<tr>
<td>0,1,2,4</td>
<td>0.01299</td>
<td>0.05310</td>
<td>0.05805</td>
<td>0.02868</td>
<td>0.00336</td>
</tr>
<tr>
<td>0,2,4</td>
<td>0.02927</td>
<td>0.03962</td>
<td>0.00409</td>
<td>0.00246</td>
<td>0.00254</td>
</tr>
<tr>
<td>0,1,2,8</td>
<td>0.01240</td>
<td>0.03833</td>
<td>0.01100</td>
<td>0.00126</td>
<td>0.00126</td>
</tr>
</tbody>
</table>
Noninformative case (cont’d)

In the noninformative setting,

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

converges for all \( y \)’s
Casella & Moreno’s example

$$t_1(\gamma) = \sum_{i=1}^{10^6} \pi(\gamma|y, c) \pi(c)$$

| $t_1(\gamma)$  | $\sum_{i=1}^{10^6} \pi(\gamma|y, c) \pi(c)$ |
|-----------------|-------------------------------------------|
| 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 $\gamma^0$ from the uniform distribution on $\Gamma$
- At $t$, for $i = 1, \ldots, p$, draw
  $\gamma^t_i \sim \pi(\gamma_i|y, \gamma^t_1, \ldots, \gamma^t_{i-1}, \ldots, \gamma^t_{i+1}, \ldots, \gamma^t_p)$
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))

| $\gamma$                          | $\pi(\gamma|y)$ | $\pi(\gamma|y)^{GIBBS}$ |
|-----------------------------------|-----------------|--------------------------|
| 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 $\hat{\beta} = 0_{21}$ 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

\[ x_1 \text{ altitude (in meters)}, \ x_2 \text{ slope (in degrees)}, \]
\[ x_3 \text{ number of pines in the square,} \]
\[ x_4 \text{ height (in meters) of the tree at the center of the square,} \]
\[ x_5 \text{ diameter of the tree at the center of the square,} \]
\[ x_6 \text{ index of the settlement density,} \]
\[ x_7 \text{ orientation of the square (from 1 if southb’d to 2 ow),} \]
\[ x_8 \text{ height (in meters) of the dominant tree,} \]
\[ x_9 \text{ number of vegetation strata,} \]
\[ x_{10} \text{ mix settlement index (from 1 if not mixed to 2 if mixed).} \]
Model uncertainty and model choice: Bayesian tools
Compatible priors for variable selection
Application
## Bayesian regression output

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

evidence against H0: (****) decisive, (*** ) strong, (**) substantial, (*) poor
Bayesian variable selection

| $t_1(\gamma)$          | $\pi(\gamma|y, X)$ | $\hat{\pi}(\gamma|y, X)$ |
|------------------------|---------------------|---------------------------|
| 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]
Use for classification purposes of a training dataset

\[((y_i^{tr}, x_i^{tr}))_{i=1,...,n}\]

with class label \(1 \leq y_i^{tr} \leq Q\) and predictor variables \(x_i^{tr}\)
Classification

**Principle**

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

Neighbourhood based on a distance metric.
Principle

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

Neighbourhood based on a distance metric.
**Classification**

**Principle**

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

Neighbourhood based on a distance metric
**Principle**

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

Neighbourhood based on a distance metric.
Model uncertainty and model choice: Bayesian tools

$k$-nearest-neighbour classification

**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 \(x_j^{\text{te}}\) in the training set

Neighbourhood based on a distance metric
**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 \(x_{j}^{\text{te}}\) in the training set.

Neighbourhood based on a distance metric.
Model choice perspective

Choice of $k$?

Usually chosen by minimizing cross-validated misclassification rate (non-parametric or even non-probabilist!)
Formalisation thru a probability model

$k$ nearest neighbour model

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

$$P(y_{i}^{tr} = \omega | y_{-i}^{tr}, x^{tr}, \beta, k) \propto \exp \left( \beta \sum_{l \sim i}^{k} \frac{\delta_{\omega}(y_{l}^{tr})}{k} \right) \quad \beta > 0$$

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

[Holmes & Adams, 2002]
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 $x_i$),
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 $x_i^{tr}$.
Given the full conditionals

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

where $l \sim i$ 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

\[ P(y^\text{tr}_i = \omega | y^\text{tr}_{-i}, x^\text{tr}, \beta, k) \propto \exp \left( \beta \sum_{l \sim i}^k \delta_{\omega}(y^\text{tr}_l) / N(i) \right) \]

where \( l \sim i \) is the symmetrized \( k \) nearest neighbour relation, and \( N(i) \) denotes the size of the symmetrized \( k \)-nearest neighbourhood of \( x^\text{tr}_i \) there exists a corresponding joint distribution
Use for the predictive distribution of $y_{j}^{\text{te}} (j = 1, \ldots, m)$

$$
\mathbb{P}(y_{j}^{\text{te}} = \omega | x_{j}^{\text{te}}, y^{\text{tr}}, x^{\text{tr}}, \beta, k) \propto \exp \left( \beta \sum_{l \neq j}^{k} \delta_{\omega}(y_{l}^{\text{tr}}) / k \right)
$$

where $l \neq j$ 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 $y_{j}^{\text{te}}$ given $x_{j}^{\text{te}}$ ($j = 1, \ldots, m$)
Bayesian global inference

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

$$\int \mathbb{P}(y_{j}^{te} = \omega | x_{j}^{te}, y^{tr}, x^{tr}, \beta, k) \pi(\beta, k | y^{tr}, x^{tr}) d\beta dk$$

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

$[\hat{y}_{j}^{te} = \text{MAP estimate}]$
Bayesian global inference

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

$$\int P(y_{j}^{te} = \omega | x_{j}^{te}, y^{tr}, x^{tr}, \beta, k)\pi(\beta, k| y^{tr}, x^{tr})d\beta dk$$

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

$[\hat{y}_{j}^{te} = \text{MAP estimate}]$

Note

Model choice \textit{without} varying dimension because $\beta$ is the same on all models
To compute $f(y^{tr}|x^{tr}, \beta, k)$ requires a normalisation constant that is not readily available.
To compute $f(y^{tr}|x^{tr}, \beta, k)$ requires a normalisation constant that is not readily available.

**Approximation**

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

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

Even with this approximation, the computation of
\[ P(y_{j}^{te} = \omega | x_{j}^{te}, y^{tr}, x^{tr}) \] is not feasible.
Further difficulty

Even with this approximation, the computation of 
\[ P(y_{te}^j = \omega | x_{te}^j, y^{tr}, x^{tr}) \]
is not feasible.

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

\[
M^{-1} \sum_{i=1}^{M} P \left( y_{te}^j = 0 \left| x_{te}^j, y^{tr}, x^{tr}, (\beta, k)^{(i)} \right. \right)
\]

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

[Gibbs too costly]
MCMC for $k$-nearest-neighbours

Random walk $k$-nearest-neighbours

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

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

1. Generate $\log \tilde{\beta} \sim \mathcal{N}(\log \beta^{(t-1)}, \tau^2)$ and $\tilde{k} \sim \mathcal{U}(\{k - r, k - r + 1, \ldots, k + r - 1, k + r\})$
Model uncertainty and model choice: Bayesian tools

$k$-nearest-neighbour classification

Bayesian inference in $k$ mean models

MCMC for $k$-nearest-neighbours

Random walk $k$-nearest-neighbours

At time 0, generate $\beta^{(0)} \sim N(0, \tau^2)$ and $k^{(0)} \sim U\{1, \ldots, K\}$

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

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

2. Calculate Metropolis-Hastings acceptance probability $
\rho(\tilde{\beta}, \tilde{k}, \beta^{(t-1)}, k^{(t-1)})$
Model uncertainty and model choice: Bayesian tools

\textit{k}-nearest-neighbour classification

Bayesian inference in \textit{k} mean models

MCMC for \textit{k}-nearest-neighbours

Random walk \textit{k}-nearest-neighbours

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

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

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

2. Calculate Metropolis-Hastings acceptance probability $\rho(\tilde{\beta}, \tilde{k}, \beta^{(t-1)}, k^{(t-1)})$

3. Move to $(\beta^{(t)}, k^{(t)})$ by Metropolis-Hastings step
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
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.
Use of the prior

\[ \pi(\beta, k) \propto \mathbb{I}_{(0,15)}(\beta) \mathbb{I}_{\{1,\ldots,\lfloor n/2 \rfloor\}}(k) \]
Use of the prior

$$\pi(\beta, k) \propto \mathbb{I}_{(0,15)}(\beta) \mathbb{I}_{\{1, \ldots, \lfloor n/2 \rfloor\}}(k)$$
Same label allocation and same misclassification rate (8.4%) for both algorithms
Model uncertainty and model choice: Bayesian tools

\(k\)-nearest-neighbour classification

Global classification

Alternative perspective

Lack of coherence of previous predictive:

- Each testing point processed marginally
Alternative perspective

Lack of coherence of previous predictive:

- Each testing point processed marginally
- Different distribution for training and testing points
Model uncertainty and model choice: Bayesian tools

$k$-nearest-neighbour classification

Global classification

Alternative perspective

Lack of coherence of previous predictive:

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

Lack of coherence of previous predictive:
- Each testing point processed marginally
- 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

$y = (y^{tr}, y^{te}) = (y_1, \ldots, y_{n+m})$ and

$x = (x^{tr}, x^{te}) = (x_1, \ldots, x_{n+m})$

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

where $l \# i$ 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 $x_i$ ($1 \leq i \leq n + m$)
Model uncertainty and model choice: Bayesian tools

\textit{k\textendash}nearest-neighbour classification

Global classification

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 $1 \leq t \leq T$,

1. For $n + 1 \leq i \leq n + m$, compute
   \[ q_i = \mathbb{P} \left( y_i = 1 \mid y_{-i}^{(t)}, x, \beta^{(t-1)}, k^{(t-1)} \right) \]
   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 $(\beta^{(t-1)}, k^{(t-1)})$
For Ripley’s benchmark and testing sample of 1,000 points, use of prior

\[ \pi(\beta, k) \propto I_{0 \leq \beta \leq 15} \{1, \ldots, \lfloor \frac{m+n}{2} \rfloor\}(k) \]

and misclassification rate 8.3%
Model uncertainty and model choice: Bayesian tools

$k$-nearest-neighbour classification

Global classification

Benchmark illustration

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

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

and misclassification rate 8.3%
Extensions

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