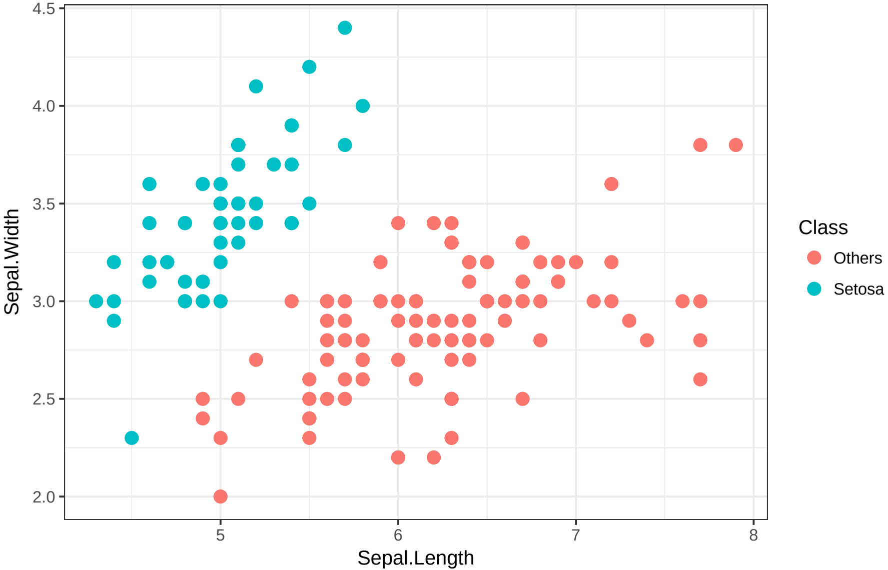
Fisher's Iris data set



Fisher's Iris data set - Model fitting

- Varitional inference for I-prior probit models implemented in R package iprobit (still lots of work to do!).

```
R> system.time(
+   (mod <- iprobit(y, X))
+ )


##
## |=================================                     |  61%
##  Converged after 6141 iterations.
##  Training error rate: 0 %
##     user  system elapsed
##   67.857   6.396  74.277
```

HJ (2017). *iprobit: Binary Probit Regression with I-priors*. R Package version 0.1.0: GitHub

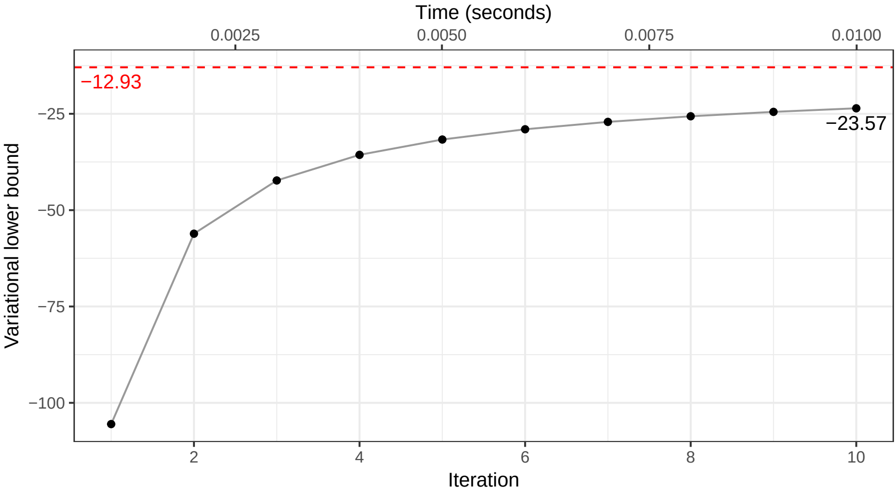
Fisher's Iris data set - Model summary

```
R> summary(mod)

##
## Call:
## iprobit(y = y, X, maxit = 10000)
##
## RKHS used: Canonical
##
##            Mean   S.E.    2.5%   97.5%
## alpha  -4.1730 0.0816 -4.3330 -4.0129
## lambda  1.2896 0.0142  1.2618  1.3175
##
## Converged to within 1e-05 tolerance. No. of iterations: 6141
## Model classification error rate (%): 0
## Variational lower bound: -12.93486
```

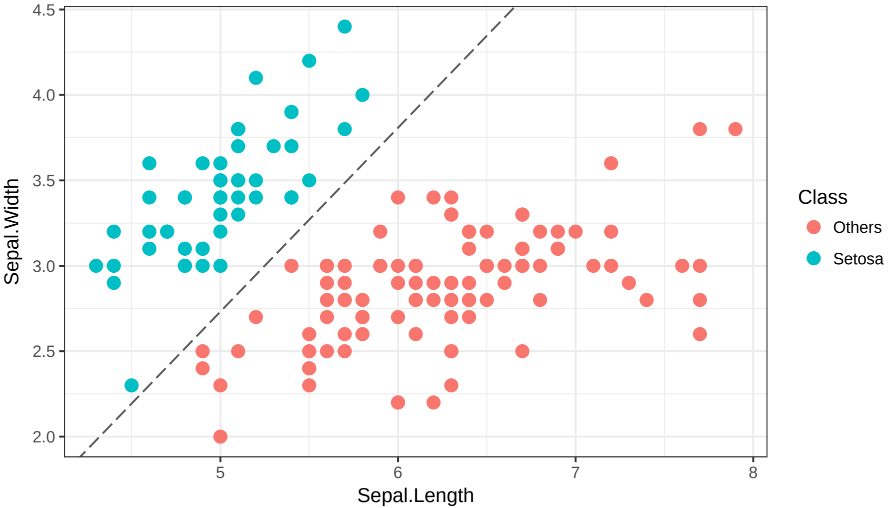
Fisher's Iris data set - Lower bound

```
R> iplot_lb(mod, niter.plot = 10)
```



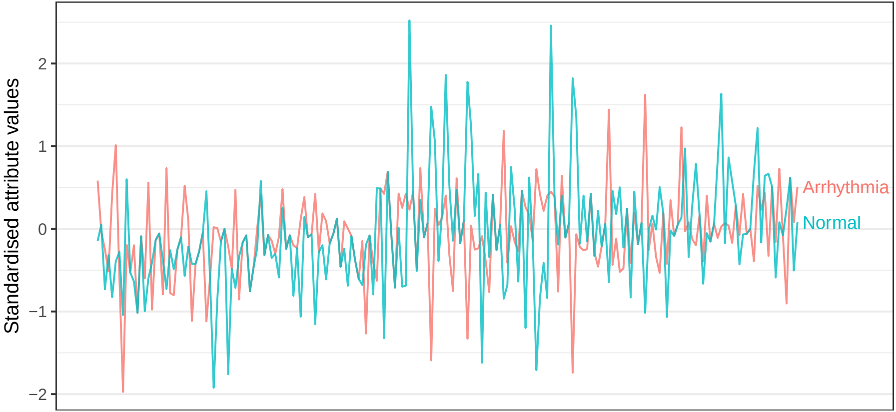
Fisher's Iris data set - Decision boundary

```
R> iplot_decbound(mod)
```



Cardiac arrhythmia data set

- Detect the presence of cardiac arrhythmia based on various ECG data and other attributes such as age and weight ($n = 451, p = 194$).



H. A. Guvenir, M. Burak Acar, and H. Muderrisoglu (1998). *UCI Machine Learning Repository: Arrhythmia Data Set*. URL: https://archive.ics.uci.edu/ml/datasets/Arrhythmia

Cardiac arrhythmia data set - Model fit

- Fit an I-prior probit model using Canonical and FBM kernel. The full data set takes about 35 seconds.

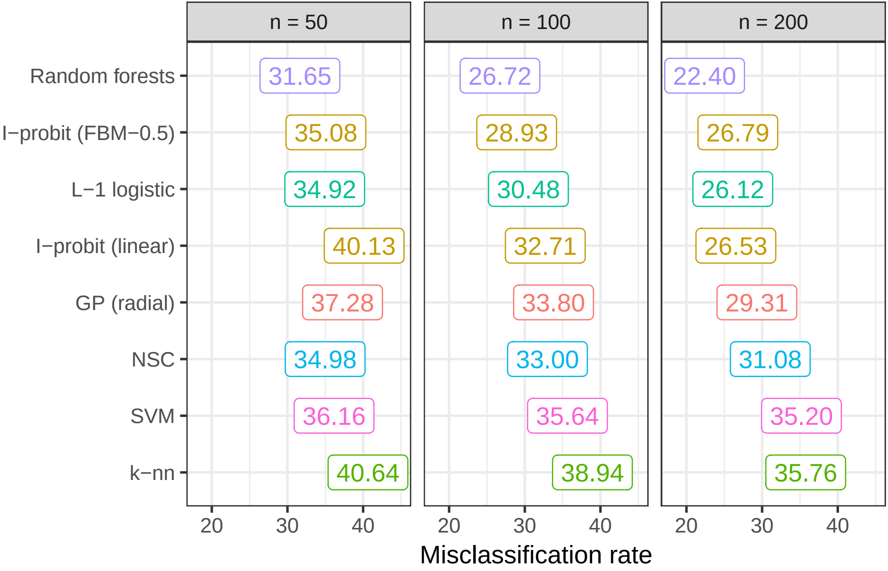
```
R> mod <- iprior(y, X, kernel = "FBM")
```

- Compare against popular classifiers: 1) *k*-nearest neighbours; 2) support vector machine; 3) Gaussian process classification; 4) random forests; 5) nearest shrunken centroids (Tibshirani, Hastie, Narasimhan, and Chu 2003); and 6) L-1 penalised logistic regression.

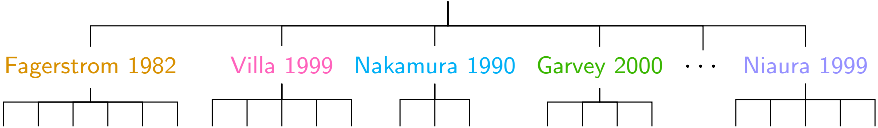
- Experiment set-up:
  - Form training set by sub-sampling $n_{\text{sub}} \in \{50, 100, 200\}$ data points.
  - Use remaining data as test set.
  - Fit model on training set and obtain test error rates.
  - Repeat 100 times.

T. I. Cannings and R. J. Samworth (2017). "Random-projection ensemble classification". *J. R. Stat. Soc. Ser. B: Stat. Methodol (w. discussion)*, to appear

Cardiac arrhythmia data set - Results



Meta-analysis of smoking cessation



- Data from 27 separate smoking cessation studies, where participants subjected to nicotine gum treatment or placed in control group.
- Some summary statistics:

	Min.	Avg.	Max.	Prop. quit	Odds quit
Control	20	101	617	0.207	0.261
Treated	21	117	600	0.320	0.470

- Raw odds ratio: 1.801.
- Random-effects analysis using a multilevel logistic model estimates this odds ratio as 1.768.

A. Skrondal and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC, §9.5

Meta-analysis of smoking cessation - model

- Let $i = 1, \dots, n_j$ index the patients in study group $j \in 1, \dots, 27$.
- Denote y_{ij} as the binary response variable indicating `Quit` (1) or `Remain` (0), and x_{ij} as patient$_{ij}$'s treatment group indicator.
- Model binary data using I-probit model

$$\begin{aligned}\Phi^{-1}(p_{ij}) &= f(x_{ij}, j) \\ &= f_1(x_{ij}) + f_2(j) + f_{12}(x_{ij}, j)\end{aligned}$$

with $f_1, f_2 \in$ Pearson RKHS, and $f_{12} \in$ ANOVA RKHS.

	Model	Lower bound	Brier score	No. of RKHS param.
1	f_1	-3210.79	0.0311	1
2	$f_1 + f_2$	-3097.24	0.0294	2
3	$f_1 + f_2 + f_{12}$	-3091.21	0.0294	2

Meta-analysis of smoking cessation - results