# Bayesian analysis of Heart Disease data

Project in Bayesian Learning course
Linköping University, Fall semester

Table of Contents

# 1 Project description

This project will analyze the cause for heart disease using Bayesian variable and model selection techniques. The goal is to find which variables give an increased risk to obtain heart disease and also to find a model to be able to describe the outcome as good as possible. Good alternatives to the best model will be found using the variable selection techniques Gibbs variable selection, the Kuo-Mallick approach, Stochastic search variable selection and the Horseshoe regression. A group of good alternative models will then be tested against each other (using the marginal likelihood as goodness criteria) to conclude which one truly is the best one.

# 2 Data

The data used in this project is the heart disease dataset available from the UCI repository, at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. There are 4 sources for the data but I only chose to analyze the data coming from V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. The other 3 data sets have a lot of missing data and I chose to not include them in order to fully concentrate on Bayesian ways to conduct variable and model selection.
I chose to analyze the impact of 11 predictors. The data with description is shown below.

**Predictors**

| *Name* | *Type* | *Description* |
|---|---|---|
| Age | Continuous | Age in years |
| Sex | Discrete | 1 = male, 0 = female |
| Cp | Discrete | Chest pain type: 1 = typical angina,<br>2 = atypical angina, 3 = non-anginal pain, 4 =asymptomatic |
| Trestbps | Continuous | Resting blood pressure (in mm Hg) |
| Chol | Continuous | Serum cholesterol in mg/dl |
| Fbs | Discrete | Fasting blood sugar > 120 mg/dl: 1 = true , 0=false |
| Restecg | Discrete | Resting electrocardiographic results:0 = normal,<br>1 = having ST-T wave abnormality, 2 =showing<br>probable or define left ventricular hypertrophy by Estes'criteria |
| Thalach | Continuous | Maximum heart rate achieved |
| Exang | Discrete | Exercise induced angina: 1 = yes, 0 = no |
| Old peak ST | Continuous | Depression induced by exercise relative to rest |
| Slope | Discrete | The slope of the peak exercise segment : 1 = up sloping, 2 = flat, 3= down sloping |
| **Response** | Discrete | 1= presence of heart disease, 0= no presence |

*Table 1: The data with description*

## 2.1 Data manipulation

In the original data there are some nominal variables which were binarized. The resulting predictor matrix X contains 15 variables.

There were 6 observations in the original data with missing values which I chose to remove in order to fully concentrate on variable selection techniques. The resulting dataset contains 303 observations.

# 3 Softwares

For the horseshoe regression and calculation of the marginal likelihood I have used the Bayesian statistical software Stan.(Stan, 2014)

For the Gibbs based variables selection techniques I have used the Bayesian statistical software OpenBUGS. (OpenBUGS, 2014)

OpenBUGS is based on the Gibbs sampler while Stan is not and this is the reason I didn't Stan for all the tasks.

# 4 Logistic regression model

Every response can be modeled as coming from the Bernoulli distribution. The Bernoulli distribution is a discrete distribution having two outcomes (e.g Presence of heart disease and no presence) where presence has probability $\pi$. Since it is reasonable to assume that the observations were sampled independently from each other one could model the whole data by simply multiplying every single Bernoulli distribution. The resulting model is the likelihood function :

$$f(y|\pi) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}$$

where

$$\pi(x) = \frac{e^{\boldsymbol{\beta X}_i}}{1+e^{\boldsymbol{\beta X}_i}} \quad .$$

In the Bayesian setting one always specifies a prior distribution for each predictor. Multiplying the priors with the likelihood function yields a expression proportional to the posterior distribution. The prior will be different depending on what variable selection technique is used and the different priors are covered in their respective section. For the regression model with no variable selection (comparing models) I choose to set weakly to non-informative priors. This is because I have no insight in the heart disease field and thus don't have any prior knowledge. Every prior were modeled as a normal distribution with mean 0 and a high standard deviation. The general setting is shown below:

$$f(\pi|y) \propto \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \times \prod_{j=0}^{p} \frac{1}{\sqrt{2\pi\sigma_j}} \exp\{-\frac{1}{2}\left(\frac{\beta_j-\mu_j}{\sigma_j}\right)^2\} \quad .$$

# 5 Gibbs based Variable selection

Gibbs based variable selection is based on effectively searching the model space for the most probable models in order to be able to estimate the posterior model probabilities for these models. The posterior model probability given data can be written as:

$$f(m_k|y) = \frac{f(m_k)(f(y|m_k)}{\sum_{m_k \in M} f(m_k)(f(y|m_k)}, \quad m_k \in M$$

where a certain model is denoted by $m_k$ and the whole set of models by *M*. During every iteration of the Gibbs sampler there will be one best model chosen using a method specific criteria . The posterior model probability for model *k* is thus the proportion of times that model is drawn in the sampling process. (Clyde & Edward, 2004)

The Gibbs based variable section techniques considered in this project will be the Gibbs variable selection, the Kuo- Mallick approach and Stochastic search variable selection. (Ntzoufras, 2009)

In variable selection one could express a model using a model indicator $\gamma$ that indicates which variables are included in a certain model. eg. $\gamma = [1, 1, 0, 1]$ indicates that the two first and the last variables are included in the model. One could now express the model as:

$$\phi = \sum_{j=0}^{p} \gamma_j X_j \beta_j \ .$$

In the following sections, describing these three techniques it will be beneficial to express $f(m_k)$as $f(\gamma)$ and to partition $\beta$ as $(\beta_\gamma, \beta_{|\gamma})$ where $\beta_{|\gamma}$ defines the $\beta$ not in the model. The prior for the model indicator were set to uninformative so that all possible combinations of $\gamma$ are considered:

$f(\gamma_j) \sim Bern(U[0, 1])$

## 5.1 Posterior inclusion probability

The posterior inclusion probability (PIP) for a variable is the proportion of times that variable is included in the best models drawn in the sampler. It can be seen as a measure of importance. More precisely:

$$PIP_j = P(\beta_j \in model|Y) = \sum_{k:\beta_j \in m_k} P(m_k|Y)$$

A PIP larger than 0.5 indicates that a variable is more times included in the best models than excluded. A model including variables with a PIP larger than 0.5 is called the median probability model and can be a candidate for the final best model.(Barbieri and Berger, 2004)

## 5.2 Gibbs variables selection

The description of the Gibbs variable selection follow Ntzoufras (2009).

In the Gibbs variable selection(GVS) the prior of $\beta$ is dependent on $\gamma$ and one can express $f(\beta_\gamma|\gamma)$ and $f(\beta_{|\gamma}|\gamma)$ as priors of $\beta$ for variables in a model and for variables not in the model, respectively.

In the GVS the full conditional posterior for $\beta$ in the model is given by:
$f(\beta_\gamma|\beta_{|\gamma},\gamma,y) \propto f(y|\gamma,\beta)f(\beta_\gamma|\gamma)$
and for $\beta$ not in the model:
$f(\beta_{|\gamma}|\beta_\gamma,\gamma,y) \propto f(\beta_{|\gamma}|\gamma)$ .

Further, one can express the model prior $f(\beta_j|\gamma_j=1,\gamma_{|j})$ and the pseudo prior $f(\beta_j|\gamma_j=0,\gamma_{|j})$ as priors of $\beta_j$ when the variable is in a model and when it is not in a model, respectively.

Now the full posterior for the model indicator $\gamma_j$ can be written as:
$\gamma_j|\beta,\gamma_j,y \sim Bern(\frac{O_j}{1+O_j})$
where
$$O_j = \frac{f(\gamma_j=1|\gamma_{|\gamma},\beta,y)}{f(\gamma_j=0|\gamma_{|\gamma},\beta,y)} = \frac{f(y|\beta,\gamma_j=1,\gamma_{|j})}{f(y|\beta,\gamma_j=0,\gamma_{|j})}\frac{f(\beta|\gamma_j=1,\gamma_{|j})}{f(\beta|\gamma_j=0,\gamma_{|j})}\frac{f(\gamma_j=1,\gamma_{|j})}{f(\gamma_j=0,\gamma_{|j})}.$$

This ratio determines how likely $\beta_j$ should be drawn to a model in the sampling process. A $O_j$ greater than 1 indicates that the probability that variable $j$ is drawn to the model is larger than 0.5. We see that this ratio is composed of a likelihood ratio (general appropriateness of having $\beta_j$ in the model), ratio between model prior and pseudo prior and a ratio between two draws from $f(\gamma_j)$ (to add some randomness to $O_j$).

Using an empirical Bayes prior the full conditional posterior distribution for $\beta_j$ becomes:

$f(\beta_j|\beta_{|j},\gamma,y) \propto f(y|\gamma,\beta)N(\bar{\mu}_j,n\bar{S}_j)$when $\gamma_j=1$
$f(\beta_j|\beta_{|j},\gamma,y) \propto N(\bar{\mu}_j,\bar{S}_j)$ when $\gamma_j=0$.

The $n$ indicates the number of observations and $\bar{\mu_j}$ and $\bar{S_j}$ are set as the MLE from a run of the full model. Now we see that the second ratio in $O_j$ (ratio of $\beta$ priors ) has a high probability of being less than 1. This acts as a penalty to variables in the model (variables with $\gamma_j=1$). The GVS thus tries to steer these variables away from dominating all drawn best models in order to search the model space and give less probable variables a chance to enter the model. When $n$ is low one can see the sampler being in the same state for too long giving less chance to less probable variables to enter the model.

## 5.3 Kuo- Mallick approach

The description of the Kuo- Mallick approach follow Ntzoufras (2009).

In the Kuo- Mallick approach the prior of $\beta$ is independent of $\gamma$. It can be seen as a simplified version of GVS where one uses the same prior for all $\beta_j$.

The full conditional posterior for $\beta_j$ is given by
$f(\beta_j|\beta_{|j},\gamma,y) \propto f(y|\gamma,\beta)f(\beta_j|\beta_{|j})$ when $\gamma_j = 1$
$f(\beta_j|\beta_{|j},\gamma,y) \propto f(\beta_j|\beta_{|j})$ when $\gamma_j = 0$

In the application I have chosen to set $f(\beta_j|\beta_{|j})$ as the weakly informative $N(0,8)$.

The full posterior for the model indicator $\gamma_j$ is:
$\gamma_j|\beta,\gamma_j,y \sim Bern(\frac{O_j}{1+O_j})$
where
$$O_j = \frac{f(\gamma_j = 1|\gamma_{|\gamma},\beta,y)}{f(\gamma_j = 0|\gamma_{|\gamma},\beta,y)} = \frac{f(y|\beta,\gamma_j = 1,\gamma_{|j})}{f(y|\beta,\gamma_j = 0,\gamma_{|j})}\frac{f(\gamma_j = 1,\gamma_{|j})}{f(\gamma_j = 0,\gamma_{|j})}$$

In the $O_j$ we see that the choice of which variables should be drawn to a model only are composed of its appropriateness (ratio of likelihoods) and a ratio of randomness. There is no mechanic to steer the appropriate variables away from dominating every drawn model except by chance. This will result in that less appropriate variables only will be considered in a model if they are lucky (if the randomness ratio is very low) and the jumping between states will happen less frequently than in GVS.

## 5.4 Stochastic search variable selection

The description of the Kuo- Mallick approach follow Ntzoufras (2009).

The difference to KM and GVS is that one excludes $\gamma_j$ from the linear expression $\phi$. This results in that the $O_j$ is calculated without the likelihood ratio and that the determination of which variable should have a high probability to be drawn (high $O_j$) only depends on the prior for $\beta_j$ and the last random ratios of $f(\gamma_j)$. The Prior

$\beta_j|\gamma_j \sim \gamma_j N(0,\Sigma_1) + (1+\gamma_j)N(0,\Sigma_2)$

includes $\gamma_j$ just as in GVS and works in the same fashion by giving high values to $\Sigma_1$ and low to $\Sigma_2$ .

Since the choice of variances plays an important role in this method I chose to compare two sets of variances to see how the results will differ. These are $\Sigma_1 = 30, \Sigma_2 = 0.003$ and
$\Sigma_1 = 10, \Sigma_2 = 0.01$. Reasonably one would think that a higher penalty to variables in the model ($\Sigma_1$) will lead to less jumping between states of the indicator $\gamma_j$.

# 6 Horseshoe regression

Horseshoe regression is a Bayesian shrinkage approach which shrinks the coefficients in the full model in order to try to explain the outcome using only the most important variables. This works by applying horseshoe priors to each $\beta_j$. The horseshoe prior can be written as:

$\beta_j|\lambda_j, \tau \sim N(1, \lambda^2\tau^2)$
$\lambda_j \sim C^+(0, 1)$
$\tau \sim C^+(0, 1)$
where $C^+$ denotes the half-Cauchy distribution and $\lambda_j$ and $\tau$ are shrinkage parameters. The prior has the characteristics that it has a high peak at center mass and heavy tails that never reach zero density. This prior is similar to the Laplace prior giving the Bayesian lasso. The difference in the priors is that the horseshoe has a even taller peak and a bit heavier tails allowing better separation between shrunk and not shrunk $\beta$. Looking at the prior one can see that it is more likely to shrink already small coefficients and leave the larger ones less or not shrunk. (Carvalho et al. ,2009).

The median of the posterior distribution of $\lambda_j$ can be obtained which can be seen as a measure of how much coefficient $j$ has been shrunk. This together with the $\tilde{\beta}_j$ will give us an indication of variable importance.

# 7 Model comparison

In the variable selection results one finds which variables and models are considered as best according to different variable selection techniques. Different variable techniques may choose different models as the best and in order to find the model that truly is the best we need to test these best models to some universal measure. This can be done by computing the marginal likelihood or evidence $P(D)$for each considered model and comparing these to each other.

By the Bayes Factor (assuming equal priors) one can say that there is evidence for a model $j$ over a model $k$ if the ratio

$$\frac{P(D|m_j)}{P(D|m_k)}$$

is larger than 1.

## 7.1 The modified harmonic mean

The modified harmonic mean is one way to compute the marginal likelihood $P(D)$ using MCMC simulations of the likelihood. It is defined as:

$$\frac{1}{P(D)} = \frac{1}{M} \sum_{\beta_j \sim p_{trunc}(w|D)} \frac{\Phi_i}{p(D|\beta_j)p(\beta_j)}$$

where $\Phi_i$ is the multivariate normal density with the i'th draw of $\beta$ as mean vector and a common covariance matrix estimated using all the $\beta$ draws. The summation is over the number of MCMC simulations $M$ with extreme draws removed in order to make the estimates more realistic. (Geweke, 2009)

# 8 Results for the variable selection

For the Gibbs based variable selection techniques I used 10000 burn in iterators and sampled the PIP for 103000 iterations. The PMP were only sampled during the last 3000 iterations. It wasn't possible to sample the model indicators for more than 3000 iterations because of the large size of the resulting data files. The software gave errors. The result from this is that the results for the PMP might be misleading since 3000 iterations might not be enough to properly discover the model space in some cases.

## 8.1 Posterior inclusion probabilities for the Gibbs based techniques

The posterior inclusion probabilities for each method are shown in table 2. One can note that Sex, Oldpeak, Cp1, Cp2, and Cp3 are in the median probability model (model including all PIP>0.5) in all the 4 methods. Slope2 has a PIP> 0.5 for GVS and KM while thalac has PIP>0.5 for only GVS and exang has PIP>0.5 for only KM.

Additionally one can note a pattern that some variables with high PIP in GVS have 0 or close to 0 PIP in the other methods. These are age, trestbps, thalac and chol.

| *Variable* | *GVS* | *SSVS(0.003,30)* | *SSVS(0.01,10)* | *KM* |
|---|---|---|---|---|
| age | 0.29 | 0.01 | 0.03 | 0.01 |
| sex | **1.00** | **1.00** | **1.00** | **1.00** |
| trestbps | 0.32 | 0.01 | 0.03 | 0.00 |
| chol | 0.27 | 0.01 | 0.03 | 0.00 |
| fbs | 0.07 | 0.07 | 0.12 | 0.19 |
| thalach | **0.51** | 0.01 | 0.03 | 0.00 |
| exang | 0.48 | 0.23 | 0.38 | **0.79** |
| oldpeak | **0.98** | **0.86** | **0.82** | **0.99** |
| cp1 | **0.99** | **0.99** | **0.99** | **0.99** |
| cp2 | **0.96** | **0.98** | **0.96** | **0.97** |
| cp3 | **1.00** | **1.00** | **1.00** | **1.00** |
| restecg0 | 0.20 | 0.09 | 0.17 | 0.36 |
| restecg1 | 0.08 | 0.18 | 0.28 | 0.37 |
| slope1 | 0.21 | 0.27 | 0.33 | 0.31 |
| slope2 | **0.53** | 0.36 | 0.49 | **0.87** |

*Table 2: Posterior inclusion probabilities for Gibbs based techniques*

The further the variances in SVSS are form each other the larger separation between PIP values is obtained as one should expect. The method with the largest separation in PIP values are KM. This is the result of the missing likelihood ratio in $O_j$ and equal prior for all $\beta$.

## 8.2 Posterior model probabilities for the Gibbs based techniques

The posterior model probabilities (PMP) were calculated for every Gibbs based technique. These values are based on only the last 3000 iterations. Table 3 reveals the three best models for every technique. The thing in common for every model is that they include sex+oldpeak+cp1+cp2+cp3. Slope is included in 6 of the models and exang is included in two of the models. GVS includes trestbps in all of its best models  while it is absent from all the other models. There are 4 models which are frequent 2 times , sex+oldpeak+cp1+cp2+cp3, sex+oldpeak+cp1+cp2+cp3+slope1, sex+oldpeak+cp1+cp2+cp3+slope2 and sex+exang+oldpeak+cp1+cp2+cp3. All models in the table will be further considered as candidates for the best models.

| GVS | PMP | SSVS (30, 0.003) | PMP |
|---|---|---|---|
| Sex+Trestbps+Thalac+Oldpeak+Cp1+Cp2+Cp3+slope2 | 0.16 | Sex+Oldpeak+Cp1+Cp2+Cp3+slope2 | 0.24 |
| Sex+Trestbps+Thalac+Oldpeak+Cp1+Cp2+Cp3 | 0.09 | Sex+Oldpeak+Cp1+Cp2+Cp3 | 0.15 |
| Sex+Trestbps+Thalac+Exang+Oldpeak+Cp1+Cp2+Cp3 | 0.07 | Sex+Exang+Oldpeak+Cp1+Cp2+Cp3 | 0.14 |

| SSVS(0.01,10) | PMP | KM | PMP |
|---|---|---|---|
| Sex+Oldpeak+Cp1+Cp2+Cp3 | 0.14 | Sex+Oldpeak+Cp1+Cp2+Cp3+slope2 | 0.11 |
| Sex+Exang+Oldpeak+Cp1+Cp2+Cp3 | 0.08 | Sex+Oldpeak+Cp1+Cp2+Cp3+slope1 | 0.09 |
| Sex+Oldpeak+Cp1+Cp2+Cp3+slope1 | 0.07 | Sex+Oldpeak+Cp1+Cp2+Cp3+Restecg1+slope2 | 0.06 |

*Table 3: Posterior model probabilities for Gibbs based techniques*

## 8.3 Results horseshoe regression

The results for the horseshoe regression are based on 10000 burn ins and 100000 drawn proper chains.

The result for the horseshoe regression is summarized in table 4 with the $\tilde{\beta}$ and $\tilde{\lambda}$ for every variable. The higher $\tilde{\lambda}$ the more important is the variable according to this method. The highest $\tilde{\lambda}$ are in falling order for sex, cp1,cp3,cp2, oldpeak, exang, slope1, slope2, restecg0, restecg1, … The horseshoe regression thus shows similar results as KM and SSVS .

| Variable | Beta median | Lambda median |
|---|---|---|
| age | 0.00 | 0.28 |
| sex | 1.62 | 3.12 |
| trestbps | 0.00 | 0.31 |
| chol | 0.00 | 0.21 |
| fbs | 0.00 | 0.68 |
| thalac | 0.00 | 0.51 |
| exang | 0.50 | 1.40 |
| oldpeak | 0.53 | 1.70 |
| cp1 | -1.95 | 2.95 |
| cp2 | -1.11 | 2.36 |
| cp3 | -1.85 | 2.79 |
| restecg0 | -0.05 | 0.83 |
| restecg1 | 0.00 | 0.70 |
| slope1 | -0.03 | 1.02 |
| slope2 | 0.15 | 0.97 |

*Table 4: Results for the Horseshoe regression*

# 9 Results model selection

The marginal likelihoods $P(D)$ for different prior distributions are shown in table 5. The choice of analyzing the impact of the prior on $P(D)$ lies on that $P(D)$ is very dependent on the prior and different priors may give different best models. Another thing to note is that the results for large priors like $N(0, 100)$ are very unstable and may give different results between runs. To deal with this problem I have chosen to compute the $P(D)$ three times for every model and prior and then average these three $P(D)$ to get a final estimate. One computation of $P(D)$ is based on 10000 burn ins and 100000 proper draws.

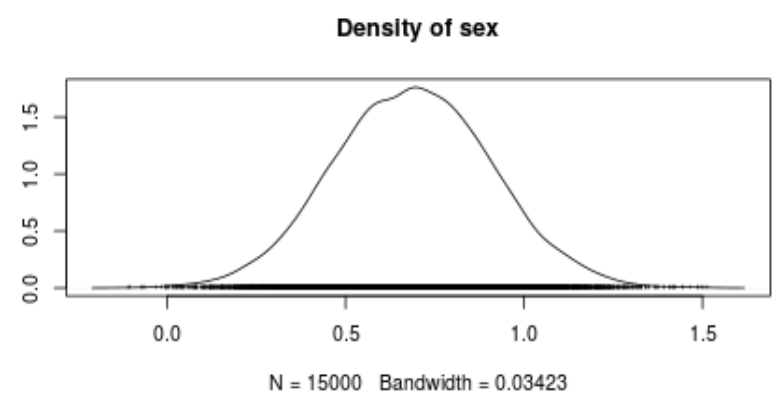
One sees that the two best models for both standard deviation include Sex+Oldpeak+Cp1+Cp2+Cp3+Thalac and the one when prior $N(0, 10)$ also includes Slope2 while the one when prior is $N(0, 100)$ includes Exang. When $N(0, 100)$ one clearly sees that every model including thalac has much lower $P(D)$ compared to when it is not included.

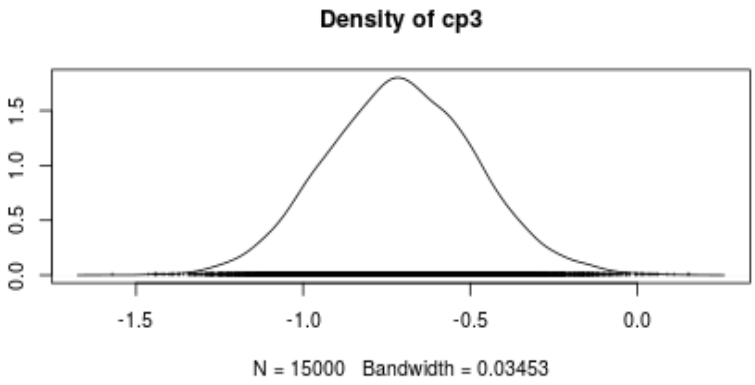
| Model | P(D) when P(beta)=N(0,10) | P(D) when P(beta)=N(0,100) |
|---|---|---|
| Sex+Oldpeak+Cp1+Cp2+Cp3 | 3.23E-74 | 1.17E-084 |
| Sex+Oldpeak+Cp1+Cp2+Cp3+slope1 | 9.29E-73 | 7.12E-084 |
| Sex+Oldpeak+Cp1+Cp2+Cp3+slope2 | 3.63E-72 | 6.12E-084 |
| Sex+Oldpeak+Cp1+Cp2+Cp3+Exang | 2.56E-71 | 1.59E-083 |
| Sex+Oldpeak+Cp1+Cp2+Cp3+Thalac | 1.90E-70 | 3.48E-077 |
| Sex+Oldpeak+Cp1+Cp2+Cp3+Thalac+Exang | 7.79E-69 | **1.62E-076** |
| Sex+Oldpeak+Cp1+Cp2+Cp3+Thalac+Slope2 | **1.17E-69** | 7.89E-077 |
| Sex+Oldpeak+Cp1+Cp2+Cp3+Exang+Slope1 | 1.84E-70 | 9.43E-083 |
| Sex+Oldpeak+Cp1+Cp2+Cp3+Exang+Slope2 | 9.23E-70 | 9.78E-083 |
| Sex+Oldpeak+Cp1+Cp2+Cp3+Slope1+Slope2 | 8.97E-72 | 3.18E-083 |
| Sex+Oldpeak+Cp1+Cp2+Cp3+Exang+Slope1+Slope2 | 1.41E-69 | 3.80E-082 |
| Sex+Oldpeak+Cp1+Cp2+Cp3+Restecg1+Slope2 | 6.22E-73 | 6.65E-084 |
| Sex+Trestbps+Thalac+Oldpeak+Cp1+Cp2+Cp3 | 5.53E-71 | 1.68E-077 |
| Sex+Trestbps+Thalac+Exang+Oldpeak+Cp1+Cp2+Cp3 | 4.37E-71 | 4.12E-077 |
| Sex+Trestbps+Thalac+Oldpeak+Cp1+Cp2+Cp3+Slope2 | 1.77E-70 | 6.84E-077 |

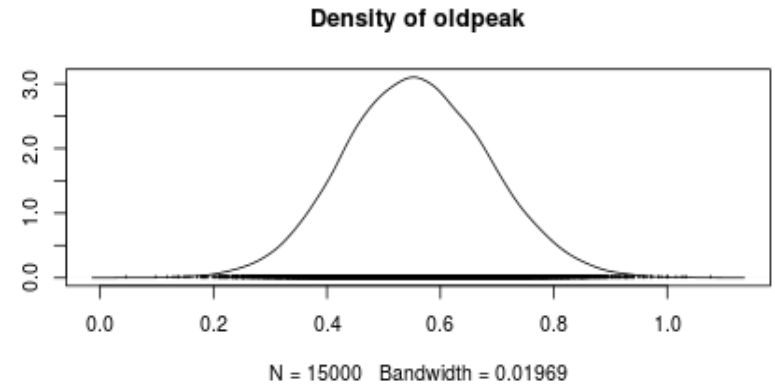
*Table 5: Marginal likelihoods for most probable models from variable selection*

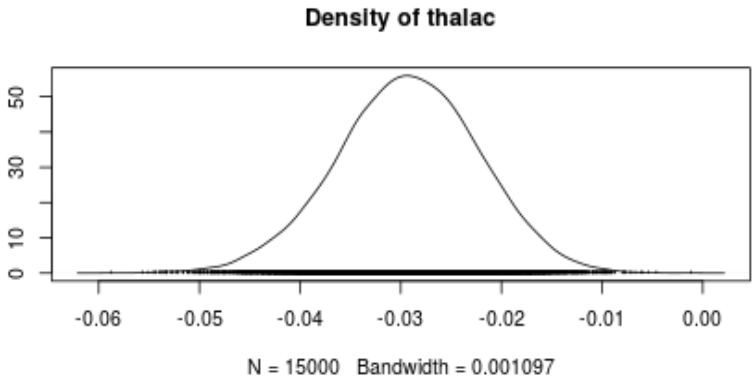
The difference between the two chosen standard deviations (10,100) are that the larger standard deviation more distinctly identifies the best models. The larger the standard deviation, the less informative is the prior and the more will the data by itself shape the results. Therefore I choose to believe the estimates with the largest standard deviation the most and accept Sex+Oldpeak+Cp1+Cp2+Cp3+Thalac+Exang as the best model.

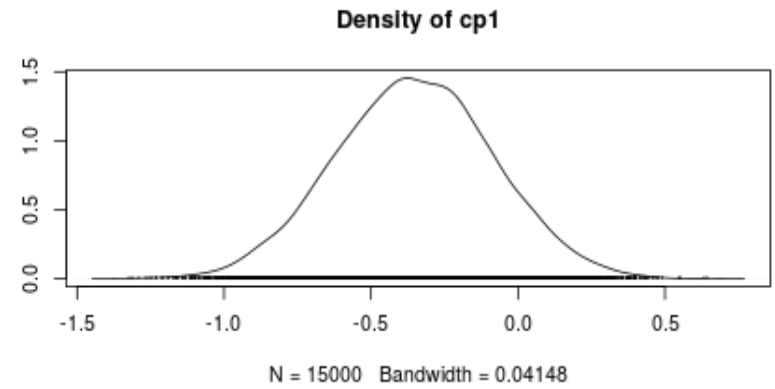
The marginal posterior distributions from the best model Sex+Oldpeak+Cp1+Cp2+Cp3+Thalac+ Exang are shown below. V

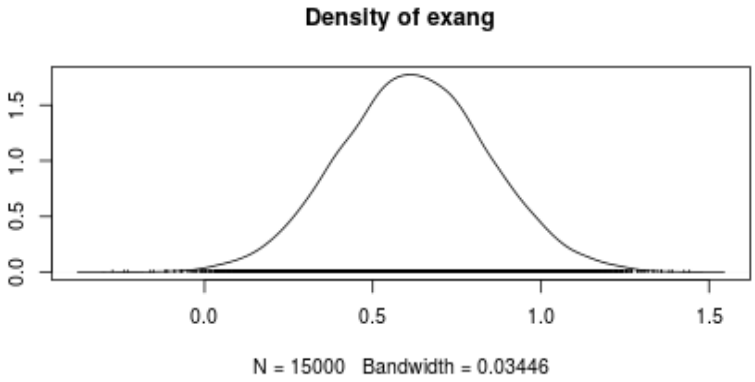


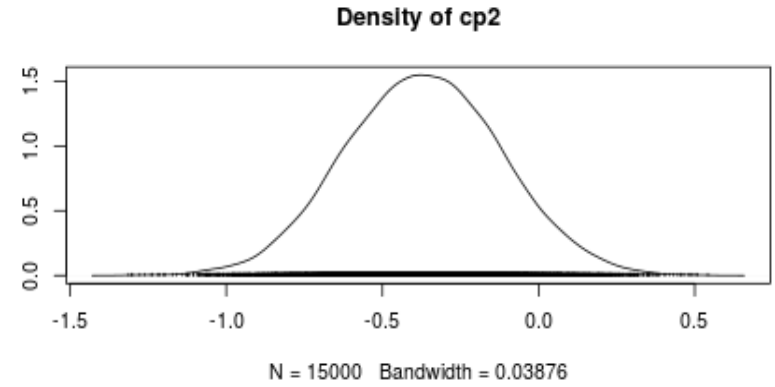




*Illustration 1: Posterior densities*

# 10 Conclusions

If one uses the marginal likelihood computed using the modified harmonic mean as a measure of model adequacy one sees that the Gibbs variable selection using empirical Bayes prior gives the best results. All the other techniques fail to identify one important variable thalac .The reason the GVS finds this one is because of the small proposal variance for this variable. All the other Gibbs based methods assume equal priors for variables in the same group (in model or not in the model) which seem to favor variables with larger variance.

One can say that the variables that have a substantial effect on the outcome of heart disease are sex, thalac, exang, oldpeak, cp and slope.

The best identified model includes Sex+Oldpeak+Cp1+Cp2+Cp3+Thalac+ Exang.

The interpretation of the best models coefficients is:

- Males have a higher risk than women to get heart disease.
- A high value of depression induced by exercise relative to rest gives a high risk to obtain heart disease compared to a small depression.
- People with exercise induced angina have a higher risk of getting a heart disease than those without.
- People with asymptomatic chest pain have higher risk of obtaing heart disease compared to if they have the three other chest pains typical angina, atypical angina, non-angina pain.

# 11 References


- Barbieri, M. M. & Berger, J. O. (2004). Optimal Predictive Model Selection, The Annals of Statistics 2004, Vol. 32, No. 3, 870–897
- Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott. “Handling Sparsity via the Horseshoe.” *Journal of Machine Learning Research* W&CP 5, no. 73–80 (2009): 111.
- Geweke, J. F. (1999): “Using Simulation Methods for Bayesian Econometric Models: Inference, Development and Communication,” Econometric Reviews, 18, 1–126.
- Clyde, M. & George, E. I.(2004) Model Uncertainty, Statistical Science, 2004, Vol. 19, No. 1, 81–94
- Ntzoufras, I. Bayesian Modeling Using WinBUGS, John Wiley & Sons, Inc, Hoboken, NJ, 2009
- Stan (2014)

http://mc-stan.org/

- OpenBUGS (2014)

http://www.openbugs.net

# Appendix

**R-CODE**

```
library("coda")
library("rstan")
setwd("~/Dropbox/Bayes/BayesProject")
cl<-read.csv("processed.cleveland.data.csv",na.strings="?",header=F)
colnames(cl)<-
c("age","sex","cp","trestbps","chol","fbs","restecg","thalach","exang","oldpeak","slope",
            "ca","thal","num")
#Code the data
#From disease 0:4, to 0,1
#Dummy for cp,restecg,slope
presence<-cp1<-cp2<-cp3<-restecg0<-restecg1<-slope1<-slope2<-rep(0,nrow(cl))
for( j in 1:nrow(cl)){
  if(cl$num[j]==0){presence[j]<-0}
  else{presence[j]<-1}
  if(is.na(cl$cp[j])){cp1[j]<-cp2[j]<-cp3[j]<-NA}
  else if(cl$cp[j]==1){cp1[j]<-1}
  else if(cl$cp[j]==2){cp2[j]<-1}
  else if(cl$cp[j]==3){cp3[j]<-1}

  if(is.na(cl$restecg[j])){restecg0[j]<-restecg1[j]<-NA}
  else if(cl$restecg[j]==0){restecg0[j]<-1}
  else if(cl$restecg[j]==1){restecg1[j]<-1}

  if(is.na(cl$slope[j])){slope1[j]<-slope2[j]<-NA}
  else if(cl$slope[j]==1){slope1[j]<-1}
  else if(cl$slope[j]==2){slope2[j]<-1}
}
cl<-cbind(cl,presence,cp1,cp2,cp3,restecg0,restecg1,slope1,slope2)
Nna<-rowSums(is.na(cl))#number count number of NA and remove those with too many NA
cl<-cl[Nna<1]#w
rmX <- names(cl) %in% c("num","thal","cp","restecg","slope","ca","presence")
```

```
y<-cl[,"presence"]
X<-cl[,!rmX]
cl<-cbind(y,X)
#write.csv(y,"~/Dropbox/Bayes/BayesProject/y.csv",row.names=F)
#write.csv(X,"~/Dropbox/Bayes/BayesProject/X.csv",row.names=F)
###Get MLE
m1<-glm(as.factor(y)~.,data=cl,family=binomial(logit))
summary(m1)
#GVS
###STAN CODE for HOrseshoe regression
blr='
data{
int N;
int<lower=0> p;
int  y[N];
matrix[N,p] X;
}
parameters {
vector [p] w;
real w0;
vector<lower=0>[p] lambda;
real<lower=0> tau;
}
model {
w0~normal(0,1);
lambda ~ cauchy(0, 1);
  tau ~ cauchy(0, 1);
for(j in 1:p){
w[j] ~ normal(0, square(lambda[j]) * square(tau));}
for(i in 1:N){
y[i]~bernoulli_logit(w0+X[i]*w);
}
}

'
```

```
dataStan<-list(N=nrow(X),p=ncol(X),X=X,y=y)

model=stan_model(model_code=blr)
fitData=sampling(model,data=dataStan,
                 warmup=10000, iter=110000, chains=1,
                 par=c("w0","w","lambda","tau"));

a<-as.mcmc(fitData)
print(fitData, pars=c("lambda","w"),digits_summary=3)
plot(fitData)

##Marginal likelihood computation
blr='
data{
int N;
int<lower=0> p;
int  y[N];
matrix[N,p] x;
cov_matrix[p] V0;
vector[p] w0;
}
parameters {
vector [p] w;
real intercept;
}
model {
w ~ multi_normal(w0, V0);
for(i in 1:N){
y[i]~bernoulli_logit(intercept+x[i]*w);

}
}
generated quantities{
real ll;
```

```
ll<-multi_normal_log(w,w0,V0);
for (i in 1:N){
ll<-ll+bernoulli_logit_log(y[i],intercept+x[i]*w);
}
}
'
model=stan_model(model_code=blr)
MLs<-vector()
x<-X[,c(2,8,9,10,11,6,3)]
p<-ncol(x)
V0<-diag(p)*0.01
w0<-rep(0,p)
dataStan<-list(N=nrow(x),p=ncol(x),x=x,y=y,V0=V0,w0=w0)
fitData=sampling(model,data=dataStan,
                 warmup=10000, iter=100000, chains=1,
                 par=c("ll","w"));
res<-extract(fitData)
MCMCsamples=cbind(res$w)

logP=res$ll
MLs<-print(mhmEstimator(MCMCsamples, logP))

mhmEstimator=function(MCMCsamples,logP){
  library(mvtnorm)
  p=dim(MCMCsamples)[2]
  N=dim(MCMCsamples)[1]
  M=colMeans(MCMCsamples);
  S=cov(MCMCsamples)
  ph=dmvnorm(MCMCsamples, M, S, log=T)
  Q=vector(length=N)
  for (i in 1:N)
    Q[i]=(MCMCsamples[i,]-M)%*%solve(S)%*%matrix(MCMCsamples[i,]-M, nrow=p)
  acc= Q<qchisq(0.9,p)
  Quote=exp(ph-logP)
  marginal=1/mean(Quote[acc])
```

```
  return(marginal)
}
```

## WINBUGS CODE GVS

```
model{
for(i in 1:303){

y[i]~dbern(p.b0[i])
p.b0[i]<-max(0,min(1,p.0[i]))
logit(p.0[i])<-
beta0+beta[1]*X[i,1]*gamma[1]+beta[2]*X[i,2]*gamma[2]+beta[3]*X[i,3]*gamma[3]+beta[4]*X[i,4]*gamma[4]+beta[5]*X[i,5]*gamma[5]+beta[6]*X[i,6]*gamma[6]+beta[7]*X[i,7]*gamma[7]+beta[8]*X[i,8]*gamma[8]+beta[9]*X[i,9]*gamma[9]+beta[10]*X[i,10]*gamma[10]+beta[11]*X[i,11]*gamma[11]+beta[12]*X[i,12]*gamma[12]+beta[13]*X[i,13]*gamma[13]+beta[14]*X[i,14]*gamma[14]+beta[15]*X[i,15]*gamma[15]

}


rho~dbeta(1,1)
beta0~dnorm (0, 0.1)

for (j in 1:15){gamma[j]~dbern(rho) }



for (j in 1:15){

beta[j]~dnorm( mb[j], taub[j])

mb[j]<-prop.mean.beta[j]

taub[j]<-(gamma[j]/303+(1-gamma[j]))/pow(prop.sd.beta[j],2)

}



for(j1 in 1:2){for(j2 in 1:2){for(j3 in 1:2){for(j4 in 1:2){for(j5 in 1:2){for(j6 in 1:2){for(j7 in 1:2){for(j8 in 1:2){for(j9 in 1:2){for(j10 in 1:2){for(j11 in 1:2){for(j12 in 1:2){for(j13 in 1:2){for(j14 in 1:2){for(j15 in 1:2){

models[j1,j2,j3,j4,j5,j6,j7,j8,j9,j10,j11,j12,j13,j14,j15]<-equals(gamma[1],j1-1)*equals(gamma[2],j2-1)*equals(gamma[3],j3-1)*equals(gamma[4],j4-1)*equals(gamma[5],j5-1)*equals(gamma[6],j6-1)*equals(gamma[7],j7-1)*equals(gamma[8],j8-1)*equals(gamma[9],j9-1)*equals(gamma[10],j10-1)*equals(gamma[11],j11-1)*equals(gamma[12],j12-1)*equals(gamma[13],j13-1)*equals(gamma[14],j14-1)*equals(gamma[15],j15-1)

}}}}}}}}}
}}}}}}


}

list(gamma=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
```

## WINBUGS CODE KM

```
model{
for(i in 1:303){

y[i]~dbern(p.b0[i])
p.b0[i]<-max(0,min(1,p.0[i]))
logit(p.0[i])<-
beta0+beta[1]*X[i,1]*gamma[1]+beta[2]*X[i,2]*gamma[2]+beta[3]*X[i,3]*gamma[3]+beta[4]*X[i,4]*gamma[4]+beta[5]*X[i,5]*gamma[5]+beta[6]*X[i,6]*gamma[6]+beta[7]*X[i,7]*gamma[7]+beta[8]*X[i,8]*gamma[8]+beta[9]*X[i,9]*gamma[9]+beta[10]*X[i,10]*gamma[10]+beta[11]*X[i,11]*gamma[11]+beta[12]*X[i,12]*gamma[12]+beta[13]*X[i,13]*gamma[13]+beta[14]*X[i,14]*gamma[14]+beta[15]*X[i,15]*gamma[15]

}


rho~dbeta(1,1)
beta0~dnorm (0, 0.1)

for (j in 1:15){gamma[j]~dbern(rho) }
```

```
for (j in 1:15){
tau[j]<-1/8;
beta[j]~dnorm( 0, tau[j])



}



for(j1 in 1:2){for(j2 in 1:2){for(j3 in 1:2){for(j4 in 1:2){for(j5 in 1:2){for(j6 in 1:2){for(j7 in 1:2){for(j8 in 1:2){for(j9 in 1:2){for(j10 in 1:2){for(j11 in 1:2){for(j12 in 1:2){for(j13 in 1:2){for(j14 in 1:2){for(j15 in 1:2){

models[j1,j2,j3,j4,j5,j6,j7,j8,j9,j10,j11,j12,j13,j14,j15]<-equals(gamma[1],j1-1)*equals(gamma[2],j2-1)*equals(gamma[3],j3-1)*equals(gamma[4],j4-1)*equals(gamma[5],j5-1)*equals(gamma[6],j6-1)*equals(gamma[7],j7-1)*equals(gamma[8],j8-1)*equals(gamma[9],j9-1)*equals(gamma[10],j10-1)*equals(gamma[11],j11-1)*equals(gamma[12],j12-1)*equals(gamma[13],j13-1)*equals(gamma[14],j14-1)*equals(gamma[15],j15-1)

}}}}}}}}}}
}}}}}}



}
```

# WINBUGS CODE SSVS

```
model{
for(i in 1:303){

y[i]~dbern(p.b0[i])
p.b0[i]<-max(0,min(1,p.0[i]))
logit(p.0[i])<-
beta0+beta[1]*X[i,1]+beta[2]*X[i,2]+beta[3]*X[i,3]+beta[4]*X[i,4]+beta[5]*X[i,5]+beta[6]*X[i,6]+beta[7]*X[i,7]+beta[8]*X[i,8]+beta[9]*X[i,9]+beta[10]*X[i,10]+beta[11]*X[i,11]+beta[12]*X[i,12]+beta[13]*X[i,13]+beta[14]*X[i,14]+beta[15]*X[i,15]

}



rho~dbeta(1,1)
beta0 ~ dnorm(0,0.1)
for(j in 1:15){

     gamma[j]~dbern(rho)

     tmp[j]<-gamma[j] + 1

     beta[j]~dnorm(0, tau[tmp[j]])

  }

tau[1]<-100

tau[2]<-0.1

for(j1 in 1:2){for(j2 in 1:2){for(j3 in 1:2){for(j4 in 1:2){for(j5 in 1:2){for(j6 in 1:2){for(j7 in 1:2){for(j8 in 1:2){for(j9 in 1:2){for(j10 in 1:2){for(j11 in 1:2){for(j12 in 1:2){for(j13 in 1:2){for(j14 in 1:2){for(j15 in 1:2){

models[j1,j2,j3,j4,j5,j6,j7,j8,j9,j10,j11,j12,j13,j14,j15]<-equals(gamma[1],j1-1)*equals(gamma[2],j2-1)*equals(gamma[3],j3-1)*equals(gamma[4],j4-1)*equals(gamma[5],j5-1)*equals(gamma[6],j6-1)*equals(gamma[7],j7-1)*equals(gamma[8],j8-1)*equals(gamma[9],j9-1)*equals(gamma[10],j10-1)*equals(gamma[11],j11-1)*equals(gamma[12],j12-1)*equals(gamma[13],j13-1)*equals(gamma[14],j14-1)*equals(gamma[15],j15-1)

}}}}}}}}}}
}}}}}}


}
```