

# Sparse Bayesian logistic regression

## Model formulation

Given design matrix  $X$  and response  $y \in \{-1, 1\}$ , the indicator logistic regression model assumes

$$P(y = 1) = \sigma(X\Gamma\beta),$$

where  $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_p)$ ,  $\beta = (\beta_1, \dots, \beta_p)^T$ . Here we introduce the latent indicator variable  $\gamma_j$  who equals to 1 if  $j$ th variable  $x_j$  is important and 0 if  $x_j$  is unimportant.

To complete the Bayesian inference, it assumes  $\beta_j | \alpha_j \sim N(0, \alpha_j^{-1})$ ,  $\alpha_j \sim \text{Gamma}(a_0, b_0)$ ,  $\gamma_j \sim \text{Bernoulli}(\rho)$ , where  $a_0, b_0, \rho$  are hyper-parameters. Generally speaking,  $a_0, b_0$  can be set to  $10^{-2}, 10^{-4}$  respectively, in the uninformative fashion. As for  $\rho$ , it controls the sparsity of the model and is tuned by BIC. Other parameters are inferred by variational Bayesian.

Reference: C. Zhang, S. Xu and J. Zhang. A Novel Variational Bayesian Method for Variable Selection in Logistic Regression Models. 2018

## Example

The following example shows how to apply this code.

1. Consider a dataset with 30 features and 100 samples. Let 6 variables be important and takes value 2, 1.5, 1, -1, -1.5, 2.

```
p = 30;
n = 100;

beta = zeros(p, 1);
beta([1,6,11]) = [-2, -1.5, -1];
beta([16,21,26]) = [1, 1.5, 2];
inx = beta~=0;

X = mvnrnd(zeros(p,1), eye(p), n);
y = random('bino', 1, logsig(X*beta));
y(y==0) = -1;
```

2. Function `vbvs_logit` can infer the model by means of the variational Bayes technique and automatically select the optimal  $\rho$  according to BIC.

```
model = vbvs_logit(X,y);
model.BICmodel.optrho
```

```
ans = 0.4298
```

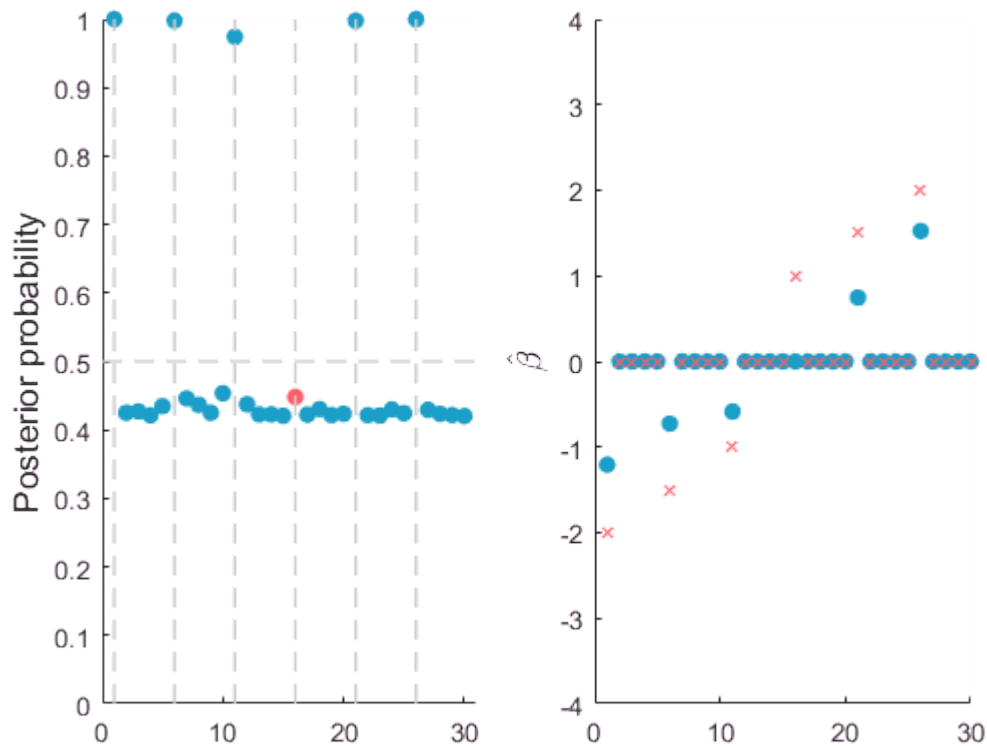
Plot the posterior probability of  $\gamma = 1$  and estimates of  $\beta$ .

```
post = model.BICmodel.optmodel.theta;
beta_hat = model.mu;
red = [1.0000    0.3686    0.4118];
blue = [0.1059    0.6196    0.7804];
cb = [red; blue];
```

```

% posterior probability
subplot(1,2,1)
hold on
scatter(1:p, post(2:end), 40, cb(((post(2:end)>0.5)==inx)+1,:), 'filled' )
for i=[1,6,11,16,21,26]; plot([i i], [0 1], '--', 'LineWidth', 1, 'color', 0.8*ones(3,1));end
plot(0:31, 0.5*ones(32), '--', 'LineWidth', 1, 'color', 0.8*ones(3,1)),xlim([0,31])
ylabel('Posterior probability', 'FontSize', 13)
% estimate of beta
subplot(1,2,2)
hold on
scatter(1:p, beta_hat(2:end), 40, repmat(cb(2,:),30,1), 'filled')
scatter(1:p, beta, 30, repmat(cb(1,:),30,1), 'x')
ylim([-4,4])
ylabel('$\hat{\beta}$', 'FontSize', 13, 'interpreter', 'latex')

```

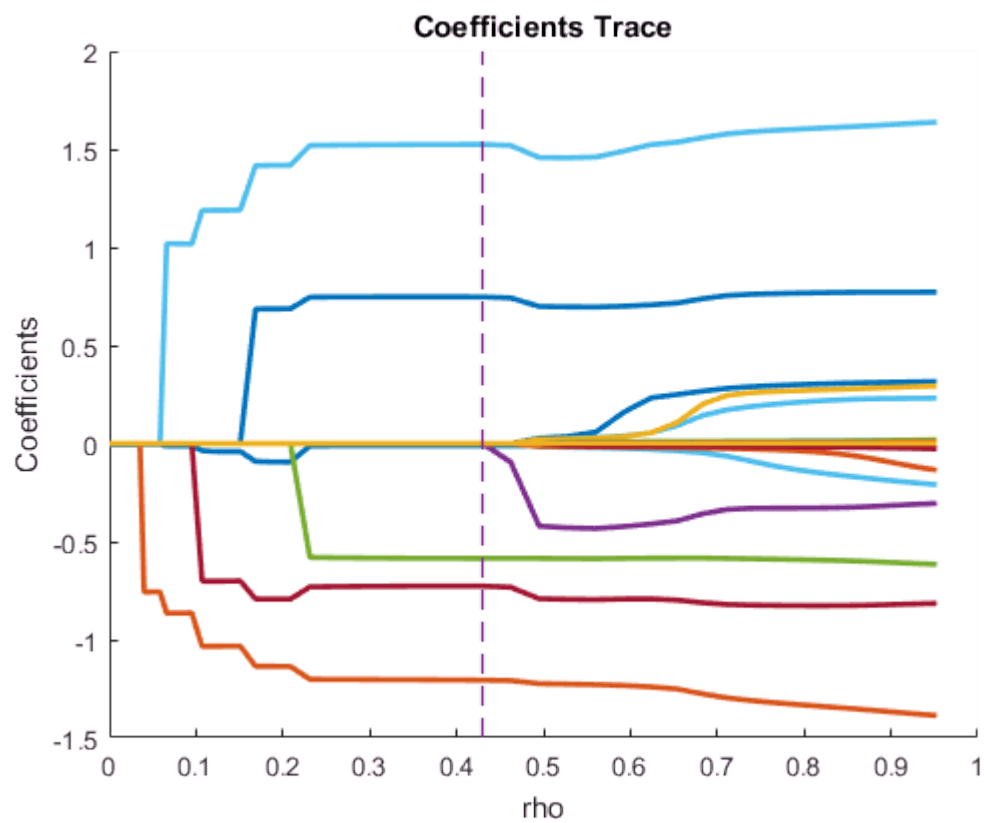


3. Plot the coefficients trace and BIC as  $\rho$  varying. The dashed line corresponds to optimal  $\rho$ .

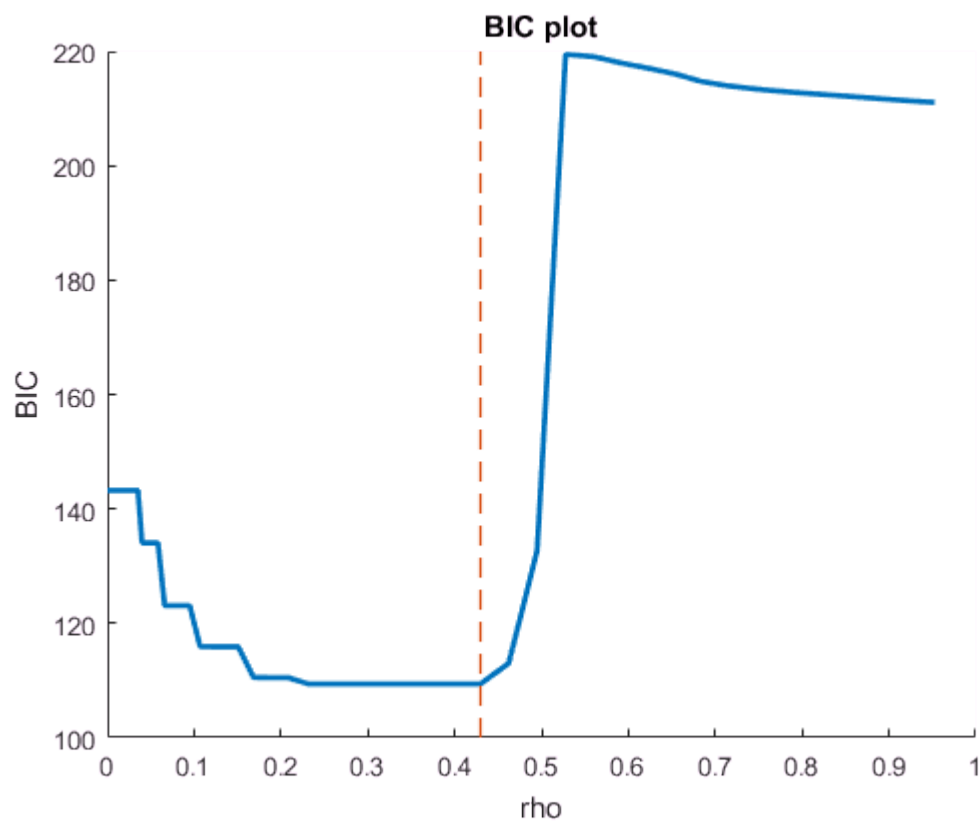
```

figure
hold on
plot(model.rho, model.allcoef, 'LineWidth', 2)
plot(model.BICmodel.optrho*[1,1], get(gca, 'YLim'), '--', 'LineWidth', 1)
hold off
xlabel('rho'), ylabel('Coefficients'), title('Coefficients Trace')

```



```
figure
hold on
plot(model.rho, model.BIC, 'LineWidth', 2)
plot(model.BICmodel.optrho*[1,1], get(gca, 'YLim'), '--', 'LineWidth', 1)
hold off
xlabel('rho'), ylabel('BIC'), title('BIC plot')
```



#### 4. Compare with traditional logistic regression.

```
% Train logistic regression
y(y==1) = 0;
[B,dev,stats] = mnrfits(X, categorical(y));
% Generate test set
Xtest = mvnrnd(zeros(p,1),eye(p),1000);
ytest = random('bino', 1, logsig(Xtest*beta));
% Make prediction
y_vbvs = [ones(1000,1),Xtest]*model.mu;
y_logit = mnrfits(B,Xtest); y_logit = y_logit(:,2);
display(['Accuracy of vbvs_logit is ',num2str(sum(y_vbvs>0.5 == ytest)/1000)])
```

Accuracy of vbvs\_logit is 0.833

```
display(['Accuracy of logistic model is ',num2str(sum(y_logit>0.5 == ytest)/1000)])
```

Accuracy of logistic model is 0.789