Bayesian variable selection and classification with control of predictive values

Eleni Vradi
Research and Early Development Statistics, Bayer AG, Berlin, Germany

////////////

Joint work: Werner Brannath (University of Bremen)
Thomas Jaki (Lancaster University)
Richardus Vonk (Bayer AG)

EFSPI Regulatory Statistics meeting
September 25, 2018

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 633567

Outline

- Motivation
- Model
- Simulation Results
- Application
- Conclusion

Motivation

Case study example

- Protein (biomarker) measurements $X_1, \ldots, X_{187}$ and $n = 53$ patients

- Q: How can one best select a subset of biomarkers to classify patients?
- A: a) Perform variable selection (e.g. penalization methods) and define a risk score

  b) Patient classification requires determination of appropriate cutoff value on the risk score
  - Youden index: $J = max_c\{sensitivity(c) + specificity(c) - 1\}$
    - *To what degree does the test reflect the true disease status?*
  - $PSI = max_c\{PPV(c) + NPV(c) - 1\}$
    - *How likely is disease given test result?*

PPV: Positive Predictive Value
NPV: Negative Predictive Value

Motivation *cont'd*

Biomarker selection and cutoff estimation

- However, in clincial practice, a target performance is required
- Simultaneously perform variable selection and cutoff estimation
- Build in the selection procedure a minimun (pre-specified) predictive value of the risk score
- Take prior information into account
- Quantify the uncertainty around the cutoff and the predictive values

Model

- Binary response $Y \in \{0,1\}$
- Biomarkers $X_1, X_2, \dots, X_d$
- A step function is used to model the probability of response
  - The cutoff and predictive values are parameters of the model
- *Model*
  - $Y|X \sim Bernoulli(p)$
  - $$p = P(Y = 1|Z = X\beta) = \begin{cases} P(Y = 1|Z \leq cp) = p_1 \\ P(Y = 1|Z > cp) = p_2 \end{cases}$$
  - $\beta \sim F$
- $p_1 \sim Uniform(0, p_2)$, $\boldsymbol{p_2 \sim Uniform(l, 1)}$ *i.e.* $l = 0.8$ and $cp \sim Uniform(a, b)$

Thresholding criteria for variable selection

- Laplace (Bayesian Lasso): $\beta_j \sim DE(0, \frac{1}{\lambda})$, $\lambda \sim Gamma(a, b)$
  - Indicator variable $\gamma_j = 1$ if β_j is included in the model and $\gamma_j = 0$ otherwise
  - incorporated in the linear predictor $\eta^* = XD_\gamma\beta$ where $D_\gamma = diag(\gamma_1, \gamma_2, \ldots, \gamma_d)$
- Spike and slab prior: $\beta_j \sim (1-\gamma_j)\delta_0 + \gamma_j N(0, \sigma^2)$, $\gamma_j \sim Bernoulli(\pi)$ and $\pi \sim Unif(0,1)$
  - By construction, γ_j indicates if β_j is included in the model
- Horseshoe prior $\beta_j \sim \mathrm{N}(0, \lambda_j{}^2\tau^2)$, with local shrinkage $\lambda_j \sim Cauchy^+(0,1)$ and global shrinkage $\tau \sim Cauchy^+(0, c^2)$ usually with $c^2 = 1$
  - Proposed by Carvalho et al. (2010) $\gamma_j \geq 0.5$ where $\gamma_j := 1 - \frac{1}{1+\lambda_j{}^2\tau^2}$
- Variable selection is *ad hoc*
  - based on the posterior inclusion probabilities $f(\gamma_j = 1 | \mathrm{y}) \geq 0.5$ (suggested by Barbieri and Berger, 2004)

Estimation of cutoff cp

MCMC Gibbs sampling, „*R2jags*“ library in R

- Fit the model with the step function
  - Estimate (marginal) posterior inclusion probabilities for each variable and select X_j by $f(\gamma_j = 1 \mid y) \geq 0.5$
  - Calculate the estimated risk score of the selected variables $X\hat{\beta}$, where $\hat{\beta}$ is taken for example as the mean of the posterior density

- Fit the model with the step function but now for fixed $\hat{\beta}$
  - From the posterior $f(cp, p_1, p_2 | X, \hat{\beta}, y)$ marginalize over cp, over p_1, over p_2

Scenario 1 (Null model): Posterior Incl Probabilities

$X \sim MVN(0, \Sigma)$, *m=10* noisy predictors, *k=0* informative predictors, *n=200*

- Generating model: logistic
- Fiting model: step

	Laplace	SpSI	HS
Average of correct selections of the null model	0.879	0.943	0.849

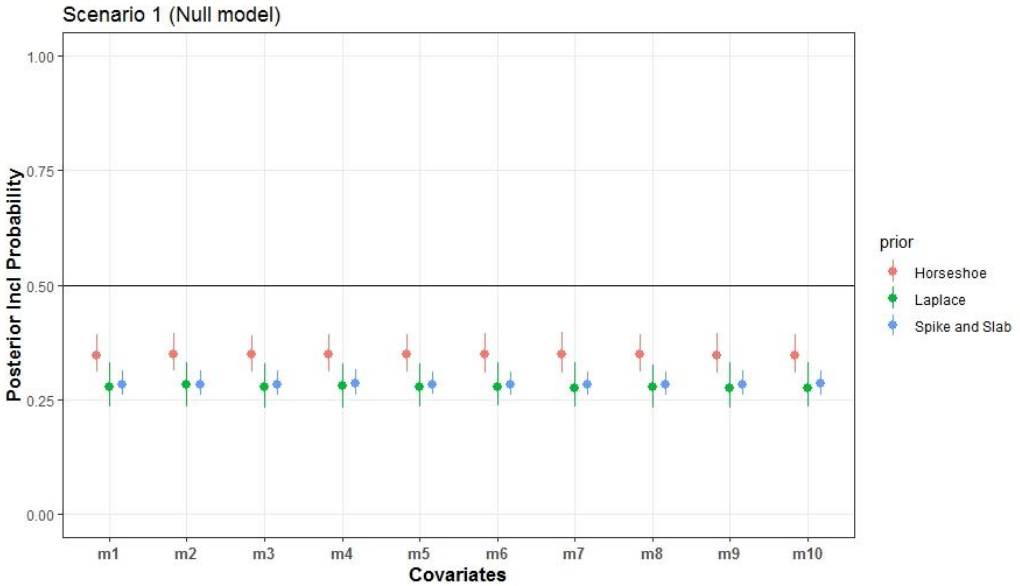


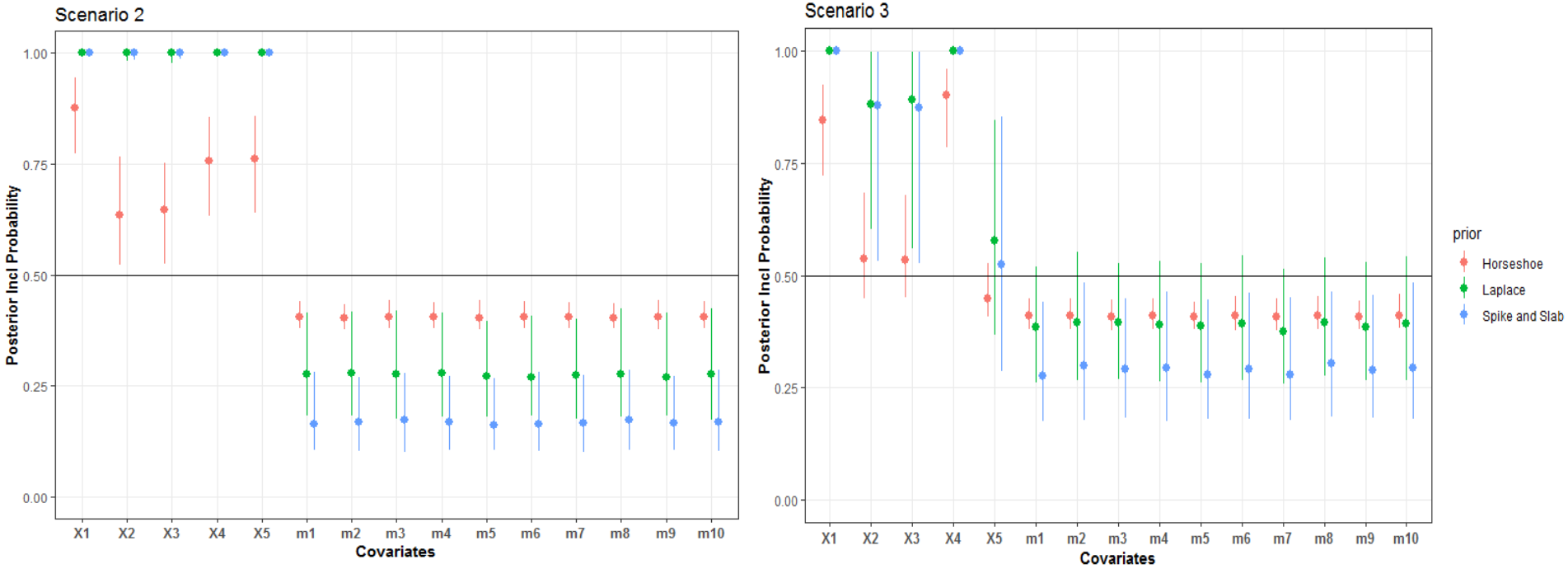
Figure: Plot of the median posterior inclusion probabilities (dots) over 1,000 simulation runs, together with the 1st and 3rd quantile. The horizontal black line corresponds to the value 0.5 that was used as a threshold for variable inclusion.

Posterior inclusion probabilities

$X \sim MVN(0, \Sigma)$, *m=10* noisy predictors , *k=5* informative predictors, *n=200*

Scenario 2: generate from a step function and fit a step model $\beta = (1.5, \mathbf{0.7}, \mathbf{0.7}, -1, -1)$

Scenario 3: generate from a logistic function and fit a step model $\beta = (1.5, \mathbf{0.7}, \mathbf{0.7}, -2, -\mathbf{0.5})$

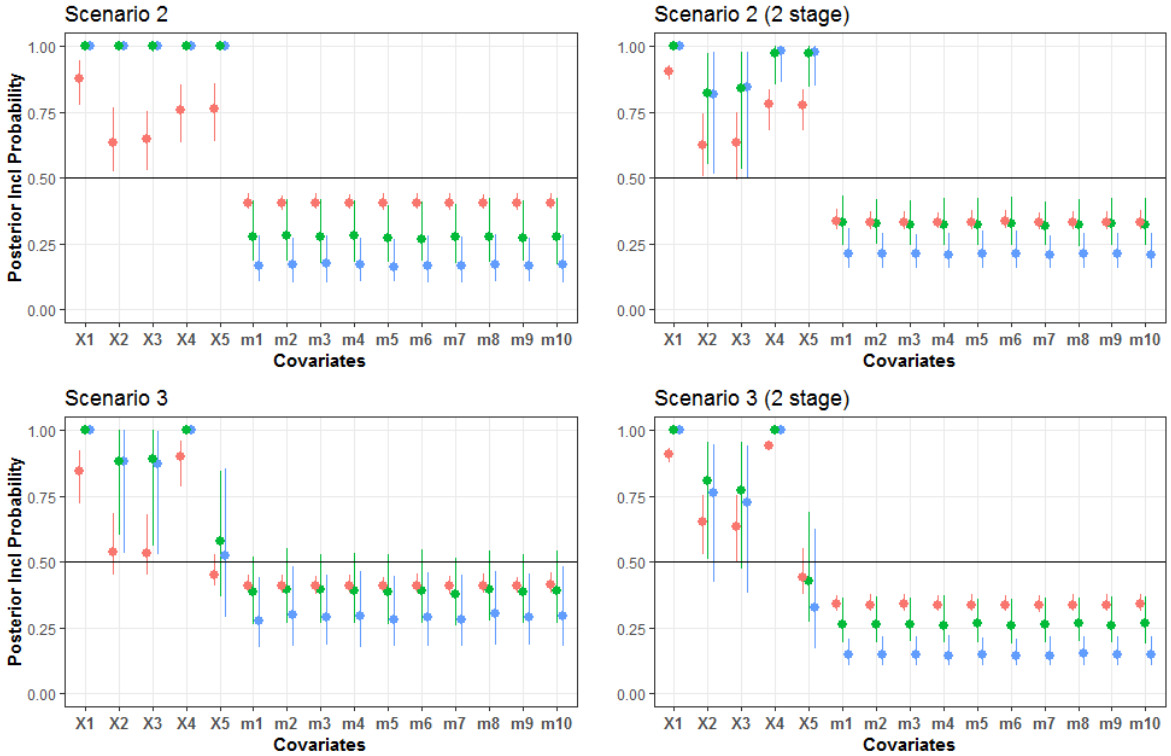


Posterior inclusion probabilities

Scenario 2: generate from a step function and fit the 2 stage approach
Scenario 3: genarate from a logistic function and fit the 2 stage approach

2 stage approach:

- at the 1st stage fit a logistic model for variable selection and
- at the 2nd stage fit a step model for cutoff estimation



Classification error

Mean of incorrectly predicted y_i on a validation dataset

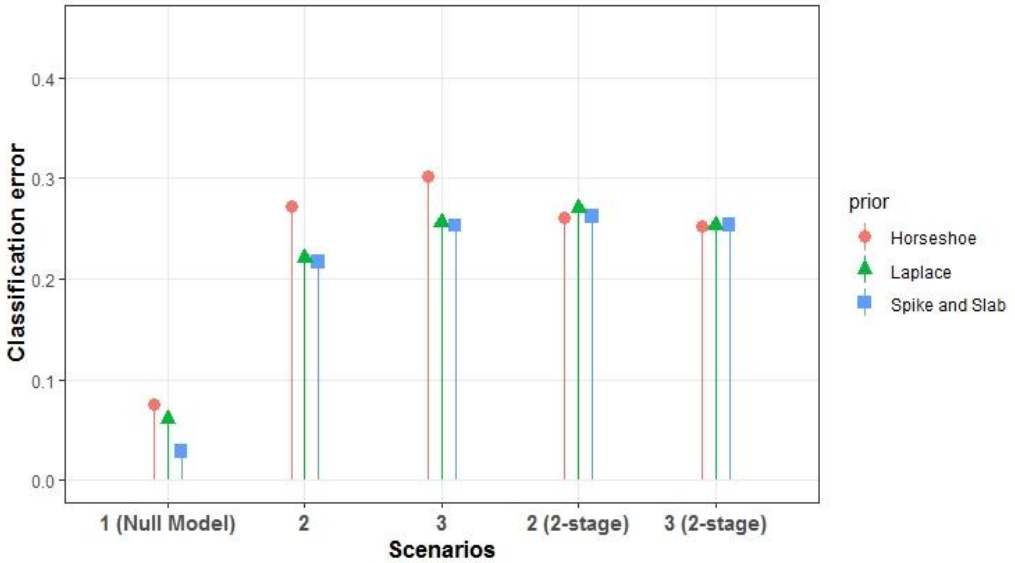


Figure: Average number over 1,000 simulation runs that the predicted $\hat{y}_i \neq y_i$

Application

n=53, d=187 protein measurements, binary response , $p_2 \sim Unif\,(0.8,1)$

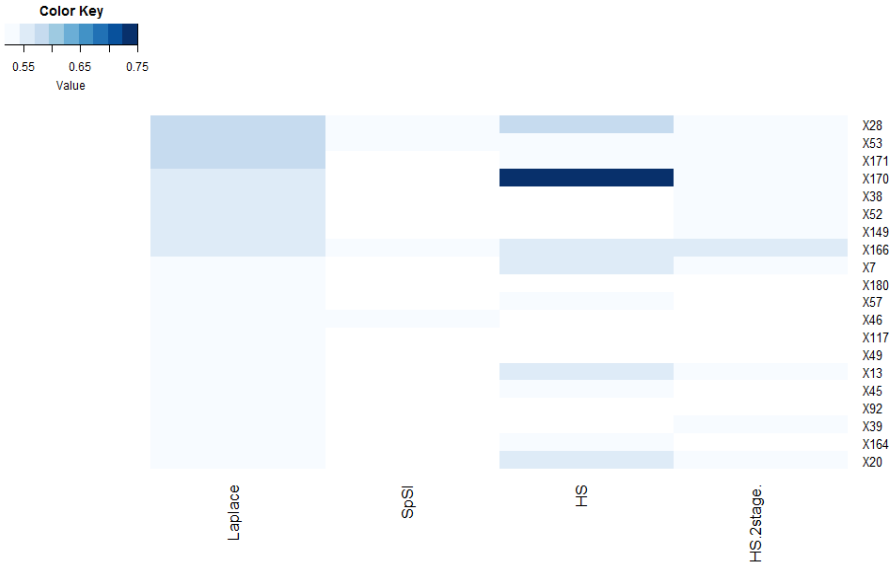


Figure: Heatmap of inclusion probabilities of the top 20 variables selected by the Laplace prior. Matched with the variables selected by the SpSI, HS and HS (2-stage) the SpSI (2 stage) and Laplace (2stage) selected the null model, i.e the posterior inclusion probabilities were below 0.5

Table: Posterior median of cp, p_1, p_2 together with the 95% credible intervals for the different priors. The second column gives the number of variables selected by each prior.

Priors	# selected variables	cutoff	p_1	p_2
Laplace	20	0.49 (0.14-56)	0.10 (0.03-0.25)	0.89 (0.81-0.99)
SpSI	5	0.64 (0.12-0.90)	0.19 (0.07-0.32)	0.87 (0.80-0.96)
HS	24	0.49 (0.32-0.84)	0.20 (0.08-0.35)	0.87 (0.80-0.97)
HS (2-stage)	18	0.32 (0.17-0.62)	0.13 (0.03-0.27)	0.86 (0.80-0.96)

Conclusion and future work

- We proposed a Bayesian method for biomarker selection and classification
  - Built-in pre-specified predictive value of the risk score (of the selected variables)
- Simulation results showed that the proposed method
  - performs well in terms of selecting the important variables
  - classification error was found on average below 30%
  - performs as well and occasionaly better that the classical 2-stage approach

- Future work
  - Extension to time-to-event data

References

- Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023-1032.

- Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465-480.

- Youden WJ. Index for rating diagnostic tests. Cancer. 1950 Jan 1;3(1):32-5

- Linn S, Grunau PD. New patient-oriented summary measure of net total gain in certainty for dichotomous diagnostic tests. Epidemiologic Perspectives & Innovations. 2006 Dec;3(1):11

- Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. *The annals of statistics*, *32*(3), 870-897.

- Vradi, E., Jaki, T., Vonk, R., & Brannath, W. (2018). A Bayesian model to estimate the cutoff and the clinical utility of a biomarker assay. *Statistical Methods in Medical Research*. In press

Thank you!

Bye-Bye

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 633567