Vanderbilt University
January 2012

Objective Bayes Model Selection in Probit Models

Luis Leon-Novelo
University of Florida

Elías Moreno
University of Granada

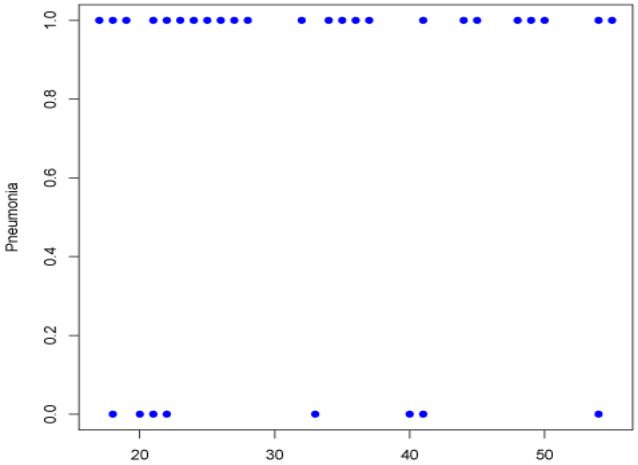
George Casella
University of Florida

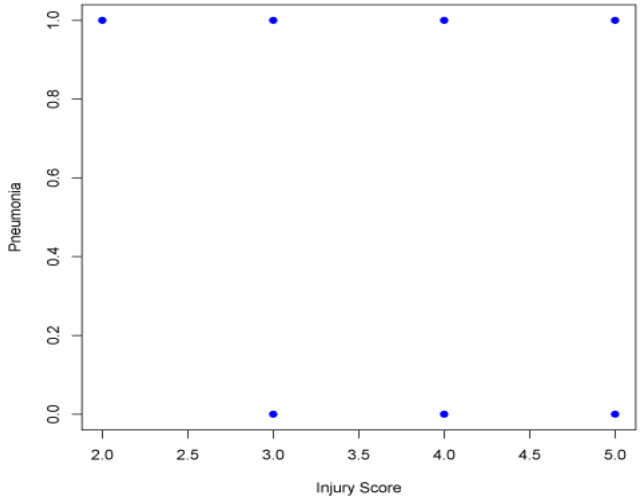
Introduction
How it Started

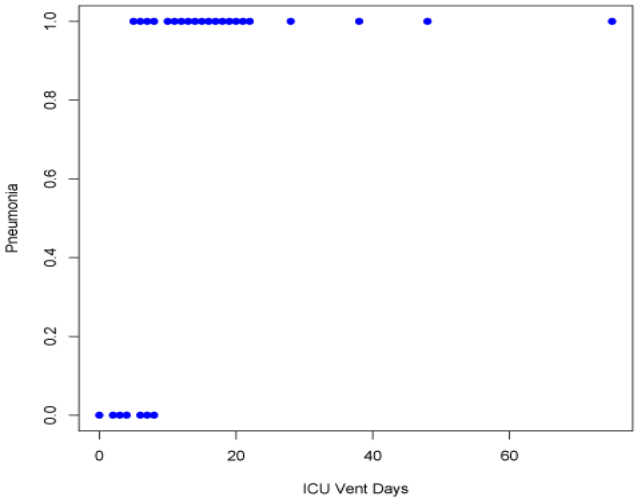
- 47 patients in an intensive care unit following trauma surgery.
- Physicians need to better manage post-operative sepsis (infection)
- Interested to see if there is association with any subset of genes.
  - Here we consider the $0-1$ endpoint "pneumonia"
  - Of the 47 patients; 39 of them exhibited pneumonia
- For each patient, expression of 296 genes measured in peripheral blood
  - Along with three clinical covariates
- This is a model selection problem
- Find best model that includes the clinical covariates and relevant genes.

Introduction

Motivation

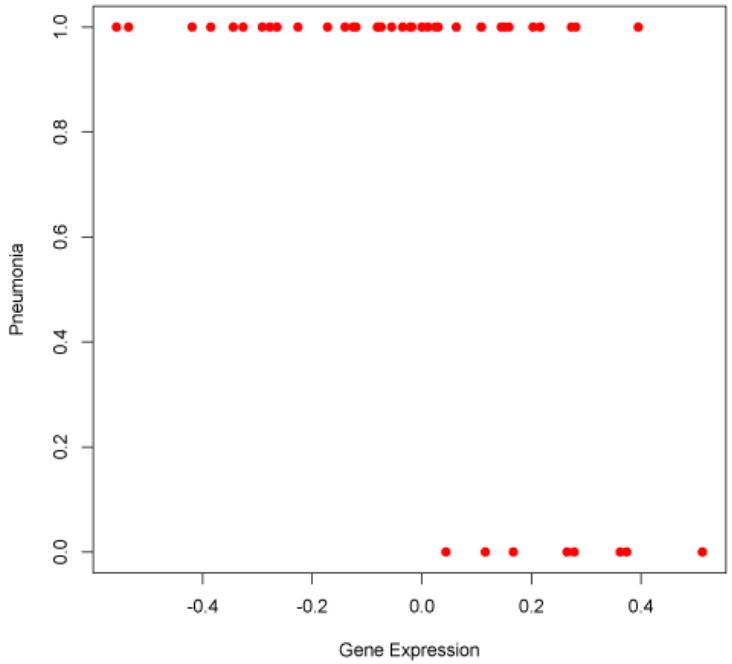






- Not much information here
- ICU Vent Days good covariate
  - Not a useful predictor

Introduction
Information in the Genes



- One of the selected genes
- Good covariate (predictor?)
- Need biological story

Outline of the Talk

- Background — Bayesian Model Selection
- The Model — Intrinsic Bayes
- Probit Regression — Computing the Bayes Factor
- Searching — Finding the Bayes Factor
- Illustrations — Simulations and Comparisons
- Implementation — Finding the Genes
- Conclusions — What we learned

Background

Bayesian Model Selection

- Let $p(\mathbf{z}|\theta_j, M_j)$ be the distribution of the sample
  - ▷ Regression model M_j
  - ▷ θ_j represents the parameters under model M_j
  - ▷ M_j belongs to a finite set of models

- $p(\mathbf{z}|M_j) = \int p(\mathbf{z}|\theta_j, M_j)\pi(\theta_j|M_j)d\theta_j$
  - ▷ the marginal distribution of the sample $\mathbf{z}$ under model M_j
  - ▷ $\pi(\theta_j|M_j)$ denotes the prior distribution for the model parameters θ_j

Background
Bayes Factors

- We compare models using the Bayes Factor

$$BF_{j1}(\mathbf{z}) = \frac{\int p(\mathbf{z}|\theta_j, M_j)\pi(\theta_j|M_j)d\theta_j}{\int p(\mathbf{z}|\theta_1, M_1)\pi(\theta_1|M_1)d\theta_1} = \text{Ratio of Marginals}$$

  - Equivalent to posterior probability

- With p regressors, we have 2^p models
  - M_1 is typically the intercept only model

- We search for models with high values of $BF_{j1}(\mathbf{z})$
- In normal regression models, intrinsic Bayes variable selection:
  - Gives consistent model selectors,
  - Has moderate Type I and Type II errors for finite sample sizes

Probit Models
Latent Variable Formulation

- Sample $\mathbf{z} = (z_1, ..., z_n)$, where z_i, $i = 1, ..., n$, is a $0 - 1$ random variable
- Under model M_j

$$z_i|\theta_i, M_j \sim \text{Bernoulli}(z_i|\theta_i) \quad \text{with} \quad \theta_i|M_j = \Phi(x_i'\beta_j),$$

  - Φ is the normal cdf, and , β_j a vector of dimension $j + 1$.
- The maximum length of the vector of covariates is $p + 1$.

The probit model is a latent normal model

- y_i follows a normal regression model
- Only the sign of y_i is observed
- We observe the variable $z_i = 1(y_i > 0)$

Probit Models
Latent Intrinsic Priors

- For $\mathbf{y} = (y_1, ..., y_n)'$ the null normal model is

$$M_1 : \{N_n(\mathbf{y}|\alpha \mathbf{1}_n, \mathbf{I}_n),\ \pi(\alpha)\}, \text{Intercept Only}$$

- A candidate model M_j with $j+1$ regressors is

$$M_j : \{N_n(\mathbf{y}|\mathbf{X}_j\beta_j, \mathbf{I}_n),\ \pi(\boldsymbol{\beta}_j)\},$$

  - $\mathbf{X}_j$ has dimension $n \times (j+1)$

- We use intrinsic methodology for the linear model
  - Starting with improper reference priors $\pi^N(\alpha)$ and $\pi^N(\beta)$
  - We obtain automatic specification of the priors $\pi(\alpha)$ and $\pi(\beta)$

Interlude
A Primer on Intrinsic Priors

- Test $H_0 : \theta = \theta_0$, $\mathbf{y} \sim f(\mathbf{y}|\theta)$
  - ▷ Improper reference prior $\pi(\theta|\theta_0)$
  - ▷ $\mathbf{y}_{\min}$ = Minimal Training Sample

$$\pi(\theta|\theta_0, \mathbf{y}_{\min}) = \frac{f(\mathbf{y}_{\min}|\theta)\pi(\theta|\theta_0)}{\int_\Theta f(\mathbf{y}_{\min}|\theta)\pi(\theta|\theta_0)} \qquad \text{Proper Prior}$$

Intrinsic Prior= Average over all theoretical training samples

- Model dependent, not data dependent
- Centered at H_0
- Could be improper, but not a problem.
  - ▷ Nested hypotheses $\Rightarrow$ unknown constants cancel

Probit Models
Bayes Factor

- Marginal Distributions

$$m_1(\mathbf{y}) = \int N_n(\mathbf{y}|\alpha\mathbf{1}_n, \mathbf{I}_n)\pi^N(\alpha)d\alpha,$$
$$m_j(\mathbf{y}) = \int\int N_n(\mathbf{y}|\mathbf{X}_j\beta_j, \mathbf{I}_n)\pi^I(\beta|\alpha)\pi^N(\alpha)d\alpha d\beta.$$

  - $BF_{j1}^{IP}(\mathbf{y}) = m_j(\mathbf{y})/m_1(\mathbf{y})$ is a consistent model selector

- The Probit marginals are

$$m_j(\mathbf{z}) = \int_{A_1\times\cdots\times A_n} m_j(\mathbf{y})d\mathbf{y}, \quad A_i = \begin{cases} (0,\infty) & \text{if } z_i = 1, \\ (-\infty, 0) & \text{if } z_i = 0, \end{cases}$$

- The Probit intrinsic Bayes factor is $BF_{j1}^{IP}(\mathbf{z}) = m_j(\mathbf{z})/m_1(\mathbf{z})$.

Probit Models
Computing the Bayes Factor

- Model M_j for the variable y includes j covariates plus the intercept.
  - The minimal training sample size is $j+1$
- The references priors for α and β are proportional to 1
- For example

$$\pi^I(\beta \mid \alpha) = N_{j+1}\left(\beta|\alpha \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \frac{2\,n}{j+1}(\mathbf{X}_j'\mathbf{X}_j)^{-1}\right).$$

  - $\mathbf{X}_j$ is the corresponding submatrix.
  - $\mathbf{X}_j'\mathbf{X}_j$ must be invertible, so we need $j+1 \leq n$.
- We can compute the intrinsic prior
  - When covariates + intercept $\leq n$, the sample size.
- Oh oh! Here it comes, $p >>> n$! (47 patients, 296 genes)

Probit Models
Marginals for the Bayes Factor

- Integrating out α and β

$$m_j(\mathbf{y}) = \frac{c}{(2\pi)^{(n-1)/2}|\mathbf{1}'\Sigma_j^{-1}\mathbf{1}|^{1/2}|\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}\mathbf{y}'\Lambda_j\mathbf{y}\right\},$$
$$m_1(\mathbf{y}) = \frac{c}{n^{1/2}(2\pi)^{(n-1)/2}} \exp\left\{-\frac{1}{2}ns_y^2\right\},$$

  - $\Sigma_j = \mathbf{I}_n + 2\,[n/(j+1)]\,\mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'$
  - $\Lambda_j = \Sigma_j^{-1} - \Sigma_j^{-1}\mathbf{1}(\mathbf{1}'\Sigma_j^{-1}\mathbf{1})^{-1}\mathbf{1}'\Sigma_j^{-1}$
  - Fairly standard calculations

- For the Probit Model

$$m_j(\mathbf{z}) = \int_A m_j(\mathbf{y})d\mathbf{y} = \int_{-\infty}^{\infty}\int_A N_n(\mathbf{y}|\alpha 1, \Sigma_j)d\mathbf{y}\; d\alpha.$$

  - Implemented with `pmvnorm` in the `R` package `mvtnorm`

Controlled Dimension Stochastic Search
Introduction

- We search for models with high values of $BF_{j1}(\mathbf{z})$
  - Can only calculate $BF_{j1}(\mathbf{z})$ if $p \leq n$

We use a hybrid random walk

- Through models with $q \leq n-1$ covariates
- q is selected by the researcher

- We identify the models with a vector $\gamma \in \{0,1\}^p$
  - $\mathcal{M}_\gamma$ includes the covariate j only if $\gamma_j = 1$.
  - The intercept is always included; it is not considered in γ explicitly
  - For $\gamma = (0,1,1,0,\ldots,0)$, $\mathcal{M}_\gamma$ = intercept and covariates 2 and 3.

Controlled Dimension Stochastic Search
Defining the Model Space

- There are 2^p such models, and we denote this full model space by $\mathcal{M}_{p:p}$
- The feasible model space is $\mathcal{M}_{p:q}$
  - ▷ There are $\sum_{j=0}^{q} \binom{p}{j}$ such models.
- MCMC algorithm
  - ▷ Stationary distribution proportional to $BF_{\gamma 1}(\mathbf{z})$ for $\gamma \in \mathcal{M}_{p:q}$.
- Three Pieces

- $\delta \in \mathcal{M}_{p:p}$, $0-1$ vector
- $A = (a_1, \ldots, a_p)$, $0-1$ vector
- $\gamma \in \mathcal{M}_{p:q}$, $\gamma = \delta \star A$

- δ is any model
  - ▷ Even more than q covariates
- A has *active covariates*
  - ▷ $\sum_{j=1}^{p} a_j = q$
- γ is the current model

Controlled Dimension Stochastic Search
Metropolis-Hastings Update

$(A^{(t)}, \delta^{(t)}, \gamma^{(t)})$

$\downarrow$

Get δ^{cand} $\rightarrow \gamma^{\text{cand}} = \delta^{\text{cand}} \star A^{(t)}$ $\rightarrow$ accept with probability $\min\{1, m_{\gamma^{\text{cand}}}(\mathbf{z})/m_{\gamma^{(t)}}(\mathbf{z})\}$

$\downarrow$

$(A^{(t)}, \delta^{(t+1)}, \gamma')$ $\gamma' = \delta^{(t+1)} \star A^{(t)}$

$\downarrow$

Get A^{cand} $\rightarrow \gamma^{\text{cand2}} = \delta^{(t+1)} \star A^{\text{cand}}$ $\rightarrow$ accept with probability $\min\{1, m_{\gamma^{\text{cand2}}}(\mathbf{z})/m_{\gamma'}(\mathbf{z})\}$

$\downarrow$

$(A^{(t+1)}, \delta^{(t+1)}, \gamma^{t+1})$ $\gamma^{t+1} = \delta^{(t+1)} \star A^{(t+1)}$

- This is a Markov chain on the set $\mathcal{M}_{p:q}$ of feasible models
  - It has stationary distribution proportional to the Bayes factor.

Simulations
Setting it Up

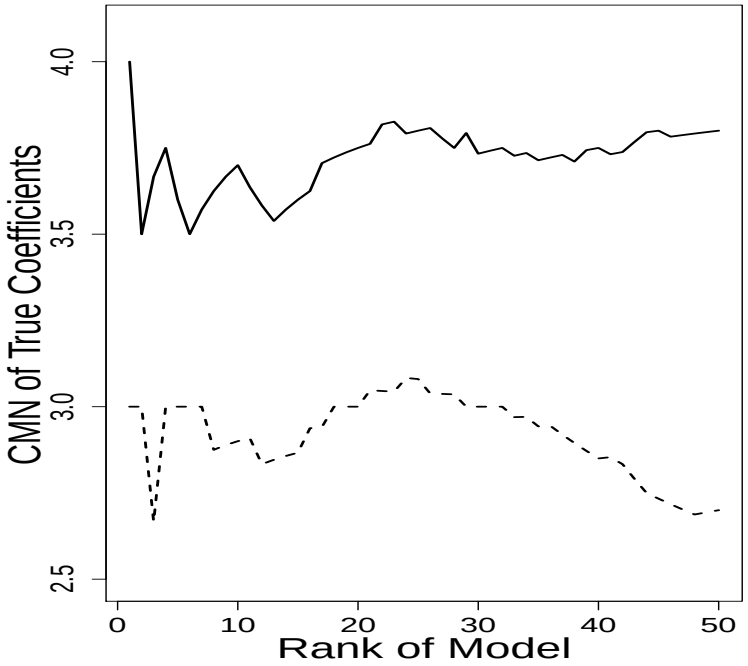
- Compare with Hu/Johnson (2009 JRSSB) -use their simulation
- $\beta_i = 0.5i$ for $i = 0, \ldots, 6$ and $\beta_i = 0$ for $i = 7, \ldots, 15$.
- Evaluate all $2^{15} = 32,768$ models
- No search

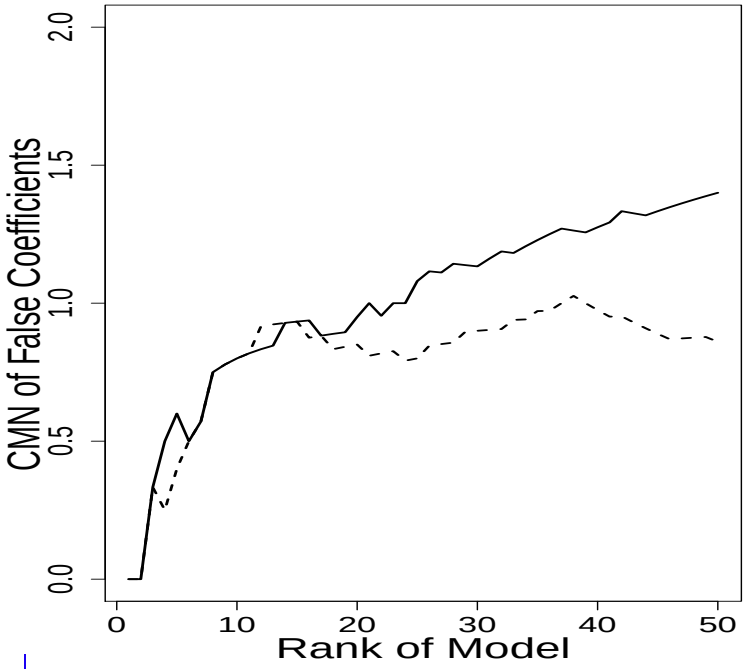
- True Model is Bad
- After $\beta_3 - \beta_6$, no others enter
- Cannot overcome dimension penalty

We find good models,
not true models

Simulations

Selecting the Coefficients





- Cumulative mean number (CMN)
  - ▷ True coefficients selected
- Intrinsic selects more true coefficients

- Cumulative mean number
  - ▷ False coefficients selected
- Intrinsic selects more false coefficients
  - ▷ In lower ranking models

Simulations
Loss of Information from Dichotomizing

- This one was really surprising
- We simulate the latent normal data y
- Examine the performance for both data sets, the y and the z

Model	True Model Ranked Number 1 (%)					
True	y			z		
Coefficients	BFIP	BIC	H&J	BFIP	BIC	H&J
-1,0,0,0,0,0	72	66	78	31	41	54
1,1,0,0,0,0	83	74	85	61	34	42
-1,1,-1,0,0,0	92	77	86	70	44	45
-1,1,-1,1,0,0	97	84	92	51	22	20

- H&J best when all coefficients 0
- Intrinsic and H&J similar for y
- Intrinsic rocks for z

Application
Pneumonia in the ICU

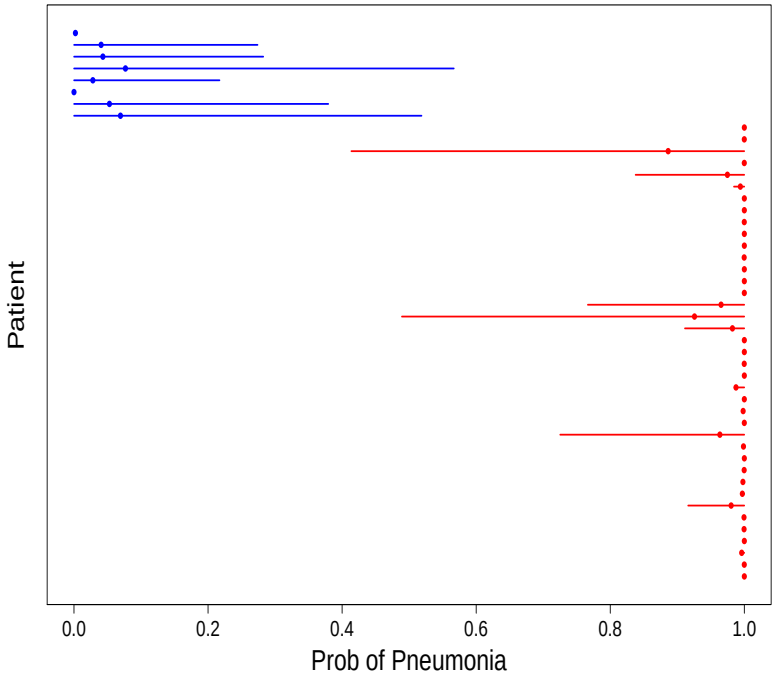
- Recall
  - 47 patients in an intensive care unit following trauma surgery.
  - For each patient, expression of 296 genes measured in peripheral blood
  - Along with three clinical covariates
- Find best model that includes the clinical covariates and relevant genes.

Rank	Number of Genes	Genes		
1	3	ARL10	ERICH1	OR4D1
2	3	GCLM	OLFM1	TEP1
3	3	ERICH1	OLFM1	TEP1
4	3	BCL3	ERICH1	TMEM56
5	3	C8orf34	ERICH1	WDR26

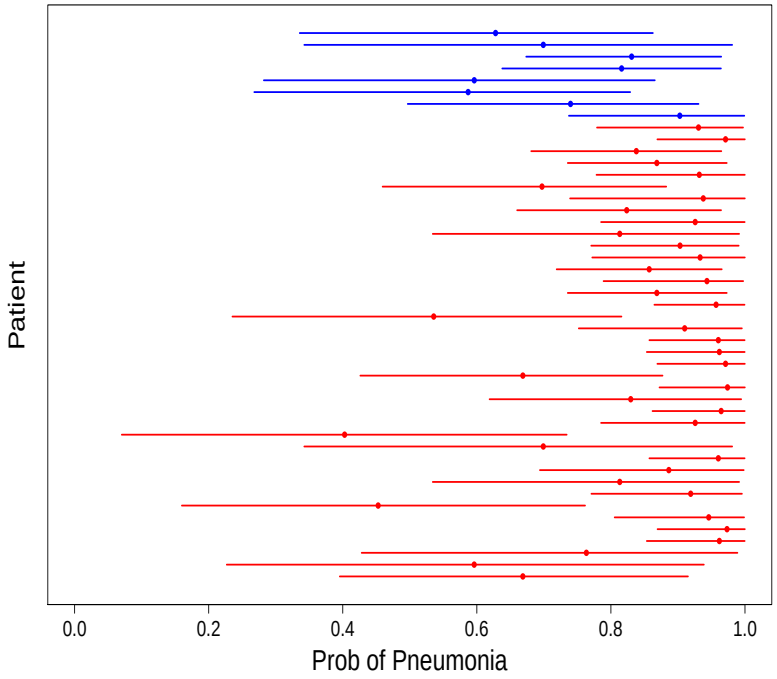
- Search limited to $q \leq 10$ genes
- Clinical covariates always in
- Top 20 models had ≤ 4 genes
- Polygenic search
- Biological story?

Application
Do the Genes Make a DIfference?

- In sample prediction, 95% credible intervals
  - ▷ Red: Patient had Pneumonia ▷ Blue: Patient did not have Pneumonia



- Clinical covariates + Genes



- Clinical covariates only

Conclusions

What We Did

- ▶ Variable Selection
  - Intrinsic Bayes Factors
  - Latent Normal Formulation
- ▶ Stochastic Search
  - Hybrid Metropolis-Hastings
  - Controlled Dimension
  - Addresses $p >> n$
- ▶ Examples
  - Intrinsic Better for z information
- ▶ `R` Package
  - `varselectIP` is on CRAN

Conclusions
Final Remarks

- Intrinsic Bayes is Automatic
  - Model dependent, not data dependent
- Variable Selection Cures Multicollinearity
  - Will not select SNPs in LD
- Find GOOD models
  - Forget about finding the true model
- Polygenic Search
  - "GWAS don't work"
- Selected Inference
  - Need to account for model uncertainty

Thank You for Your Attention

Luis Leon-Novelo
luis@stat.ufl.edu

Elías Moreno
emoreno@ugr.es

George Casella
casella@ufl.edu

Selected References
All on my web page

- Leon-Novelo, L., Moreno, E., and Casella, G. (2011). Objective Bayes Model Selection in Probit Models. *Statistics in Medicine*. To appear.
- Gopal, V, Leon-Novelo, L, Casella, G. `varSelectIP: Objective Bayes Model Selection in Linear Regression and Probit models`., 2011. URL http://CRAN.R-project.org/package=varSelectIP, r package version 0.1-4.
- Casella, G. and Moreno, E. (2006) Objective Bayes Variable Selection. *Journal of the American Statistical Association* **101** 157-167.
- Casella, G., Girón, F.J., Martínez, M.L. and Moreno E. (2009). Consistency of Bayesian procedure for variable selection. *Annals of Statistics*, **37,** 3, 1207-1228.
- Girón, F.J., Moreno, E., Casella, G. and Martínez, M.L. (2010). Consistency of objective Bayes factors for nonnested linear models and increasing model dimension. *RACSAM* **104** (1), 61–71.
- Moreno, E., Bertolino, F. and Racugno,W. (1998). An intrinsic limiting procedure for model selection and hypothesis testing. *J. Amer. Statist. Assoc.*, **93**, 1451-1460.
- Moreno, E., Girón, F.J. and Casella, G. (2010). Consistency of objective Bayesian tests as the model dimension increases. *Annals of Statistics* **38** 1937-1952