

BAYESIAN VARIABLE SELECTION FOR PROBIT MIXED MODELS

Selection of probesets characterising the estrogen receptor
hormonal status

Meili Baragatti

Institut de Mathématiques de Luminy (IML)

Ipsogen SA

January 2010

Introduction

Probit Mixed Model for Gene Selection

Bayesian Gene Selection via Gibbs Sampling

Application

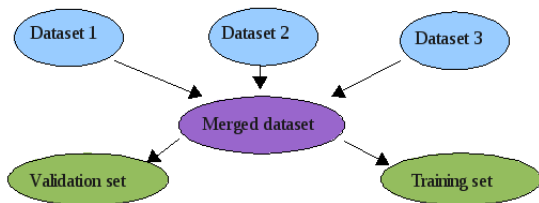
Discussion

What has been done in variable selection

- ▶ **Univariate tests** to identify differentially expressed probesets in different classes (Dudoit et al., 2002).
- ▶ **Model-based approaches** :
 - ▶ **Support Vector Machine** : Recursive feature elimination method of Guyon et al (2002).
 - ▶ **Bayesian variable selection in a linear model** : George and McCulloch (1993) and Chipman and George (2001).
 - ▶ **Bayesian variable selection for binary responses** : Lee et al. (2003), Sha et al. (2004), Zhou et al. (2004) for probit regression, Chen and Dey (2003), Tüchler (2008) for logistic regression.

Application in Genetics

- ▶ Identify probesets which distinguish ER-positive from ER-negative patients having breast cancer.
- ▶ ≈ 54000 probesets measured per patient.
- ▶ Meta analysis :



GOAL

- ▶ To select a few important variables among 1000's.
- ▶ Explain a binary variable of interest : probit regression was selected over logistic regression.
- ▶ ACCOUNT FOR DATA DESIGN, with random effects : mixed model.

⇒ Variable selection in a probit mixed model.

Tüchler (2008) : used approximations, only few dozens of predictors.

Introduction

Probit Mixed Model for Gene Selection

The Hierarchical Model

Prior distributions

Bayesian Gene Selection via Gibbs Sampling

Application

Discussion

The Hierarchical Model (1/4)

- ▶ Y_i the n binary events.
- ▶ **probit mixed model** : The Y_i have binary distributions, and they are not independent.

$$P(Y_i = 1|U, \beta) = \Phi(X_i' \beta + Z_i' U)$$

- ▶ β the fixed-effect coefficients, of dimension p .
- ▶ U the random-effect coefficients, of dimension q .
- ▶ X ($n \times p$) and Z ($n \times q$) the design matrices associated with the fixed and random effects.

The Hierarchical Model (2/4)

Concerning our application : $P(Y_i = 1|U, \beta) = \Phi(X_i'\beta + Z_i'U)$

- ▶ Only one random effect : the dataset.
- ▶ X_{ij} corresponds to the measurement of the expression level of the j^{th} probeset for the i^{th} observation.
- ▶ $Z_{ij} = 1$ if the i^{th} observation is from the j^{th} dataset, and 0 otherwise.

The Hierarchical Model (3/4)

n latent variables L_1, \dots, L_n , $L_i \sim \mathcal{N}(X_i' \beta + Z_i' U, 1)$, (Albert and Chib (1993) and Lee et al. (2003)).

$$Y_i = \begin{cases} 1 & \text{if } L_i > 0 \\ 0 & \text{if } L_i < 0 \end{cases}$$

Then

$$P(Y_i = 1 | U, \beta) = P(L_i > 0 | U, \beta) = \Phi(X_i' \beta + Z_i' U)$$

In multivariate notation, $L | U, \beta \sim \mathcal{N}_n(X\beta + ZU, I_n)$

The Hierarchical Model (4/4)

To perform the variable selection : γ a vector of indicator variables of length p is introduced :

$$\gamma_j = \begin{cases} 1 & \text{if } \beta_j \neq 0, & \text{variable } j \text{ selected} \\ 0 & \text{if } \beta_j = 0, & \text{variable } j \text{ not selected} \end{cases}$$

- ▶ Given γ , β_γ is the vector of all nonzero elements of β
- ▶ X_γ the matrix X with only the columns corresponding to the elements of γ that are equal to 1.

Prior distributions (1/2)



$$\forall j \in \{1, \dots, K\}, \quad U_j | A_j \sim \mathcal{N}_{q_j}(0, A_j)$$

The A_j are known symmetric design matrices. The random effects are supposed independent,

$$U|D \sim \mathcal{N}_q(0, D), \quad D = \text{diag}(A_j)$$

- ▶ $A_j \sim \mathcal{W}^{-1}(\Psi, m)$ (conjugate prior to the multivariate Gaussian).
- ▶ Remark : by convenience, the A_j are often diagonal, $A_j = \sigma_j^2 I$, and a priori $\sigma_j^2 \sim \mathcal{IG}(a, b)$.

Prior distributions (2/2)

- ▶ Prior distribution for β_γ :

$$\beta_\gamma | \gamma \sim \mathcal{N}_d(0, c(X'_\gamma X_\gamma)^{-1})$$

With c a positive scale factor specified by the user. It corresponds to the g-prior of Zellner (1986).

- ▶ The γ_j are assumed to be independent Bernoulli variables, with

$$P(\gamma_j = 1) = \pi_j, \quad 0 \leq \pi_j \leq 1$$

- ▶ No use of prior knowledge to favor some probesets
 $\implies \forall j, \quad \pi_j = \pi.$

Introduction

Probit Mixed Model for Gene Selection

Bayesian Gene Selection via Gibbs Sampling

The conditional distributions

The Metropolis-within-Gibbs algorithm

Application

Discussion

Use of the Gibbs sampling algorithm

- ▶ Goal : select relevant variables for the probit regression
 - ⇒ the **posterior distribution** of γ is of interest.
 - ⇒ use of the **Gibbs sampling** algorithm to approximate this posterior distribution and search for high probability γ values.
- ▶ All the **full conditional distributions** must be simulated from :
 $f(L|Y, \beta, U)$, $f(\beta|L, U, \gamma)$, $f(U|L, \beta, D)$, $f(\gamma|L, U, \beta)$ and $f(D|U)$.

The full conditional distributions (1/2)

- ▶ $L_i | \beta, U, Y_i$: truncated gaussian, depending on the value of Y_i .
- ▶ $\beta_\gamma | L, U, \gamma$: multivariate gaussian.
- ▶ $U | L, \beta, D$: multivariate gaussian.
- ▶ $D | U$: $D = \text{diag}(A_j)$ and a posteriori the A_j follow inverse-Wishart distributions.
If the A_j are diagonal, $A_j = \sigma_j^2 I$, the σ_j^2 follow a posteriori inverse-Gamma distributions.

The full conditional distributions (2/2)

► $\gamma \mid \beta_\gamma, L, U$:

$$f(\gamma \mid \beta_\gamma, L, U) \propto \exp \left[-\frac{1}{2} \left(-L' X_\gamma \beta_\gamma - \beta_\gamma' X_\gamma' L + \beta_\gamma' X_\gamma' Z U + U' Z' X_\gamma \beta_\gamma + \beta_\gamma' V_\gamma^{-1} \beta_\gamma \right) \right] \times |c(X_\gamma' X_\gamma)^{-1}|^{-\frac{1}{2}} \prod_{j=1}^p \pi_j^{\gamma_j} (1 - \pi_j)^{1-\gamma_j}$$

DEPEND ON β_γ !

Use of the collapsing technique (1/2)

- ▶ The full conditional distribution of γ is not a standard, easy to simulate distribution.
- ▶ A Metropolis-Hastings algorithm can be used to simulate it
⇒ Metropolis-within-Gibbs algorithm.
- ▶ Even with a Metropolis-Hastings algorithm, the full conditional distribution of γ is difficult to obtain :
 - ▶ It depends on the actual value of β_γ .
 - ▶ The acceptance rate for a candidate γ^* will depend both on the actual $\gamma^{(t)}$ and $\beta_{\gamma^{(t)}}$, and on the proposed γ^* and β_{γ^*} .
 - ▶ β_{γ^*} is unknown : Problem !

Use of the collapsing technique (2/2)

- ▶ **SOLUTION** Combine the Metropolis-within-Gibbs algorithm with the collapsing technique of Liu (1994) : using an "integrated" distribution, instead of a full conditional distribution.
- ▶ Improves the algorithm, and facilitates the convergence of the Markov chain (Liu et al., 1994).

Use of the collapsing technique combined with the Metropolis-within-Gibbs algorithm

- ▶ Integrate $\beta\gamma$ out in the full conditional distribution of γ .
- ▶ The "integrated" distribution obtained for γ will be easily simulated by a Metropolis-Hastings algorithm.

Integrating β_γ out in the full conditional distribution of γ

The full conditional distribution of γ is given by $f(\gamma|\beta_\gamma, L, U)$. The "integrated" density $f(\gamma|L, U)$ is :

$$f(\gamma|L, U) \propto \exp \left[-\frac{1}{2} \left\{ (L - ZU)'(L - ZU) - \frac{c}{1+c} (L - ZU)' X_\gamma (X_\gamma' X_\gamma)^{-1} X_\gamma' (L - ZU) \right\} \right] \times \prod_{j=1}^p \pi_j^{\gamma_j} (1 - \pi_j)^{1-\gamma_j}$$

This distribution is easier to simulate with a Metropolis-Hastings algorithm than the full conditional distribution.

The Metropolis-Hastings algorithm to simulate γ

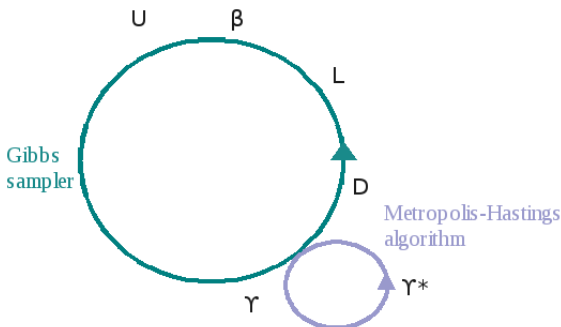
The number of selected variables should not increase between iterations, and for computational reasons

⇒ the number of variables to be selected at each iteration is fixed.

- ▶ The proposed $\gamma^{(t+1)}$ corresponds to $\gamma^{(t)}$ for which r components have been changed, but in such a way that the number of components whose values are 1 is invariant : $r/2$ components among the 1 values, and $r/2$ components among the 0 values are chosen at random and switched.

The Metropolis-within-Gibbs algorithm combined with the collapsing technique

At each iteration L , U , β and D (or σ^2) are simulated from their full conditional distributions, and γ from its "integrated" distribution with the Metropolis-Hastings algorithm.



Selection of variables

$\gamma^{(t)} = \{(\gamma_1^{(t)}, \dots, \gamma_p^{(t)})\}, t = 1, \dots, m$ after the burn-in period.

- ▶ The relevant variables appear most frequently in the simulated observations from the posterior distribution of γ .
- ▶ They can be identified as the γ values that are most often equal to 1.

Introduction

Probit Mixed Model for Gene Selection

Bayesian Gene Selection via Gibbs Sampling

Application

Discussion

Description of the datasets (1/2)

- ▶ **Three different datasets** : a dataset from the Institut Paoli Calmettes, the GEO series GSE2109 dataset, and the GEO series GSE5460 dataset.
- ▶ A microarray experiment has been done for each patient, and for each the expression measurements of ≈ 54000 **probesets** are obtained.
- ▶ Each dataset has been previously **preprocessed** : convolution background correction, normalization and summarization of the data with the RMA procedure of Irizarry et al. (2003).

Description of the datasets (2/2)

- ▶ Split and merge \implies a **training set** (497 patients) and a **validation set** (88 patients).
- ▶ A variable corresponding to the dataset is the **random effect**.
- ▶ **Two filters** applied on the probesets : to eliminate the noise and the invariants $\implies \approx 19\ 000$ **probesets**.

Goal

Select few probesets related to the ER status of the patients, by taking into account the different experimental conditions between the different merged datasets.

Choice of a priori values

- ▶ 30 probesets selected at each iteration of the Gibbs sampler.
- ▶ 10 of them changed at each iteration of the Metropolis-Hastings algorithm (5 zeros and 5 ones, $r = 10$).
- ▶ One random effect of length 3 (3 datasets). D is assumed diagonal : $D = \sigma^2 I$.

Results of our variable selection method

- ▶ 40 probesets selected at least once from the last 30000 iterations.
- ▶ Selected probesets used in a more classical way : a probit mixed model fitted on the training set with a stepwise selection and using the AIC and BIC criteria.
- ▶ Six models appeared to fit : one will be presented here.

Model fitted

Probeset in gene	Coefficient	Pvalue
Intercept	-9.12074	1.92e-05
228241_at in AGR3	0.45046	1.12e-15
205862_at in GREB1	0.77639	4.18e-08
202376_at in SERPINA3	0.37965	0.000149
216222_s_at in MYO10	-0.63551	0.004967
1568760_at in MYH11	0.42742	0.050219

Tab. 1: Probesets selected in the model and associated coefficients.

The estimated random effects of this model were reasonable :
 -0.284, 0.199 and 0.087.

Predictions on the validation set

Two kinds of predictions :

1. Using patient-dataset membership information.
2. Patient coming from an undefined or new dataset.

Same results for the two kinds of predictions.

Results quite good (1 misclassification among 88). The specificity is 0.973 and the sensitivity is 1.

With an "undetermined or grey zone" , no false predictions anymore, but 10 undetermined (11.4%).

Predictions on new datasets

- ▶ Datasets freely available from the NCBI GEO site.
- ▶ Patients coming from new datasets.

	GSE6532 dataset		GSE12763 dataset	
	ER negative	ER positive	ER negative	ER positive
Negative prediction	1	0	8	0
Positive prediction	0	85	1	20

Tab. 2: Predictions by the model for two new datasets : the GSE6532 dataset and the GSE12763 dataset.

Introduction

Probit Mixed Model for Gene Selection

Bayesian Gene Selection via Gibbs Sampling

Application

Discussion

Discussion concerning the method

- ▶ **Use of merged datasets** : variables selected in a mixed framework.
 - ▶ More observations.
 - ▶ datasets split in training and validation sets.
 - ▶ Avoid bias due to any one particular dataset.
 - ▶ Many diverse datasets freely available from the NCBI GEO site <http://www.ncbi.nlm.nih.gov/geo/>.
- ▶ **Fix number of variables** to be selected at each iteration of the Gibbs sampler.
- ▶ **Convenient to implement.**
- ▶ Once the algorithm has converged, estimates of the random effects can be used for analysis.

Discussion concerning the results

- ▶ The probesets selected by our method \implies good predictions.
- ▶ Using a SVM method on 1 dataset, Ipsogen selected 3 genes among the 5 used in our model.
- ▶ 3 genes among the 5 in our fitted model associated with ER pathways and breast cancer : GREB1, SERPINA3 and MYH11.
- ▶ Efficient and feasible, even for very large datasets with around 20000 variables.
- ▶ R package in preparation.

This approach has a clear advantage over other selection methods which handle less variables or which do not take into account random effects.

Acknowledgements

IML, Equipe Statistique et applications

- Pr Denys Pommeret
- MC Agnès Grimaud

Institut Paoli Calmettes, Marseille

- Dr Daniel Birnbaum
- Pr François Bertucci

Ipsogen Bioinformatics unit

- Sabrina Carpentier
- Rebecca Tagett
- Virginie Fasolo



For Further Reading



J.H. Albert and S. Chib.

Bayesian analysis of binary and polychotomous response data.
JASA, 88 (422), 669-679, 1993.



E.I. George and R.E. McCulloch.

Variable selection via Gibbs sampling.
JASA, 88 (423), 881-889, 1993.



K.E. Lee and N. Sha and E.R. Dougherty and M. Vannucci and B.K. Mallick

Gene selection : a bayesian variable selection approach.
Bioinformatics, 19 (1), 90-97, 2003.



J.S. Liu and W.H. Wong and A. Kong.

Covariance structure and convergence rate of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes.
Biometrika, 81, 27-40, 1994.



J.S. Liu.

The collapsed Gibbs sampler in bayesian computations with application to a gene regulation problem.
JASA, 89 (427), 958-966, 1994.



M. Baragatti.

Bayesian variable selection for probit mixed models.
Submitted.