LASSO-Patternsearch for Multivariate Bernoulli (MVB) Observations with Applications

Bin Dai

Department of Statistics,
University of Wisconsin Madison

September 16th 2010

Outline

Outline

Why we need LASSO for multivariate Bernoulli

- Correlated Bernoulli outcomes come from many applications, such as systolic blood pressure (BP) and intraocular pressure (IOP) in medical studies.
- Both biological variables (SNPs) and environmental variables (smoke, age) were proved to be important in a sparse manner so variable selection approach is of great need.
- LASSO is a powerful and efficient variable selection tool, and it has been already applied to various models.

Outline

- Let $Y = (Y_1, \ldots, Y_K)$ be a K-dimensional vector of possibly correlated Bernoulli random variables (binary outcomes) and let $y = (y_1, \ldots, y_K)$ be a realization of Y. The most general form $p(y_1, \ldots, y_K)$ of the joint density is (Whittaker, 1990)

$$p(y_1, \ldots, y_K) = p(0, 0, \ldots, 0)^{[\pi_{j=1}^{K}(1-y_j)]} p(1, 0, \ldots, 0)^{[y_1 \pi_{j=2}^{K}(1-y_j)]} \ldots p(1, 1, \ldots, 1)^{[\pi_{j=1}^{K} y_j]} \quad (1)$$

or we can write this in a simpler form

$$p(y) = p_{0,0,\ldots,0}^{[\pi_{j=1}^{K}(1-y_j)]} \, p_{1,0,\ldots,0}^{[y_1 \pi_{j=2}^{K}(1-y_j)]} \ldots p_{1,1,\ldots,1}^{[\pi_{j=1}^{K} y_j]} \quad (2)$$

- The special form of $K = 2$ can be written as

$$p(y_1, y_2) = p_{00}^{(1-y_1)(1-y_2)} \, p_{01}^{(1-y_1)y_2} \, p_{10}^{y_1(1-y_2)} \, p_{11}^{y_1 y_2} \quad (3)$$

The Log-linear model

- Let the probabilities depend on some attribute vector $X = (X_1, \dots, X_p)$, which is a subset of $\mathcal{R}^p$. By using the natural parameters, the negative log likelihood can be written as

$$-L(y, \mathbf{f}(x)) = -[\sum_{j=1}^{K} f^j(x)B_j(y) + \sum_{1 \le j_1 < j_2 \le K} f^{j_1 j_2}(x)B_{j_1 j_2}(y) + \dots + f^{12\dots K}(x)B_{12\dots K}(y) - b(\mathbf{f}(x))] \quad (4)$$

where $B_{j_1 j_2 \dots j_r}(y) = y_{j_1} y_{j_2} \dots y_{j_r}$ and $\mathbf{f} = (f^1, f^2, \dots, f^{12\dots K})^T$.

$$b(\mathbf{f}(x)) = log(1 + \sum_j e^{S^j(x)} + \sum_{1 \le j_1 < j_2 \le K} e^{S^{j_1 j_2}(x)} + \sum_{1 \le j_1 < j_2 < j_3 \le K} e^{S^{j_1 j_2 j_3}(x)} + \dots + e^{S^{12\dots K}(x)})$$

where

$$S^{j_1 j_2 \cdots j_r}(x) = \sum_{1 \le s \le r} f^{j_s}(x) + \sum_{1 \le s < t \le r} f^{j_s j_t}(x) + \dots + f^{j_1 j_2 \cdots j_r}(x)$$

LEMMA (Parameter transformation). *For multivariate Bernoulli model, the general parameters and natural parameters have the following relationship.*

$$\exp(f^{j_1 j_2 \cdots j_r}) = \tag{5}$$

$$\frac{\prod p(\textit{even number zeros among } j_1, \ldots, j_r \textit{ positions and other K-r positions are all zero})}{\prod p(\textit{odd number zeros among } j_1, \ldots, j_r \textit{ positions and other K-r positions are all zero})}$$

in addition

$$\exp(S^{j_1 j_2 \cdots j_r}) = \frac{p(j_1, \ldots, j_r \text{ positions are one, others are zero})}{p(0, 0, \ldots, 0)} \tag{6}$$

- PROPOSITION (Conditional Covariance). *In the multivariate Bernoulli model, f^{jk} is related to the conditional variance of two outcomes, without loss of generality, just take $j = 1$ and $k = 2$*

$$\exp(f^{12}) = cov(Y_1, Y_2 | Y_3 = 0, \ldots, Y_K = 0) \tag{7}$$

- What's more in the bivariate Bernoulli,
COROLLARY *When $K = 2$ for multivariate Bernoulli distribution*

$$\begin{aligned} \exp(f^{12}) &= p_{11}p_{00} - p_{01}p_{10} \\ &= cov(Y_1, Y_2) \end{aligned} \tag{8}$$

and $f^{12} = 0$ if and only if Y_1 and Y_2 are uncorrelated.

First order derivative.

- Direct calculation or shows that

$$\frac{\partial - l(y, \mathbf{f}(x))}{\partial f^{j_1 j_2 \dots, j_r}(x)} = -B_{j_1 j_2 \dots, j_r}(y) + \frac{\sum_{\tau \in \mathcal{T}(j_1, j_2, \dots, j_r)} e^{S^{\tau}(x)}}{e^{b(\mathbf{f}(x))}}$$
$$= -B_{j_1 j_2 \dots, j_r}(y) + \mu^{j_1 j_2 \cdots j_r}(x) \quad (9)$$

where $\mathcal{T}(j_1, j_2, \dots, j_r)$ is the collection of interaction indexes which include $j_1, j_2, \dots j_r$ and $\mu^{j_1 j_2 \cdots j_r}(x) = E\left(B_{j_1 j_2 \dots j_r}(Y) | \mathbf{f}(x)\right)$, which is the conditional mean.

- For instance in $K = 2$, the first derivative with respect to f^1 is

$$\frac{\partial - l(y, \mathbf{f}(x))}{\partial f^1(x)} = -B_1(y) + \frac{e^{S^1} + e^{S^{12}}}{e^{b(\mathbf{f}(x))}}$$
$$= -y_1 + \frac{e^{f^1} + e^{f^1 + f^2 + f^{12}}}{e^{b(\mathbf{f}(x))}} \quad (10)$$

Second order derivative.

- From the first order derivative, we can derive that

$$\frac{\partial^2 - l(y, \mathbf{f}(x))}{\partial f^{j_1 j_2 \dots, j_r}(x) \partial f^{h_1 h_2 \dots, h_s}(x)} = Cov\left(B_{j_1 j_2 \dots j_r}(Y), B_{h_1 h_2 \dots h_s}(Y) \mid \mathbf{f}(x)\right) \tag{11}$$

Hence the Hessian with respect to f is

$$\frac{\partial^2 - l(y, \mathbf{f}(x))}{\partial \mathbf{f}(x) \partial \mathbf{f}(x)^T} = Var\left(B(Y) | \mathbf{f}(x)\right) \tag{12}$$

which is exactly the conditional covariance matrix.

The negative log-likelihood for Bivariate Bernoulli log linear model can be written as follows:

$$
\begin{aligned}
L(y, f) &= -\frac{1}{n}\sum_{i=1}^{n}\Big[y_1(i)f^1(x(i)) + y_2(i)f^2(x(i)) + y_1(i)y_2(i)f^{12}(x(i)) - b(f(x(i)))\Big] \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{\tau=1,2,12} f^{\tau}(x(i))B^{\tau}(y(i)) - b(f(x(i)))\right] \qquad (13)
\end{aligned}
$$

here the index i refers to the subjects, with range $1, \ldots, n$. The f functions are formulated as the so-called linear predictors, for instance the f^1 function can be represented by:

$$f^1(x) = c_0^1 + x_1 c_1^1 + \ldots + x_p c_p^1 \qquad (14)$$

The target function

In most cases of real applications, the dimension of the genetic data p is large but only a small portion of covariates have important effects on the responses, so the l_1 penalty can be applied to impose sparsity. The target function can be formulated as:

$$\mathcal{I}_\lambda(y, f) \quad = \quad L(y, f) + J_\lambda(f), \tag{15}$$

where the penalty function is defined to be sum of l_1 penalty:

$$J_\lambda(f) = \lambda_1 \sum_{j=1}^{p} |c_j^1| + \lambda_2 \sum_{j=1}^{p} |c_j^2| + \lambda_{12} \sum_{j=1}^{p} |c_j^{12}|, \tag{16}$$

Outline

First-order step

The basic (first-order) step at iteration k is obtained by forming a simple model of the objective by expanding around current iterate $\mathbf{c}^k$ ($\mathbf{c}$ is the coefficients vector) as follows:

$$\mathbf{d}^k = arg \min_{\mathbf{d}} L(\mathbf{c}^k) + \bigtriangledown L(\mathbf{c}^k)^T \mathbf{d} + \frac{1}{2}\alpha_k \mathbf{d}^T \mathbf{d} + \lambda^T ||\mathbf{c}^k + \mathbf{d}||_1 \tag{17}$$

where α_k is a positive scalar and $\mathbf{d}^k$ is the proposed step. The subproblem (17) is separable in the components of $\mathbf{d}$ and therefore trivial to solve in closed form, in $O(3p)$ operations.

Active and inactive set

The solution $\mathbf{d}^k$ can be examined to obtain an estimate of the active set:

$$\mathcal{A}_k = \{j = 1, 2, \ldots, 3p | (\mathbf{c}^k + \mathbf{d}^k)_j = 0\} \tag{18}$$

The definition of the "inactive set" estimate $\mathcal{I}_k$ is the complement of the active set estimate, that is:

$$\mathcal{I}_k = \{1, 2, \ldots, 3p\} \setminus \mathcal{A}_k \tag{19}$$

Newton step

We enhance the step by computing the restriction of the Hessian $\bigtriangledown^2 L(\mathbf{c}^k)$ to the set $\mathcal{I}_k$ (denoted by $\bigtriangledown^2_{\mathcal{I}_k\mathcal{I}_k} L(\mathbf{c}^k)$) and then computing a Newton-like step in the $\mathcal{I}_k$ components as follows:

$$(\bigtriangledown^2_{\mathcal{I}_k\mathcal{I}_k} L(\mathbf{c}^k) + \delta_k I)\mathbf{p}^k_{\mathcal{I}_k} \quad = \quad -\bigtriangledown_{\mathcal{I}_k} L(\mathbf{c}^k) - \lambda^T \omega_{\mathcal{I}_k} \tag{20}$$

where δ_k is a small damping parameter that goes to zero as $\mathbf{c}^k$ approaches the solution, and $\omega_{\mathcal{I}_k}$ captures the gradient of the term $||\mathbf{c}||_1$ at the nonzero components of $\mathbf{c}^k + \mathbf{d}^k$.

Algorithm

The first-order step is cheaper to calculate than the Newton step, the general iterative steps of the algorithm therefore can be summarized as follows:

1. Evaluate the current first-order step $\mathbf{d}^k$ with a proper α_k.
2. Calculate the Newton step $\mathbf{p}^k_{\mathcal{I}_k}$, only if the inactive size is less than a predefined threshold.
3. Take the better step between first-order and Newton.
4. Check optimal condition, repeat if not satisfied.

There are some improvement to the algorithm omitted here.

Outline

Tuning Criterion

So far, all smoothing parameters are considered fixed. However, the choice of the tuning parameters is crucial and 4 different criterion are considered

- AIC, aimed at prediction, and the degrees of freedom can be approximated by the number of nonzero coefficients.
- BIC, used for variable selection, is the Bayesian version of AIC but achieving more sparsity.
- GACV (generalized approximate cross-validation) used to minimize the comparative Kullback-Leibler (CKL) distance.
- BGACV the Bayesian version of GACV criteria, analogous to BIC.

The augmented response for the ith subject $y(i) = (y_1(i), y_2(i))$ is defined by

$$\mathcal{Y}(i) = (y_1(i), y_2(i), y_1(i)y_2(i))^T \tag{21}$$

the augmented covariate $\mathcal{X}$ can be similarly defined, then the vector form can be constructed as follows:

$$\begin{aligned}\vec{f}(x) &= (f^1(x(1)), f^2(x(1)), \ldots, f^{12}(x(n)))^T \\ \vec{\mathcal{Y}} &= (\mathcal{Y}(1), \mathcal{Y}(2), \ldots, \mathcal{Y}(n))^T\end{aligned}$$

Leaving-out-one-augmented-subject lemma.

For fixed i and a new augmented response $\tilde{\mathcal{Y}}$, let $h_\lambda[i, \tilde{\mathcal{Y}}]$ be the minimizer of

$$-\sum_{k \neq i} l(y(k), \mathbf{f}(x(k))) - \tilde{\mathcal{Y}}^T \mathbf{f}(x(i)) + b(\mathbf{f}(x(i))) + n\mathbf{J}_\lambda(\mathbf{f}) \tag{22}$$

Then $h_\lambda \left[i, \mu_\lambda^{[-i]}(x(i)) \right] = \mathbf{f}_\lambda^{[-i]}$. Here $\mu_\lambda^{[-i]}(x(i)) = E[\mathcal{Y} | \mathbf{f}_\lambda^{[-i]}(x(i))]$.

Augmented Linear Predictor

The vector form of the linear predictor $\vec{f}(x)$ can be formulated as:

$$\vec{f}(x) = \mathcal{D}\beta$$

where the corresponding design matrix and the coefficients to be estimated are

$$\begin{aligned} \mathcal{D} &= \begin{pmatrix} x(1) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & x(1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & x(1) \\ x(2) & \mathbf{0} & \mathbf{0} \\ \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & x(n) \end{pmatrix} \\ \beta &= \left(c_1^1, c_2^1, \dots, c_n^{12}\right)^T \end{aligned}$$

GACV

- Let $\hat{\beta}_\lambda$ be the estimated β for a specific tuning parameter λ, and denote the number of nonzero elements in $\hat{\beta}_\lambda$ to be s and $\mathcal{D}^*$ is the sub-matrix of $\mathcal{D}$ with columns corresponding to nonzero elements in $\hat{\beta}_\lambda$. Define the H matrix

$$H = {\mathcal{D}^*}^T \left(\mathcal{D}^* W(f_\lambda)(\mathcal{D}^*)^T\right)^{-1} \mathcal{D}^*$$

  where $W(f_\lambda) = \text{Var}(\mathcal{Y}|\vec{f}_\lambda)$.

- The GACV score can therefore be evaluated:

$$\text{GACV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[-\mathcal{Y}(i)^T f_\lambda(x(i)) + b(f_\lambda(x(i))\right] + \frac{tr(H)}{n}\frac{\sum_{i=1}^{n}\mathcal{Y}(i)^T(\mathcal{Y}(i) - \vec{\mu})}{n-s} \quad (23)$$

  here $\vec{\mu} = E\left[\mathcal{Y}|\mathbf{f}(x)\right]$

Setup

- In this simulation, the sample size is set to 500 ($n = 25$), and 25 ($p = 25$) independent binary predictor variables ($X_1, X_2, \ldots, X_{25}$) are generated. The true model is

$$\begin{aligned} f^1(X) &= -4 + 2X_1 + 2X_2 + 1.5X_6 \\ f^2(X) &= -3 + 2X_3 + 1.5X_4 + 1.5X_7 \\ f^{12}(X) &= -3 + 2X_5 \end{aligned}$$

- Thus there are in total 78 candidate patterns in the model and only 10 of them are nonzero patterns in the true model.
- 100 independent data sets were generated and fitted by the LASSO in bivariate Bernoulli model.

Simulation Result

f^1	-4	$2X_1$	$2X_2$	$1.5X_6$
GACV	100	100	100	87
BGACV	100	95	94	69
AIC	100	100	100	85
BIC	100	100	100	82
f^2	-3	$2X_3$	$1.5X_4$	$1.5X_7$
GACV	100	100	80	88
BGACV	100	99	57	66
AIC	100	100	65	77
BIC	100	98	65	70
f^{12}	-3	$2X_5$	Average Noise	
GACV	100	100	19.15	
BGACV	100	98	9.34	
AIC	100	99	16.3	
BIC	100	87	2.56	

Table: The number of true patterns captured in 100 simulations.

Outline

Beaver Dam Eye Study

Introduction to the data set

- The Beaver Dam Eye Study (BDES) is an ongoing population-based study of age-related ocular disorders including cataract, age-related macular degeneration, visual impairment and refractive errors.
- 2061 patient with 4886 SNPs information with missing observations.
- Pedigree information available for a few families
- Measurements of environmental variables (blood pressure, intraocular pressure, etc.) as follow-up data collected every 4 to 5 years.

What do we want to find

- Both continuous and discrete variables that contribute to main effects and interactions of BP and IOP
- Whether the influence of the continuous variables is linear to the outcomes
- The improvement of the accuracy of the model with pedigree information.

Outline

Summary

1. LASSO penalty is a powerful tool in model selection, it can be applied to multivariate Bernoulli models.
2. The LASSO-Patternsearch algorithm can efficiently handle large scale convex problems with l_1 penalty.
3. The tuning scores such as GACV, BGACV, AIC and BIC has superior performance than 10-fold cross validation in terms of runtime and achieving sparsity.

Future Development

- n gets larger.
- p gets larger.
- K gets larger.
- Relax linearity assumption of f.