Bayesian Analysis of Risk for Data Mining Based on Empirical Likelihood

Yuan Liao Wenxin Jiang

Northwestern University

Presented at:

Department of Statistics and Biostatistics

Rutgers University

November 3, 2010

Binary Classification Problem

- $Y \in \{0, 1\}$ is a target variable to be predicted, associated with covariates X.
- Classification rule: $C(X, \theta) \in \{0, 1\}$. θ: action parameter
  Example: $C(X, \theta) = I(X^T\theta > 0)$
- Classification risk: $E(l(Y, C(X, \theta))|\theta)$. We consider absolute loss: $l(Y, C(X, \theta)) = |Y - C(X, \theta)|$.
- Let $r = E|Y - C(X, \theta)|$: risk parameter.
- Observe i.i.d. data $D = (Y_1, X_1, ..., Y_n, X_n)$

In This Paper

- Classical parametric approach assumes that $P(Y = 1|X)$ has a parametric form, i.e., logistic regression.
- This paper: does not specify a parametric form of $P(Y = 1|X)$, to avoid mis-specification
- $C(X, \theta)$ is fixed.
- The only information we have is

$$E|Y - C(X, \theta)| = r$$

- We would like to control r, and answer questions like: In order for $r \leq 0.1$, what action θ should be taken?

Our Bayesian Approach

- We construct the posterior $P(\theta, r|Data)$.
- Once the posterior is obtained, it allows us to look at:
  - $P(\theta|r \leq r_0, Data)$
  - $P(r|\theta, Data)$
- When X is multi-dimensional: we can do model selection.
  - $M_1 = (X_1, X_2)$
  - $M_2 = (X_2, X_3, X_4)$
  - etc.
- We can look at $P(M|r \leq r_0, Data)$

Outline

Empirical Likelihood

- As the functional form of $P(Y = 1|X)$ is not specified, we construct the likelihood nonparametrically, based on

$$E|Y - C(X, \theta)| = r$$

- Empirical likelihood (Owen (1990)):

$$L_{EL}(\theta, r) = \max_{p_1,\ldots,p_n} \prod_{i=1}^{n} p_i$$

$$s.t. \qquad p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i |Y_i - C(X_i, \theta)| = r$$

Empirical Likelihood Posterior

- Using Lagrange's multiplier, we obtain (Qin and Lawless (1994)):

$$\log L_{EL}(\theta, r) = -\sum_{i=1}^{n} \log\{1+\mu(\theta, r)[|Y_i - C(X_i, \theta)| - r]\} - n \log n$$

where $\mu(\theta, r)$ solves

$$\sum_{i=1}^{n} \frac{|Y_i - C(X_i, \theta)| - r}{1 + \mu(\theta, r)[|Y_i - C(X_i, \theta)| - r]} = 0$$

- EL-posterior:

$$P_{EL}(\theta, r | Data) \propto L_{EL}(\theta, r)\pi(\theta, r)$$

Bayesian Interpretation of EL-posterior

EL has not formally been shown to have a well-defined probabilistic interpretation that would justify its use in Bayesian inference.

Informal justification:

- Monahan and Boos (1992) proposed a definition of validity of a “posterior” $P_a(.|Data)$ resulting from alternative likelihood:
  - Recall that if $\Lambda \sim P(\lambda|Data)$, with posterior cdf F, then $F(\Lambda|Data) \sim Uniform[0, 1]$
  - valid ”posterior”: $\int_{-\infty}^{\Lambda} P_a(\lambda|Data)d\lambda \sim Uniform[0, 1]$
- Lazar (2003)

Back to our framework: $E|Y - C(X, \theta)| = r$.

Define "empirical risk"

$$\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} |Y_i - C(X_i, \theta)|$$

Theorem 1

$\log L_{EL}(\theta, r) = -nK(\hat{R}(\theta), r)$, *where*

$$K(p,q) = \begin{cases} p \ln(p/q) + (1-p)\ln\{(1-p)/(1-q)\}, \text{ if } p, q \in (0,1) \\ +\infty, \quad \text{if } p \in (0,1], q = 0, \text{ or } p \in [0,1), q = 1 \\ 0 \quad \text{if } q \in [0,1), p = 0, \text{ or } q \in (0,1], p = 1. \end{cases}$$

Interpretation of $\pi(\theta, r)$

(θ, r)

- θ is the action parameter in $C(X, \theta)$, which is NOT the model parameter in $P(Y = 1|X)$. It can be ANY action that the decision makers can take.
- r is the resulting risk after an action is taken.

$\pi(\theta, r)$

- Can assume $\pi(\theta, r) = \pi(\theta)\pi(r)$: (θ, r) are *a priori* independent: data can tell their relationship
- $\pi(\theta)$: Distribution of All possible actions: decision makers' "prior preference" before looking at the data

Posterior Consistency under Partial Identification

$E|Y - C(X,\theta)| = r$ does not point identify (θ, r): $P_{EL}(\theta, r|Data)$ does not de-generate to any point mass. We can show the folllowing "partially identified" version of posterior consistency:

Theorem 2

Let $R(\theta) = E|Y - C(X,\theta)|$*, and* $\eta(\theta, r) = \min\{R, 1-R, r, 1-r\}$*.*

(i) $\pi(|R - r| \leq \delta, \eta \geq \tau) > 0 \ \forall \delta > 0 \forall \tau \in (0,1)$*;*

(ii) $\sup_{\theta \in \Theta} |\hat{R}(\theta) - R(\theta)| \to^p 0$

Then $\forall \epsilon > 0$*,*

$$P_{EL}(R(\theta) - \epsilon \leq r \leq R(\theta) + \epsilon | D) \to^p 1.$$

Hence $P_{EL}(\theta, r|D)$ clusters around $\{(\theta, R(\theta)) : \theta \in \Theta\}$ as $n \to \infty$.

Corollary 2.1

Suppose that $P_{EL}(r \leq r_0 | D) > \xi$ for some constant $\xi > 0$, then for any $\epsilon > 0$,

$$P_{EL}(R(\theta) \leq r_0 + \epsilon | r \leq r_0, Data) \rightarrow^{P} 1$$

This corollary implies: if $\theta \sim P_{EL}(\theta | r \leq r_0, Data)$, then the true risk $E|Y - C(X, \theta)| \leq r_0$ with very high posterior probability.

A Numerical Example

- Model: $Y = I(3X > \epsilon), \quad X \sim N(0,1) \perp \epsilon \sim N(0,3)$
  Generated 2000 data points $(Y_1, X_1), ..., (Y_n, X_n)$.
- Classification rule: $C(X, \theta) = I(X > \theta)$
- $E|Y - C(X,\theta)| = E_X\{[1 - \Phi(\sqrt{3}X)]I_{(X>\theta)} + \Phi(\sqrt{3}X)I_{(X\leq\theta)}\}$
- $\pi(\theta) \sim N(0,1)$: my "prior preference" of taking action.
- $P(\theta, r|Data) \propto \pi(\theta)\pi(r)\exp(-nK(\hat{R}(\theta), r))$:
  $\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq I(X_i > \theta))$

Figure: Plot of $R(\theta) = E|Y - C(X,\theta)|$ and MCMC draws

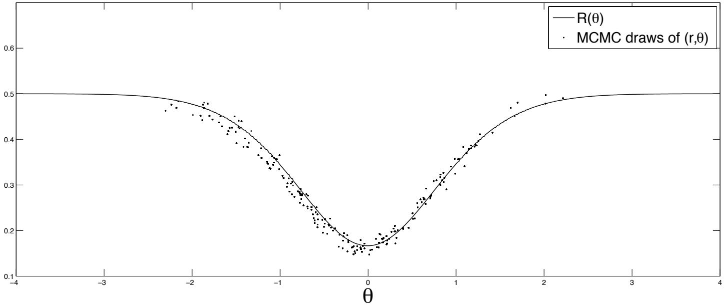


Figure: $P_{EL}(\theta|D, r \leq$ 5th percentile of MCMC draws)

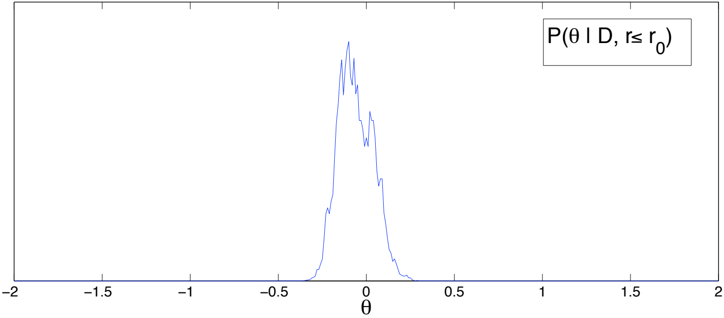


Figure: $P_{EL}(r|D), P_{EL}(r|D, \theta = \arg\min \hat{R}(\theta))$

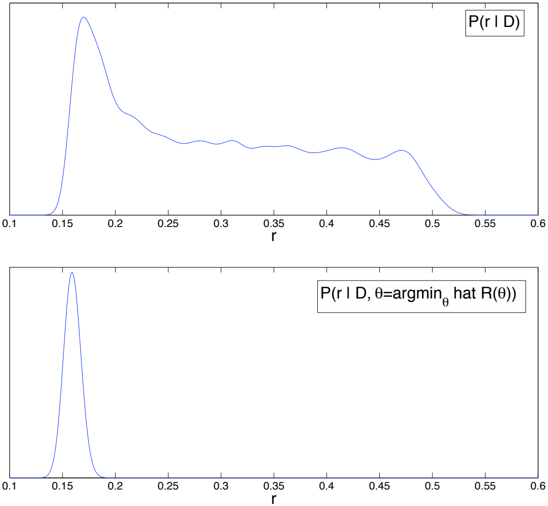
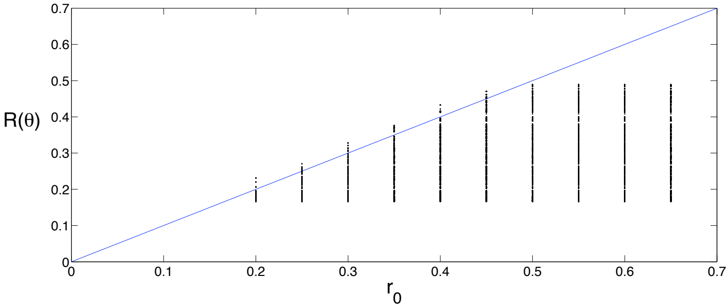


Figure: Scatter plot of $(R(\theta_i), r_0)$

$$R(\theta_i) = E|Y - C(X, \theta_i)|$$



The dots with the same horizontal coordinate r_0 represents $R(\theta_i)$ where $\theta_i \sim P_{EL}(\theta|r \leq r_0, D)$. When $r_0 \leq 0.5$, about 97.3% dots are below the identical line $R(\theta) = r_0$.

More general loss functions

- Symmetric loss:

  $l(Y, C) = I(Y = 1, C = 0) + I(Y = 0, C = 1) = |Y - C|$

- Asymmetric loss:

  $l(Y, C) = I(Y = 1, C = 0) + aI(Y = 0, C = 1)$, for $a \neq 1$.

  Example: $Y = 0$: good/bad credit card user. $C = 1$: issue/not credit card

- EL-posterior based on: $El(Y, C(X, \theta)) = r$.

  For general $l(Y, C)$, there is no explicit expression of $L_{EL}(\theta, r)$.

German credit data: an application

- Data set comes from Asuncion and Newman (2007), which consists of 1000 past applicants
- Y : credit rating (Good/ Bad); X : demographic data, etc.
- $C(X,\theta) = I(X_1 + \theta_1 + \sum_{i=2}^{24} X_i\theta_i > 0)$

Table: Cost Matrix

		Classification	
		GOOD	BAD
Y	GOOD	0	1
	BAD	5	0

Variable Selection

- $C(X, \theta(\psi)) = I(X_1 + \theta_1 + \sum_{i=2}^{24} X_i \theta_i \psi_i > 0)$:

  $\psi_i = 1/0$ if θ_i is selected/not selected

- $L_{EL}(\theta, r, \psi)$ is based on:

  $E[I_{Y=G, C(X,\theta(\psi))=B} + 5 I_{Y=B, C(X,\theta(\psi))=G}] = r$

- Priors: $\theta(\psi)|\psi \sim N(0, 10I)$, $\psi_i \sim Bino(1, 0.4)$

  $r \sim Uniform[0, 5]$

- $P_{EL}(\theta, r, \psi | Data) \propto \pi(\theta, r, \psi) L_{EL}(\theta, r, \psi)$

Figure: Estimated $P(M|r \leq r_0, D)$ versus r_0

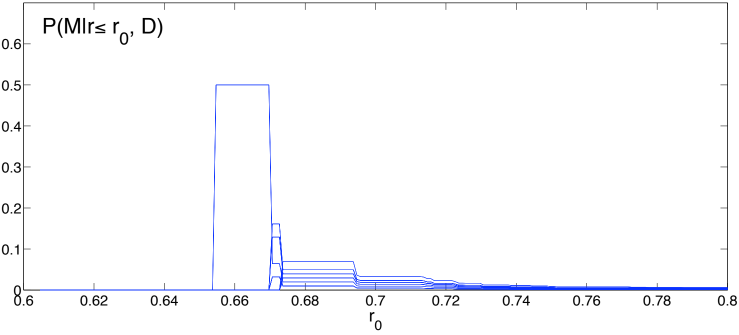


Table: The estimated posteriors of the sampled models when $r_0 \leq 0.674$

Model	M_1	M_2	M_3	M_4	$M_5 \sim M_7$	Another five
$r_0 \in (.655, .670)$	.50	.50	0	0	0	0
$r_0 \in (.670, .674)$	.07	.07	.16	.16	.13	.03

Model Assessment

- To assess the performance of the predictive models: divide dataset into training (2/3) and validation (1/3).
- Generate a new set of MCMC $\{(\theta_i, r_i)\}_{i=1}^{8,000} \sim P_{EL}(\theta, r | M_i, \text{Training})$.
- Choose $r_0 = $ 1st, 3rd, 5th, and 10th percentiles of $P_{EL}(r | M_i, \text{Training})$.
- $S(r_0) = \{(\theta_i, r_i) \in \{(\theta_i, r_i)\}_{i=1}^{8,000} : r_i \leq r_0\}$. Calculate

$$\hat{R} = \frac{1}{\#S(r_0)} \sum_{(\theta_j, r_j) \in S(r_0)} \frac{1}{n_v} \sum_{i=1}^{n_v} l(Y_i, C(X_i; \theta_j)),$$

$\approx E[\frac{1}{n_v} \sum_{i=1}^{n_v} l(Y_i, C(X_i; \theta)) | r \leq r_0, M, \text{Training Data}]$

Model	M_1	M_2	M_3	M_4	M_5	M_6	M_7
# variables	10	10	11	11	12	12	12
r_0	0.674	0.685	0.692	0.687	0.686	0.619	0.677
$\hat{R}$	0.683	0.678	0.680	0.679	0.681	0.787	0.787
$\frac{J}{8,000}$	1.0%	1.2%	1.15%	1.1%	1.1%	1.0%	1.38%
r_0	0.698	0.699	0.697	0.692	0.694	0.682	0.708
$\hat{R}$	0.683	0.680	0.681	0.679	0.680	0.793	0.781
$\frac{J}{8,000}$	2.9%	3.2%	3.7%	3.0%	3.0%	3.8%	3.3%
r_0	0.700	0.706	0.702	0.701	0.697	0.688	0.711
$\hat{R}$	0.680	0.679	0.693	0.679	0.680	0.771	0.768
$\frac{J}{8,000}$	5.3%	5.3%	5.1%	5.3%	5.2%	5.3%	5.5%
r_0	0.710	0.737	0.719	0.719	0.713	0.703	0.723
$\hat{R}$	0.681	0.678	0.696	0.687	0.681	0.742	0.770
$\frac{J}{8,000}$	10.0%	11.7%	10.2%	10.4%	10.0%	10.1%	10.5%

Model	Variables	
M_1	Other Debtors/ Guarantors	Duration of Credit
	Real Estate Property	Credit Amount
	Present Employment Since	Credit History
	Num. of Existing Credits at Bank	Credit Purpose
	Num. of People Being Liable	Age
M_2	M_1/ Credit Purpose	Telephone

Performance with symmetric loss

$$l(Y, C(X,\theta)) = |Y - C(X,\theta)|$$

- First run MCMC for model selection: the 1st percentile of $\{r_i\}_{i=1}^{B}$ is 0.275, achieved by only one model.
- Then split data into training and validation sets. Generate new $\{\theta_i\}$ from $P_{EL}(\theta|r \leq 0.275, \textit{training}, \textit{Model})$
- We can obtain the average of $\{\frac{1}{n_v}\sum_{i \in Validation}|Y_i - C(X_i,\theta_j)|\}, j = 1, ..., B.$

Performance with symmetric loss

- It is more satisfactory to use θ_i such that $P_{EL}(\theta_i | r \leq 0.275, \mathit{training}, \mathit{Model})$ is high than to use all the generated θ_i's.
- Let $f(\theta) = P(\theta | r \leq 0.275, \mathit{training}, M)$. Define $A(\alpha) = \{\theta : f(\theta) > \alpha \text{th percentile of } \{f(\theta_i)\}\}$
- Average $\{\frac{1}{n_v} \sum_{i \in \mathit{Validation}} |Y_i - C(X_i, \theta_j)|\}$ over $\theta_i \in A(\alpha)$

Table: Comparison of $\hat{R}_\alpha$ and the risk of logistic regression

α	$\hat{R}_\alpha$	logistic
5	0.2782	0.2733
30	0.2667	
50	0.2613	
95	0.2613	

- Our method is designed to provide a new language to make robust inference on the risk and the corresponding actions.
- It can still perform comparably with other well-established methods, when used for risk reduction.

Discussion

- We provide a new language for probabilistic inference on the relationship between risk-action.
  - $P_{EL}(\theta | r \leq r_0, Data)$
  - $P_{EL}(r | \theta, Data)$
- $P_{EL}(.|Data)$ is based on EL$\Leftarrow El(Y, C(X, \theta)) = r$.
  $P_{EL}(.|Data)$ does not degenerate to a point, but clusters around the curve $\{(\theta, r) : El(Y, C) = r\}$
- No need to specify a full probability model on $P(Y = 1|X)$.
- Need to be more rigorous on the relationship between EL-posterior and exact posterior: not fully understood yet.