LA-UR- 11-03165

Approved for public release;
distribution is unlimited.

Title:	Bayesian Model Selection for Good Prediction of Future Reliability Using a Generalized Linear Model
Author(s):	Adam Pintar Christine Anderson-Cook Huaiquig Wu
Intended for:	Quality and Productivity Research Conference Roanoke, VA June 8-10, 2011

Form 836 (7/06)

Bayesian Model Selection for Good Prediction of Future Reliability Using a Generalized Linear Model

Adam L. Pintar[1], Christine M. Anderson-Cook[2], Huaiqing Wu[3]

[1]National Institute of Standards & Technology
[2]Los Alamos National Laboratory
[3]Iowa State University

Generalized linear models such as probit and logit regression are important tools for assessing reliability; however, what explanatory variables to include is an important consideration. In the Bayesian paradigm, if unimportant explanatory variables are included, posterior distributions will have inflated variance. If important explanatory variables are excluded, posterior distributions can be biased and miss their (true, but unknown) target. Several model selection methodologies currently exist for this setting, including selection of the model with smallest deviance information criterion, the model with largest posterior probability, or the model containing all terms with posterior probability greater than 0.5. A common theme to all of these methodologies is that they consider only the observed data. However, if one is interested in predicting future reliability, a different strategy is suggested because it is possible that the best model for prediction is dependant on age range. We propose a model selection methodology that focuses on good prediction over a user-specified distribution on the covariate space. The methodology quantifies the prediction ability of all models under consideration at covariate points sampled from the user-specified distribution. Then, a best model is identified by graphically comparing the distributions of prediction abilities. The methodology is illustrated via an example, and a simulation study highlighting its performance is presented.

KEY WORDS: Model Selection; Reliability; Bayesian Information Criterion; Deviance Information Criterion; Posterior Probability; Bayesian Model Averaging

Bayesian Model Selection for Good Prediction of Future Reliability Using a Generalized Linear Model

Adam L. Pintar[1] Christine M. Anderson-Cook[2] Huaiqing Wu[3]

[1]National Institute of Standards and Technology

[2]Los Alamos National Laboratory

[3]Iowa State University

June 10, 2011

Outline

- Motivation
- Related work
- Overview of the procedure
- Missile example
- Avoiding high predictions
  - Missile example revisited
- Simulation study
- Conclusion

Motivation

- When the prediction of reliability over a particular part of the covariate space is the goal of model building, it should influence the selection process
- Consider a population of missiles:
  - Will the population of missiles require maintenance in the next 5 years or not?
  - Answering that question requires prediction the reliability of the missiles 5 years into the future
  - Because extrapolating inflates the prediction variance, a model with less terms may be preferred
- Statisticians know well the dangers of extrapolation
- When possible, extrapolation should be based on underlying scientific or engineering understanding
- The methodology is not restricted to extrapolation

Related Work

- Model selection
  - Deviance information criterion
  - Stochastic search variable selection
  - Median probability model
  - Rank with posterior probabilities
- Model Averaging
- Graphical tools used in experiment design literature
  - Boxplots
  - Fraction of design space plots

Procedure Overview

1. Characterize the relationship between covariates, and use that characterization, as well as the study goal, to select the covariate distribution of interest (DI)
2. Randomly sample new points from the DI
3. Calculate a statistic (presented on the next slide), on which comparisons are based, at all newly sampled points for all models considered
4. Compare models numerically and graphically using those statistics

The Measure of Prediction Ability I

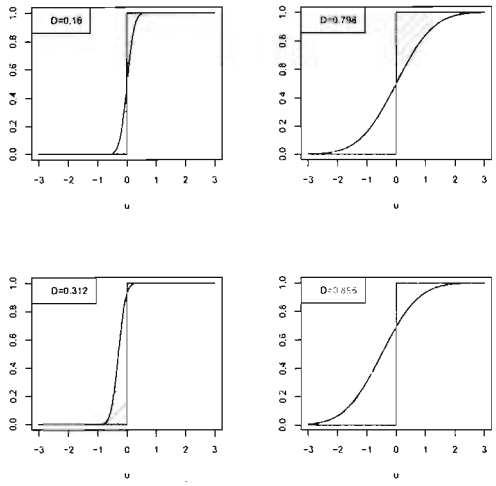
- The best possible posterior distribution for prediction is a point mass at $p(\boldsymbol{x}_{new})$ the true reliability, if it is known
  - $\boldsymbol{x}_{new}$ is a sampled point from the DI
- Let $F_{\boldsymbol{x}_{new}}$ be a cumulative distribution function (cdf) representing a point mass at $p(\boldsymbol{x}_{new})$
  - $F_{\boldsymbol{x}_{new}}$ steps from 0 to 1 at $p(\boldsymbol{x}_{new})$
- Let $F^m_{\boldsymbol{x}^m_{new}}$ be the posterior cdf of $p^m(\boldsymbol{x}^m_{new})$
- The discrepancy between the cdfs can be quantified by the following expression

$$D^k_m(\boldsymbol{x}^m_{new}) = \left\{ \int_{-\infty}^{\infty} \left| \frac{F^m_{\boldsymbol{x}^m_{new}}(u) - F_{\boldsymbol{x}_{new}}(u)}{\sqrt{V(\boldsymbol{x}_{new})}} \right|^k du \right\}^{\frac{1}{k}}$$

- Need a surrogate for $p(\boldsymbol{x}_{new})$

The Measure of Prediction Ability II

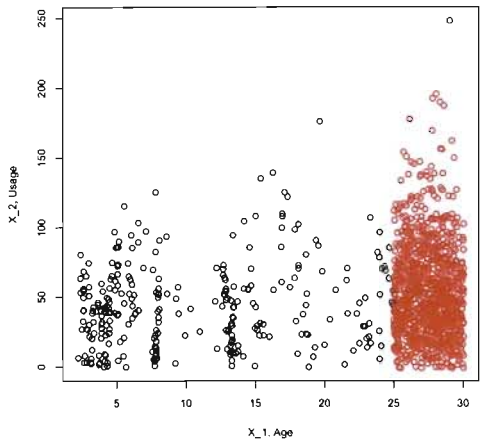
- Illustrations with $k = 1$ and $V(\boldsymbol{x}_{new}) = 1$



Example Introduction

- Response
  - Y = pass/fail of test coded as 1/0
- Two Covariates
  - X_1 = Age in years
  - X_2 = Usage in time in ready mode
- Goal
  - The observed X_1 values are between 2.23 and 24.97
  - Predict future reliability for $X_1 \in [25, 30]$
- Full model (probit model)
  - $Y \sim bernoulli(p)$
    - $V = p(1-p)$
  - $p(\boldsymbol{X}) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2)$
  - Number of models, $N_{mod} = 2^5 = 32$

Covariate Distribution of Interest



- Black circles depict the observed (X_1, X_2) pairs
- Red circles depict points sampled from the DI
- Increasing spread and a slightly positive trend

The Posterior cdf of Reliability for Model m I

- The prior distribution
  - Can use an informative prior if prior information is available
  - A conjugate prior is not available
  - We use the non-informative $g(\beta^m) \propto 1$ in the absence of prior information
- The posterior distribution of β^m is unavailable in closed form, so we use a Markov Chain Monte Carlo (MCMC) algorithm
- Let $(\beta^m)^{(1)}, (\beta^m)^{(2)}, \cdots, (\beta^m)^{(N_{MCMC})}$ be a sample from the posterior distribution of β^m

The Posterior cdf of Reliability for Model m II

- A sample from the posterior distribution of reliability for model m at $\boldsymbol{x}_{new}$ is then

$$\begin{aligned}(p^m(\boldsymbol{x}^m_{new}))^{(1)} &= \Phi(\boldsymbol{x}'_{new}(\beta^m)^{(1)})\\ (p^m(\boldsymbol{x}^m_{new}))^{(2)} &= \Phi(\boldsymbol{x}'_{new}(\beta^m)^{(2)})\\ &\vdots\\ (p^m(\boldsymbol{x}^m_{new}))^{(N_{MCMC})} &= \Phi(\boldsymbol{x}'_{new}(\beta^m)^{(N_{MCMC})})\end{aligned}$$

- Letting $(p^m(\boldsymbol{x}^m_{new}))^{[1]}$, $(p^m(\boldsymbol{x}^m_{new}))^{[2]}$, $\cdots$, $(p^m(\boldsymbol{x}^m_{new}))^{[N_{MCMC}]}$ be the ordered posterior draws of reliability,

$$\hat{F}^m_{\boldsymbol{x}^m_{new}}(u) = \sum_{i=1}^{N_{MCMC}} (p^m(\boldsymbol{x}^m_{new}))^{[i]} I(u \leqslant (p^m(\boldsymbol{x}^m_{new}))^{[i]})$$

approximates the posterior cdf of reliability

A Surrogate for $p(\boldsymbol{x}_{new})$

- Let $P(M = m|\boldsymbol{y})$ be the posterior probability of model m
- Let $\hat{p}^m(\boldsymbol{x}^m_{new})$ be a point prediction from the posterior distribution of $p^m(\boldsymbol{x}^m_{new})$
- The weighted average

$$\hat{p}(\boldsymbol{x}_{new}) = \sum_{i=1}^{N_{mod}} P(M = i|\boldsymbol{y})\hat{p}^m(\boldsymbol{x}^m_{new})$$

can be used as a surrogate for $p(\boldsymbol{x}_{new})$

- Options for calculating $P(M = m|\boldsymbol{y})$
  - The BIC approximation

$$P(M = m) \approx \frac{\exp\{\frac{-1}{2}\mathrm{BIC}_m\}}{\sum_{i=1}^{N_{mod}} \exp\{\frac{-1}{2}\mathrm{BIC}_i\}}$$

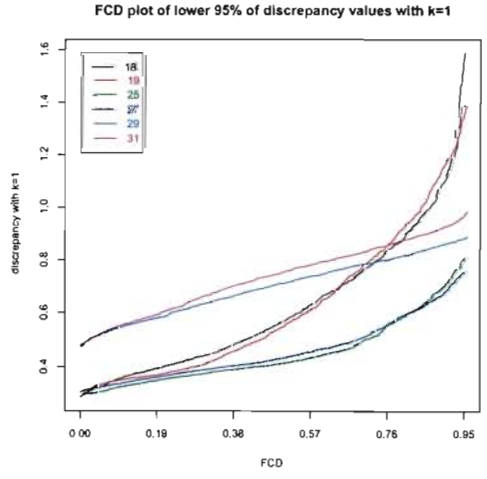
  - Use an MCMC algorithm
    - Carlin and Chib (1995)
    - Dellaportas et al. (1998)
    - Reversible Jump MCMC

Table of Summary Statistics

# terms	Model	95th Percentile value	95th Percentile rank	Mean value	Mean rank
0	(1)	1.73898	1(20)	0.9757	1(15)
1	X_1 (17)	1.73076	3(19)	0.7332	1(7)
1	X_2^2 (3)	1.77032	4(21)	1.01064	2(17)
2	X_1, X_2 (25)	0.80833	1(2)	0.46939	1(1)
2	X_1, X_2^2 (19)	1.39035	3(7)	0.67146	2(3)
2	X_1, X_1X_2 (18)	1.58853	6(13)	0.69645	3(4)
2	X_2, X_1^2 (13)	1.20248	2(5)	0.8725	4(11)
3	X_1, X_2, X_2^2 (27)	0.75447	1(1)	0.48818	1(2)
3	X_1, X_2, X_1^2 (29)	0.88409	2(3)	0.70358	2(5)
3	X_1, X_2^2, X_1X_2 (20)	1.49895	5(12)	0.71121	3(6)
3	X_1, X_2, X_1X_2 (26)	2.262	8(25)	0.79937	4(10)
4	X_1, X_2, X_1^2, X_2^2 (31)	0.98176	1(4)	0.76852	1(8)
4	X_1, X_2, X_2^2, X_1X_2 (28)	1.68413	2(17)	0.79353	2(9)
5	X_1, X_2, X_1^2, X_2^2, X_1X_2 (32)	1.8545	1(22)	1.02201	1(18)

- The values are summaries of $D_m^k(\boldsymbol{x}_{new})$ across the sample of points from the DI
- Local ranks (with the same number of terms): outside parentheses
- Global ranks (over all models): inside parentheses
- Ordered locally according to mean

FCD Plot



- Choose 25 (X_1 and X_2) as best

Results From Other Selection Procedures

# terms	model	DIC
3	X_1, X_2, X_1X_2 (26)	157.72
4	X_1, X_2, X_1^2, X_1X_2 (30)	159.67
2	X_1, X_2 (25)	159.84
4	X_1, X_2, X_2^2, X_1X_2 (28)	160.08
3	X_1, X_2, X_1^2 (29)	160.21
		Posterior Probability
2	X_1, X_2 (25)	0.51
3	X_1, X_2, X_1X_2 (26)	0.19
2	X_2, X_1^2 (13)	0.14
3	X_1, X_2, X_1^2 (29)	0.05
3	X_1, X_2, X_2^2 (27)	0.04

term	X_1	X_2	X_1^2	X_2^2	X_1X_2
Posterior Probability	0.824	0.998	0.229	0.064	0.229

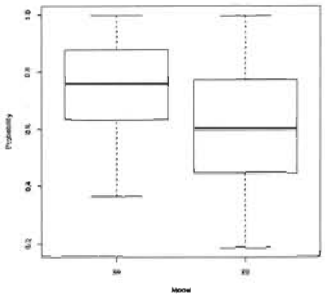
- Model 26 leads to the smallest DIC
- Model 25 has the highest posterior probability
- The median probability model is 25
- There are similarities and differences between the 4 selection algorithms

When Direction of Error Matters

- In the missile example, a prediction of reliability that is too high may be viewed as more costly than a prediction of reliability that is too low
  - A prediction that is too high can lead to too many malfunctions in the field
  - A prediction that is too low can lead to unnecessary maintenance expense
- Avoiding high (or low) predictions can be incorporated into the selection algorithm
- D_m^k treats discrepancy between the posterior cdf for model m and ideal cdf, below the true reliability, equal to the discrepancy above the true reliability
- To discourage high predictions, the discrepancy above the true reliability is penalized more harshly than the discrepancy below

Modification of D_m^k and Results

$$M_m^k(x_{new}) = \left\{ \int_{\mathbb{R}} a_c I[u \geq \hat{\mu}(x_{new})] \left| \frac{F_{x_{new}^m}^m(u) - F_{x_{new}}(u)}{\sqrt{V(x_{new})}} \right|^k + b_c I[u < \hat{\mu}(x_{new})] \left| \frac{F_{x_{new}^m}^m(u) - F_{x_{new}}(u)}{\sqrt{V(x_{new})}} \right|^k du \right\}^{1/k}$$

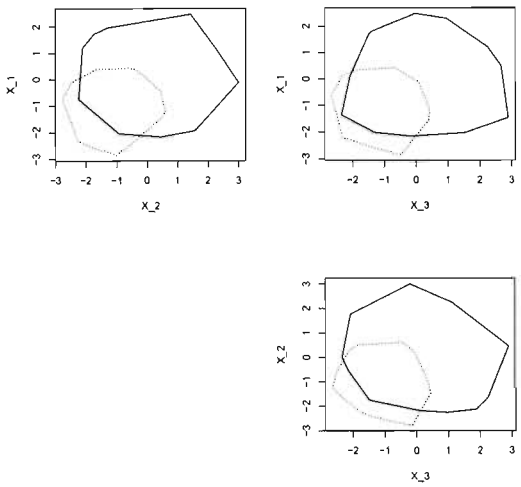


- Model 25 (X_1 and X_2) is highlighted as best when (a_c, b_c) = (1, 1) and (2, 1), and model 29 (X_1, X_2, and X_1^2) is highlighted as best when (a_c, b_c) = (3, 1)
- the boxplots are over the sampled points from the DI
- The largest predicted reliability for both models is about the same
- The median of the predicted reliabilities for model 29 is less than that of model 25

Overview

- 9 simulation scenarios in all
- Commonalities across scenarios
  - Three covariates, X_1, X_2, X_3
  - The full model includes the main effects, the two factor interactions, and the three factor interaction, so $2^7 = 128$ possible models
  - $N_o = 200$ observations
  - $N_{new} = 1,000$ sampled points from the DI
- Differences between scenarios
  - The data generating model (3 models)
    - $p(\boldsymbol{X}) = \Phi(0.75X_1)$
    - $p(\boldsymbol{X}) = \Phi(0.75X_1 - 0.2X_2)$
    - $p(\boldsymbol{X}) = \Phi(0.75X_1 - 0.2X_1X_2)$
  - The value of b_c (3 values)
    - $b_c = 1, 2, 3$

Observations and DIs



Procedure

- For a combination of b_c and a true model
  - Generate a data set from the true model and observed $\boldsymbol{X}$'s
  - Find the best model according to the new algorithm, the model with the highest posterior probability, the model with the lowest DIC, and the MPM
    - Find the best model according to our algorithm with a proxy to the graphical approach
  - Calculate $\hat{p}^m(\boldsymbol{x}^m_{new})$ at each $\boldsymbol{x}_{new}$ sampled from the DI for the selected models
  - Calculate the proportion of times that the new algorithm leads to a $\hat{p}^m(\boldsymbol{x}^m_{new})$ that is lower than $p(\boldsymbol{x}_{new})$, which is known
  - Calculate the proportion of times that the new algorithm leads to a $\hat{p}^m(\boldsymbol{x}^m_{new})$ that is lower than the $\hat{p}^m(\boldsymbol{x}^m_{new})$'s from the other selected models
  - Calculate $\frac{1}{N_{new}} \sum_{\boldsymbol{x}_{new}} [\hat{p}^m(\boldsymbol{x}^m_{new}) - p(\boldsymbol{x}_{new})]^2$ for all of the selected models

Results I

	PBMSM		HPPM		DIC		MPM	
	Simulation 1	Simulation 2	Simulation 1	Simulation 2	Simulation 1	Simulation 2	Simulation 1	Simulation 2
				$b_c = 1$				
Average	**0.0054**	**0.0052**	0.0059	0.0058	0.0094	0.0087	0.0058	0.0056
Median	**0.0032**	**0.0030**	**0.0032**	**0.0030**	0.0048	0.0044	0.0033	0.0031
90th Percentile	**0.0124**	**0.0118**	0.0139	0.0136	0.0240	0.0206	0.0136	0.0130
95th Percentile	**0.0174**	**0.0165**	0.0203	0.0200	0.0336	0.0305	0.0197	0.0191
				$b_c = 2$				
Average	**0.0063**	**0.0056**	0.0064	0.0060	0.0097	0.0089	**0.0063**	0.0057
Median	**0.0033**	**0.0030**	**0.0033**	0.0031	0.0048	0.0049	**0.0033**	**0.0030**
90th Percentile	0.0147	**0.0134**	0.0147	0.0146	0.0232	0.0202	**0.0146**	0.0135
95th Percentile	0.0220	**0.0189**	0.0223	0.0216	0.0336	0.0304	**0.0216**	0.0196
				$b_c = 3$				
Average	0.0062	0.0069	0.0058	0.0062	0.0086	0.0092	**0.0056**	**0.0061**
Median	**0.0031**	0.0034	**0.0031**	**0.0030**	0.0045	0.0045	0.0032	0.0031
90th Percentile	0.0146	0.0164	0.0132	0.0150	0.0211	0.0235	**0.0128**	**0.0144**
95th Percentile	0.0218	0.0256	0.0197	0.0223	0.0299	0.0351	**0.0191**	**0.0220**

Results II

	Simulation 1	Simulation 2
	$b_c = 1$	
Average	0.5110	0.5126
Median	0.5323	0.5303
90th Percentile	0.9947	0.9943
95th Percentile	0.9997	0.9993
	$b_c = 2$	
Average	0.4063	0.4368
Median	0.3193	0.4173
90th Percentile	0.9568	0.9590
95th Percentile	0.9947	0.9944
	$b_c = 3$	
Average	0.3980	0.3818
Median	0.3127	0.2730
90th Percentile	0.9433	0.9136
95th Percentile	0.9897	0.9745

Results III

	HPPM		DIC		MPM	
	Simulation 1	Simulation 2	Simulation 1	Simulation 2	Simulation 1	Simulation 2
$b_c = 1$						
Average	0.0870	0.0805	0.3504	0.3320	0.0845	0.0772
Median	0.0000	0.0000	0.1945	0.1922	0.0000	0.0000
90th Percentile	0.4402	0.3201	0.9529	0.9441	0.5210	0.2010
95th Percentile	0.6762	0.7722	0.9737	0.9641	0.6755	0.8599
$b_c = 2$						
Average	0.0356	0.0334	0.2544	0.2467	0.0295	0.0267
Median	0.0000	0.0000	0.0252	0.0272	0.0000	0.0000
90th Percentile	0.0537	0.0497	0.9230	0.9198	0.0497	0.0489
95th Percentile	0.1225	0.1041	0.9616	0.9608	0.0989	0.1007
$b_c = 3$						
Average	0.0260	0.0274	0.1535	0.1560	0.0231	0.0229
Median	0.0000	0.0000	0.0177	0.0207	0.0000	0.0000
90th Percentile	0.0620	0.0674	0.6819	0.6470	0.0625	0.0638
95th Percentile	0.0927	0.0990	0.8929	0.9088	0.0904	0.0954

Conclusion

- The focus of the new method is good prediction over a user-specified distribution of interest (DI) on the covariate space
- The DI should match the study goal
- Different models may be preferred over different DI's
- General four-step algorithm
  - Select the DI
  - Randomly sample points from the DI
  - Calculate the measure of prediction ability at each sampled location for all model under consideration
  - Compare models numerically and graphically based on the measures of prediction ability
- The measure of prediction ability can be modified to avoid high or low predictions
- The simulation study highlighted the strengths of the new methodology