On Bayesian predictive methods for high-dimensional covariate selection

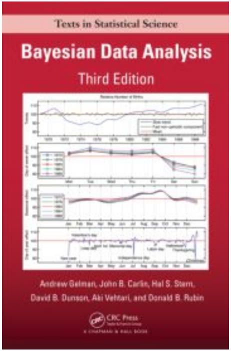
University of Warwick, 27 Feb 2014

Aki Vehtari

Aki.Vehtari@aalto.fi

http://becs.aalto.fi/~ave/

Department of Biomedical Engineering and Computational Science (BECS)
Aalto University


Texts in Statistical Science
Bayesian Data Analysis
Third Edition
Andrew Gelman, John B. Carlin, Hal S. Stern,
David B. Dunson, Aki Vehtari, and Donald B. Rubin
CRC Press

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari and Donald B. Rubin (2013). Bayesian Data Analysis, Third Edition. Chapman and Hall/CRC.

Outline

- Identifying disease risk factors
- Big data?
- Why model selection?
- Toy example
- Traffic speed cameras
- Selection induced bias
- Bayesian predictive model selection
- Reference predictive approaches

Identifying disease risk factors

- Motivation: “bioinformatics and medical applications”
- Predict risk of CVD, diabetes, cancers
  - biomarkers: lipids, growth hormones, etc.
  - genetic markers

Big data?

- We have
  - people in studies $n \sim 200 - 8000$
  - clinical covariates $p < 20$
  - biomarkers $p < 200$
  - genetic markers $p \sim 10 - 1e6$
  - survival models with latent linear, sparse linear or Gaussian process model

Why model selection?

- Assume a model rich enough capturing lot of uncertainties
  - e.g. Bayesian model average (BMA) or non-parametric
  - model criticism and predictive assessment done
  - → if we are happy with the model, no need for model selection
  - Box: “All models are wrong, but some are useful”
  - there are known unknowns and unknown unknowns
- Model selection
  - what if some smaller (or more sparse) or parametric model is practically as good?
  - which uncertainties can be ignored?
  - → reduced measurement cost, simpler to explain (e.g. less biomarkers, and easier to explain to doctors)

Previously

- University of Warwick, 2010, CRiSM Workshop: Model uncertainty and model selection
    - I talked about Bayesian predictive model selection

So many predictive papers

- Predictive model selection
- Predictive variable selection in generalized linear models
- A predictive model selection criterion
- A predictive approach to model selection
- Optimal predictive model selection
- Bayesian predictive model selection
- A Bayesian predictive approach to model selection
- A Bayesian predictive semiparametric approach to variable selection and model comparison in regression
- A generalized predictive criterion for model selection
- Some Bayesian predictive approaches for model selection
- Model determination using predictive distributions
- Model choice: A minimum posterior predictive loss approach
- etc.

Previously

- Aki Vehtari and Janne Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. In Statistics Surveys, 6:142-228.
- Andrew Gelman, Jessica Hwang and Aki Vehtari (2014). Understanding predictive information criteria for Bayesian models. Statistics and Computing, in press. Published online 20 August 2013.
- This talk is about how the reviewed methods behave in high dimensional cases

Toy example

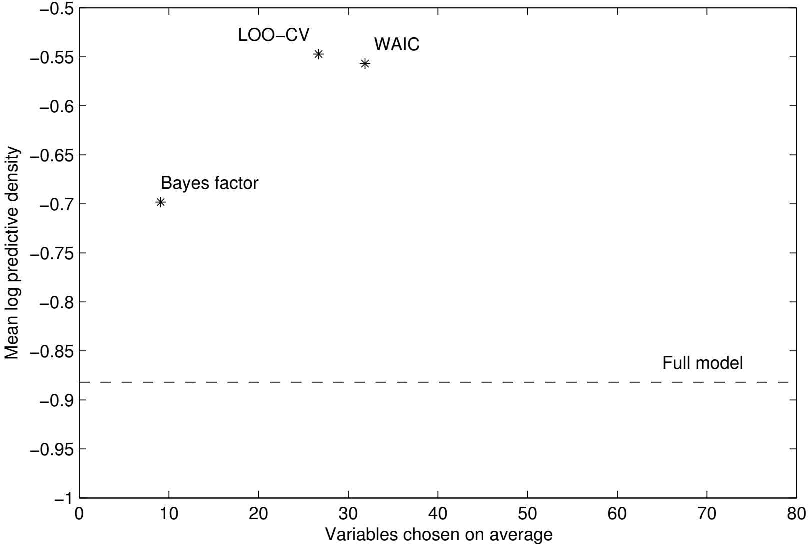
- Toy data with $n = 20$, 200 replications

$$\begin{aligned} z_1, z_2, z_3, z_4 &\sim \mathrm{U}(-1.73, 1.73) \\ x_{1,2,3,4} &\sim \mathrm{N}(z_1, .05^2) \\ x_{5,6,7,8} &\sim \mathrm{N}(z_2, .05^2) \\ x_{9,10,11,12} &\sim \mathrm{N}(z_3, .05^2) \\ x_{13,\ldots,100} &\sim \mathrm{N}(z_4, .05^2) \\ y &= z_1 + .5z_2 + .25z_3 + \epsilon \\ \epsilon &\sim N(0, 0.5^2), \end{aligned}$$

  that is, x's are noisy observations of z so that there are four groups of correlated covariates and 88 of the covariates have no effect on y

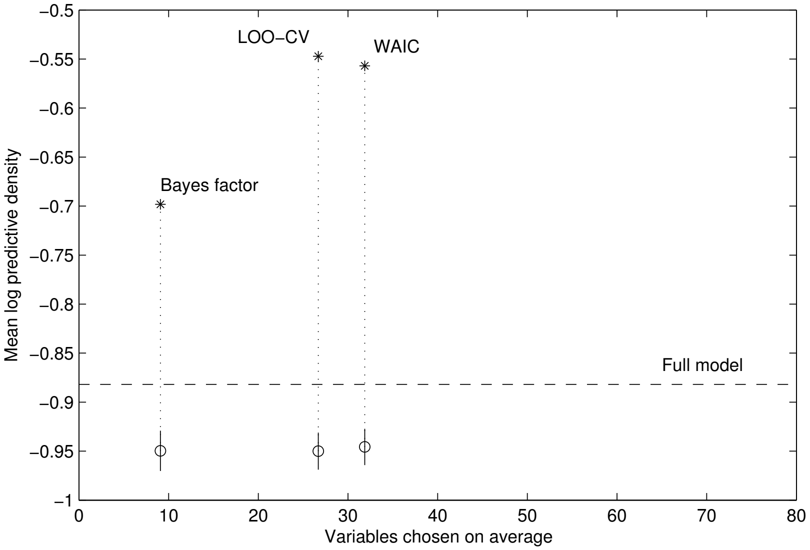
- Linear model with prior on weights

Toy example



LOO−CV
WAIC
Bayes factor
Full model
Mean log predictive density
Variables chosen on average
−0.5
−0.55
−0.6
−0.65
−0.7
−0.75
−0.8
−0.85
−0.9
−0.95
−1
0
10
20
30
40
50
60
70
80

Toy example



LOO−CV
WAIC
Bayes factor
Full model
Mean log predictive density
Variables chosen on average
−0.5
−0.55
−0.6
−0.65
−0.7
−0.75
−0.8
−0.85
−0.9
−0.95
−1
0
10
20
30
40
50
60
70
80

Traffic speed cameras

- In UK there have been a lot of discussion about the effectiveness of speed cameras to reduce the number of traffic accidents
- Illustration with dice
  - throw a bunch of dice
  - choose the dice showing six dots
  - re-throw the chosen dice
  - note that the re-thrown dice show the same or reduced number of dots

Selection induced bias

- Even if the original model performance estimate is unbiased (like bias corrected cross-validation) selecting a model with a better estimate can lead to overfitting

- Illustration with covariate selection
  - adding an irrelevant covariate to a model is like throwing a die
  - just by chance an irrelevant covariate can improve the model fit
  - if the model with best fit is chosen, it is likely that it does not fit so well the future data
  - the problem is not solved by penalising complexity as we have same phenomenon when comparing models with equal number of covariates

Selection induced bias

- Even if the original model performance estimate avoids double use of data (like cross-validation),
  the model selection step uses the data again
- We could use two-layer / nested cross-validation to obtain unbiased estimate for the effect of model selection
  - this does not fix the problem of getting worse predictions

Bayes factor

- Marginal likelihood in Bayes factor is also a predictive criterion
    - chain rule

$$p(y|M_k) = p(y_1|M_k)p(y_2|y_1, M_k), \ldots, p(y_n|y_1, \ldots, y_{n-1}, M_k)$$

How to avoid double use of data in model selection

- Decision theory helps

Predictive model

- $p(\tilde{y}|\tilde{x}, D, M_k)$ is the posterior predictive distribution
  - $p(\tilde{y}|\tilde{x}, D, M_k) = \int p(\tilde{y}|\tilde{x}, \theta, M_k) p(\theta|D, \tilde{x}, M_k) d\theta$
  - $\tilde{y}$ is a future observation
  - $\tilde{x}$ is a future random or controlled covariate value
  - $D = \{(x^{(i)}, y^{(i)}); i = 1, 2, \ldots, n\}$
  - M_k is a model
  - θ denotes parameters

Predictive performance

- Future outcome $\tilde{y}$ is unknown (ignoring $\tilde{x}$ in this slide)
- With a known true distribution $p_t(\tilde{y})$, the expected utility would be

$$\bar{u}(a) = \int p_t(\tilde{y}) u(a; \tilde{y}) d\tilde{y}$$

  where u is utility and a is action (in our case, a prediction)
- Bayes generalization utility

$$BU_g = \int p_t(\tilde{y}) \log p(\tilde{y}|D, M_k) d\tilde{y}$$

  where $a = p(\cdot|D, M_k)$ and $u(a; \tilde{y}) = \log(a(\tilde{y}))$
  - a is to report the whole predictive distribution
  - utility is the log-density evaluated at $\tilde{y}$

Bayesian predictive methods

- Many ways to approximate

$$BU_g = \int p_t(\tilde{y}) \log p(\tilde{y}|D, M_k) d\tilde{y}$$

  for example

  - Bayesian cross-validation
  - WAIC
  - reference predictive methods

- Many other Bayesian predictive methods estimating something else, e.g.,
  - DIC
  - *L*-criterion, posterior predictive criterion
  - projection methods
- See our survey for more methods

$\mathcal{M}$-open,-closed,-completed

- Following Bernardo & Smith (1994), there are three different approaches for dealing with the unknown p_t
  - $\mathcal{M}$-open
  - $\mathcal{M}$-closed
  - $\mathcal{M}$-completed

$\mathcal{M}$-open

- Explicit specification of $p_t(\tilde{y})$ is avoided by re-using the observed data D as a pseudo Monte Carlo samples from the distribution of future data
- For example, Bayes leave-one-out cross-validation

$$\mathrm{LOO} = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i | x_i, D_{-i}, M_k)$$

Cross-validation

- Bayes leave-one-out cross-validation

$$\mathrm{LOO} = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|x_i, D_{-i}, M_k)$$

  - different part of the data is used to update the posterior and assess the performance
  - almost unbiased estimate for a single model

$$\mathrm{E}[\mathrm{LOO}(n)] = \mathrm{E}[BU_g(n-1)]$$

expectation is taken over all the possible training sets

Selection induced bias

- Selection induced bias in LOO-CV
  - same data is used to assess the performance and make the selection
  - the selected model fits more to the data
  - the LOO-CV estimate for the selected model is biased
  - recognised already, e.g., by Stone (1974)
- Same holds for many other methods, e.g., DIC/WAIC
- Performance of the selection process itself can be assessed using two level cross-validation, but it does not help choosing better models
- Bigger problem if there is a large number of models as in covariate selection

$\mathcal{M}$-closed and $\mathcal{M}$-completed

- Explicit model for $p_t(\tilde{y})$
- $\mathcal{M}$-closed
    - possible to enumerate all possible model candidates $\{M_k\}_{k=1}^K$
    - belief that one of the candidate models is “true”
    - set a prior distribution $p(M_k)$ and compute $p_{\mathrm{BMA}}(\tilde{y}|D)$
- $\mathcal{M}$-completed
    - suitable when $\mathcal{M}$-closed can not be assumed
    - rich enough model M_* whose predictions are considered to best reflect the uncertainty in the prediction task

$\mathcal{M}$-closed and $\mathcal{M}$-completed

- Actual belief model M_*
  - a rich enough model, describing well the knowledge about the modeling problem and capturing the essential prior uncertainties
  - could be, for example
    - encompassing model
    - Bayesian model averaging model
    - flexible non-parametric model
  - the predictive distribution of the actual belief model $p(\tilde{y}|\tilde{x}, D, M_*)$ is a quantitatively coherent representation of our subjective beliefs about the unobserved future data

Reference model and reference predictive approach

- Reference model
  - a model used to asses the predictive performance of other models
  - natural choice is the actual belief model M_*
- Reference predictive approach
  - predictive model assessment using a reference model

Unknown $\tilde{x}$

- $\mathcal{M}$-open for both $p(\tilde{y}|\tilde{x})$ and $p(\tilde{x})$
- Reference model for both $p(\tilde{y}|\tilde{x})$ and $p(\tilde{x})$
- Reference model for $p(\tilde{y}|\tilde{x})$ and $\mathcal{M}$-open for $p(\tilde{x})$

see our survey for discussion about fixed and deterministic x

Reference predictive approach

- Reference model for both $p(\tilde{y}|\tilde{x}, D, M_*)$ and $p(\tilde{x}|D, M_*)$
  - good model for $\tilde{x}$ may often be difficult to construct
- Lindley (1968)
  - use of linear Gaussian model for $y|x$ and squared error cost function made computations simpler
  - only first moments of x were needed

Reference predictive approach

- Reference model for $p(\tilde{y}|\tilde{x})$ and simple $\mathcal{M}$-open for $p(\tilde{x})$

$$\bar{u} \approx \bar{u}_*(M_k) = \frac{1}{n}\sum_{i=1}^{n}\int \log p(\tilde{y}|\dot{x}_i, D, M_k) p(\tilde{y}|\dot{x}_i, D, M_*) d\tilde{y}$$

- San Martini & Spezzaferri (1984) used BMA model as the reference model

Reference predictive approach

- Reference model for $p(\tilde{y}|\tilde{x})$ and CV for $p(\tilde{x})$

$$\bar{u} \approx \bar{u}_*(M_k) = \frac{1}{n}\sum_{i=1}^{n}\int \log p(\tilde{y}|x_i, D_{-i}, M_k) p(\tilde{y}|x_i, D_{-i}, M_*) d\tilde{y}$$

  - better assessment of the out-of-sample predictive performance

Reference predictive approach

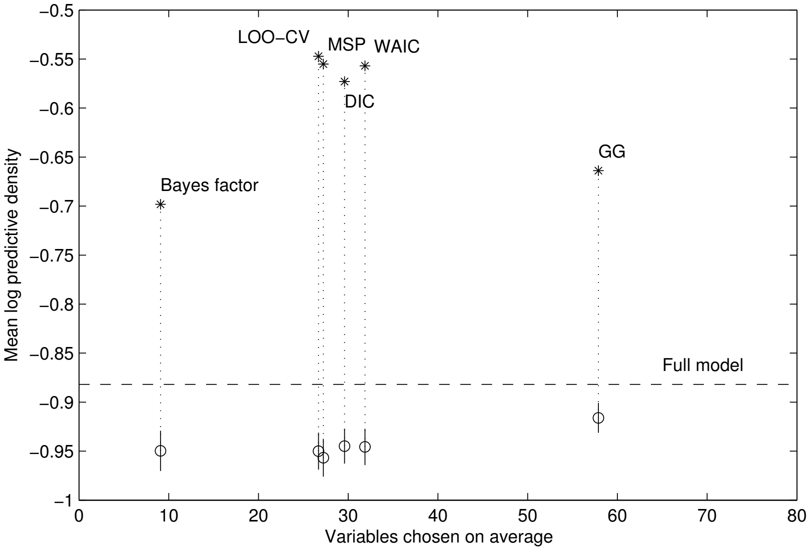
- Reference predictive model selection using log-score corresponds to minimizing the KL-divergence from the reference predictive distr. to the submodel predictive distr.
  - divergence is minimized by the reference model itself
  - requires additional cost term or calibration of acceptable divergence from the reference
  - no selection induced bias, since data has been used only once to create the reference model, and selection process fits towards the reference model
  - bias depends on the reference model and is generally unknown
  - variance is reduced as model is used for $p(\tilde{y})$ instead of n pseudo Monte carlo samples
  - reduced variance helps discriminating good models from the others

Experimental results

Methods compared

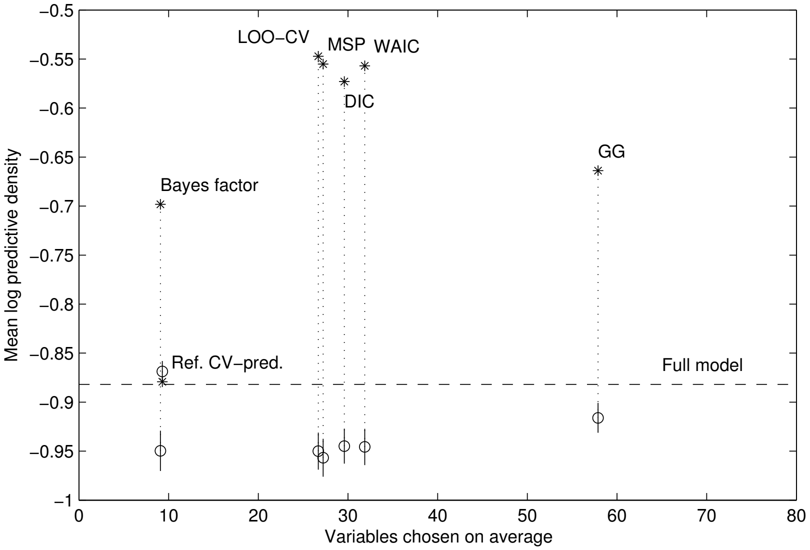
- reference predictive for $y|x$ (1% inform. loss)
- reference predictive for $y|x$ + CV for x (1% inform. loss)
- LOO-CV
- DIC (Spiegelhalter et al, 2002)
- WAIC (watanabe 2010)
- posterior predictive loss = GG (Gelfand & Ghosh, 1998)
- cross-validation predictive loss = MSP (Marriot et al, 2001)
- Bayes factor

Toy example



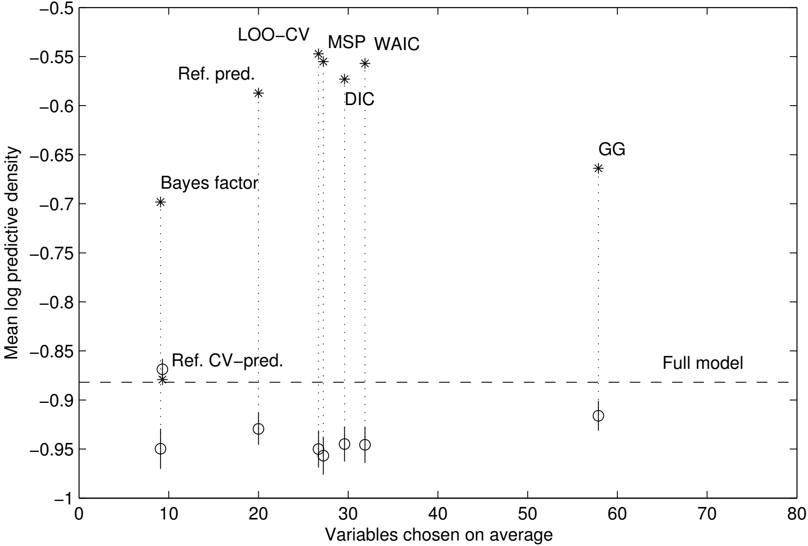
LOO−CV
MSP
WAIC
DIC
GG
Bayes factor
Full model
Mean log predictive density
−0.5
−0.55
−0.6
−0.65
−0.7
−0.75
−0.8
−0.85
−0.9
−0.95
−1
0
10
20
30
40
50
60
70
80
Variables chosen on average

Toy example



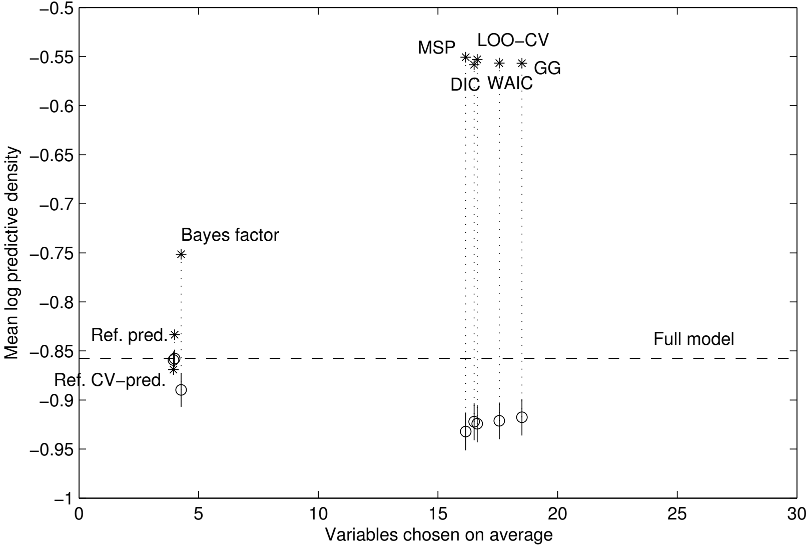
LOO−CV
MSP
WAIC
DIC
GG
Bayes factor
Ref. CV−pred.
Full model
Mean log predictive density
Variables chosen on average
−0.5
−0.55
−0.6
−0.65
−0.7
−0.75
−0.8
−0.85
−0.9
−0.95
−1
0
10
20
30
40
50
60
70
80

Toy example

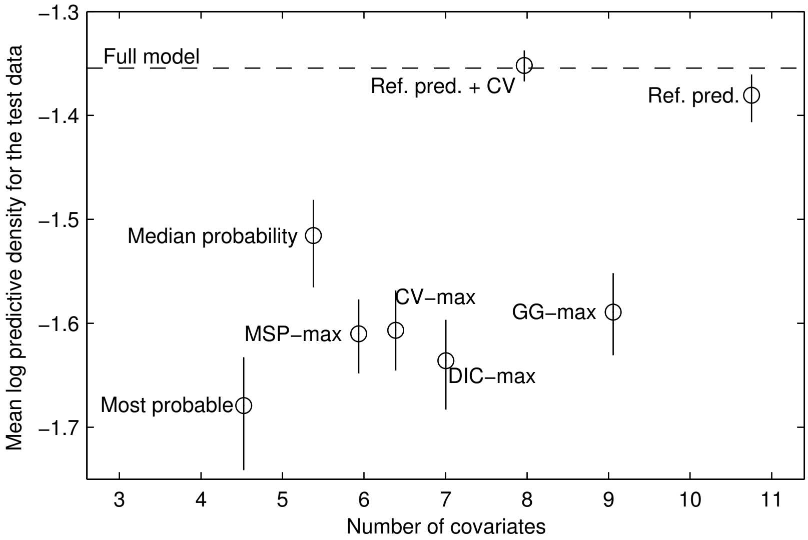


Mean log predictive density
−0.5
−0.55
−0.6
−0.65
−0.7
−0.75
−0.8
−0.85
−0.9
−0.95
−1
0
10
20
30
40
50
60
70
80
Variables chosen on average
LOO−CV
MSP
WAIC
Ref. pred.
DIC
GG
Bayes factor
Ref. CV−pred.
Full model

Toy example - alternative prior



Toy example - noisier small data



Computational issues

- The presented reference approach requires computation of KL divergences of the posterior predictive densities
  - generally no closed form equation
  - quadrature can be used for pointwise predictions
  - some models are easy, like binary classification
- Traversing model space
  - forward selection
  - branch and bound
  - greedy selection
  - stochastic search

Conclusions

- Selection induced bias is a problem when there are many models (e.g. in covariate selection)
- Reference predictive approach with CV for x avoids selection induced bias

Extras

- Gibbs score
- Projection methods