# Bayes for model search and representing uncertainty

presented at

"Building Statistical Methodology and Theory 2014"

In honor of Jeff Wu's 65th birthday

**Hugh Chipman**

Acadia University

July 8, 2014

# Outline

Motivating example: Glucose data

Principles to guide model search

Bayesian subset selection

Illustration with glucose example

Comments

Closing remarks

# Outline

Motivating example: Glucose data

Principles to guide model search

Bayesian subset selection

Illustration with glucose example

Comments

Closing remarks

# Motivating Example

Blood Glucose Experiment

| design | | | | | | | | mean |
|---|---|---|---|---|---|---|---|---|
| A | G | B | C | D | E | F | H | reading |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 97.94 |
| 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 83.40 |
| 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 95.88 |
| 1 | 2 | 1 | 1 | 2 | 2 | 3 | 3 | 88.86 |
| 1 | 2 | 2 | 2 | 3 | 3 | 1 | 1 | 106.58 |
| 1 | 2 | 3 | 3 | 1 | 1 | 2 | 2 | 89.57 |
| 1 | 3 | 1 | 2 | 1 | 3 | 2 | 3 | 91.98 |
| 1 | 3 | 2 | 3 | 2 | 1 | 3 | 1 | 98.41 |
| 1 | 3 | 3 | 1 | 3 | 2 | 1 | 2 | 87.56 |
| 2 | 1 | 1 | 3 | 3 | 2 | 2 | 1 | 88.11 |
| 2 | 1 | 2 | 1 | 1 | 3 | 3 | 2 | 83.81 |
| 2 | 1 | 3 | 2 | 2 | 1 | 1 | 3 | 98.27 |
| 2 | 2 | 1 | 2 | 3 | 1 | 3 | 2 | 115.52 |
| 2 | 2 | 2 | 3 | 1 | 2 | 1 | 3 | 94.89 |
| 2 | 2 | 3 | 1 | 2 | 3 | 2 | 1 | 94.70 |
| 2 | 3 | 1 | 3 | 2 | 3 | 1 | 2 | 121.62 |
| 2 | 3 | 2 | 1 | 3 | 1 | 2 | 3 | 93.86 |
| 2 | 3 | 3 | 2 | 1 | 2 | 3 | 1 | 96.10 |

Analysis based on linear model.

Design features:

- 18 runs
- *A* discrete, *B* - *H* continuous
- Some continuous settings unevenly spaced.
- Complex aliasing ⇒ interactions and polynomial terms can be considered.

# Blood Glucose Example

**What model terms?**

- Standard: $A, B, B^2, \ldots H, H^2$ (15 terms)
- Interactions: $AB, AB^2, \ldots G^2H^2$ (98 terms)
- Total: 113 terms
- There are $\sum_{i=0}^{17} \binom{113}{i} = 7.65 \times 10^{19}$ possible models.

With so many possible terms and only 18 runs, assumptions will need to be made.

# Outline

Motivating example: Glucose data

Principles to guide model search

Bayesian subset selection

Illustration with glucose example

Comments

Closing remarks

# What are reasonable assumptions about the space of models?

Hamada & Wu (1992), Wu and Hamada book (2000):

- Effect hierarchy: main effects more likely than interactions.
- Effect sparsity: only a few effects are important.
- Effect heredity*: when a two-factor interaction is active, at least one corresponding main effect should be active.

(with extensions to polynomials and polynomial interactions)

Hamada and Wu (1992) used these principles to motivate a stepwise model search algorithm.

* name suggested by Randy Sitter

# Hamada-Wu (1992) search

Stepwise search algorithm, described with main effects and 2fi's:

1. Select significant effects from main effects and 2fi's orthogonal to main effects.
2. Search over effects from step 1 and 2fi's with at least one active main effect in 1.
3. Search with forward stepwise over main effects and interactions related to those identified in 2.
4. Steps 2 & 3 repeated to convergence.

- Search employs "weak heredity": an interaction can enter with one corresponding main effect, e.g. $A, AB$ active, but $B$ inactive.
- More thorough search is also proposed as an alternative.

# Is the search good enough?

- Hamada-Wu stepwise search explores only a small subset of models permitted under heredity.
- H-W can miss important terms in some circumstances (e.g. $Y = A + 2AB + 2AC + \varepsilon$, larger interactions than main effects).
- But conventional all-subsets searches do not respect the three principles.
- What we really want is a thorough search.

# Outline

Motivating example: Glucose data

Principles to guide model search

Bayesian subset selection

Illustration with glucose example

Comments

Closing remarks

# Bayesian Model Search

Chipman (1996) and Chipman Hamada and Wu (1997) develop a Bayesian formulation that:

- Incorporates hierarchy, sparsity, and heredity in the prior distributions
- Uses MCMC for stochastic search ("SSVS", George and McCulloch 1993)
- Quantifies model uncertainty via posterior distribution on models.

# Model Specification:

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I)$$

Additional parameter vector $\delta$ specifies which terms are included in the model.

Example: $$\begin{array}{cccccc} A & B & C & AB & AC & BC \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \delta_A & \delta_B & \delta_C & \delta_{AB} & \delta_{AC} & \delta_{BC} \end{array}$$

Each $\delta$ element is 0 or 1

- $\{\delta_A = 0\} \Rightarrow$ A not in model
- $\{\delta_A = 1\} \Rightarrow$ A in model

$\delta = (1\ 0\ 0\ 1\ 0\ 0) \Leftrightarrow$ model has terms $A$ and $AB$ only.

# What prior for the model (i.e. $\delta$)?

Independent Bernoullis: $\pi(\delta) = \prod_{i=1}^{p} p_i^{\delta_i}(1-p_i)^{1-\delta_i}$

This violates heredity; instead use conditional structure:

$$P(\delta_{AB} = 1 | \delta_A, \delta_B) = \begin{cases} p_{00} & \text{if } (\delta_A, \delta_B) = (0,0) \\ p_{01} & \text{if } (\delta_A, \delta_B) = (0,1) \\ p_{10} & \text{if } (\delta_A, \delta_B) = (1,0) \\ p_{11} & \text{if } (\delta_A, \delta_B) = (1,1) \end{cases}$$

- Weak heredity: $(p_{00}, p_{01}, p_{10}, p_{11}) = (0, 0.10, 0.10, 0.25)$
- Strong heredity: $(p_{00}, p_{01}, p_{10}, p_{11}) = (0, 0, 0, 0.25)$
- Relaxed (weak/strong) heredity: change 0's to 0.01's.
- Ideas generalize to higher order terms and extend to categorical predictors with $\geq 3$ levels ("effect grouping").

# Example prior calculation

Consider a simple example with 5 main effects ($A ... E$),
5 quadratics ($A^2 ... E^2$), 10 2fi's ($AB, ..., DE$):

Prior probability of inclusion:

- 0.25 for main effects
- (0.01, 0.25) for quadratics
- (0.01, 0.10, 0.10, 0.25) for interactions

$$
\begin{array}{llll}
\multicolumn{4}{l}{\Pr(A, B, C, D, E) =} \\
= & (.25^5) & \times(.75^5) & \times(.75^{10}) \\
 & (A ... E \text{ active }) & (A^2 ... E^2 \text{ inactive }) & (AB ... DE \text{ inactive }) \\
= & .000013 & &
\end{array}
$$

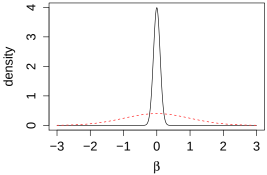
# Prior on $\beta, \sigma$:

Prior factored as $\pi(\beta, \sigma, \delta) = \pi(\beta, \sigma | \delta)\pi(\delta)$.

Various priors possible, here we use George & McCulloch, 93/97

$$\nu\lambda/\sigma^2 \sim \chi^2_\nu$$

$$\beta_i | \delta_i \sim \begin{cases} N(0, \tau_i^2) & \text{if } \delta_i = 0 \\ N(0, (c_i\tau_i)^2) & \text{if } \delta_i = 1 \end{cases}$$



where $c_i > 1$

- Posteriors obtained by the Gibbs sampler (stochastic search)
- Important variant: conjugate priors, enabling $\beta, \sigma$ to be analytically integrated out of the posterior.
    - Enables evaluation of $\Pr(\delta | Y)$ up to a normalizing constant.

# MCMC model search

## Example assuming strong heredity:

| $\delta$ values | | | | | |
|---|---|---|---|---|---|
| $A$ | $B$ | $C$ | $AB$ | $AC$ | $BC$ |
| $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ |
| $\delta_A$ | $\delta_B$ | $\delta_C$ | $\delta_{AB}$ | $\delta_{AC}$ | $\delta_{BC}$ |
| 1 | 1 | 0 | 1 | 0 | 0 |
| — | — | — | — | — | — |

- Gibbs sampler updates $\delta$ vector one element at a time, Bernoulli draws.
- Update for $\delta_A$ will depend on value of $\delta_{AB}$.
- Similarly $\delta_{AB}$ depends on $\delta_A, \delta_B$.
- Gibbs is a stochastic stepwise search algorithm.

# MCMC model search

## Example assuming strong heredity:

| $\delta$ values | | | | | |
|---|---|---|---|---|---|
| $A$ | $B$ | $C$ | $AB$ | $AC$ | $BC$ |
| $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ |
| $\delta_A$ | $\delta_B$ | $\delta_C$ | $\delta_{AB}$ | $\delta_{AC}$ | $\delta_{BC}$ |
| 1 | 1 | 0 | 1 | 0 | 0 |
| $\underline{1}$ | _ | _ | _ | _ | _ |

- Gibbs sampler updates $\delta$ vector one element at a time, Bernoulli draws.
- Update for $\delta_A$ will depend on value of $\delta_{AB}$.
- Similarly $\delta_{AB}$ depends on $\delta_A, \delta_B$.
- Gibbs is a stochastic stepwise search algorithm.

# MCMC model search

## Example assuming strong heredity:

| $\delta$ values | | | | | |
|---|---|---|---|---|---|
| $A$ | $B$ | $C$ | $AB$ | $AC$ | $BC$ |
| $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ |
| $\delta_A$ | $\delta_B$ | $\delta_C$ | $\delta_{AB}$ | $\delta_{AC}$ | $\delta_{BC}$ |
| 1 | 1 | 0 | 1 | 0 | 0 |
| $\underline{1}$ | $\underline{1}$ | _ | _ | _ | _ |

- Gibbs sampler updates $\delta$ vector one element at a time, Bernoulli draws.
- Update for $\delta_A$ will depend on value of $\delta_{AB}$.
- Similarly $\delta_{AB}$ depends on $\delta_A, \delta_B$.
- Gibbs is a stochastic stepwise search algorithm.

# MCMC model search

## Example assuming strong heredity:

| $\delta$ values | | | | | |
|---|---|---|---|---|---|
| $A$ | $B$ | $C$ | $AB$ | $AC$ | $BC$ |
| $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ |
| $\delta_A$ | $\delta_B$ | $\delta_C$ | $\delta_{AB}$ | $\delta_{AC}$ | $\delta_{BC}$ |
| 1 | 1 | 0 | 1 | 0 | 0 |
| $\underline{1}$ | $\underline{1}$ | $\underline{1}$ | _ | _ | _ |

- Gibbs sampler updates $\delta$ vector one element at a time, Bernoulli draws.
- Update for $\delta_A$ will depend on value of $\delta_{AB}$.
- Similarly $\delta_{AB}$ depends on $\delta_A, \delta_B$.
- Gibbs is a stochastic stepwise search algorithm.

# MCMC model search

## Example assuming strong heredity:

| $\delta$ values | | | | | |
|---|---|---|---|---|---|
| $A$ | $B$ | $C$ | $AB$ | $AC$ | $BC$ |
| $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ |
| $\delta_A$ | $\delta_B$ | $\delta_C$ | $\delta_{AB}$ | $\delta_{AC}$ | $\delta_{BC}$ |
| 1 | 1 | 0 | 1 | 0 | 0 |
| $\underline{1}$ | $\underline{1}$ | $\underline{1}$ | $\underline{0}$ | _ | _ |

- Gibbs sampler updates $\delta$ vector one element at a time, Bernoulli draws.
- Update for $\delta_A$ will depend on value of $\delta_{AB}$.
- Similarly $\delta_{AB}$ depends on $\delta_A, \delta_B$.
- Gibbs is a stochastic stepwise search algorithm.

# Outline

Motivating example: Glucose data

Principles to guide model search

Bayesian subset selection

Illustration with glucose example

Comments

Closing remarks

# Priors for glucose example:

$\pi(\delta)$: (relaxed weak heredity)

2 parents:

$(p_{00}, p_{01}, p_{10}, p_{11}) = (0.01, 0.10, 0.10, 0.25)$

1 parent:

$(p_0, p_1) = (0.01, 0.25)$

0 parents:

$p = 0.25$

$\pi(\sigma)$:

use $S_y/5$ as guess for $\mathrm{E}(\sigma)$

put $99^{th}$ quantile near $S_y$.

$S_y = 10.06$ gives $(\nu, \lambda) = (2, 1.29)$.

# Glucose example - posterior

Results: Most probable models

| model | prob | $R^2$ |
|---|---|---|
| $BH^2, B^2H^2$ | 0.183 | 0.7696 |
| $B, BH^2, B^2H^2$ | 0.080 | 0.8548 |
| $B, BH, BH^2, B^2H^2$ | 0.015 | 0.8601 |
| $F, BH^2, B^2H^2$ | 0.014 | 0.7943 |
| $GE, BH^2, B^2H^2$ | 0.013 | 0.8771 |
| $AH^2, BH^2, B^2H^2$ | 0.009 | 0.8528 |
| $G^2D, BH^2, B^2H^2$ | 0.009 | 0.8517 |
| $A, BH^2, B^2H^2$ | 0.008 | 0.7938 |

- Marginal probabilities also available: $\Pr(B) = .33, \Pr(BH^2) = .927, \Pr(B^2H^2) = .907$
- Changing prior 0.01's to 0.0's (relaxed weak heredity → weak heredity) makes the model $B, BH, BH^2B^2H^2$ most probable.
- With independence priors, most probable model has mass $\approx 0.0003$

# Parametrization and variable selection

Comment:

- In this case, products and powers of B (volume), H (dilution) seem most important.
  - Suggests that in fact “amount of material” may really be the important factor.
- Be careful to ensure the right parametrization.
  - (related to sliding factors - Hamada and Wu 1995, Cheng, Wu and Huwang (2006))
- Variable selection priors concentrate prior mass on $\beta$ values near the axes (i.e., some elements 0).

# Parametrization and variable selection, continued

Related issue: Why strong heredity may be desirable:

- Peixoto (1990): strong heredity guarantees selection of same terms under linear transformations of predictors (e.g., $A \rightarrow (A - 1.2)$ and $A^2 \rightarrow (A - 1.2)^2$).

# Outline

Motivating example: Glucose data

Principles to guide model search

Bayesian subset selection

Illustration with glucose example

Comments

Closing remarks

# Value of a Posterior Distribution on Models

What can you do with a posterior on models?

- Pick most probable model, knowing how much (or little) support it has.
- Incorporate model uncertainty in “downstream” decisions.

# Value of a Posterior Distribution on Models

“Downstream” decision example: robust parameter design optimization (Shoemaker, Tsui, Wu 1991; Tan and Wu 2013):

- Model response as a function of control and noise factors (“Response model approach”)
- Assume distribution for noise factors, giving response mean and variance functions as “performance measures”.
- Posterior on parameters $(\beta, \sigma)$ and models leads to uncertainty of performance measures (Chipman 1997).
- Accounting for uncertainty can change effectiveness of different adjustment variables.

# Priors as penalty functions

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

$$\log(\text{Posterior}) \propto \log(\text{Likelihood}) + \log(\text{Prior})$$

Posterior is like a penalized likelihood, with prior = penalty.

Example: 5 variables (A, B, C, D, E)
Full second order model (20 terms):
$A, B, C, D, E, A^2, B^2, C^2, D^2, E^2$
$AB, AC, AD, AE, BC, BD, BE, CD, CE, DE$
Probability of inclusion:
0.25 for main effects
(0.01, 0.25) for quadratics
(0.01, 0.10, 0.10, 0.25) for interactions

# Priors as penalty functions

Probability of inclusion:

0.25 for main effects

(0.01, 0.25) for quadratics

(0.01, 0.10, 0.10, 0.25) for interactions

$$
\begin{array}{llll}
\multicolumn{4}{l}{\Pr(A, B, C, D, E) =} \\
= & (.25^5) & (.75^5) & (.75^{10}) \\
 & (A ... E) & (A^2 ... E^2) & (AB ... DE) \\
= & .000013 & & \\
\multicolumn{4}{l}{\Pr(A, B, A^2, B^2, AB)} \\
= & (.25^2)(.75^3) & (.25^2)(.99^3) & (.25)(.9^6)(.99^3) \\
 & (A ... E) & (A^2 ... E^2) & (AB ... DE) \\
= & .000206 & &
\end{array}
$$

$$
\frac{\Pr(A, B, A^2, B^2, AB)}{\Pr(A, B, C, D, E)} = \frac{.000206}{.000013} = 15.79
$$

The main-effect-only model is less probable than a polynomial model in 2 factors!

# Other uses of heredity principles & Bayes:

Three principles and/or Bayes formulation can be used in a variety of “regression” contexts:

- Other designs: screening designs, hard-to-control factors, supersaturated designs
- Other responses: binary, ordinal, censored, Poisson, circular, ...
- Part of overall framework (Wu and Hamada book).
- Tan and Wu (2013) and Goh and Bingham (2014) extended the SSVS idea with heredity to split plot experiments and robust design experiments.
    - Different approaches to search - G&B utilize MCMC, T&W develop a stochastic search that exploits ability to evaluate marginal posterior $\Pr(\delta|Y)$.

# Other uses of heredity principles & Bayes:

## Model Selection in Design

- Construction and analysis of 3-level designs incorporating the 3 principles:
  - Cheng & Wu 2001 strategy of selection, projection, fitting interactions.
  - Xu, Cheng & Wu 2004 for optimal design.
- Design for model discrimination: Meyer, Steinberg and Box (1996), Bingham and Chipman (2007).
  - Average of design criterion over a prior placed on models, or over a posterior (for a followup design).

# Isn't variable selection old-fashioned?

What about the Lasso? Or many other "modern" sparse regression methods? Isn't model selection old-fashioned these days?

- Principles have been incorporated into Lasso (Yuan, Joseph & Lin 2007), Garrotte (Yuan, Joseph & Zou 2009).
- L1 and other penalized regression methods are solving a somewhat different problem: Selection of one model, without quantification of uncertainty.

# Closing remarks

- Uncertainty quantification is central to statistics.
- In industrial settings, scarce data and/or complex models often lead to statistical uncertainty.
- Model uncertainty can be easily overlooked.
- “UQ” is central to computer experiments.
- Hallmark of Jeff’s research is the appropriate quantification of uncertainty, combined with efficient and imaginative algorithms to design and analyze statistical studies.

**Thank you**

# Beyond Regression

Model search and uncertainty in other models:

- Ensemble models

  $$y = f_1(x) + f_2(x) + \dots + f_m(x) + \varepsilon$$

  Where each $f_j$ is a decision tree model, with its own set of parameters.
- Similar tools for quantifying uncertainty as in regression:
  - regression coefficient ⇔ terminal node output
  - model uncertainty ⇔ uncertainty in tree structure.
  - MCMC used to compute the posterior.
- Uncertainty in $f$'s translates to uncertainty in the functional form of response.
- Sequential design or simply uncertainty quantification.