Simulated Annealing Model Search (SAMS) for Subset Selection in Screening Experiments

Mark Wolters
Statistical and Actuarial Sciences
University of Western Ontario

Derek Bingham
Statistics and Actuarial Science
Simon Fraser University

ASQ Fall Technical Conference

October 14, 2011

Outline

- Screening Experiments and Nonregular Designs
- Example: Plackett-Burman designs
- Model Selection from Ensembles of Oversized Models
  - Raster plot
  - Link plot
- SAMS for Generating Model Sets
- Demonstrations
- Conclusions

Screening Experiments and Nonregular Designs

Screening Experiments

- **Goal of screening:** select the few important predictors from the many available ones.
  - Small number of runs
  - Many variables
  - Not very interested in inference or prediction

- Designs of interest: ***nonregular factorial designs***.
  - regular designs ➡ Either have no aliasing, or complete aliasing.
  - nonregular designs ➡ Exhibit ***complex aliasing***.

- **Benefit:** can consider main effects and interactions.

- **Cost:** huge model set, model selection problem.

Principles of Analysis

- **Effect sparsity:** Only a few effects are important.
  - Justifies the use of a screening experiment (few runs).

- **Effect heredity:** If interactions are active, then main effects should be active too.
  - Different ways to specify.
  - Heredity specification defines “interpretable” models.
  - Usual default: ***weak heredity***. AB can only be active if A or B are active too.

- **Effect hierarchy:** Higher-order effects are less likely to be active than lower-order ones.

Model Selection is Difficult in Screening Experiments

- Key aspects of problem:
  - Small n, large k
  - Huge model set
  - Complex aliasing
  - Heredity requirement

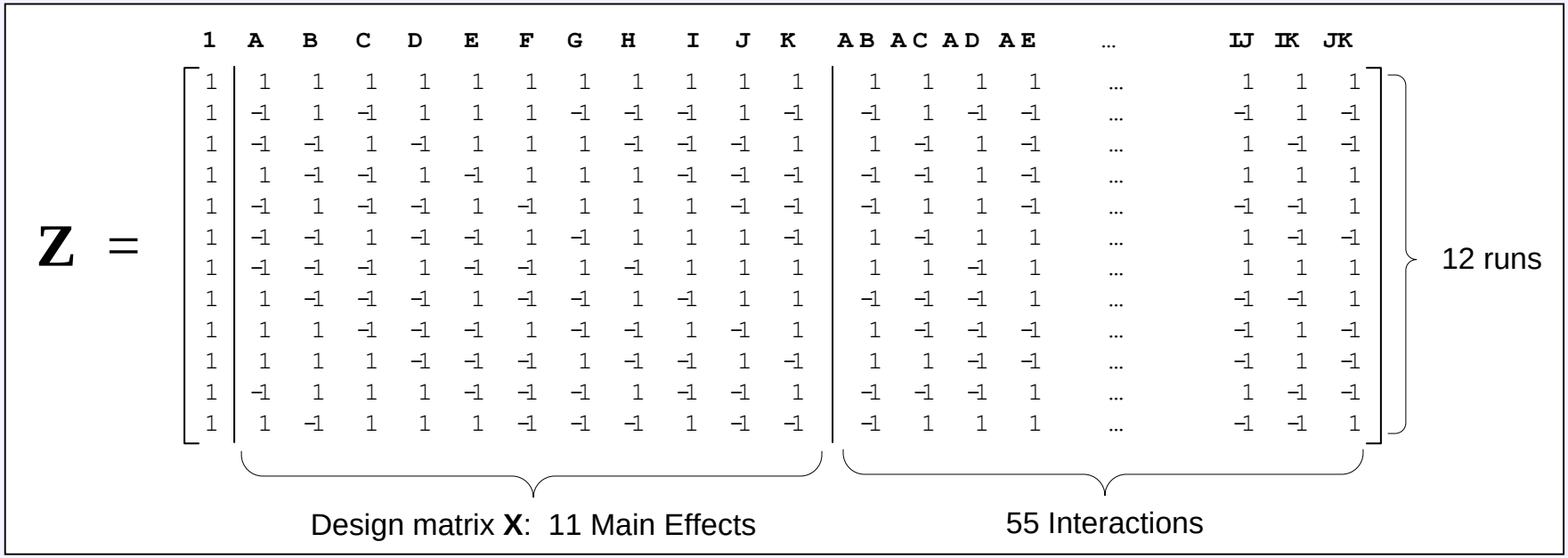
- Why is model selection difficult in this situation?
  - High ***model selection uncertainty*** (aka ***model aliasing***).
  - Exhaustive search may not be possible.
  - Strong tendency toward overfitting.
  - Common search methods don't respect heredity.

- Opinion: model aliasing is not given enough attention in this context. Usually multiple competing models should be identified.

Example: Plackett-Burman Designs

Plackett-Burman Designs

- Design matrix is orthogonal array
- Add 2-way interactions, lose orthogonality
- Smallest cases: 12-run and 20-run
  - PB_{12}: 11 main effects, 55 interactions k = 66 variables
  - PB_{20}: 19 main effects, 171 interactions k = 190 variables



Plackett-Burman Designs

- With interactions included, ***finding active variables becomes a model selection problem***.

p	PB_{12}		PB_{20}	
	All Models	**Hereditary**	**All Models**	**Hereditary**
1	66	11	190	19
2	2145	165	1.796×10^{4}	513
3	4.576×10^{4}	1705	1.125×10^{6}	9861
4	7.207×10^{5}	1.551×10^{4}	5.260×10^{7}	1.705×10^{5}
5	8.937×10^{6}	1.252×10^{5}	1.957×10^{9}	2.680×10^{6}
6	9.086×10^{7}	9.026×10^{5}	6.033×10^{10}	3.857×10^{7}
7	7.788×10^{8}	5.894×10^{6}	1.586×10^{12}	5.147×10^{8}

Summary of Available Methods

	Pros	Cons
Stepwise and variants	-Easy to use	-No heredity -Greedy searches -Not theoretically sound -Returns 1 model
Model selection criteria	-Optimality in some sense	-No heredity -Multiple criteria -Doesn't adjust for size of model set
Bayesian Methods[1]	-Handles heredity -Returns distribution of models	-Sensitive to priors -Interpretation of output
Groupwise LARS[2]	-Handles heredity	-Returns 1 model -Need stopping rule

[1]George and McCullogh 1993; Chipman, Hamada, and Wu 1997 [2]Yuan, Joseph, and Lin 2007

Model Selection from Ensembles of Oversized Models

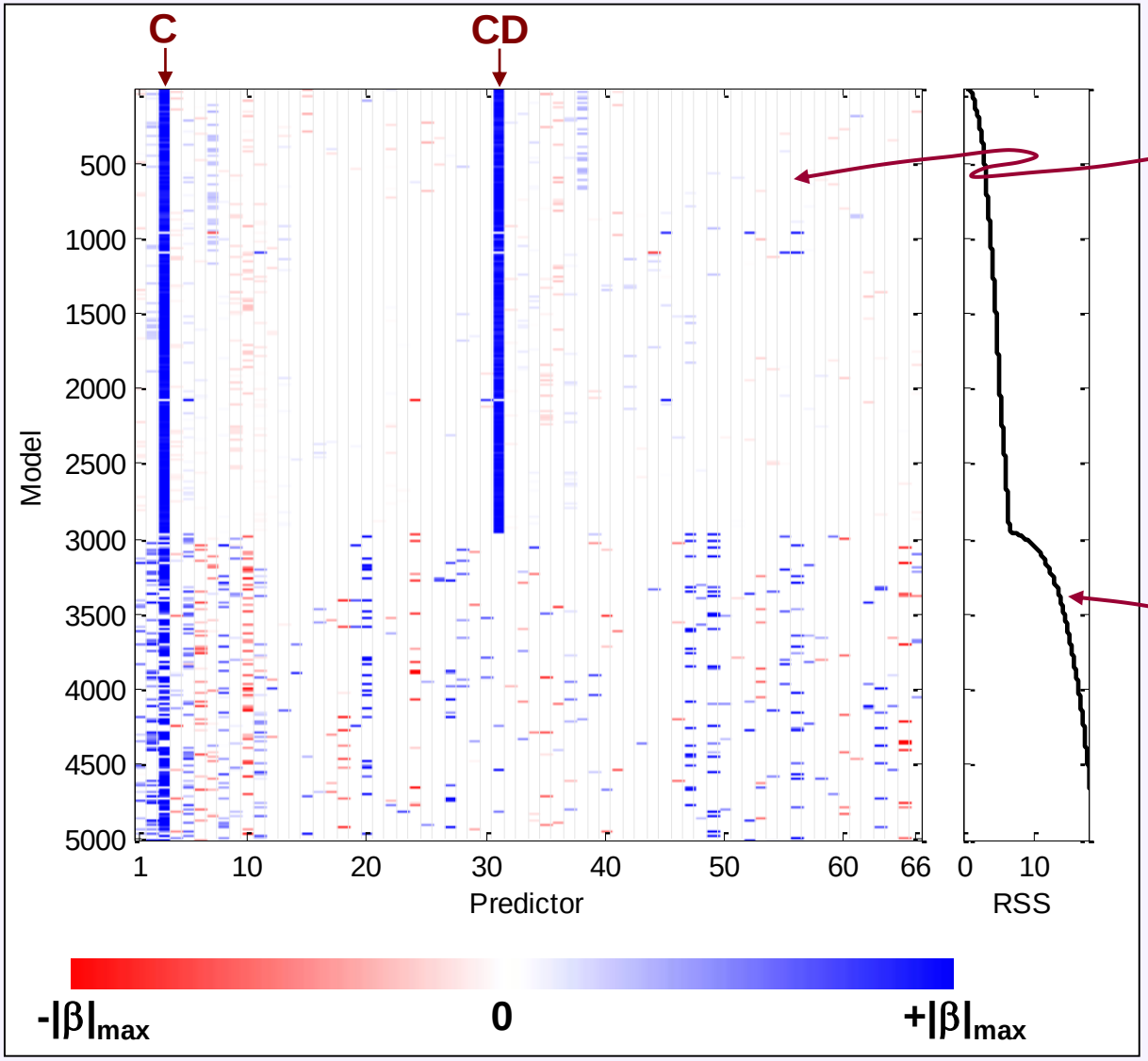
Small-Scale Example

Goal: introduce the idea and the key graphs.

Avoid the model search problem for now to illustrate ideas.

- PB_{12} design
- Generate $Y = 1 + 2C + 1.5CD + \varepsilon, \quad \varepsilon \sim N(0,1)$
- Do exhaustive search of models size p = 5 (125202 models)
- Sort all models by residual sum of squares (RSS).
- Expect most of the top models to contain (C, CD)
  - There are 2904 such models
- Plot coefficients of top models… can the truth be detected?

Raster Plot



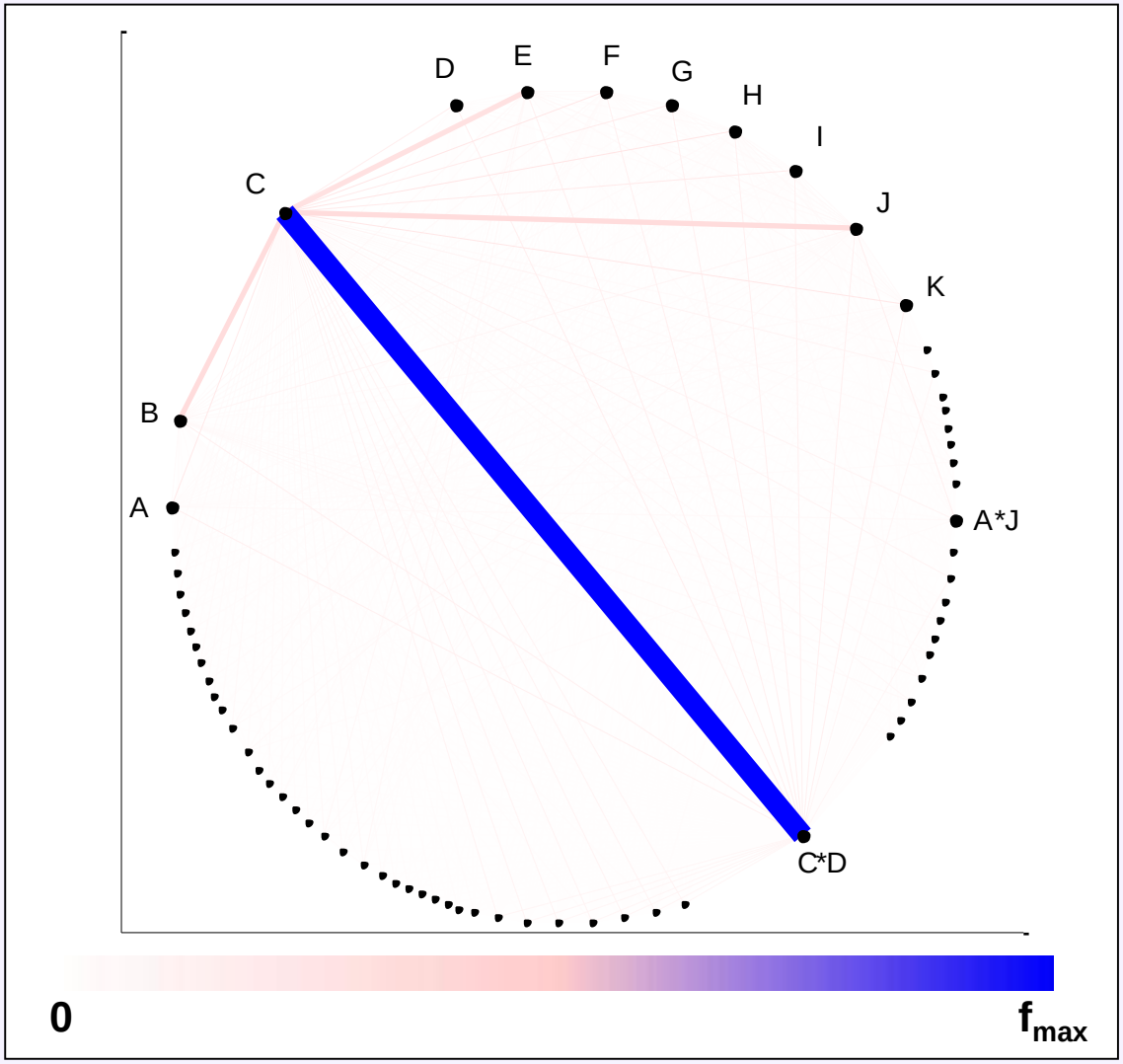
-Coefficients of top 5000 models represented as a matrix.

-Each model plotted as a row of pixels.

-Colors scaled to coefficient sign and magnitude.

-Also plot RSS of models

Link Plot



-Each effect represented by a point.

-Each pair of variables represented by a line.

-Line color and thickness represents occurrence frequency in the good-model set.

-Color and thickness scaled relative to most frequent pair.

-More frequent variables have points larger and more spaced out.

SAMS for Generating Model Sets

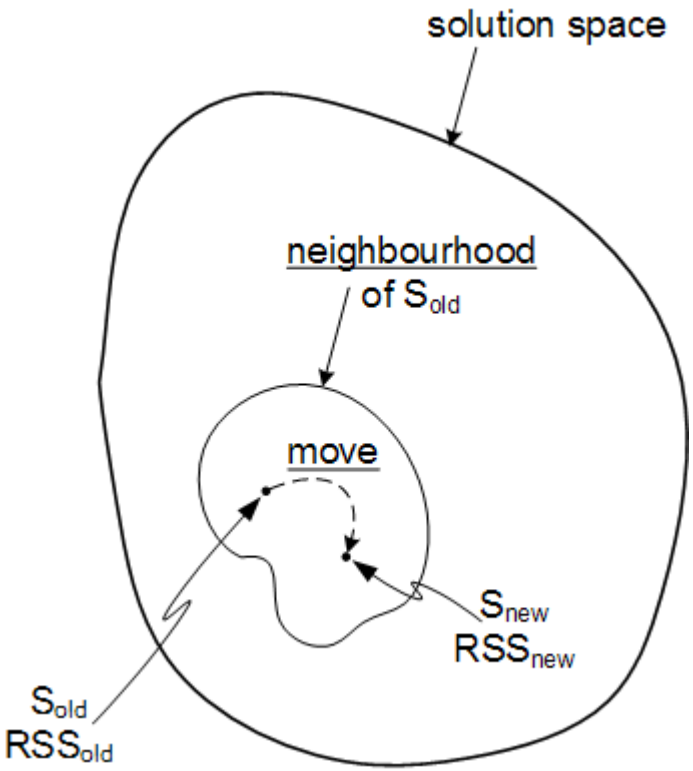
Summary of the Method

Search for lots of good, oversized models, then extract most common variable combinations from them.

1. Choose maximum truth size, s_{max}, and large model size, $p > s_{max}$.
2. Use a search heuristic to get 5000-10000 well-fitting models of size p.
   - Use RSS as goodness measure.
3. Use plots to visualize the model set and extract good small models.

The search heuristic used is based on simulated annealing, with two important changes.

Simulated Annealing



$$\delta = RSS_{new} - RSS_{old}$$

$\delta \leq 0$: accept move

$\delta > 0$: accept move w/ prob. $exp(-\delta/T)$

T = temperature (control parameter)

Reduce T on a schedule to cause convergence.

Problems:

- No heredity considerations
- Convergent search

SA Innovation 1: Hereditary Move

Hereditary move

1. Randomly choose one variable in M_{old}.
2. Drop the chosen variable.
3. Drop any other variables that no longer respect heredity.
4. Build the model back up to size p:
   a) Make a list of admissible variables.
   b) Randomly choose one and add to the model.
   c) repeat a) and b) until size = p.

Move	Model			
1	*A*	*B*	*C*	*BD*
2	*A*	*B*	*AB*	*BD*
3	*A*	*C*	*AB*	*AE*
4	*A*	*C*	*AJ*	*AE*
5	*A*	*AK*	*AJ*	*AE*
	D	*H*	*J*	*DG*

SA Innovation 2: Adaptive Temperature Control

- ***Premise: reduce T on every accepted move, increase T on every rejected move***
- Improving moves always accepted; this rule influences uphill moves.
  - Accept: set $T := \rho T$ $(0 < \rho < 1)$ ➡ makes it harder to accept bad moves
  - Reject: set $T := \alpha T$ $(\rho < \alpha < 1)$ ➡ makes it easier to accept bad moves
- Control by setting ρ and κ, where $\rho/\alpha^{\kappa} = 1$
  - Call κ the "search depth" – approx. κ rejections per acceptance.
  - Larger κ, more thorough local search.
- In the code, also put a lower bound on P(accept). Call this P_{min}.
- **Main idea**: this control makes the search ***non-convergent***. Keep generating models indefinitely.

Simulated Annealing Model Search

- Call the modified algorithm SAMS.
- Run until n_{gen} models are accepted.
- Default parameters: $\rho = 0.95$, $\kappa = 4$, $P_{min}=0.01$, $n_{gen}=10000$.
  - Performs well across wide range of problems.
- Characteristics:
  - Will generate good models indefinitely
  - May re-visit models
  - Fast (~20sec for $n_{gen}= 10000$ with PB_{20})

Demonstrations

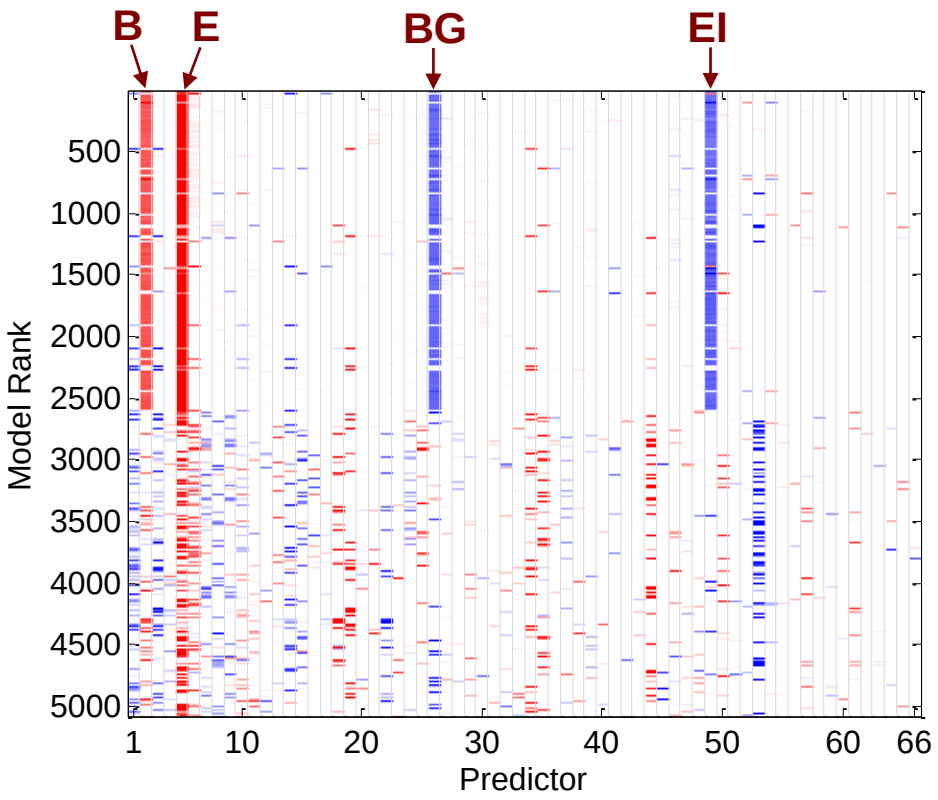
DEMO 1: Another PB_{12} Example

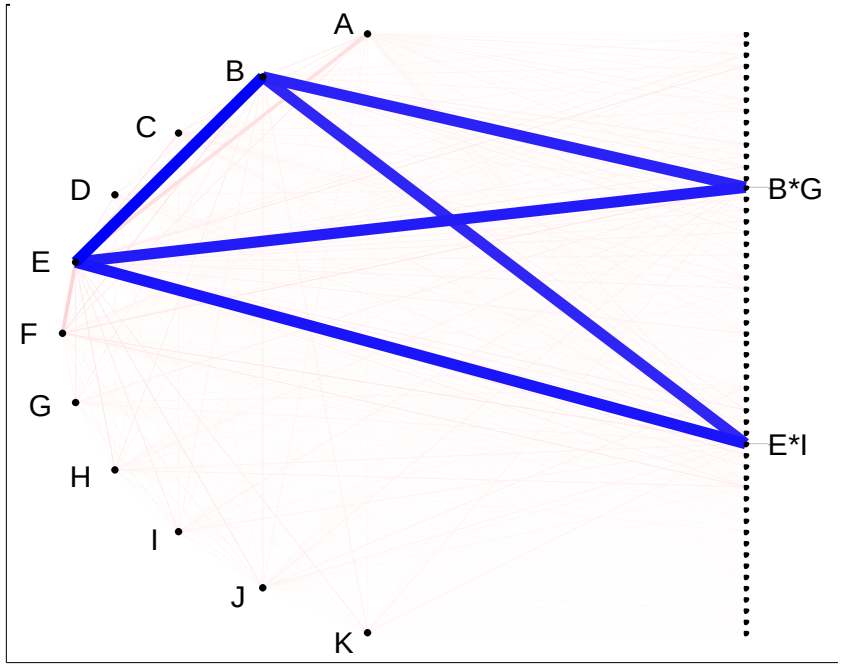
- Choose true model $Y = 0.5 + 1.1\mathbf{B} + 1.5\mathbf{E} + 0.9\mathbf{BG} + 1\mathbf{EI} + \varepsilon$.
  - Response 1: $\varepsilon \sim N(0,0.25)$ ➡ Low model selection uncertainty
  - Response 2: $\varepsilon \sim N(0,0.75)$ ➡ High model selection uncertainty

- Set $\tau = 4$, $p = 7$.
- Set parameters: $\rho = 0.95$, $\kappa = 4$, $P_{min}=0.001$, $n_{gen}=10000$

- Run SAMS and view raster and link plots.

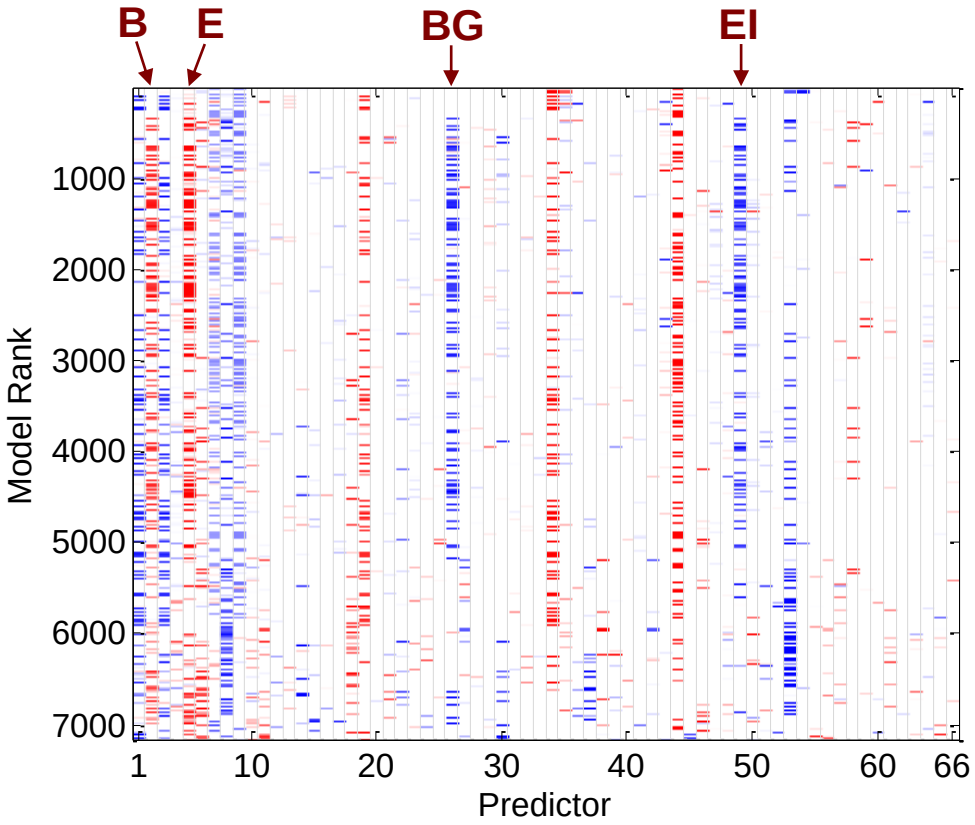
Low Model Selection Uncertainty

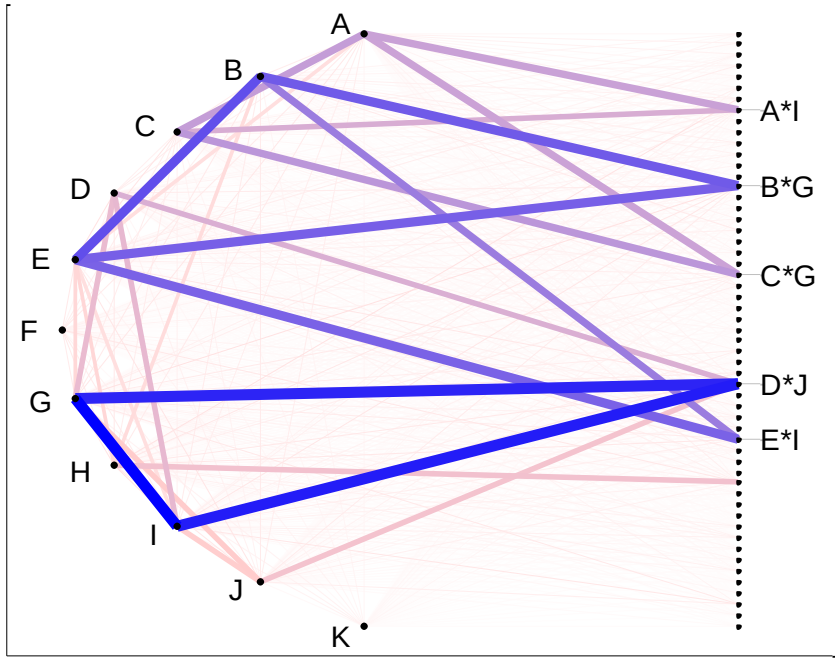




Can optionally plot interactions on a line to emphasize heredity relationships

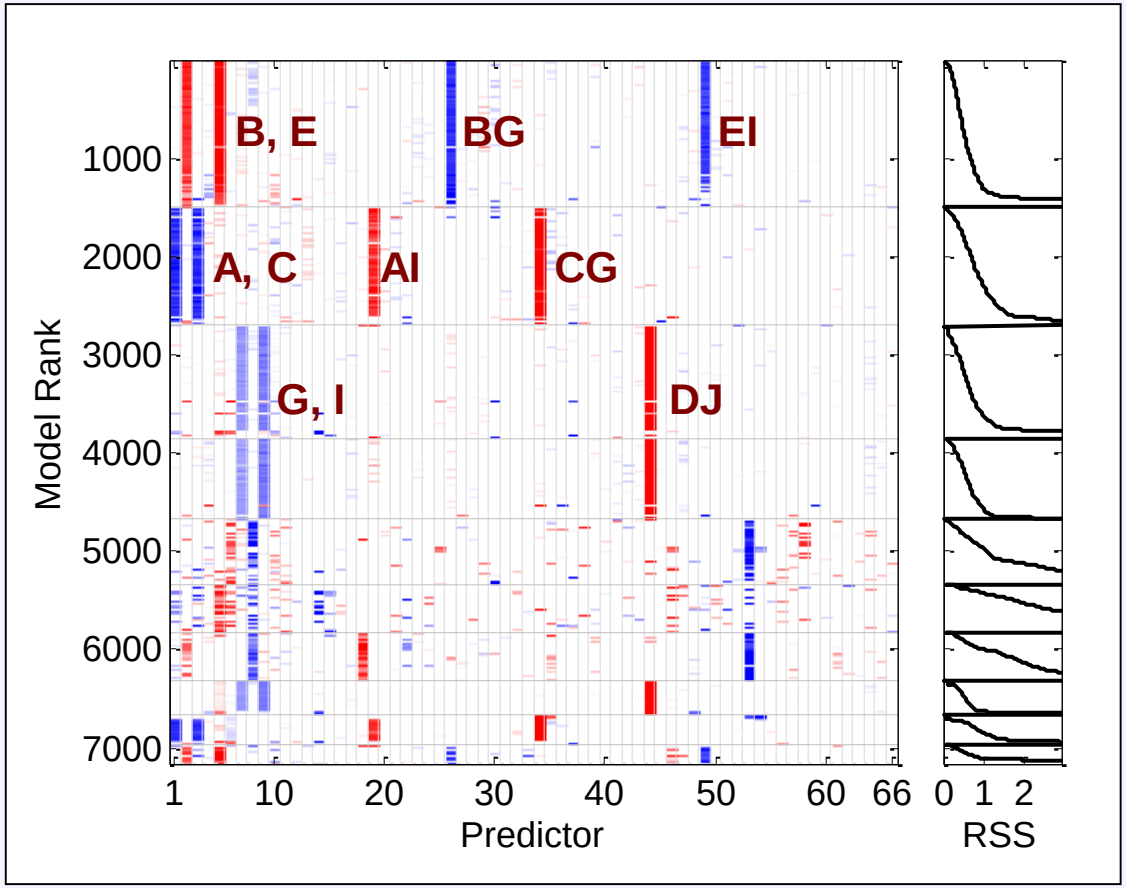
High Model Selection Uncertainty





Clustered Raster Plot

- Perform K-means clustering on rows to make common models more visible:

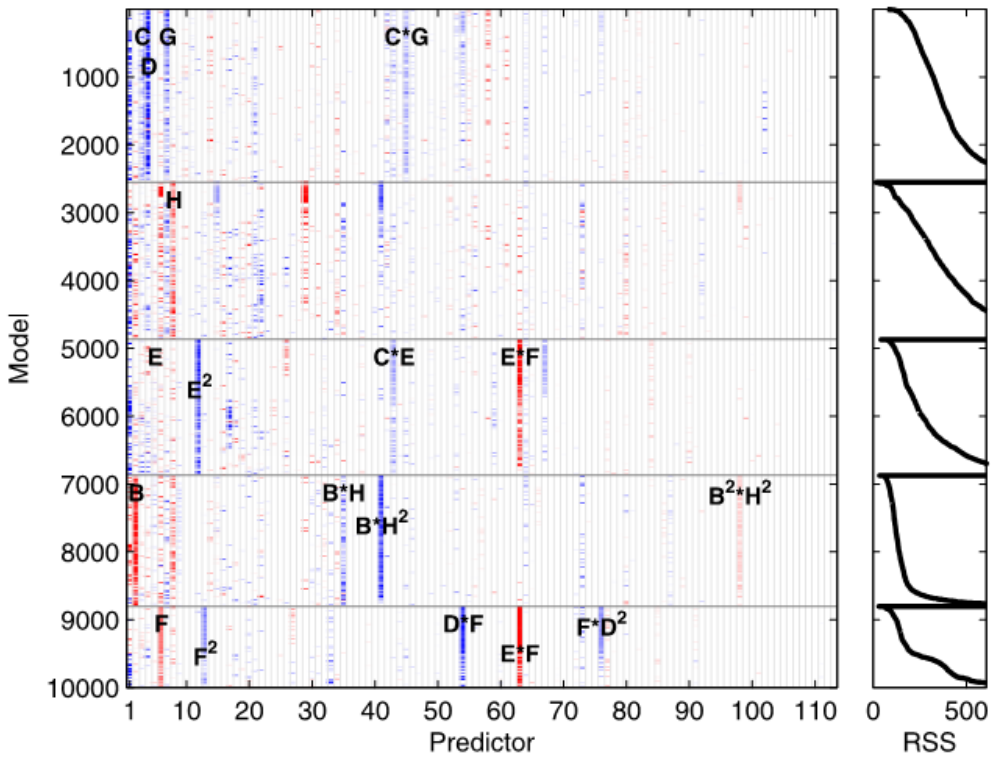


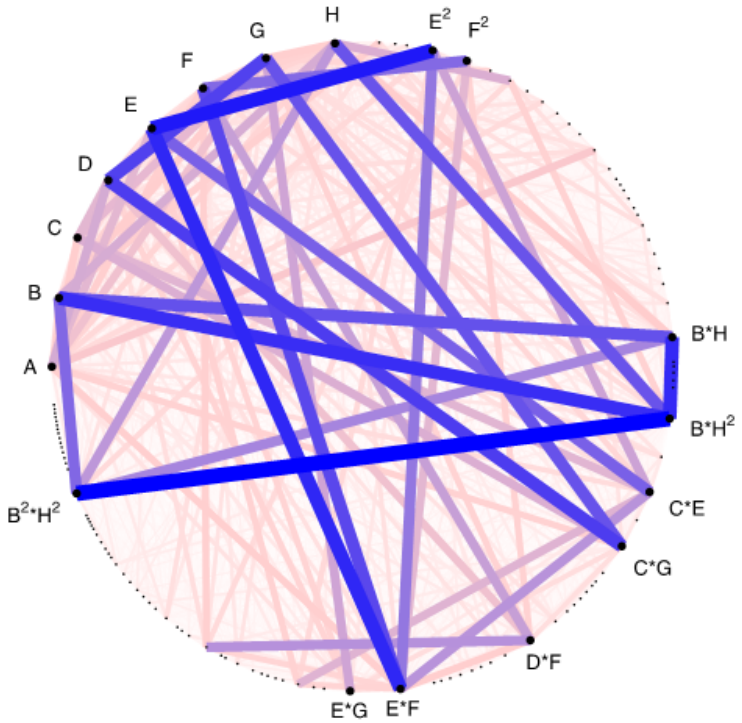
DEMO 2: The Blood Glucose Study

- Mixed 2- and 3-level design, **18 runs**.
- Factor A: 2 levels.
- Factors B-H: 3 levels. Use linear & quadratic terms.
- Include all valid 2-way interactions: **113 factorial effects**.

- More complicated heredity specification due to quadratic terms, e.g.
  - B^2C has parents BC and B^2

DEMO 2: The Blood Glucose Study





- SAMS results agree with previous studies
- Model selection uncertainty particularly clear with SAMS
- Only graphical analysis required.

DEMO 3: The Ozone Data

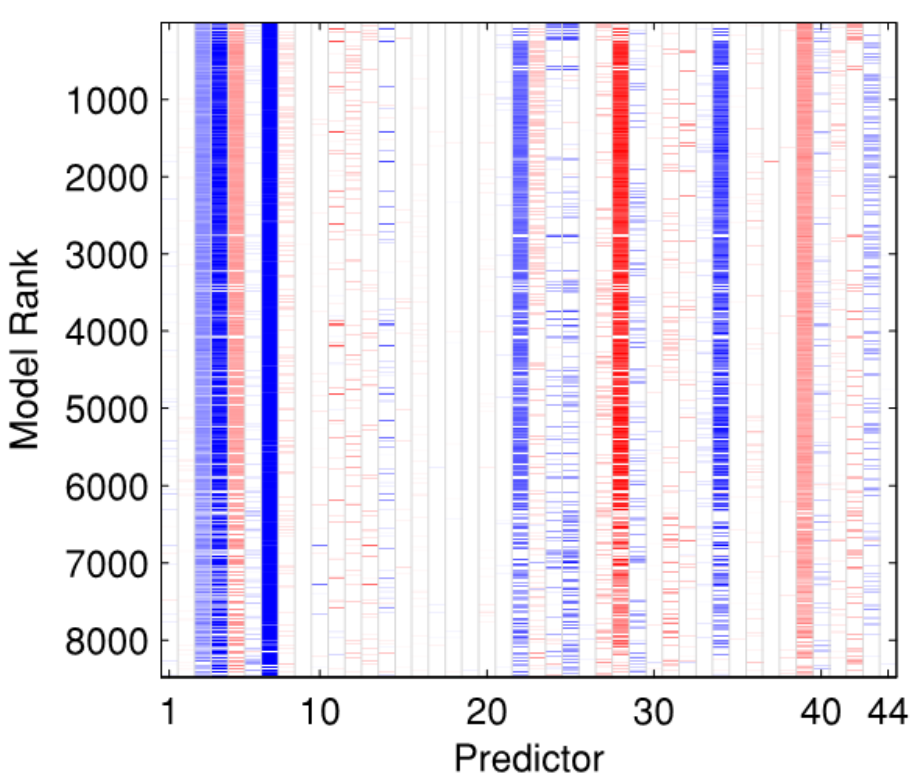
- Response: ozone concentration (**n = 330**).
- Predictors: eight meteorological variables.
- Include all interactions and squared terms (**k =44**).
- Observational data ➡ this is a ***regression problem***.

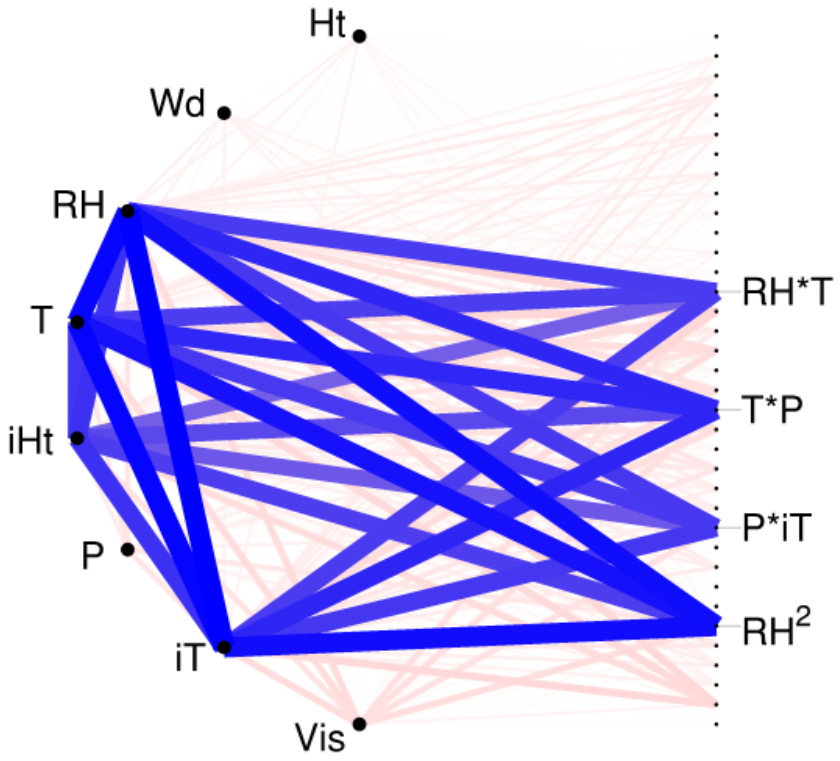
- Large n, could consider larger models.
- Can our methodology choose important variable combinations?

Variable	Quantity	Label
Y	ozone concentration (ppm)	–
X_1	500 millibar height (m)	Ht
X_2	wind speed (mph)	Wd
X_3	relative humidity (%)	RH
X_4	surface temperature (°F)	T
X_5	inversion height (ft)	iHt
X_6	pressure gradient (mmHg)	P
X_7	inversion temperature (°F)	iT
X_8	visibility (mi)	V

DEMO 3: The Ozone Data

- Run SAMS with 13-variable models.
- A particular 8-variable model stands out.





Conclusions

Conclusions

- Summary
  - Developed a heredity-respecting, non-convergent search heuristic
  - Developed graphical displays for model selection
  - Can find good small models from an ensemble of good large ones

- Advantages
  - Good performance
  - Easy to use, graphical approach
  - Heredity built in
  - Reduces the connection between GOF and choice of model size

- ***The SAMS code is available as a supplement to the paper.***

Future Work

- More investigation of regression problems
  - Effect of collinearity, variable scaling, etc. on performance.
- Handle a wider range of heredity specifications
  - Currently each effect can have any other effect(s) as parents.
  - No strong heredity
  - Can’t handle grouped variables (if A is in model, B must be as well)
  - Can handle exclusions (if A is in model, B must not be)
- Data-driven choice of m (size of ensemble of good models)

Supporting Slides

Auto-Extraction of Best-Guess Model

- Problem: can't use graphs in a simulation study
- Steps:

1. Find the 5 most frequent models of each size 1,…,τ in the good model set (use ***branch and bound*** method)
2. Calculate the entropy criterion for each.

➡ For a model with a main effects and b interactions:

$$\begin{aligned}
\text{Occurrence frequency of M in good set} &= f \\
\text{Population count of models containing M} &= N_p(a,b) \\
\text{Total population size} &= N_p \\
\text{Population proportion models containing M} &= \pi = \frac{N_p(a,b)}{N_p} \\
\text{Entropy criterion: } H(M) &= f\log_2\left(\tfrac{f}{\pi}\right) + (1-f)\log_2\left(\tfrac{1-f}{1-\pi}\right)
\end{aligned}$$

3. Take maximum-entropy model as best

A Simulation Study

Simulation Setup

- Compare performance of methods that return a single model.

- Methods:

  Oracle. Assume true model known; include each effect only if hypothesis test is significant.

  Modified stepwise selection. Method of Wu and Hamada, 2000

  AIC_c criterion. Select best model of size s_{max} or smaller.

  SAMS + entropy criterion.

- Test cases:
  - Use both PB_{12} and PB_{20} designs (s_{max} = 4 and 6, resp.).
  - Generate 5000 true models with random active factors, random coefficients.
  - Set coefficients large enough to give the model some chance to be detected, but small enough to keep R^2 to realistic levels.

Results

- Any selected model will be in one of five categories:

  The truth (T). All variables correctly included; no extras.

  Underfitted (U). Only correctly included variables, but some omitted.

  Overfitted (O). All true variables included, with some extras.

  Partial-truth (P). Some (not all) true variables, some extras.

  Wrong (W). None of the true variables included.

Partitioning of selected models (%)

	PB_{12} design					PB_{20} design				
Method	$\mathcal{T}$	$\mathcal{U}$	$\mathcal{O}$	$\mathcal{P}$	$\mathcal{W}$	$\mathcal{T}$	$\mathcal{U}$	$\mathcal{O}$	$\mathcal{P}$	$\mathcal{W}$
Oracle	62.8	37.2	0.0	0.0	0.0	61.4	38.6	0.0	0.0	0.0
Step	11.1	3.7	41.2	17.6	26.4	2.0	0.6	42.3	37.2	17.9
SAMS	43.3	16.2	15.8	15.0	9.7	35.7	30.4	10.0	17.6	6.3
AIC_c	7.2	0.7	53.8	25.5	12.8	0.5	0.0	52.7	41.0	5.8

Results

- Alternatively, count the total number of errors made in a model selection:
  - Excluding a truly-active variable.
  - Including a spurious variable.

Average number of errors made

	PB_{12} design				PB_{20} design			
Size	oracle	step	AIC_c	SAMS	oracle	step	AIC_c	SAMS
1	0.01	1.91	2.84	0.69	0.00	4.13	5.01	0.70
2	0.15	2.09	2.04	0.46	0.05	3.87	4.19	0.35
3	0.72	3.54	2.69	1.60	0.22	4.47	3.83	0.73
4	1.99	5.31	4.97	3.97	0.61	5.89	4.23	1.99
5					1.31	7.81	6.17	3.96
6					2.32	8.86	7.69	5.47