The Importance of the Sparsity Assumption in Screening

Bradley Jones

Quality & Productivity Research Conference
June 2015

Statistical Discovery. From SAS®

Outline

1. Introduction & Motivation
2. Standard Analytical Approaches in Screening
3. Simulation Studies and Demonstrations
4. Summary

Fisher's DOE Principles

1. Factorial principle
2. Randomization
3. Blocking
4. Replication

R.A. Fisher

DOE – Problem solving methodology for efficiently identifying cause-and-effect relationships.

Here we limit consideration to factor screening

We start with little prior knowledge and a large initial set of potential factors influencing the response

Our purpose is to identify the smaller set of active factors.

Due to cost considerations many industrial experiments have no true replication

Important Screening Assumption

Sparsity of effects

Operational definition:

fewer than half of the factors will be active.

Notation and terminology

m factors, *n* runs

Linear main effect model (ME)

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \varepsilon_i \qquad i = 1, \ldots, n$$

Model selection for unreplicated factorial designs

Any orthogonal main effects plan works for factor screening if:

1. Main effects are >> σ
2. No 2FIs are active
3. The number of active factors < $n/2$

If the number of active terms > n/2, automated model selection procedures tend to break down.

Why Lenth's method fails

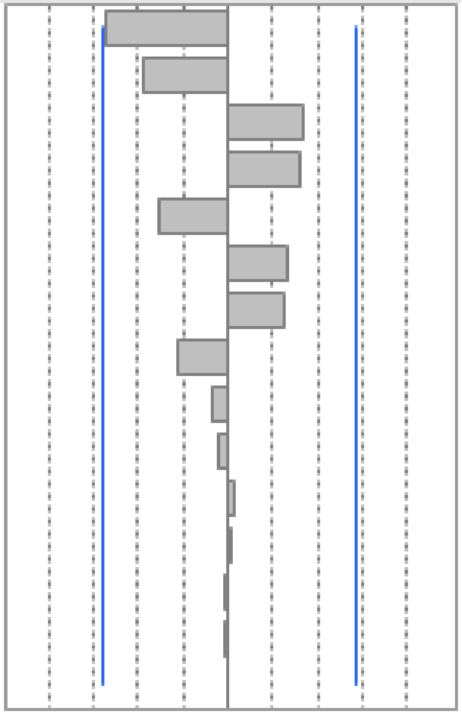
Lenth's PSE (an estimate of σ) is based on the half of the estimated effects that are smallest in magnitude.

If more than half of the effects are active the estimate of σ is biased high making all the effects seem less significant.

JMP Demonstration

Contrasts

Term	Contrast	Lenth t-Ratio	Individual p-Value	Simultaneous p-Value
C	-5.26127	-1.65	0.1104	0.7071
A	-3.57469	-1.12	0.2470	0.9756
D	3.15063	0.99	0.3048	0.9949
N	3.03315	0.95	0.3228	0.9972
F	-2.94920	-0.92	0.3357	0.9980
K	2.48290	0.78	0.4147	1.0000
L	2.44733	0.77	0.4208	1.0000
M	-2.12956	-0.67	0.5002	1.0000
J	-0.62987	-0.20	0.8565	1.0000
I	-0.46775	-0.15	0.8928	1.0000
G	0.30719	0.10	0.9315	1.0000
H	0.12815	0.04	0.9700	1.0000
E	-0.11686	-0.04	0.9730	1.0000
B	-0.09465	-0.03	0.9779	1.0000
O	0.04300	0.01	0.9896	1.0000



Automated Lenth’s method for $2^{(15-11)}$ with 8 active factors fails.

Why Forward Stepwise regression fails

When many effects are large, the effect that is largest in magnitude may fail to enter because the remaining effects are contributing to the estimate of the error variance.

JMP Demonstration

Stepwise Regression Control

Stopping Rule: Minimum AICc

Direction: Forward

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
687.186	12	7.5673972	0.5779	0.4724	.	4	121.5663	119.4292

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	-0.1376301	1	0	0.000	1
☐	☑	A	4.03398344	1	260.3684	4.547	0.05433
☐	☐	B	0	1	0.63856	0.010	0.92125
☐	☐	C	0	1	98.7273	1.845	0.20151
☐	☐	D	0	1	0.60943	0.010	0.92306
☐	☑	E	4.40726005	1	310.7831	5.427	0.0381
☐	☐	F	0	1	63.49916	1.120	0.31262
☐	☐	G	0	1	157.2336	3.264	0.09824
☐	☐	H	0	1	0.02294	0.000	0.98505
☐	☐	I	0	1	6.698942	0.108	0.74828
☐	☐	J	0	1	117.1257	2.260	0.1609
☐	☑	K	4.80776681	1	369.8339	6.458	0.02588
☐	☐	L	0	1	61.72559	1.086	0.31981
☐	☐	M	0	1	0.197097	0.003	0.95621
☐	☐	N	0	1	0.080258	0.001	0.97205
☐	☐	O	0	1	180.6275	3.922	0.07321

Automated Forward stepwise for $2^{(15-11)}$ with 9 active factors fails.

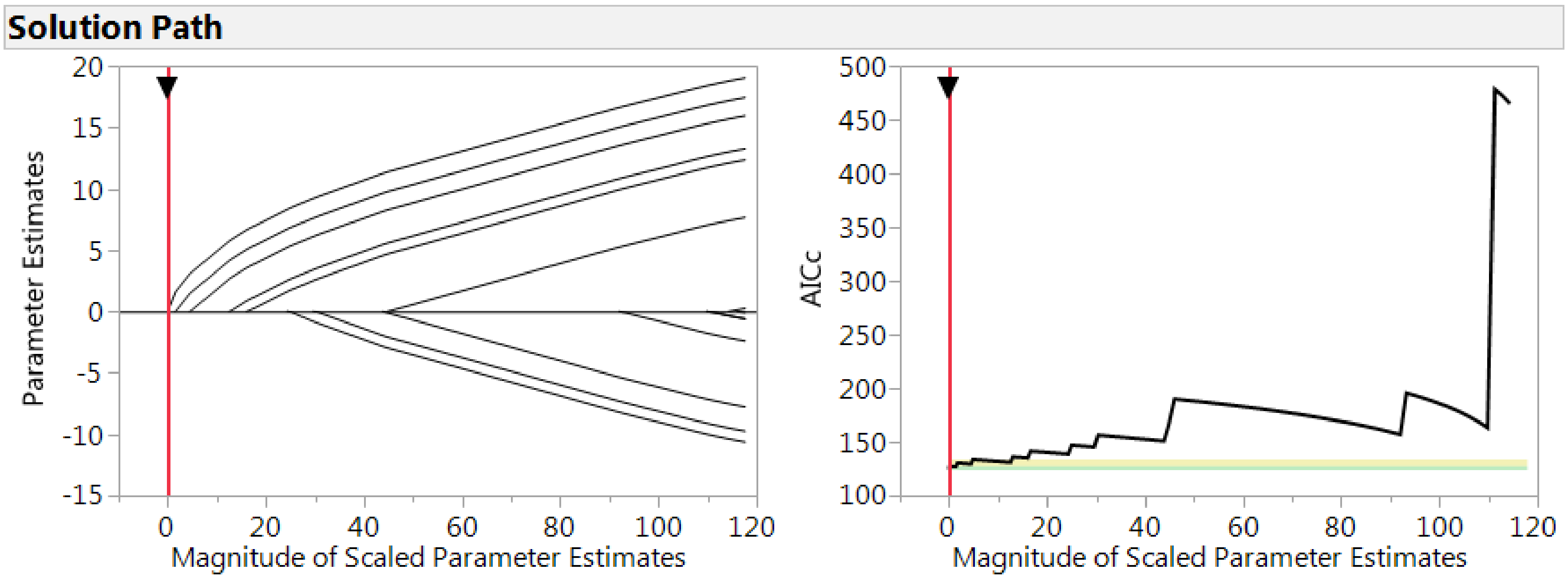
Why the Lasso fails

The Lasso works best when the number of runs far exceeds the number of model terms and there is substantial correlation among the predictors.

The analysis of designed experiments is clearly far from the above scenario.

Using generalized cross-validation criteria like AICc can over-penalize larger models when effect sparsity does not hold.

JMP Demonstration



Automated Lasso for $2^{(15-11)}$ with 9 active factors fails.

Adding a few replicated runs works…

1. Even if all the effects are active.
2. Even though the resulting design is not orthogonal for the main effects.

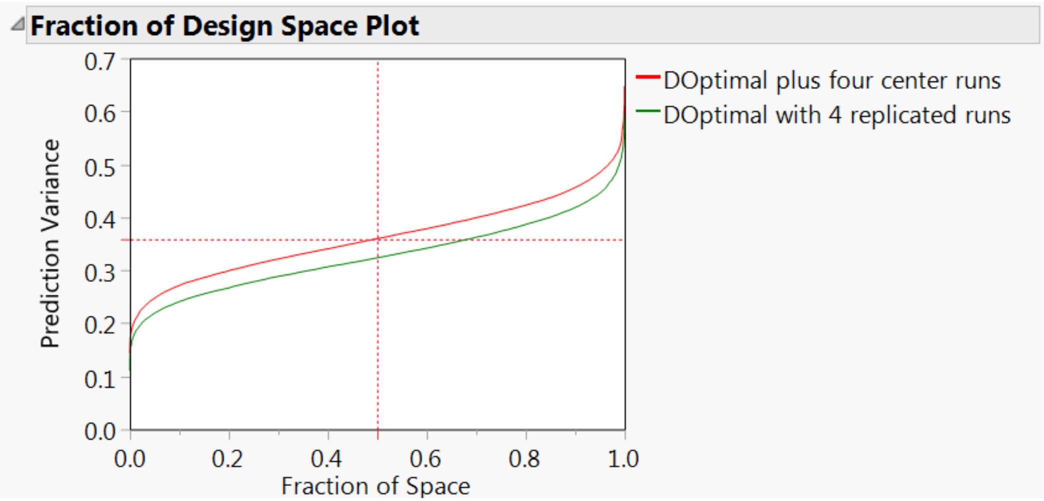
Power Analysis

Significance Level 0.05

Anticipated RMSE 0.666

Term	Anticipated Coefficient	Power
Intercept	1	0.996
X1	1	0.996
X2	1	0.996
X3	1	0.996
X4	1	0.996
X5	1	0.996
X6	1	0.996
X7	1	0.996
X8	1	0.996
X9	1	0.996
X10	1	0.996
X11	1	0.996
X12	1	0.996
X13	1	0.996
X14	1	0.996
X15	1	0.996

Replicating Vertices vs. Center Runs



Relative Estimation Efficiency

Term	Efficiency of DOptimal with 4 replicated runs Relative to Reference Design
Intercept	0.95618
X1	1.06904
X2	1.06904
X3	1.06904
X4	1.06904
X5	1.06904
X6	1.06904
X7	1.06904
X8	1.06904
X9	1.06904
X10	1.06904
X11	1.06904
X12	1.06904
X13	1.06904
X14	1.06904
X15	1.06904

Replicating vertices is more efficient than center runs for both parameter estimation and response prediction.

Conclusion

Unreplicated screening designs depend on the assumption of sparsity of effects.

An operational definition of effect sparsity is that the number of active effects is less than half the number of runs.

Replicating a few runs can allow for the detection of all active effects (SNR > 2) even when the sparsity assumption fails.

References

1. Box, G. E. P. and J. S. Hunter (1961). The 2^{k-p} *fractional* factorial designs. *Technometrics* **3**, pp. 449–458.

2. Lenth, Russell V. (1989). “Quick and Easy Analysis of Unreplicated Factorials” *Technometrics* **31** pp. 469-473.

3. Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society, Series B* **58** (1): 267–288.