LA-UR-00-1638

Statistical Theory & Methods for Evaluating Computer Models: *Quantifying Prediction Uncertainty*

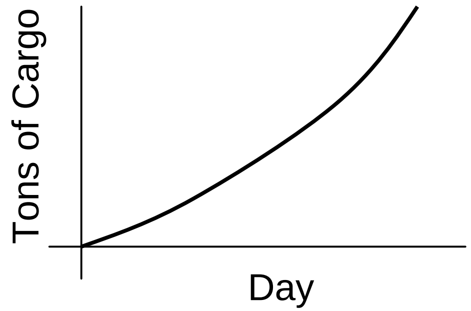
Michael D. McKay
Statistical Sciences Group
Los Alamos National Laboratory
mdm@lanl.gov

in collaboration with
Richard Beckman, Mark Fitzgerald, Todd Graves,
Lisa Moore, Vincent Thomas

Earlier version presented at the Engineering and Physical Sciences Research Council (EPSRC) Workshop
Statistical Analysis of Computer Code Outputs, Gregynog, Wales, UK, April 10-14, 2000.

Introduction to *input uncertainty* by way of a simple example with a discrete event simulation of movement of cargo by aircraft

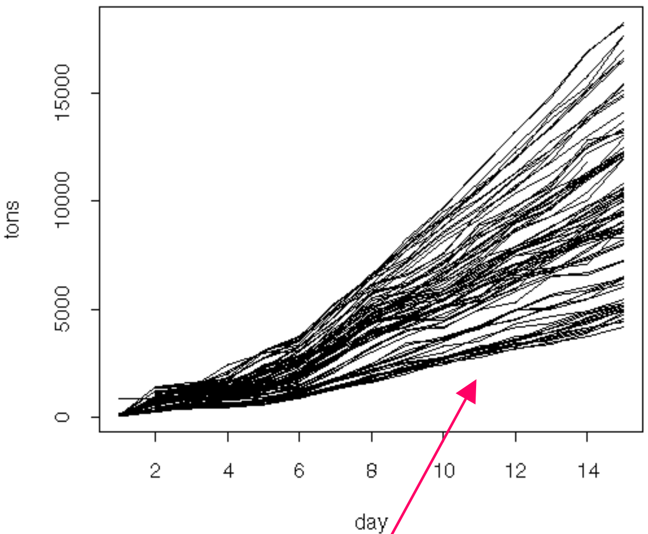
- Output variable is Cumulative Tons of Cargo Delivered
- Input variables (8) include Use Rate, Fuel Flow, ….



See next slide for examples

- A sample of plausible alternative input values generates prediction-uncertainty band.
- Discover effect of setting one of the inputs (Use Rate) to its nominal value in the runs.
- Discover that a combination of two inputs (Use Rate and Fuel Flow) controls variability.

Focus: prediction-uncertainty bands and important inputs

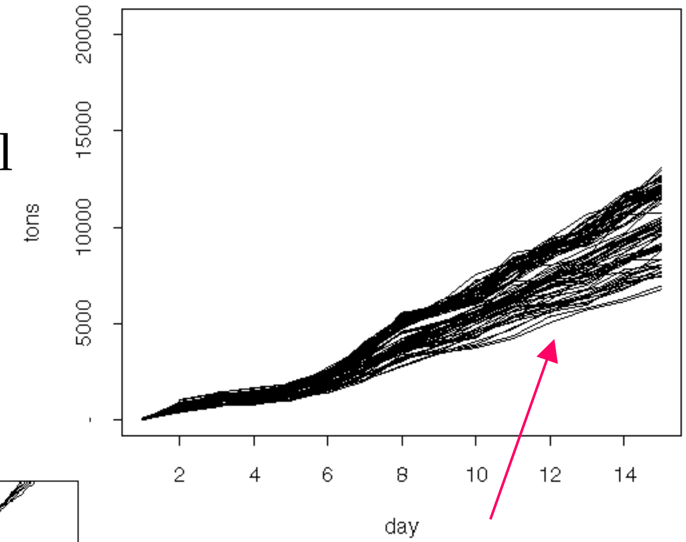


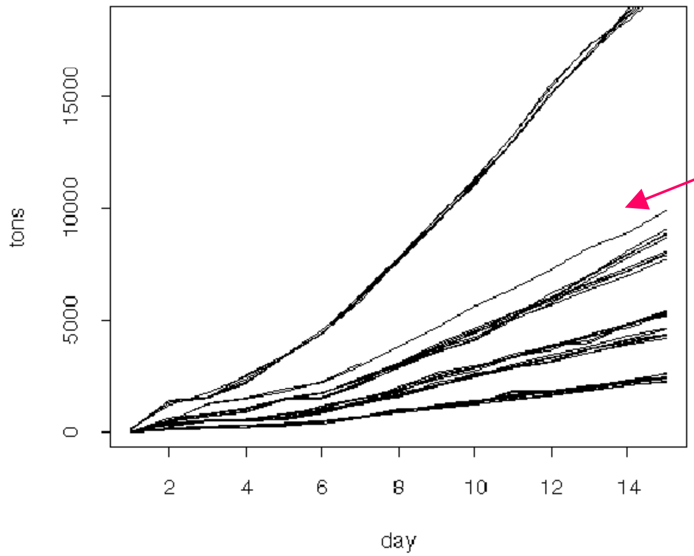
Full prediction-uncertainty band

All 8 inputs vary

1 input fixed at nominal

2 inputs fixed at
2 x 2 = 4 values





Reduced (conditional) uncertainty bands

We want to find inputs that control spread.

Introductory mathematical formulation

- System response is w.
- Physical conditions are u.
- Nature's rules denoted by $w = M(u)$.

Footnotes:

- Realistically, system response is a complex quantity W for which y models features $w = \Gamma(W)$.
- u may not be completely known (or knowable) in advance.
- More than just u might be needed to determine w.

- Model prediction is y.
- Model inputs are x.
- Calculation is denoted by $y = m(x)$.

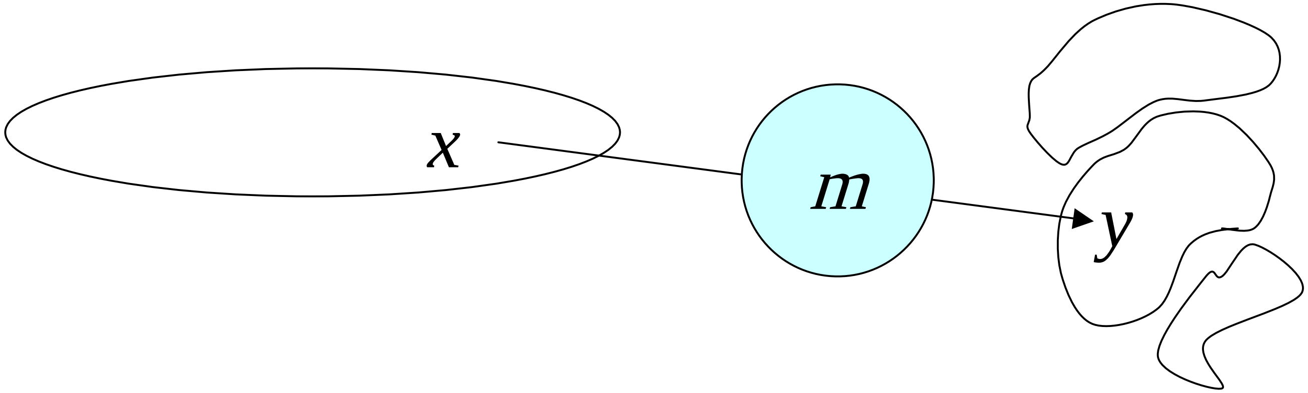
Footnotes:

- Not knowing how to match x to u, we treat x as a random variable.
- Finding a suitable probability distribution for x is usually difficult.

Towards assessing quality of prediction : *Uncertainty quantification* before *model validation*

- Objective of *uncertainty quantification* is determination of how far apart w (real outcome) and y (predicted outcome) are likely to be at a specific prediction point (x^*, u^*) in light of *evidence* V at other, specific data points (x^v, u^v).
- Some reasons why w^* and y^* might be expected to differ:
  - Principles (rules) assumed to produce w and/or the ways they are incorporated in m are incomplete (*modeling uncertainty)*.
  - (ongoing research) Specification of a single value for model inputs x in the mathematical world does not adequately characterize actual conditions u in the physical world (*input uncertainty*).
  - (future research) Degree of agreement between w^* and y^* is warranted by strength of evidence V (*combining information: input uncertainty and observed data)*.

General description for input uncertainty



Plausibility region for model inputs: D_x

$x \sim f_x(x),\ x \in D_x$

Uncertainty region for model prediction: D_y

$y = m(x) \sim f_y(y),\ y \in D_y$

Probability function $f_x(x)$ describes input uncertainty and $f_y(y)$ describes output or prediction uncertainty. For now, we are looking at a single model.

Why statistical methods?

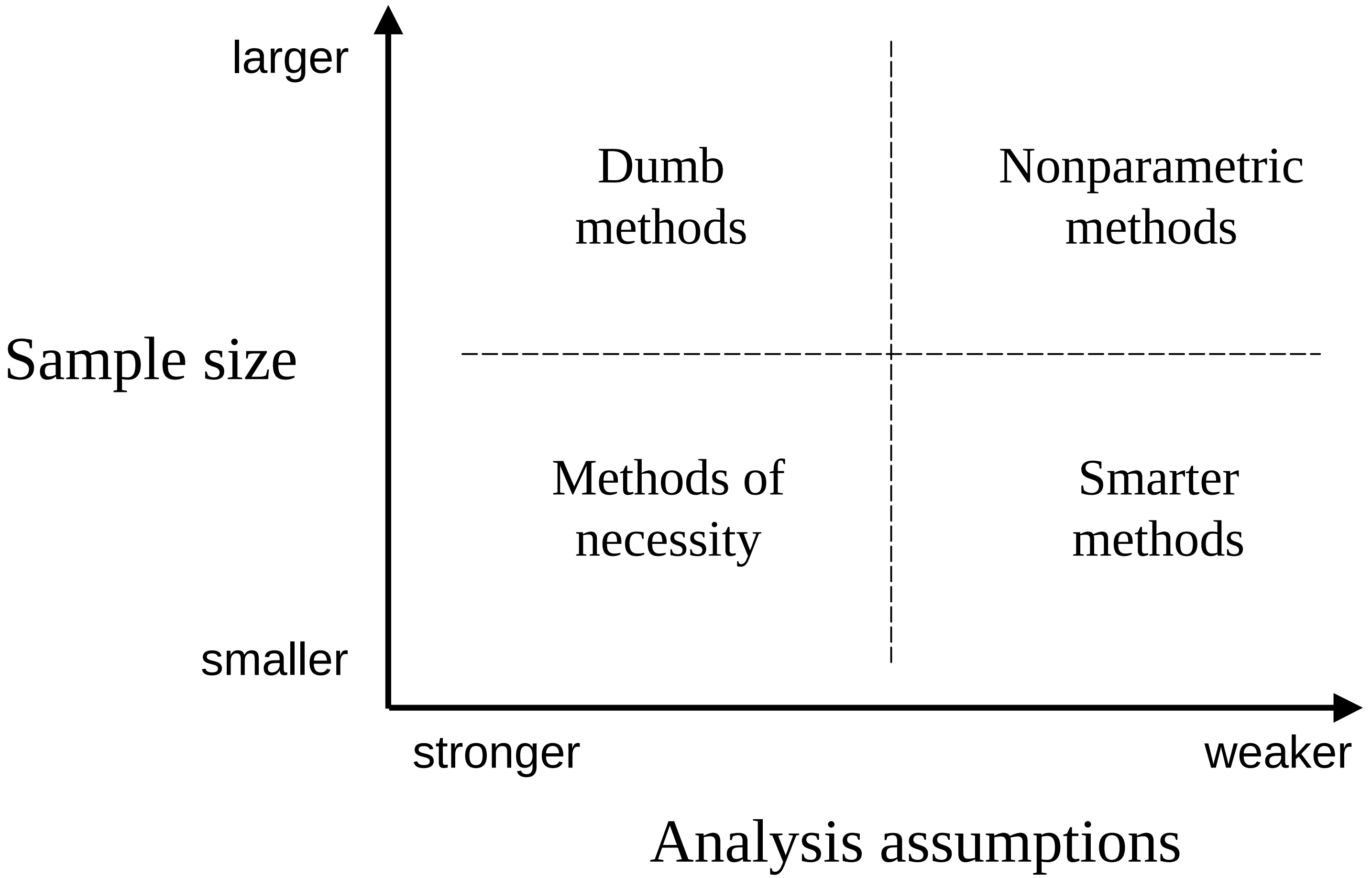
- The space of possibilities that generates uncertainties is too big to be enumerated.
- Suppose uncertainties are due to plausible alternative values of p inputs defined on sets (intervals) characterized by I values (low, high, etc.)

p inputs	*I* values	# points in input space
30	2	10^9
30	5	10^{21}
84	5	10^{58}

Three types of experiments

- *Laboratory Experiments*. Factors affecting response variables are controllable to within physical and budgetary limitations. Number of experiments is often small.
- *Field Experiments*. (e.g., clinical trials) Combinations of factors are usually only selectable. Number of experimental units is frequently quite large.
- *Computer Experiments*. Factors are completely controllable, values are numerical quantities. Number of runs may range from very few to very many.

Methodology grid



Question I:
Quantifying uncertainty

Estimate where *y* is likely to be and characteristics of its probability distribution, for example:

- $\hat{D}_y$, mean value μ_y, variance σ_y^2.
- Tolerance bounds $\left(\hat{a},\hat{b}\right)$ that have probability content p with confidence level $(1-\alpha)\times 100\%$.
- Density function or empirical distribution function $\hat{F}_y(t) = \text{ Est. } \Pr\{y \le t\}$.

Question II:

Uncertainty importance (McKay 1997 *Reliability Engineering and System Safety*)

- Full model prediction with (all) x:

$$y(x) = m(x) \text{ with } x \sim f_x(x)$$

- Partition $x = x^s \cup x^{\bar{s}}$ where $s \subset \{1, 2, \cdots, p\}$ selects a subset of input variables.
- Restricted prediction with only x^s:

$$\begin{aligned} \tilde{y}(x^s) &= E(y \mid x^s) \text{ with } x^s \sim f_s(x^s) \\ &= \int m(x) f_{x^{\bar{s}}}(x^{\bar{s}} \mid x^s) dx^{\bar{s}} \end{aligned}$$

Uncertainty importance (continued)

- How does *knowing* x^s reduce uncertainty, or
- How close is $\tilde{y}$ to y (on average)?
- Measure uncertainty importance of x^s by

$$E(\tilde{y} - y)^2 = \mathrm{Var}(y) - \mathrm{Var}(\tilde{y})$$

and

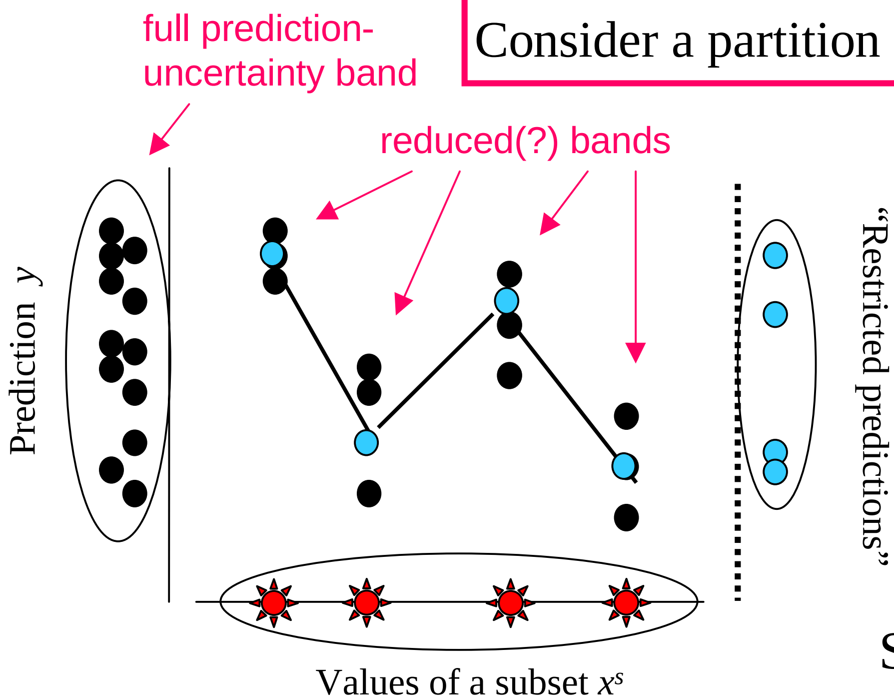
$$\text{Correlation ratio } \eta^2 = \mathrm{Var}(\tilde{y}) / \mathrm{Var}(y)$$

- η^2 must be estimated from a sample of runs.

The correlation ratio (Pearson 1903 *Proceedings of the Royal Society of London*) as an importance measure

Input uncertainty is described by the probability function $f(x)$.

Consider a partition of $x = x^s \cup x^{\bar{s}}$. x^s may be a single input.



● $= y$

● $= \tilde{y} = E(y \,|\, x^s)$

Spread of ●s $= \mathrm{Var}(y)$

Spread of ●s $= \mathrm{Var}(\tilde{y})$

Spread of ●s around ● $= \mathrm{Var}(y \,|\, x^s)$

Note the I = 4 groups of J = 3 values of y.

$$\mathrm{Var}(y) = \mathrm{Var}(\tilde{y}) + E[\mathrm{Var}(y \,|\, x^s)]$$

$$\text{Correlation ratio } \eta^2 = \mathrm{Var}(\tilde{y}) / \mathrm{Var}(y)$$

Sample estimates of components of correlation ratio, η^2

I = number of and J = number of ● for each

For each of I sample values x_i^s, let $x_{ij}^{\bar{s}}$ be J sample values of the other inputs and y_{ij} be the $N = I \times J$ corresponding output values.

$$\text{Let } \bar{y}_i = \frac{1}{J}\sum_{j=1}^{J} y_{ij} \text{ and } \bar{y}_{..} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \ .$$

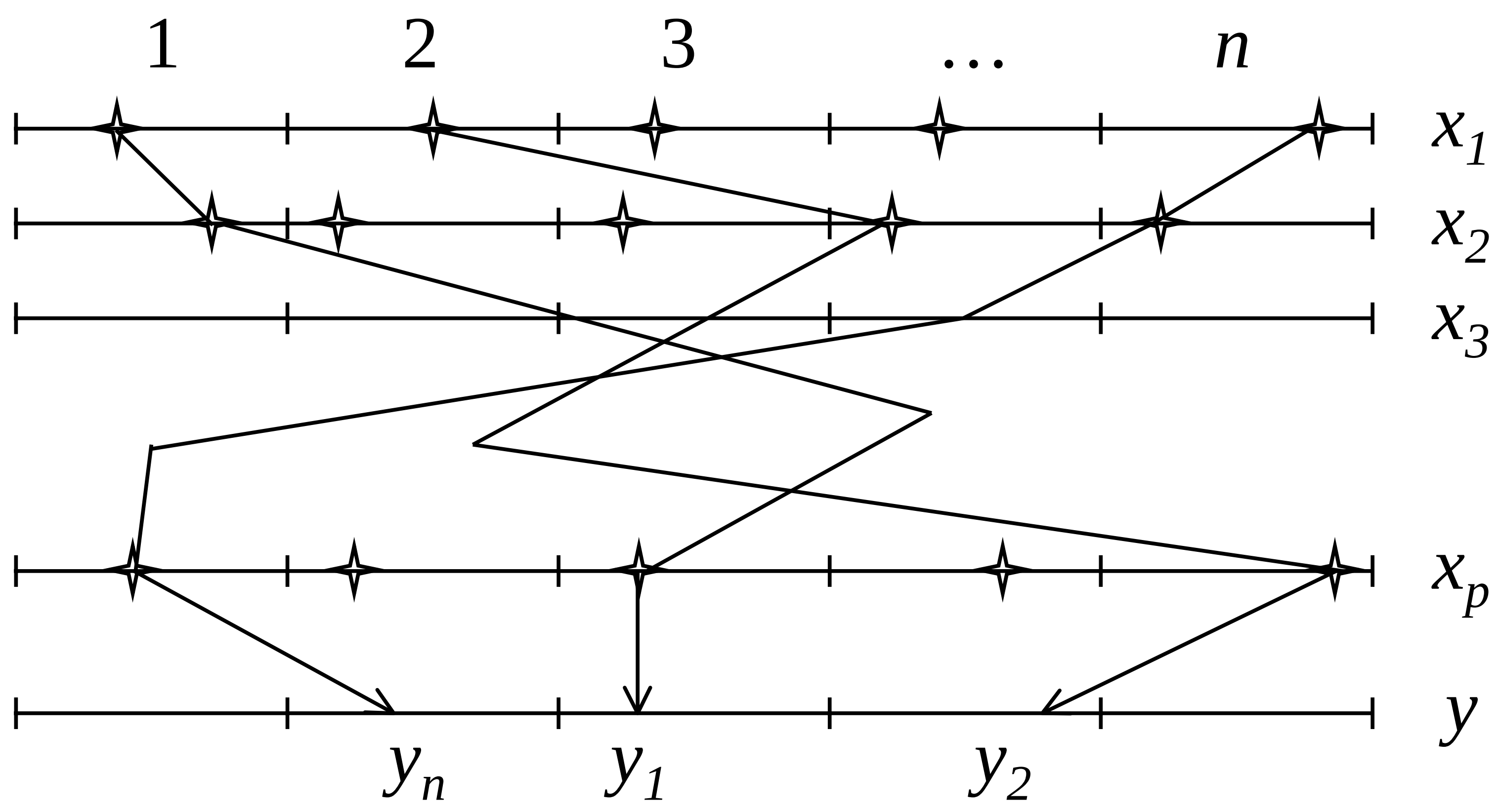
$$\text{Estimate } \eta^2 \text{ with } R^2 = \frac{\displaystyle\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\bar{y}_i - \bar{y}_{..}\right)^2 \approx \text{est Var}(\tilde{y})}{\displaystyle\sum_{i=1}^{I}\sum_{j=1}^{J}\left(y_{ij} - \bar{y}_{..}\right)^2 \approx \text{est Var}(y)}$$

Latin hypercube sampling (LHS)
(McKay, Conover and Beckman 1979 *Technometrics*)

- Range of each input is divided into n equal probability intervals.
- Each interval is (conditionally) sampled once.
- Values are combined at random across input variables.

Property: each input variable is represented by n distinct values that cover its range.

Parallel coordinates plot (Wegman 1990 *Journal of the American Statistical Association*) of an LHS

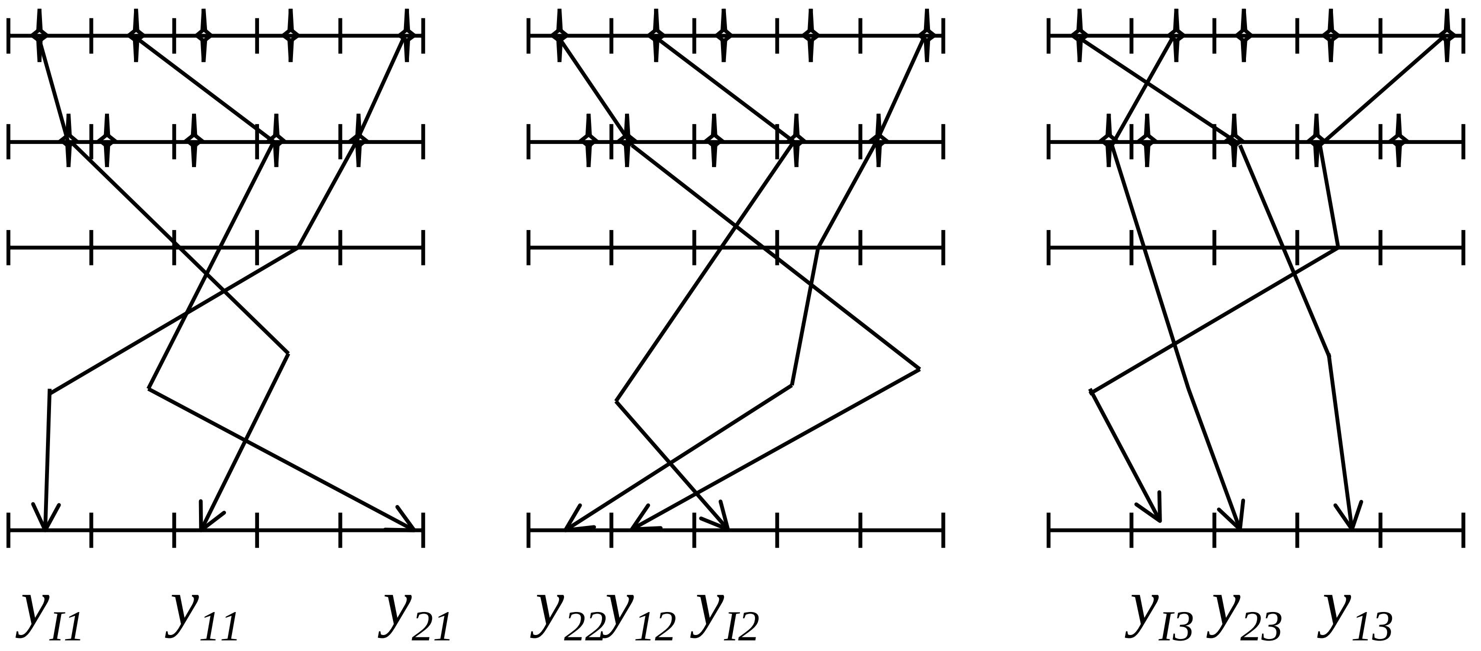


n points in D_x from an LHS often produce better estimates than a random sample, depending on model and statistic.

Replicated rLHS: *J* samples of size *I*

Alternatively, orthogonal array designs

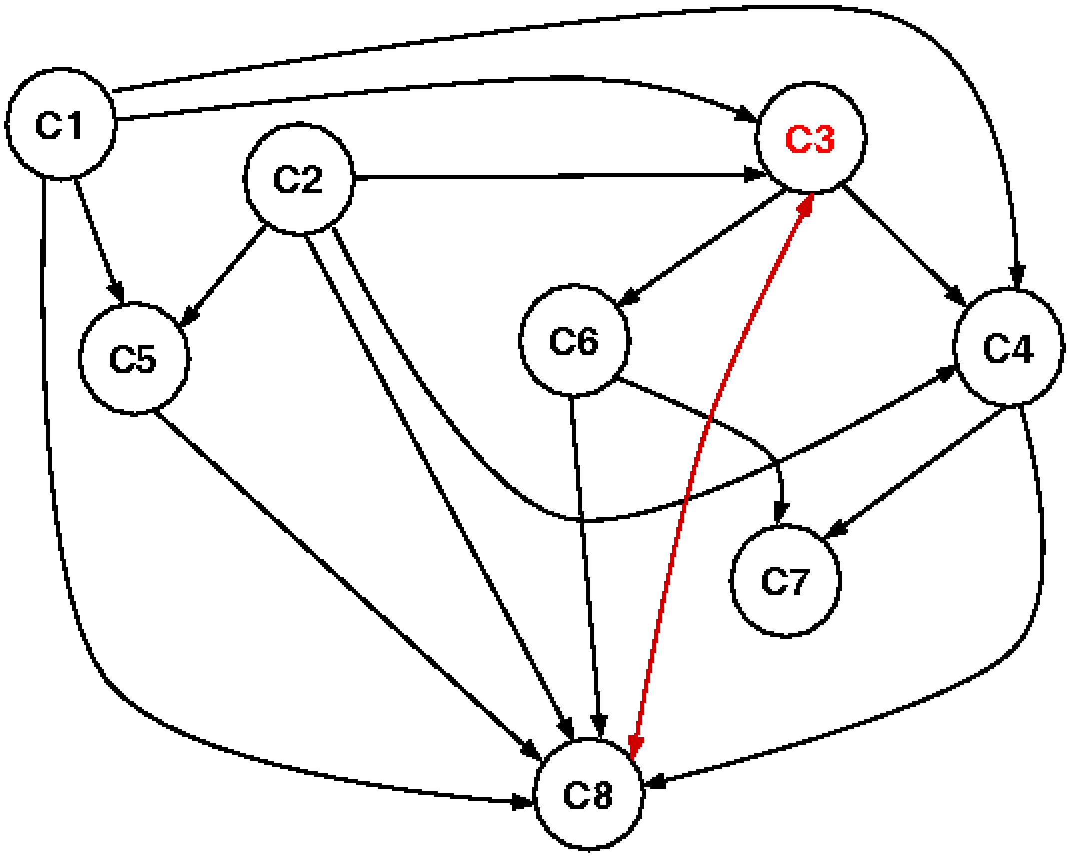
Same values ✦ on each axis but different combinations.



Practice:
Summary of a procedure

- Identify inputs.
- Define range of values and probability distributions.
- Generate sample values for inputs.
- Make computer runs and record output values.
- Analysis: estimate output probability function and calculate R^2 values for each input.

Case study: finding a small subset of inputs that drives the calculation

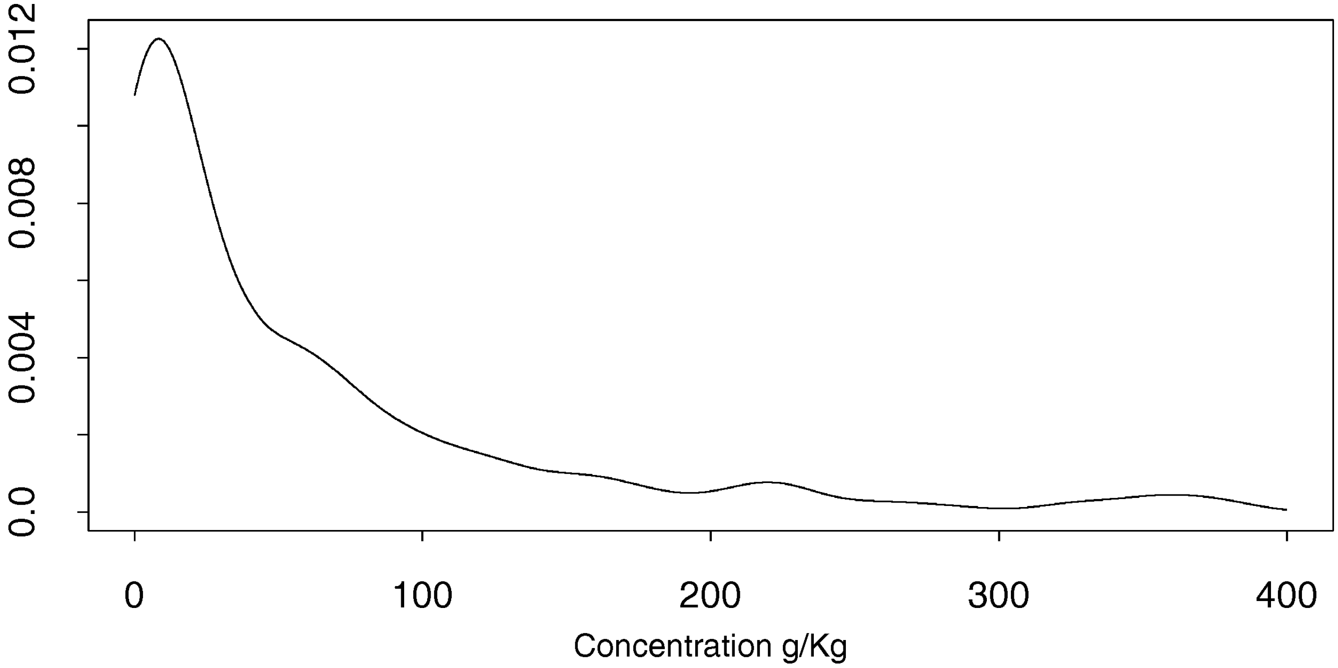


We first look at single inputs

Compartmentalized environmental transport model
(84 real plus 16 fictitious input variables)

Prediction band:

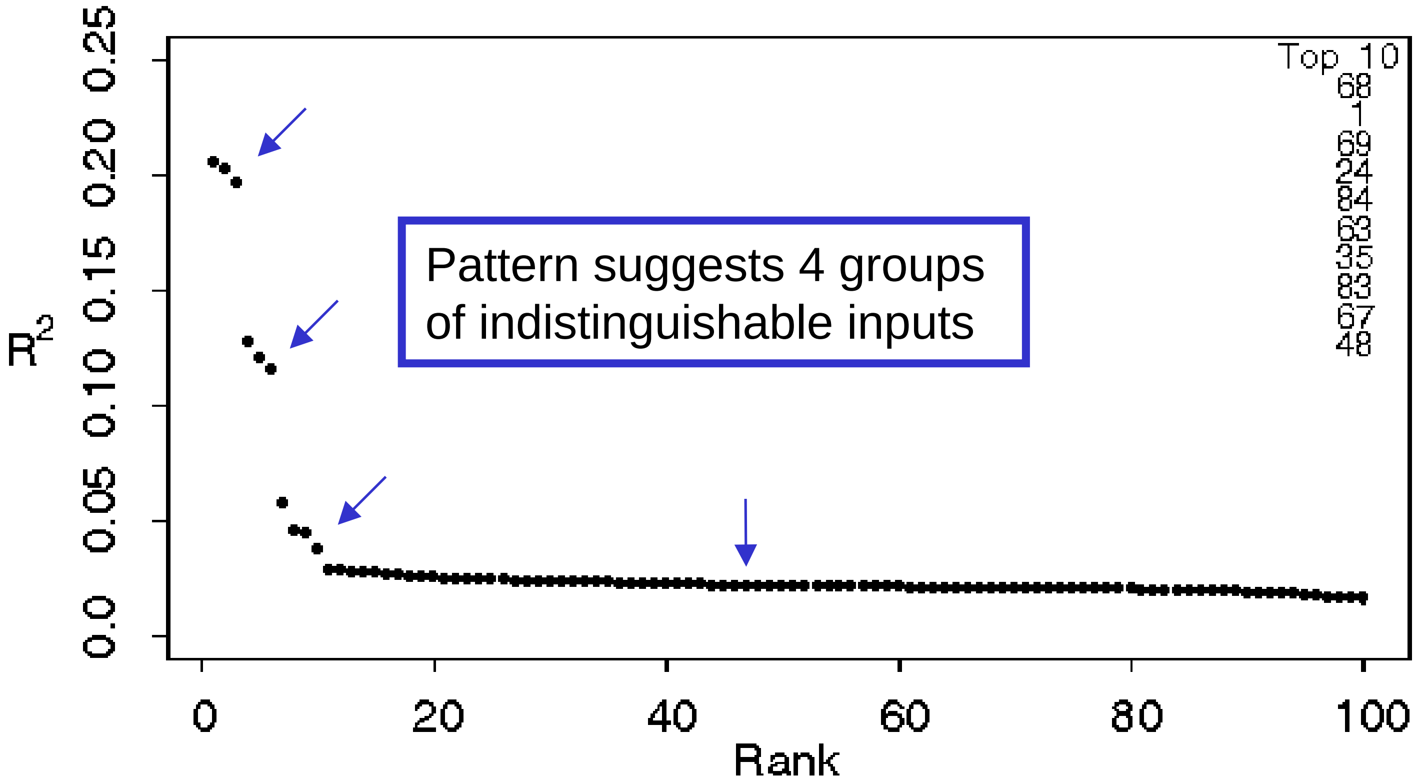
Estimated density function of *y*



In the best of worlds:

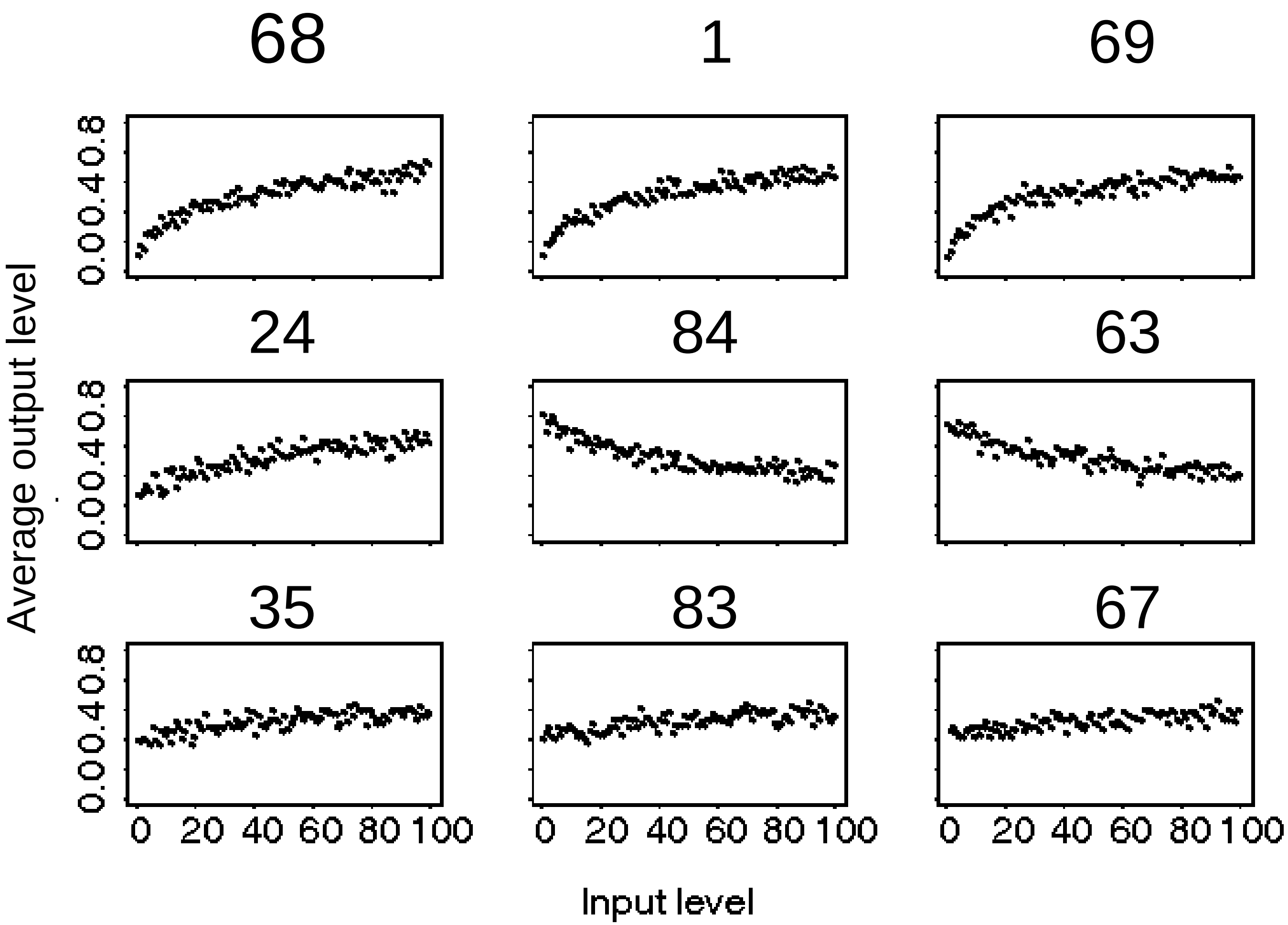
Pattern of ordered R^2 values for 100 inputs

(from a sample of size 5000)



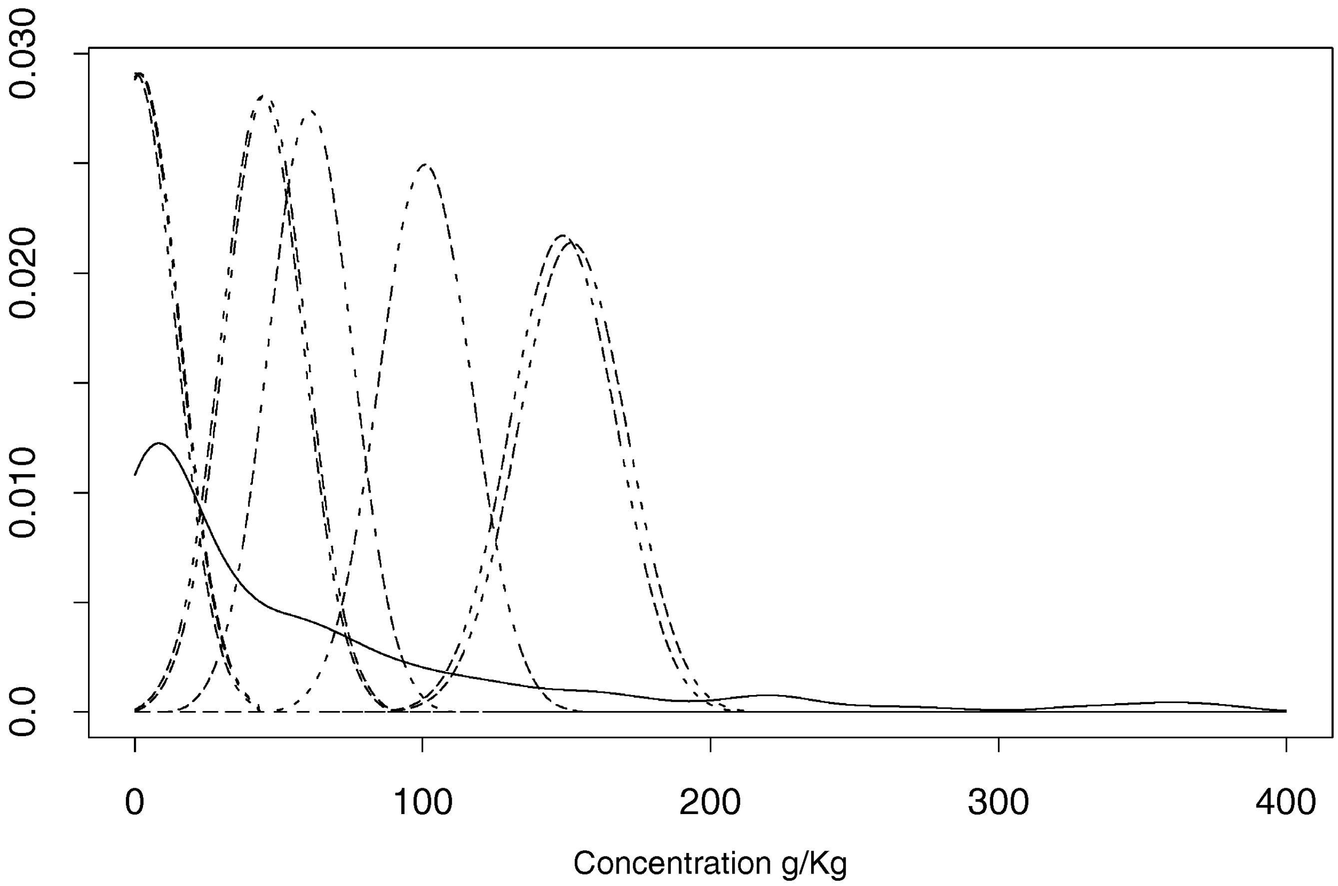
What inputs do:

Patterns of average *y* (●) for top 9 inputs



Reduced prediction bands:

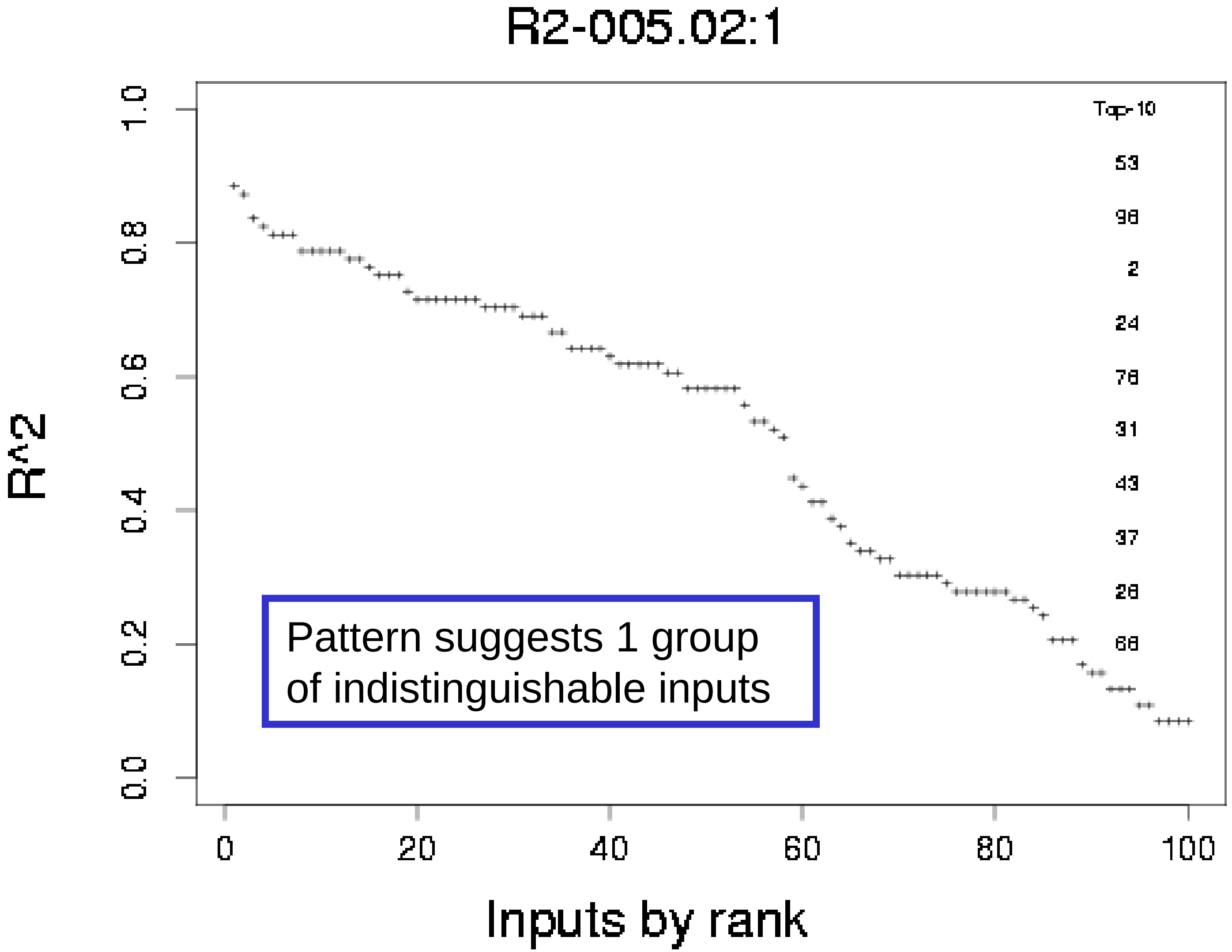
Conditional densities with 10 inputs fixed



In the not-so-best of worlds:

Pattern of ordered R^2 values for 100 inputs

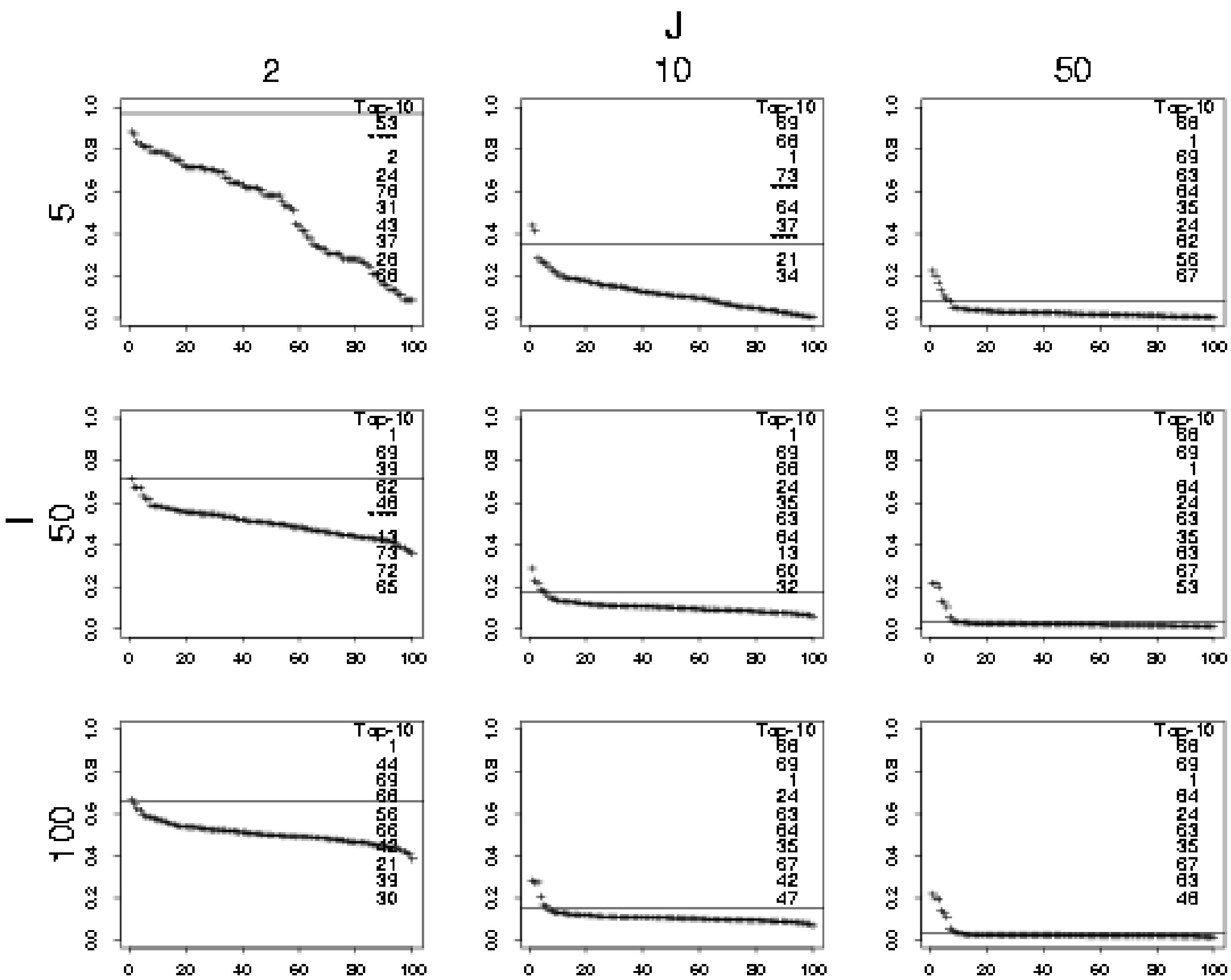
(from a sample of size 10!)



Finding significantly large values of R^2

- Critical value of R^2 for testing H_0: $\eta^2 = 0$ can be derived from the F distribution under assumptions of normality and independent random sampling of inputs.
- Because we look simultaneously at many (p = 100) inputs, we set the "experiment-wide" alpha level. That is, we want the probability of falsely detecting one or more "important" inputs (out of 100) to be alpha.
- For a choice of α and p variables (tests), we use the critical value corresponding to $\alpha^* = 1 - (1 - \alpha)1/p$. For example, for $\alpha = .05$ and $p = 100$, $\alpha^* = 1 - (1 - .05)1/100 = .000513$
- Since $\eta^2 = 0$ is not likely for model-input variables, we interpret the test as one of distinguishing among (sets of) inputs with different values of η^2.

R^2s and critical value (horizontal line) for 84 inputs plus 16 fictitious variables (*) in 9 experimental designs (I x J)



Fictitious variables

- In the analysis, *fictitious variables* have I values each repeated J times, like those of real inputs. However, the values are assigned *at random* to the computer runs, and have nothing at all to do with computations.

- If fictitious variables appear among the top K with largest R^2, we would have reason to question whether those K are *statistically different* from the rest.

- A statistical test based on the number k out of f fictitious variables appearing in the top K out of $n = p + f$ variables in total can be constructed from the of the hypergeometric distribution.

- Theoretical work is in progress. Some observations follow.

What might happen:

Sample-to-sample and design variability

Columns are top $K = 5$ inputs with largest R^2. There are 4 samples from each of 2 different designs, with $f = 16$ and $p = 84$ variables.

I=5 and J=10				I=50 and J=10			
#1	#2	#3	#4	#1	#2	#3	#4
69	77	68	37	1	1	69	69
68	68	55	85	69	69	68	1
1	55	24	69	68	68	1	68
75	82	63	24	24	24	63	84
90	84	6	72	35	84	84	35
…	…	…	…	…	…	…	…

Average number of fictitious variables in top K = 10 from (only!) 9 simulations with p = 84 real and f = 16 fictitious variables

Value of I	Value of J		
	2	10	50
5	2.0	0.9	0.7
50	1.9	0.8	0.1
100	1.3	0.1	0.0

Sample-to-sample variability

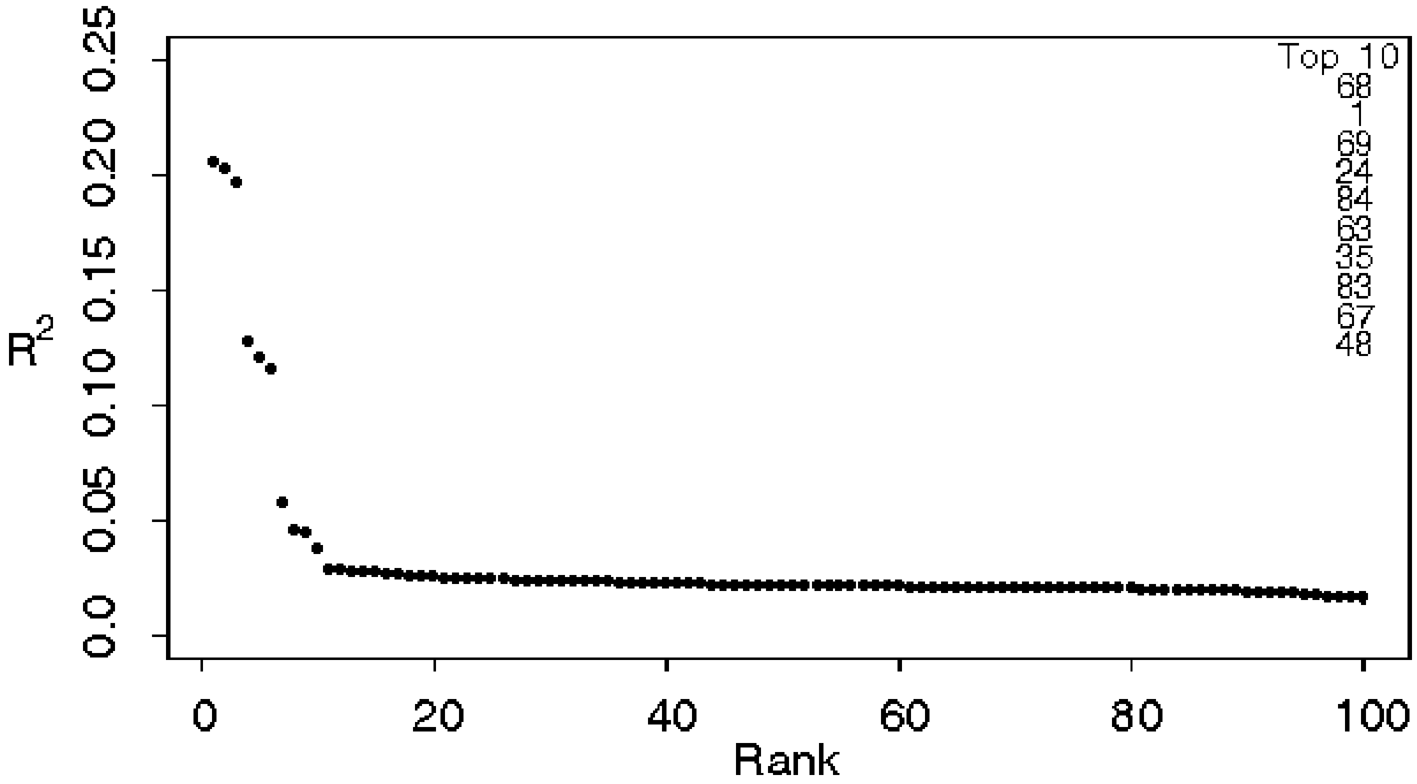
Top 3 inputs are the same for the 4 samples

I=50 and J=10			
#1	#2	#3	#4
1	1	69	69
69	69	68	1
68	68	1	68
24	24	63	84
35	84	84	35
…	…	…	…

Inputs that stand out:

Ordered R^2 values for 100 inputs

(from a sample of size 5000)



A story in patterns?

Conjecture---With the I=100 and J=50 design, the R^2 statistic identifies 4 groups of equivalent input variables of sizes 3, 3, 4, and the remaining 74. Within each group, inputs are not distinguishable.

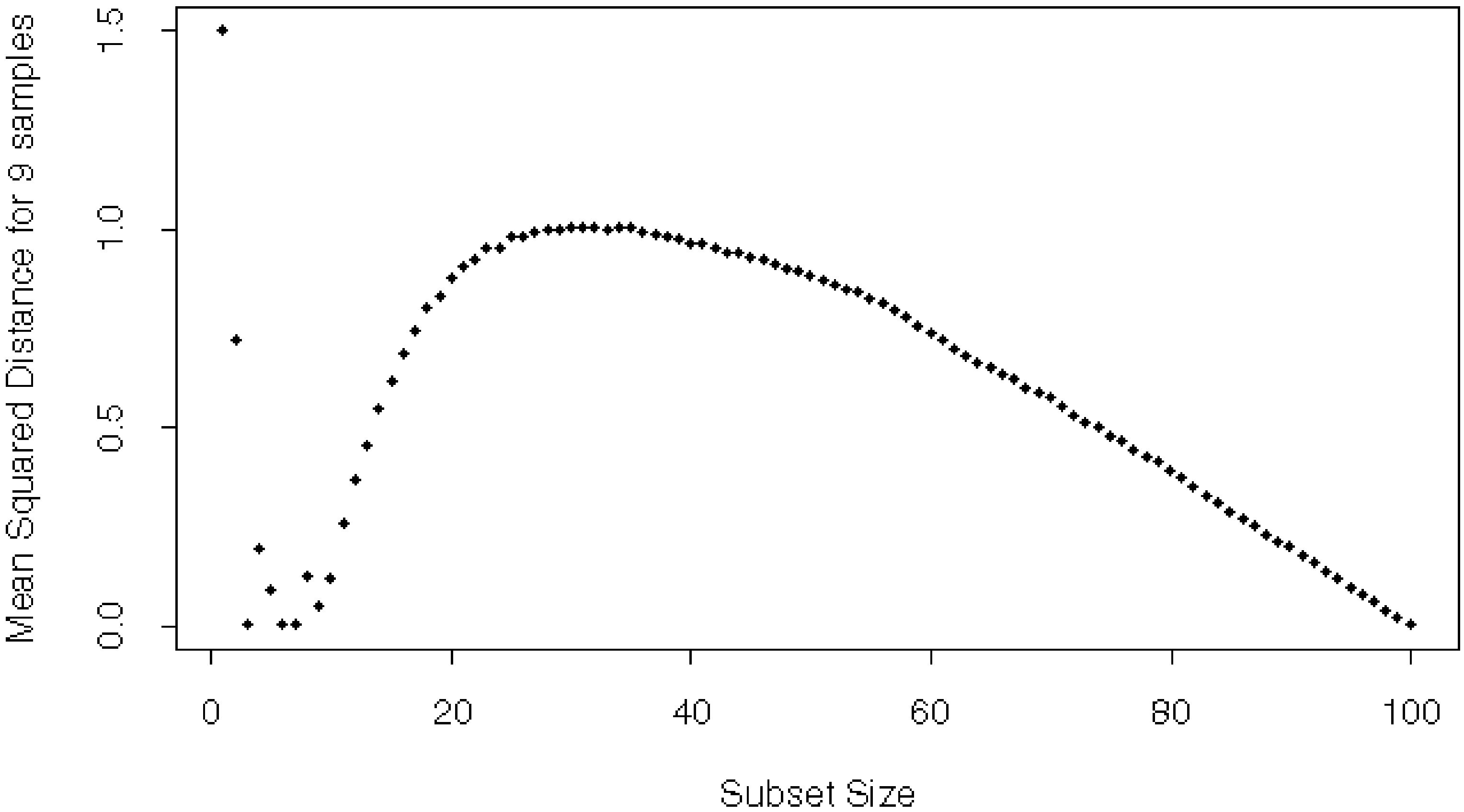
Test---Examine the consistency of composition of sets of top s =1, 2, 3 inputs with largest R^2, for replicated independent samples.

Mean Squared Distance details

- Let $s(k)$ be a vector of length p (=100) with k (=1,2,…) 1s and (p-k) 0s. Typically, (0,1,1,…,1,0) indicates a set of k out of p inputs.
- $d_{ij}(k) = \frac{1}{\sqrt{k}}\left[s_i(k) - s_j(k)\right]$ is the (normalized) vector between sets i and j, and $d^T{}_{ij}(k)\, d_{ij}(k)$ is the squared distance between them.
- $MSD(k) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d^T{}_{ij}(k)\, d_{ij}(k)$ is the average or mean squared distance among n (=9) such sets.
- $E\left[MSD(k)\right] = 2\left(1 - \frac{k}{p}\right)$ if the 1s are assigned *at random*, i.e., the inputs are not distinguishable.

Groupings of important inputs:

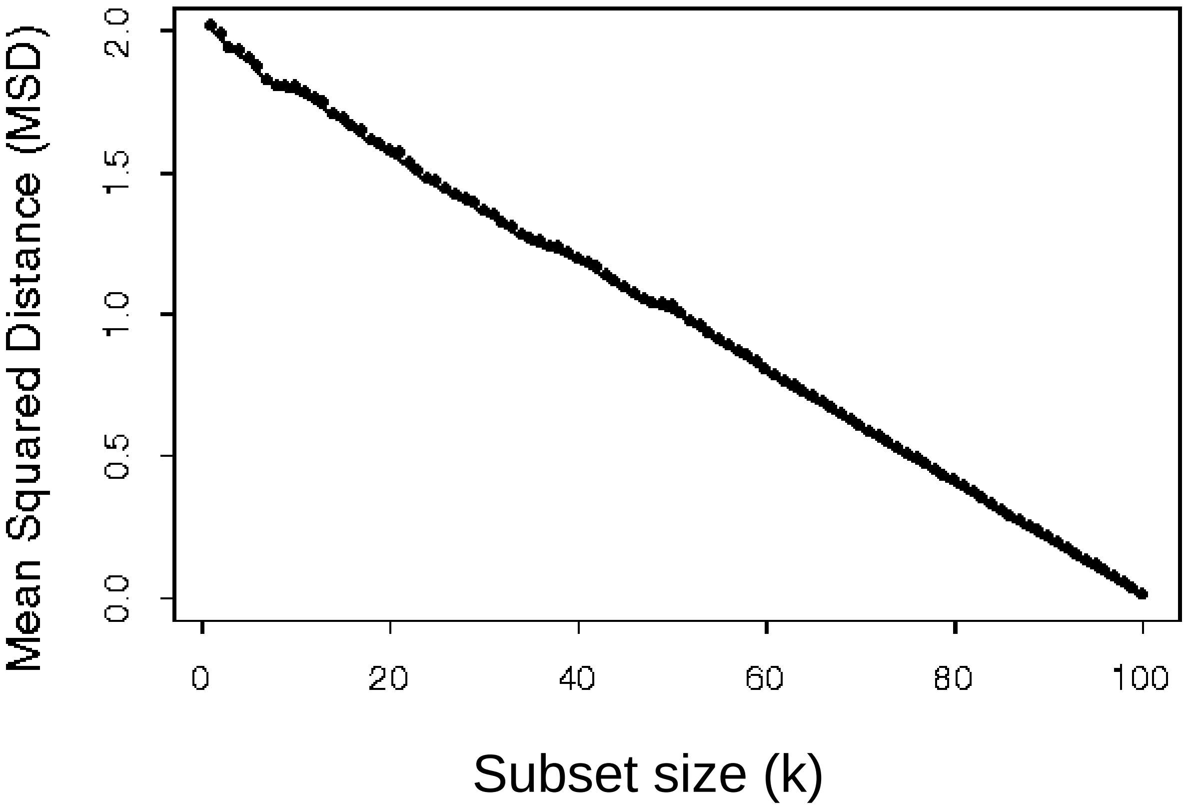
MSD for *I*=100 and *J*=50 sample



Remaining power & inconsistency:

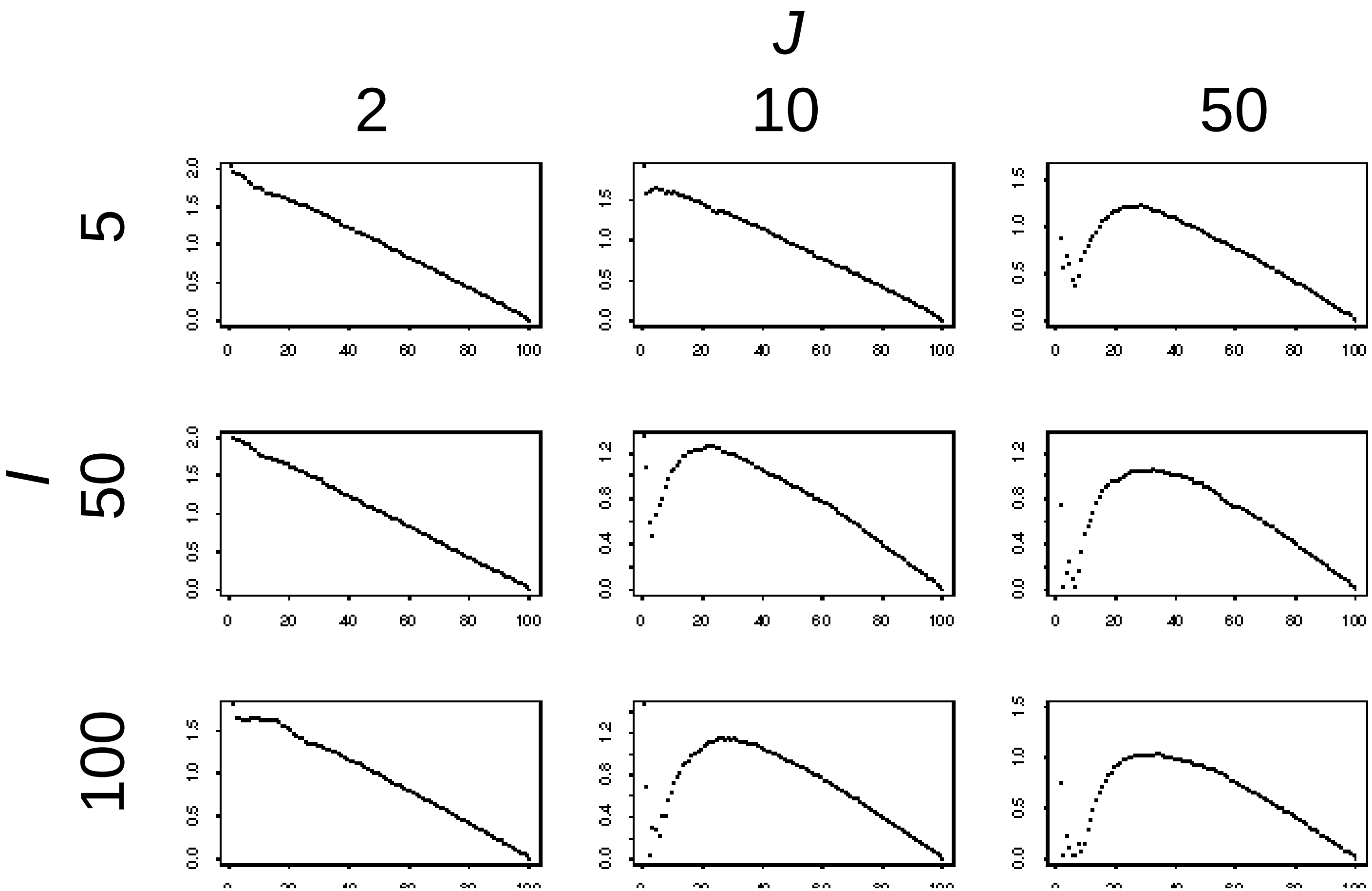
MSD without top 10 inputs

(for *I=100* and *J=50*)



How many runs are needed?

Patterns of *MSD* for 9 designs



Summary

- *Variance* of *y* is a measure of *prediction uncertainty* induced by inputs.
- *Uncertainty importance* of a subset of inputs refers to their contribution to prediction uncertainty. It can be measured by a variance ratio called the *correlation ratio*.
- The ratio of sums-of-squares called R^2 is proportional to a (biased) estimator of the correlation ratio.
- *Sample* or *statistical variability* of R^2 depends on the design parameters *I* and *J*. It causes false indications of importance as well failure to distinguish among inputs.
- Use of *critical values* with experiment-wide alpha level as well as *fictitious variables* can help control and point out ill effects due to sample design.

Concluding Remarks

Characterizing causes of prediction uncertainty is only a small part of a complex process called “model evaluation.” This talk focused on an initial step of a procedure to search for a small subset of model inputs that accounts for a significant fraction of prediction uncertainty. We saw how conclusions based on a statistic (R^2) about importance of inputs can vary significantly depending on sample size and design. We saw how a mathematical understanding of certain patterns could be used as a diagnostic tool to assess the reliability of the analysis.