Statistical methods research done as science rather than math:

Estimates on the boundary in random regressions

This lecture is about how we study statistical methods.

It uses as an example a problem that arises in analyzing data with the so-called random regressions model.

I'll begin by describing the example.

The random regressions model

Data are grouped in clusters i; j indexes observations within clusters.

The outcome y_{ij} is presumed to arise as

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \epsilon_{ij}, i = 1, \ldots, N, j = 1, \ldots, s,$$

For today, x_{ij} is scalar

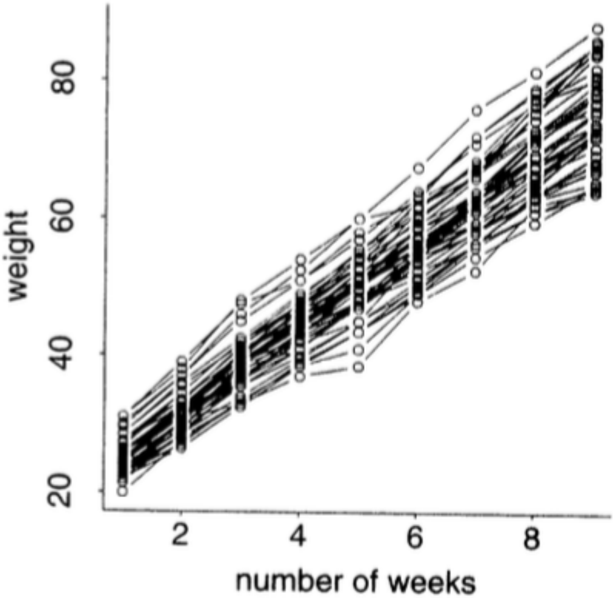
ϵ_{ij} are iid $N(0, \sigma_e^2)$

$(\beta_{0i}, \beta_{1i})'$ are iid $N((b_0, b_1)', \Sigma)$, with

$$\Sigma = \left[\begin{array}{cc} \sigma_c^2 & \rho\sigma_c\sigma_s \\ \rho\sigma_c\sigma_s & \sigma_s^2 \end{array} \right]$$

Subscripts "c", "s" are for inter$\underline{\text{c}}$epts β_{0i}, $\underline{\text{s}}$lopes β_{1i} respectively.

Example: Pig weights over 9 weeks (RWC)



The RL-maximizing estimates $(\hat{\sigma}_c^2, \hat{\sigma}_s^2, \hat{\rho})$ can be on the boundary of legal values:

$$\hat{\rho} = -1 \text{ or } +1, \text{ or } \hat{\sigma}_c^2 = 0, \text{ or } \hat{\sigma}_s^2 = 0$$

In my experience this happens often enough to be a real problem.

This research was motivated by a lot of datasets giving $\hat{\rho} = \pm 1$.

Is it even a problem that $\hat{\rho} = \pm 1$? Yes.

In the pig-weight example, $\hat{\rho} = -1$ means

“the slope and intercept for each piglet are perfectly anti-correlated in the population of pigs”

which is nonsense.

Isn't this just a software problem? No.

When $\hat{\rho} = \pm 1$, standard software gives useless or misleading information.

OK then ... examine the profiled log-likelihood or do a Bayesian analysis.

But we have no idea *why* boundary estimates happen. Multiplying the same function by a prior doesn't change that.

We need more understanding, not just more convenient software.

Why do we have so little understanding of our methods?

Our primary tool is math:

- Most of us are trained in math.
- You can't beat a theorem that establishes a useful fact.

But that's asking a lot:

- You must be able to prove a theorem, and
- it must establish a useful fact.

Asymptotic theorems do not establish useful facts.
Are finite-sample theorems even possible?

There are other ways to learn about our methods

This talk considers a tool for opening our black-box methods

modeled explicitly on

the approach molecular biologists use to open Nature's black boxes.

This tool would be a complement to math

As a matter of strategy,

we are better off doing something relatively simple (the molecular-biological approach) and learning something quickly

compared to

betting we can produce useful theorems in the long run.

A reasonable strategy would do some of each.

Here's what our colleagues in molecular biology do

1. Capture the phenomenon in a simple model system, e.g., an animal or cell-culture model.
2. Hypothesize about the phenomenon in terms of the model system.
3. Do experiments with the model system to test those hypotheses.
4. Iterate; revise the model system and hypotheses as needed.
5. Test the revised hypotheses in a more realistic *in vivo* system.

I'll demonstrate this approach by asking, for the RR model, what conditions make it likely that $\hat{\rho} = \pm 1$.

I am not the first person to suggest this

We have big designed experiments to measure operating characteristics:

- Larntz (JASA 1978): Massive simulation study of small-sample properties of three goodness-of-fit statistics for categorical data.
- JL Adams (1990): “Evaluating regression strategies” using split-plot designs, response-surface models, optimal design.

It’s harder to find simulation experiments testing explicit hypotheses:

- RE Schapire (2013 Vapnik Festschrift, 2015 JSM talk) “Explaining AdaBoost”, work with Leo Breiman.

Outline

I'll talk about each of the five steps in turn:

- Demonstrate it.
- Talk about its intellectual content.

I'll do this to gain some understanding of what conditions make it likely that $\hat{\rho} = \pm 1$ in the random-regressions model.

Step 1. The model system

For observation j in cluster i, model

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, s,$$

$$\epsilon_{ij} \sim \text{iid } N(0, \sigma_e^2)$$

$$(\beta_{0i}, \beta_{1i})' \sim \text{iid } N((b_0, b_1)', \Sigma),$$

$$\Sigma = \left[\begin{array}{cc} \sigma_c^2 & \rho\sigma_c\sigma_s \\ \rho\sigma_c\sigma_s & \sigma_s^2 \end{array} \right],$$

Cluster size is $s = 2m + 1$ for m a positive integer

In each cluster, x_{ij} takes the values in

$$\mathbf{h} = (-1, -(m-1)/m, \ldots, 0, \ldots, (m-1)/m, 1)'$$

The RL is straightforward, with these assumptions.

It's *relatively* simple but still too complicated for intuition.

Thus, I'm not going to show it.

To develop hypotheses, I simplified some more to produce a *predictor* of when $\hat{\rho} = -1$.

Intellectual content: As simple as possible . . .

In molecular biology:

- Benefits: Control inputs, measure outputs, isolate causal effects.
- Cost: Need to hedge on interpretation.

In statistical methodology:

- Benefits: More explicit derivations; simpler, faster computing.
- Cost: Maybe it omits important things.

Simplifications here:

- One regressor.
- All clusters have the same design matrix, with orthogonal columns.

Step 2. Generating hypotheses

A necessary condition for $\hat{\rho} = -1$ is:

$\partial/\partial\rho$ log RL or log profiled RL, evaluated at $\rho = -1$, < 0.

Using this condition, I derive a *predictor* of $\hat{\rho} = -1$, with these steps:

- simplify the log RL,
- profile out one unknown,
- get $\partial/\partial\rho$ of the profiled simplified log RL at $\rho = -1$,
- make more simplifications.

This has no intrinsic interest; it's just a device for generating hypotheses.

Simplify: Let $\sigma_c^2 = \sigma_s^2 \equiv \sigma_r^2$.

Profile out σ_r^2. The profiled log RL is a function of ρ and $r = \sigma_e^2/\sigma_r^2$.

Take $\partial/\partial\rho$ of the profiled simplified log RL at $\rho = -1$.

Simplify: Replace functions of the data by their expected values.

The predictor for $\hat{\rho} = -1$ ($\hat{\rho} = +1$ has a very similar predictor):

$$\left(\frac{Ns - N - 1}{(1 + r/s)(1 + r/q)}\right)\left[1 - \left(\frac{Ns - 2}{Ns - N - 1}\right)\frac{1 - \frac{N-1}{N(s-2)}\rho}{1 + \frac{2(N-1)}{N(s-2)}\frac{(1+r/s)(1+r/q)+\rho}{(1+r/s)(1+r/q)-1}}.\right]$$

Note $s = 2m + 1$, $q = (2m^2 + 3m + 1)/3m$ are about cluster size.

Easy to show: predictor > 0 for all legal N, s, r, and $\rho \in (-1, 1)$.

Generating hypotheses using the predictor

$$\left(\frac{Ns-N-1}{(1+r/s)(1+r/q)}\right)\left[1-\left(\frac{Ns-2}{Ns-N-1}\right)\frac{1-\frac{N-1}{N(s-2)}\rho}{1+\frac{2(N-1)}{N(s-2)}\frac{(1+r/s)(1+r/q)+\rho}{(1+r/s)(1+r/q)-1}}.\right]$$

Easy to prove:

- Given N, s, and ρ, as $r=\sigma_e^2/\sigma_r^2$ increases the predictor $\to 0$.
- Given ρ and r, as N or s increases, the predictor $\to \infty$.
- Given N, s, and r, as ρ goes to -1, the predictor $\to 0$.

Easy to show with simulations: small predictor $\Rightarrow$ high chance $\hat{\rho}=-1$
large predictor $\Rightarrow$ small chance $\hat{\rho}=-1$.

$\Rightarrow$ We have three hypotheses about $\hat{\rho}=-1$.

Developing more quantitative hypotheses

Let's exercise the predictor a little harder:

- Draw 1000 sets of (N, s, ρ, r).
- Compute $\log_{10}$ predictor for those 1000 sets.
- Analyze the results using main effects and interactions.

This gives some quantitative hypotheses:

- The effect of multiplying r by 2.5 is nullified by multiplying s by about 3 and N by about 5.
- Some changes in r are so large that no change in ρ can un-do them.

Intellectual content: The difference between good and brilliant scientists . . .

Hypothesis generation is one of the central creative activities of science.

Molecular biologists:

- Generate & test hypotheses by, e.g., using gene-knockout organisms.

Us:

- Generate hypotheses by using approximations to manipulate our equations and algorithms.

The math approach to methodology does generate hypotheses . . .

. . . they're called unproven conjectures.

Step 3. Test the hypotheses in simulation experiments

(1) To derive the predictor, I set $\sigma_c^2 = \sigma_s^2$.

The experiments *simulated* data by setting $\sigma_c^2 = \sigma_s^2$, but I *fit* models that allow different σ_c^2 and σ_s^2.

(2) I identified predictor values giving *some* "bad" estimates; this is a good region for testing hypotheses.

(3) Datasets simulated from the RR model: $\boldsymbol{\beta} = (0, 0)'$, $\sigma_c^2 = \sigma_s^2 = 1$.

(4) All analyses used the lmer function in the R package lme4.

A bit more about the experiments

(Recall $r = \sigma_e^2/\sigma_c^2 = \sigma_e^2/\sigma_s^2$ in the simulated data, tho' not the fit.)

The experiments are in three groups:

1. Increasing r produces bad estimates.
2. Trading off r against N, s, and ρ.
3. The effect of the true ρ.

Increasing *r* produces bad estimates

Experiment A				predictor		% with $\hat{\rho}$			
N	s	ρ	r	-1	+1	-1	+1	NaN	Bad
500	21	0	10^1	3.6e+2	3.6e+2	0	0	0	0
500	21	0	10^2	6.4e+0	6.4e+0	11	7	1	19
500	21	0	10^3	8.0e–2	8.0e–2	21	33	30	84
500	21	0	10^4	8.2e–4	8.2e–4	19	29	40	88
500	21	0	10^5	8.2e–6	8.2e–6	15	36	32	83

Experiment B				predictor		% with $\hat{\rho}$			
N	s	ρ	r	-1	+1	-1	+1	NaN	Bad
500	21	0.95	10^1	6.8e+2	1.5e+1	0	25	0	25
500	21	0.95	10^2	1.3e+1	2.6e–1	0	54	3	57
500	21	0.95	10^3	1.5e–1	3.1e–3	17	31	29	77
500	21	0.95	10^4	1.6e–3	3.2e–5	22	37	31	90
500	21	0.95	10^5	1.6e–5	3.2e–7	21	36	29	86

Points re Experiments A, B – 100 datasets/setting

$N = 500$ and $s = 21$ are a bit large, in my experience.

Given N, s, and r, the chance of a “bad” estimate is minimized by setting $\rho = 0$ (Expt A).

BUT for large r, ρ has no effect on $\hat{\rho} = -1$ (Expts A & B).

Trading off r against N, s, and ρ

Experiments C through F all have the same structure:

- Setting 1: r was chosen to give some “bad” estimates.
- Setting 2: r was changed by a factor of $10^{0.4}$.
- Settings 3, 4, and 5: Choose N, s, and ρ to change the predictor of $\hat{\rho} = -1$ back to its value in Setting 1.

400 datasets/setting except Experiment D, which had 600/setting.

Experiments C and D: Moderate N and s

Experiment C					predictor		% with $\hat{\rho}$			
setting	N	s	rho	r	-1	+1	-1	+1	NaN	Bad
1	100	9	-0.8	6.3	8.2	67	18	0	0	18
2	100	9	-0.8	15.8	1.7	15	39	0	2	41
3	500	9	-0.8	15.8	8.4	74	17	0	0	17
4	100	25	-0.8	15.8	8.7	76	20	0	0	20
5	100	9	0.0	15.8	8.2	8.2	8	6	17	30

Experiment D					predictor		% with $\hat{\rho}$			
setting	N	s	rho	r	-1	+1	-1	+1	NaN	Bad
1	100	9	-0.8	15.8	1.66	14.60	42	1	1	44
2	100	9	-0.8	6.3	8.23	67.47	19	0	0	19
3	21	9	-0.8	6.3	1.67	13.69	48	1	1	50
4	100	3	-0.8	6.3	1.83	15.14	42	0	0	42
5	100	9	-0.96	6.3	1.66	72.82	47	0	0	47

Experiments E and F: N, s have one large, one small

Experiment E					predictor		% with $\hat{\rho}$			
setting	N	s	rho	r	-1	+1	-1	+1	NaN	Bad
1	20	25	-0.8	6	9.88	79.92	23	0	0	23
2	20	25	-0.8	15	1.85	16.02	42	2	2	45
3	104	25	-0.8	15	10.00	86.64	15	0	0	15
4	20	63	-0.8	15	9.66	83.30	23	0	2	25
5	20	25	0.0	15	9.06	9.06	9	10	4	23

Experiment F					predictor		% with $\hat{\rho}$			
setting	N	s	rho	r	-1	+1	-1	+1	NaN	Bad
1	1000	3	-0.8	9	10.05	86.15	24	0	0	24
2	1000	3	-0.8	23	1.88	16.77	38	0	0	38
3	5350	3	-0.8	23	10.08	89.81	21	0	0	21
4	1000	9	-0.8	23	8.78	77.92	19	0	2	20
5	1000	3	0.0	23	9.36	9.36	10	6	0	16

Points re Experiments C, D, E, F

- I used $\rho = -0.8$ because $\rho < 0$ is more plausible than $\rho > 0$.

- These chosen N and s offset multiplying r by 2.5, $\approx$ as predicted:
  - s: multiple of 2.8, 3, 2.5, 3.
  - N: multiple of 5, 4.8, 5.2, 5.4.

- ρ does not trade off against r as the predictor predicted.
  - Predictor: Increasing ρ from -0.8 to 0 offsets the change in r.
  - Experiments: Fewer $\hat{\rho} = -1$, but *more* $\hat{\rho} = +1$ or NaN.

Experiment G: More on the effect of ρ

All settings have the same N and s.

Each block of 5 settings has one r and ρ ranging from -0.95 to $+0.95$.

400 datasets/setting

Experiment G – smallish r

					predictor		% with $\hat{\rho}$			
setting	N	s	rho	r	-1	+1	-1	+1	NaN	Bad
1	500	21	-0.95	53	1.00	38.65	45	0	7	52
2	500	21	-0.50	53	9.96	29.78	10	0	10	20
3	500	21	0.00	53	19.89	19.89	3	1	1	5
4	500	21	0.50	53	29.78	9.96	0	12	1	13
5	500	21	0.95	53	38.65	1.00	0	47	1	48

But look what happens when r increases

Experiment G – larger r

					predictor		% with $\hat{\rho}$			
setting	N	s	rho	r	-1	+1	-1	+1	NaN	Bad
6	500	21	-0.95	271	0.05	1.94	45	8	11	64
7	500	21	-0.50	271	0.50	1.50	30	17	15	62
8	500	21	0.00	271	1.00	1.00	23	27	9	58
9	500	21	0.50	271	1.50	0.50	13	40	8	61
10	500	21	0.95	271	1.94	0.05	6	52	5	64
11	500	21	-0.95	3000	4.4e-4	1.7e-2	23	30	31	83
12	500	21	-0.50	3000	4.4e-3	1.3e-2	21	33	30	84
13	500	21	0.00	3000	8.9e-3	8.9e-3	22	31	31	83
14	500	21	0.50	3000	1.3e-2	4.4e-3	22	34	29	84
15	500	21	0.95	3000	1.7e-2	4.4e-4	20	35	25	80

Settings 16-20 use $r = 100{,}000$ and are effectively identical to $r = 3{,}000$.

Comments on Experiment G

For large enough r, the true ρ doesn't matter.

There's an oddity:

- When r is large, we'd expect $\hat{\rho} = -1$ and $\hat{\rho} = +1$ about equally often, but $\hat{\rho} = +1$ is more frequent.

Intellectual content: Hypothesis-driven simulation experiments

Explicit hypotheses about statistical methods suggest non-standard simulation designs.

“Attractive ideas, after all, are cheap and much of the stuff of scientific genius is devising tests” (Judson 1979)

- e.g., Meselson & Stahl, separating macromolecules by buoyant density.

Perhaps we don’t see the creative potential of simulation experiments because we make such limited use of them.

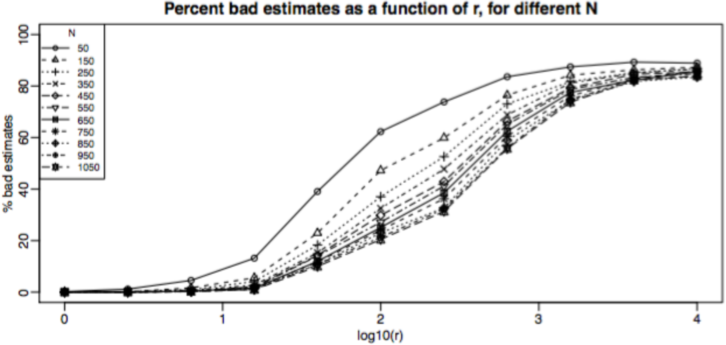
Step 4. Iterate; revise the model system and hypotheses.

Re Hypothesis 7, “Changes in ρ have a smaller effect on $\hat{\rho} = -1$ than changes in *N* or *s*”, the experiments say “you asked the question poorly”.

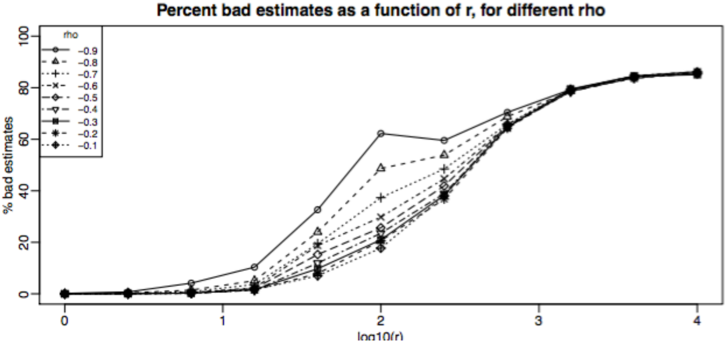
Our original question, “What conditions make it likely that $\hat{\rho} = -1$?”, was a red herring.

The disease is poor resolution; the different kinds of bad estimate are merely different symptoms.

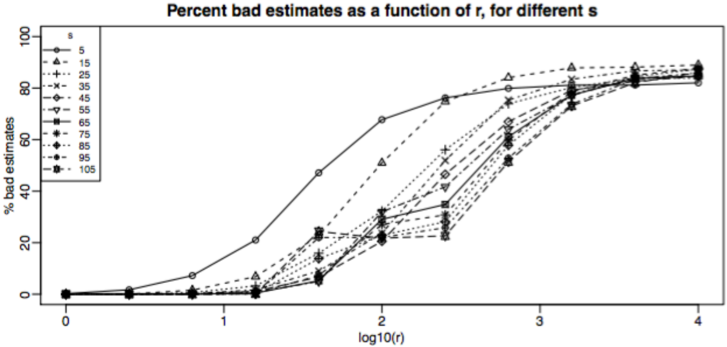
Final experiment: N-by-r interaction



ρ-by-r interaction



s-by-*r* interaction



Intellectual content: The most important step?

In the theory used to teach us, a hypothesis is stated and tested and the story ends.

This does not describe scientific practice.

Experiments designed to test particular hypotheses often motivate a reformulation of the structure of hypotheses.

This may be the most important result of a set of experiments.

Step 5. "An *in vivo*" experiment

Can the experimental results can be reproduced in a natural situation?

HMO data (Hodges 1998): y_{ij} = premium for health plan j in state i

Fit the model

$$\begin{aligned} y_{ij} = \quad & b_0 + b_1(\log_{10} \text{families})_{ij} + \ [1, (\log_{10} \text{families})_{ij}](u_{i0}, u_{i1})' \\ & + b_{0E} \ (\text{expenses per admission})_i + b_{0N}(\text{New England})_i + \epsilon_{ij} \end{aligned}$$

$$(u_{i0}, u_{i1}) \sim N(\mathbf{0}, \Sigma), \text{ and } \epsilon_{i,j} \sim \text{iid } N(0, \sigma_e^2)$$

$$(\hat{b}_0, \hat{b}_1, \hat{b}_{0E}, \hat{b}_{0N}) = (180, -2.2, 4.8, 16)$$

$$(\hat{\rho}, \hat{\sigma}_e^2, \hat{\sigma}_c^2, \hat{\sigma}_s^2) = (0.12, 487, 98, 5.34)$$

This differs from the model system

- The fit has non-zero fixed effects and “extra” fixed effects.
- Within-state sample size varies: n_i ranges from 1 to 31, median 5.
- State-specific design matrices vary.
- $(\log_{10}\text{families})_{ij}$ isn’t scaled to make $\sigma_c^2 \approx \sigma_s^2$.
- ϵ_{ij} actually have non-constant variance and are right skewed.

Test our findings by inflating the error variance

Define the artificial datum $y(\phi)_{ij}$ as

$$\begin{aligned} y(\phi)_{ij} = &\quad \text{fit}_{ij} + \phi \hat{\epsilon}_{ij} \\ \hat{\epsilon}_{ij} = &\quad y_{ij} - \text{fit}_{ij} \\ \text{fit}_{ij} = &\quad \hat{b}_0 + \hat{b}_1 (\log_{10} \text{families})_{ij} + \ [1, (\log_{10} \text{families})_{ij}](\hat{u}_{i0}, \hat{u}_{i1})' \\ &\quad + \hat{b}_{0E} \text{ (expenses per admission)}_i + \hat{b}_{0N} \text{ (New England)}_i \end{aligned}$$

$(\hat{u}_{i0}, \hat{u}_{i1})$ are the EBLUPs.

$\phi = 1$ is the real data; $\phi > 1$ is fake data with inflated errors.

Increasing error variance has the predicted effect

ϕ	$\hat{\rho}$	$\hat{\sigma}_e^2$	$\hat{\sigma}_e^2/\phi^2$	$\hat{\sigma}_c^2$	$\hat{\sigma}_s^2$
1.0	0.115	487	487	97.73	5.39
1.1	0.164	590	488	94.13	5.25
1.2	0.230	704	489	88.99	4.97
1.3	0.320	828	490	82.36	4.55
1.4	0.444	963	491	74.32	3.98
1.5	0.626	1108	492	65.05	3.26
1.6	0.920	1263	493	54.77	2.39
1.7	1.000	1427	494	44.68	2.68
1.8	1.000	1599	494	34.73	3.21
1.9	1.000	1781	493	25.14	3.58
2.0	1.000	1972	493	16.36	3.62
2.1	1.000	2171	492	8.97	3.16
2.2	1.000	2379	492	3.54	2.02
2.3	1.000	2594	490	0.23	0.22
2.4	NaN	2814	489	0	5×10^{-12}
2.5	NaN	3043	487	0	4×10^{-13}

Conclusions about the random regressions model

Bad estimates are a symptom of large error variation.

The number of clusters N and cluster size s matter when σ_e^2 is middling.

Implications:

- Experimental design: Increasing s pays more than increasing N.
- Model choice: A bad estimate suggests omitting the random effect.

"Too many" $\hat{\rho} = \pm 1$ is probably an artifact of a very flat RL.

Statistical methods research done as science

Results like this could never be discovered using asymptotic methods.

Only simulation experiments could reveal "too many" $\hat{\rho} = \pm 1$.

We produced useful facts with math and computing exercises that could be executed by a capable Master's student under faculty supervision.

The results ***are*** useful; the design of each step can be a contribution.

If utility has merit, empirical studies of methods merit publication as much as theorems.