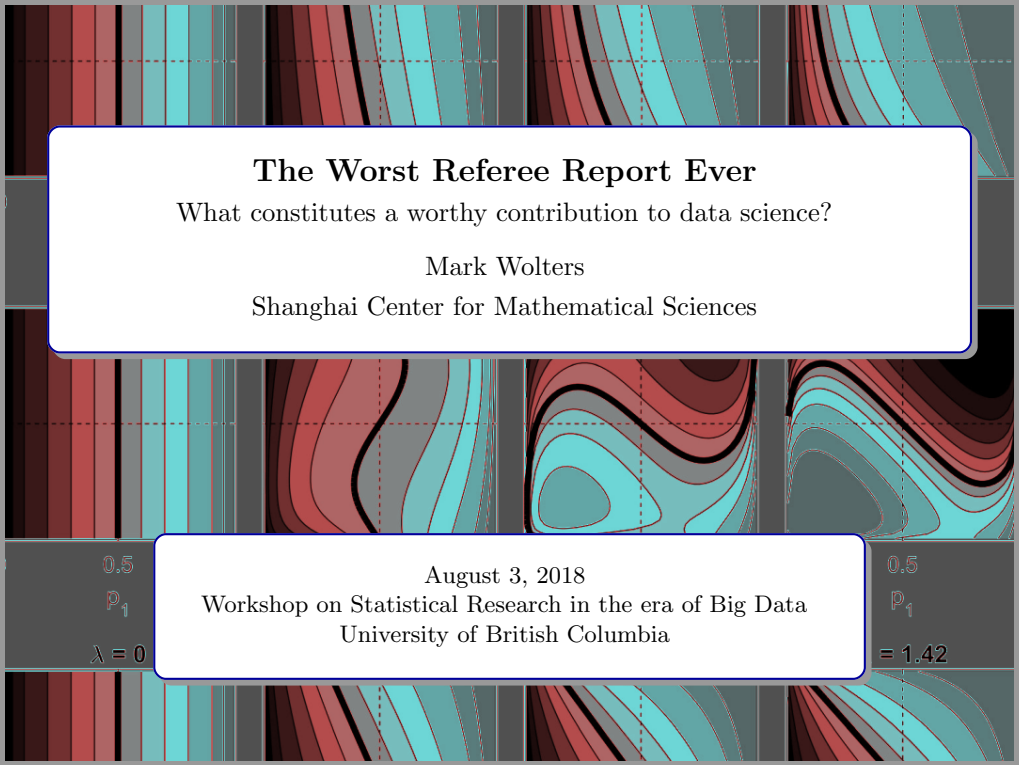


The Worst Referee Report Ever

What constitutes a worthy contribution to data science?

Mark Wolters

Shanghai Center for Mathematical Sciences

August 3, 2018
Workshop on Statistical Research in the era of Big Data
University of British Columbia

The paper

frontiers in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 27 November 2017
doi: 10.3389/fams.2017.00024

Better Autologistic Regression

*Mark A. Wolters**

Shanghai Center for Mathematical Sciences, Fudan University, Shanghai, China

Autologistic regression is an important probability model for dichotomous random variables observed along with covariate information. It has been used in various fields for analyzing binary data possessing spatial or network structure. The model can be viewed as an extension of the autologistic model (also known as the Ising model, quadratic exponential binary distribution, or Boltzmann machine) to include covariates. It can also be viewed as an extension of logistic regression to handle responses that are not independent. Not all authors use exactly the same form of the autologistic regression model. Variations of the model differ in two respects. First, the variable coding—the two numbers used to represent the two possible states of the variables—might differ. Common coding choices are (zero, one) and (minus one, plus one). Second, the model might appear in either of two algebraic forms: a standard form, or a recently proposed centered form. Little attention has been paid to the effect of these differences, and the literature shows ambiguity about their importance. It is shown here that changes to either coding or centering in fact produce distinct, non-nested probability models. Theoretical results, numerical studies, and analysis of an ecological data set all show that the differences among the models can be large and practically significant. Understanding the nature of the differences and making appropriate modeling choices can lead to significantly improved autologistic regression analyses. The results strongly suggest that the standard model with plus/minus coding, which we call the symmetric autologistic model, is the most natural choice among the autologistic variants.

Keywords: probabilistic graphical models, Markov random fields, logistic regression, correlated binary random variables, spatial statistics

OPEN ACCESS

Edited by:
Hau-tieng Wu,
Duke University, United States

Reviewed by:
John Hughes,
University of Colorado Denver, United States
Antonio Calcagnì,
University of Trento, Italy

****Correspondence:***
Mark A. Wolters
mwolters@fudan.edu.cn

Specialty section:
This article was submitted to Mathematics of Computation and Data Science, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *22 March 2017*
Accepted: *10 November 2017*
Published: *27 November 2017*

Citation:
Wolters MA (2017) Better Autologistic Regression. Front. Appl. Math. Stat. 3:24. doi: 10.3389/fams.2017.00024

1. INTRODUCTION

The autologistic (AL) model is a probabilistic graphical model for multivariate binary data. It was introduced to the statistical literature by Besag [1, 2] and has also been developed by Kaiser and Cressie [3]. The same model appeared much earlier in statistical physics, where it is known as the Ising model (see e.g., [4, 5]). It has been used extensively in image processing (e.g., [6–8]), and is also the model underlying the Boltzmann machine [9]. The same model has also been called the quadratic exponential binary distribution [10, 11], and, under that name, it has been described as the binary-variable analog of the multivariate normal distribution ([12]; see also [13]). As such, one may anticipate that the autologistic model will become increasingly useful as the number of complex, graph-structured data sets continues to grow.

When binary responses are observed along with covariate information, the autologistic model may be extended to become the autologistic regression (ALR) model. This model can be viewed as a natural extension of ordinary logistic regression to handle cases where responses are not independent. Under the ALR model, the responses follow an autologistic distribution, and the

2017. Better Autologistic Regression, *Frontiers in Applied Mathematics and Statistics*

What is "data science?"

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2017, VOL. 26, NO. 4, 745–766
https://doi.org/10.1080/10618600.2017.1384734

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

50 Years of Data Science

David Donoho

Department of Statistics, Stanford University, Standford, CA

ABSTRACT

More than 50 years ago, John Tukey called for a reformation of academic statistics. In "The Future of Data Analysis," he pointed to the existence of an as-yet unrecognized science, whose subject of interest was learning from data, or "data analysis." Ten to 20 years ago, John Chambers, Jeff Wu, Bill Cleveland, and Leo Breiman independently once again urged academic statistics to expand its boundaries beyond the classical domain of theoretical statistics; Chambers called for more emphasis on data preparation and presentation rather than statistical modeling; and Breiman called for emphasis on prediction rather than inference. Cleveland and Wu even suggested the catchy name "data science" for this envisioned field. A recent and growing phenomenon has been the emergence of "data science" programs at major universities, including UC Berkeley, NYU, MIT, and most prominently, the University of Michigan, which in September 2015 announced a $100M "Data Science Initiative" that aims to hire 35 new faculty. Teaching in these new programs has significant overlap in curricular subject matter with traditional statistics courses; yet many academic statisticians perceive the new programs as "cultural appropriation." This article reviews some ingredients of the current "data science moment," including recent commentary about data science in the popular media, and about how/whether data science is really different from statistics. The now-contemplated field of data science amounts to a superset of the fields of statistics and machine learning, which adds some technology for "scaling up" to "big data." This chosen superset is motivated by commercial rather than intellectual developments. Choosing in this way is likely to miss out on the really important intellectual event of the next 50 years. Because all of science itself will soon become data that can be mined, the imminent revolution in data science is not about mere "scaling up," but instead the emergence of scientific studies of data analysis science-wide. In the future, we will be able to predict how a proposal to change data analysis workflows would impact the validity of data analysis across all of science, even predicting the impacts fields-by-field. Drawing on work by Tukey, Cleveland, Chambers, and Breiman, I present a vision of data science based on the activities of people who are "learning from data," and I describe an academic field dedicated to improving that activity in an evidence-based manner. This new field is a better academic enlargement of statistics and machine learning than today's data science initiatives, while being able to accommodate the same short-term goals. *Based on a presentation at the Tukey Centennial Workshop, Princeton, NJ, September 18, 2015.*

ARTICLE HISTORY
Received August 2017
Revised August 2017

KEYWORDS
Cross-study analysis; Data analysis; Data science; Meta analysis; Predictive modeling; Quantitative programming environments; Statistics

1. Today's Data Science Moment

In September 2015, as I was preparing these remarks, the University of Michigan announced a $100 million "Data Science Initiative" (DSI)[1], ultimately hiring 35 new faculty.

The university's press release contains bold pronouncements:

> *"Data science has become a fourth approach to scientific discovery, in addition to experimentation, modeling, and computation," said Provost Martha Pollack.*

The website for DSI gives us an idea what data science *is*:

> *"This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and inter-disciplinary applications."*

This announcement is not taking place in a vacuum. A number of DSI-like initiatives started recently, including

(A) Campus-wide initiatives at NYU, Columbia, MIT, ...
(B) New master's degree programs in data science, for example, at Berkeley, NYU, Stanford, Carnegie Mellon, University of Illinois, ...

There are new announcements of such initiatives weekly.[2]

2. Data Science "Versus" Statistics

Many of my audience at the Tukey Centennial—where these remarks were originally presented—are applied statisticians, and consider their professional career one long series of exercises in the above "*....collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of ....applications.*" In fact, some presentations at the Tukey Centennial were exemplary narratives of "*....collection, management, processing, analysis,*

CONTACT David Donoho donoho@stanford.edu
[1]For a compendium of abbreviations used in this article, see Table 1.
[2]For an updated interactive geographic map of degree programs, see http://data-science-university-programs.silk.co

Donoho (2017, JCGS), ***50 Years of Data Science***:

- "***The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly.***" [Quoting Cleveland (2001)]
- "***. . . a litmus test re Statistical theorists: do they 'care about the data analyst' or do they not?***"

What is "data science?"

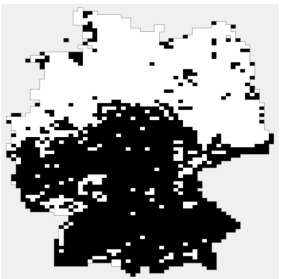
Donoho's 6 components of "greater data Science"

The activities of GDS are classified into six divisions:

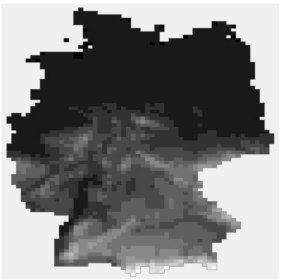
1. *Data Gathering, Preparation, and Exploration*
2. *Data Representation and Transformation*
3. *Computing with Data*
4. *Data Modeling*
5. *Data Visualization and Presentation*
6. *Science about Data Science*

The paper: autologistic regression (1/13)

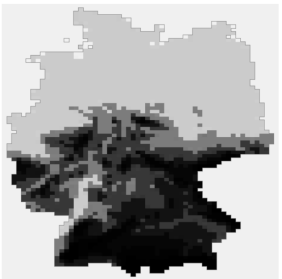
Example: *H. vulgaris* data
(Carl & Kühn, 2007, *Ecological Modeling*; Bardos et al. 2015 arXiv)



$\mathbf{z}$:
presence/absence



$\mathbf{x}_1$:
altitude



$\Pr(Z_i = 1|x_i)$,
logistic regression

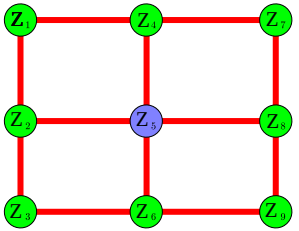
The paper: autologistic regression (2/13)

- Dichotomous data with predictors
- Local/spatial association
- The applications involve data on a grid (more generally, graph)

The ***autologistic regression (ALR) model*** is a pairwise ***Markov random field*** (MRF) of dichotomous random variables, with a linear predictor.

Applications: ecology, computer vision, dentistry, anthropology, materials science, ...

The paper: autologistic regression (3/13)



Let $\mathbf{Z}$ be a vector of dichotomous random variables.

- Autologistic (AL) model is an MRF.
- Undirected graph with adjacency matrix $\mathbf{A}$.
- PMF:

$$f_{\mathbf{Z}}(\mathbf{z}) \propto \exp\left(\underbrace{\mathbf{z}^T\boldsymbol{\alpha}}_{\textbf{unary term}} + \underbrace{\frac{1}{2}\mathbf{z}^T\boldsymbol{\Lambda}\mathbf{z}}_{\textbf{pairwise term}}\right)$$

- Let $\pi_i = \Pr(Z_i = \text{high} \mid \text{neighbours})$. Conditional form of the model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_i + \sum_{j\sim i}\lambda_{ij}z_j$$

- Let $\boldsymbol{\alpha} = \mathbf{X}\boldsymbol{\beta}$ ➡ autologistic *regression*
- Let $\boldsymbol{\Lambda} = \lambda\mathbf{A}$ ➡ "simple" form of the model

The paper: autologistic regression (4/13)

1. The STANDARD model: $\mathbf{Z} \in \{0,1\}^n$

general form	simple form
$f_{\mathbf{Z}}(\mathbf{z}) \propto \exp(\mathbf{z}^T\mathbf{X}\boldsymbol{\beta} + \frac{1}{2}\mathbf{z}^T\boldsymbol{\Lambda}\mathbf{z})$	$f_{\mathbf{Z}}(\mathbf{z}) \propto \exp(\mathbf{z}^T\mathbf{X}\boldsymbol{\beta} + \frac{\lambda}{2}\mathbf{z}^T\mathbf{A}\mathbf{z})$
$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta} + \sum_{j\sim i}\lambda_{ij}z_j$	$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta} + \lambda\sum_{j\sim i}z_j$

The paper: autologistic regression (5/13)

2. The CENTERED model ($\mathbf{Z} \in \{0,1\}^n$)

- Traditional model has a problem
  - Fix $\boldsymbol{\beta}$, increase λ, you will find $Z = 1$ everywhere.
  - Why? Because $\sum_{j \sim i} z_j$ is never negative.
- Caragea & Kaiser (2009): "***centered parametrization***":

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta} + \sum_{j\sim i}\lambda_{ij}(z_j - \mu_j), \quad \text{where} \quad \mu_j = \frac{e^{\mathbf{x}_j^T\boldsymbol{\beta}}}{1-e^{\mathbf{x}_j^T\boldsymbol{\beta}}}$$

- μ_j is the independence expectation of the Z_j

The paper: autologistic regression (6/13)

3. The SYMMETRIC model

- The responses are ***categorical***. Don't have to use $\{0, 1\}$ coding.
- In general, could use $\{\ell, h\}$.
- If $\mathbf{Z}$ has support $\{\ell, h\}^n$,

$$\mathbf{Y} = a\mathbf{Z} + b\mathbf{1}, \qquad \text{where} \quad a = \frac{H - L}{h - \ell}, \quad b = L - a\ell$$

has support $\{L, H\}^n$.

The ***symmetric model*** is the standard model, with $\mathbf{Z} \in \{-h, h\}^n$

- No centering
- Coding symmetric around 0

- We shouldn't change coding without thinking...

The paper: autologistic regression (7/13)

Say $\mathbf{Z} \in \{\ell, h\}^n$, with $f_{\mathbf{Z}}(\mathbf{z}) \propto g(\mathbf{z}; \boldsymbol{\theta})$,
but we want our model to use coding $\{L, H\}$.

The right way

$\mathbf{Y} = a\mathbf{Z} + b\mathbf{1} \iff \mathbf{Z} = \frac{1}{a}\mathbf{Y} - \frac{b}{a}\mathbf{1}$

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \Pr(\mathbf{Y} = \mathbf{y}) \\ &= \Pr(a\mathbf{Z} + b\mathbf{1} = \mathbf{y}) \\ &= f_{\mathbf{Z}}(\tfrac{1}{a}\mathbf{y} - \tfrac{b}{a}\mathbf{1}) \\ &\propto g(\tfrac{1}{a}\mathbf{y} - \tfrac{b}{a}\mathbf{1}; \boldsymbol{\theta}) \\ &\propto g(\mathbf{z}; \boldsymbol{\theta}). \end{aligned}$$

The tempting way

Just plug in $\mathbf{y} = a\mathbf{z} + b\mathbf{1}$.

Let the parameter be $\boldsymbol{\theta}'$.

$$\begin{aligned} f'_{\mathbf{Y}} &\propto g(\mathbf{y}; \boldsymbol{\theta}') \\ &\propto g(a\mathbf{z} + b\mathbf{1}; \boldsymbol{\theta}') \end{aligned}$$

To achieve $f'_{\mathbf{Y}} = f_{\mathbf{Y}}$, we need $\boldsymbol{\theta}'$ to compensate for linear transformation of $\mathbf{z}$.

The paper: autologistic regression (8/13)

- Derive the model for arbitrary $\{\ell, h\}$ coding, we find

$$\text{logit}\left(\Pr(Z_i = h|\mathbf{Z}_{-i})\right) = (h-\ell)\left[\mathbf{x}_i^T\boldsymbol{\beta} + \sum_{j\sim i}\lambda_{ij}(z_j - \mu_j)\right]$$

where

$$\mu_j = \begin{cases} \mathbf{0} & \text{for a standard model} \\ \dfrac{\ell e^{\ell\alpha_i} + h e^{h\alpha_i}}{e^{\ell\alpha_i} + e^{h\alpha_i}} & \text{for a centered model} \end{cases}$$

- Negpotential function:

$$Q(\mathbf{z}) = \mathbf{z}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{z}^T\boldsymbol{\Lambda}\boldsymbol{\mu} + \frac{1}{2}\mathbf{z}^T\boldsymbol{\Lambda}\mathbf{z}$$

- With this, we can study the effect of coding & centering changes.

The paper: autologistic regression (9/13)

Refer to any particular choice of coding and centering as a ***variant*** of the model. Are all variants equivalent?

Theorem 1: ***All AL variants are equivalent*** to any standard model.

Theorem 2: ***ALR variants are not equivalent***, in general.

- ➡ Many variants, all called "autologistic regression models," are ***actually different, non-nested distribution families***.
- ➡ Exception: all symmetric models are equivalent.

(***Equivalence***: parameter settings always exist that make the two models assign the same probabilities to every configuration of $\mathbf{Z}$.)

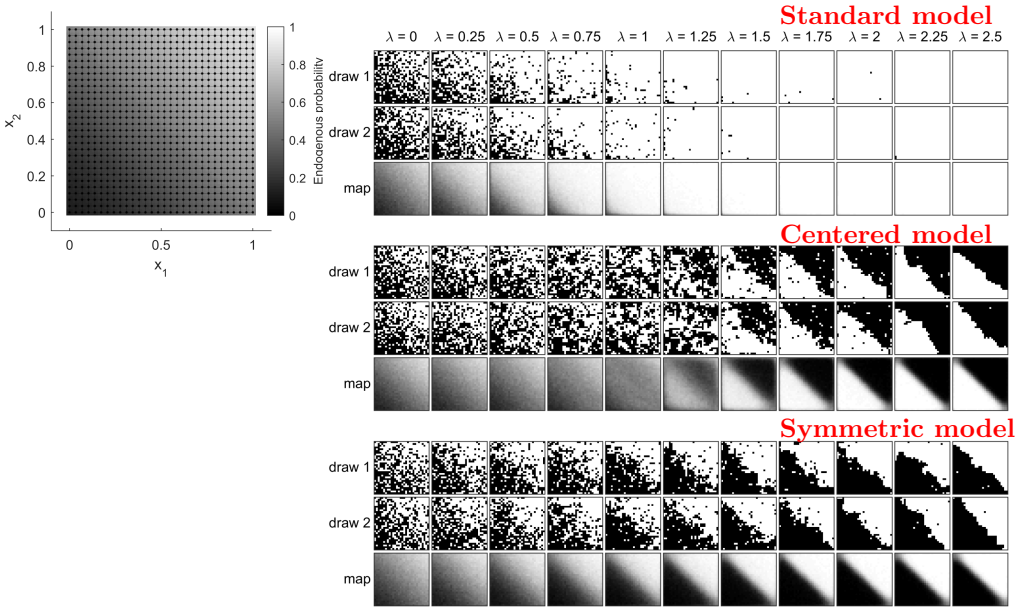
The paper: autologistic regression (10/13)

Theorem 3: *Only the symmetric models have reasonable large-association behaviour.*

"Simple" model. Let λ increase.

- Centered variants behave ***counterintuitively*** when λ large.
- Symmetric variants are the ***only ones*** with reasonable behaviour as $\lambda \to \infty$.

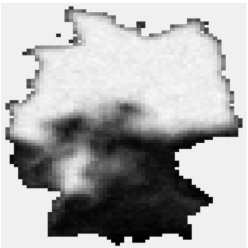
The paper: autologistic regression (11/13)



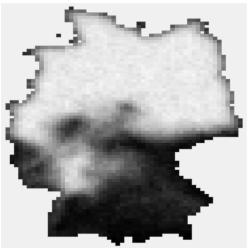
The paper: autologistic regression (12/13)

H. vulgaris fitted models' marginal probabilities

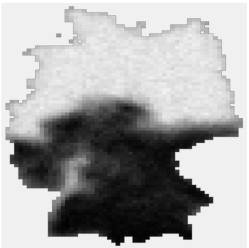
traditional



centered



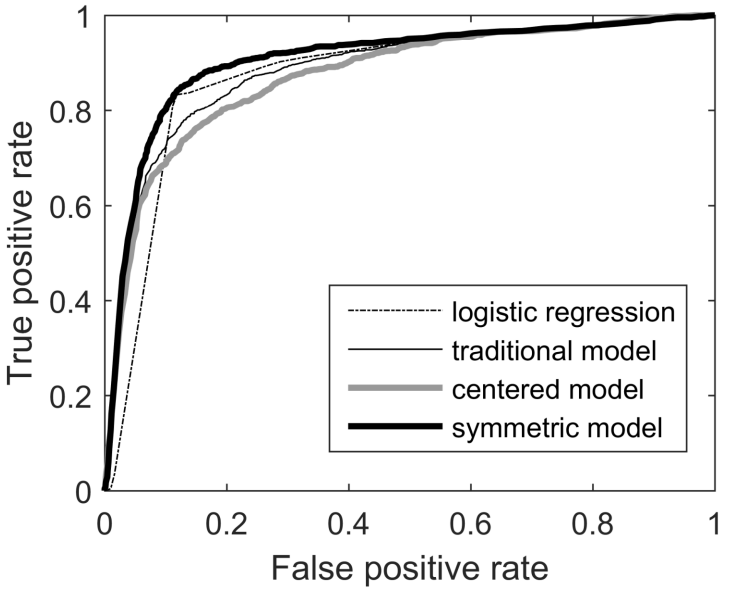
symmetric



Model	$\hat{\beta}_0$ (SE)	$\hat{\beta}_1$ (SE)	$\hat{\lambda}$ (SE)
logistic	2.78 (0.10)	−0.79 (0.028)	
traditional	−2.12 (0.22)	−0.16 (0.026)	1.43 (0.066)
centered	−1.74 (0.31)	−0.17 (0.040)	1.51 (0.050)
symmetric	0.50 (0.11)	−0.13 (0.029)	1.43 (0.071)

The paper: autologistic regression (13/13)

H. vulgaris
ROC curves



Referee reaction

Main conclusions:
- $\{-1, 1\}$ coding is much preferred over $\{0, 1\}$.
- The centered model has fundamental problems.

➡ Most prior ALR analyses are questionable.

➡ We need to change the standard of practice.

What were referee reactions?

Referee A

"Lovely"
"Surprising"...
"One of the best papers on the subject."

Referee reaction

Referee B agreed that:

- The results do not appear in the literature
- The results are correct
- The $\{-1, 1\}$ model is superior

BUT, recommended rejection ***"both for its unmotivated purpose of study and for its lack of technical sophistication."***

After revision:

- It was claimed (six times) that the paper lacked ***"intellectual merits"***.
- The use of ***"precalculus"*** math was listed (four times) as indication of the paper's low quality.
- In two places it was suggested that the paper reflects my lack of understanding of the state of the art

Referee reaction

Some quotes:

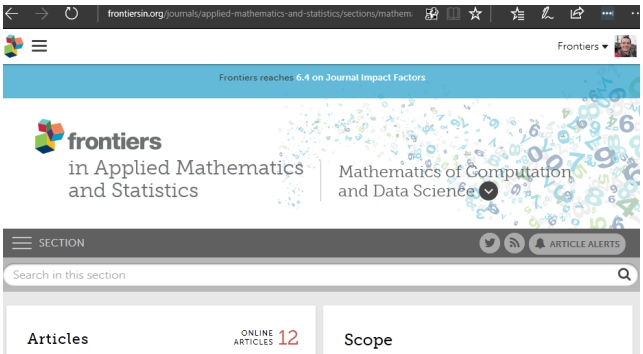
"There is a difference between creative research and a collection of ***mathematically correct but trivial, easy-to-obtain results*** that expand into the volume of a research article."

"... this paper is a ***summarization of trivial facts supported by unsophisticated mathematics*** intentionally expanded to create the illusion of undeserved mathematical complication."

"In the reviewer's career, it is rare to witness a paper with the quality of the current manuscipt published on professional academic journals. ... the reviewer considers the outcomes of publishing re-explanations of common knowledge in shallow mathematics—even without objective error due to the simplistic nature of the technical arguments—***catastrophic, since it will prevent original and innovative research from reaching their target readers."***

So how did it get published?

Fortunately, *Frontiers in Applied Mathematics and Statistics* has progressive policies.



- Open access.
- Rapid process.
- Review is structured, focused on correctness.
- Can communicate directly with reviewers.
- Reviewers' identity published with the paper.

Between two worlds?

"Big data"/"data science"/"analytics":

- CS/EE culture
- proceedings
- prediction
- software; utility

"traditional statistics":

- Math culture
- traditional journals
- estimation & inference
- theory, generality, rigor

Possible remedies?

My *personal* plan:
1. Give up trying to impress the math culture.
2. Instead, focus on actual impact, within my abilities.
3. Hope that academia's attitudes about performance catch up.

How to change focus to "actual impact"?

- Broaden publication targets. All papers are random access anyway.
  - Open access
  - Rapid, correctness-focused review
  - Newer, data-science focused journals
- Alternative outputs
  - Quality software! (all my papers: 51 citations; my 2 R packages: 200+ downloads per month)
  - Publish data sets
- Collaborative, applied work... solve actual problems.

The big challenge

Implementation is hard. Software development takes time.

The basic formula for academic reputation:

- Where you work
- How many papers
- In which journals

Moore-Sloan Data Science Environments (NYU, UCB, UW... http://msdse.org/)
"dramatically advance data-intensive scientific discovery..."

"data science in research universities requires precisely the kind of complex, long-term interdisciplinary work with methodological and engineering efforts that leads to low performance under traditional metrics and slow progress and lack of fit in existing career tracks."