Statistical Modeling: Bigger and Bigger

David Madigan

Rutgers University

stat.rutgers.edu/~madigan

“in data analysis there is no loner any problem of computation”

- Benzécri, 2005

Logistic Regression

- Linear model for log odds of category membership:

$$\log \frac{p(y=1 \mid \boldsymbol{x}_i)}{p(y=-1 \mid \boldsymbol{x}_i)} = \sum \beta_j \, x_{ij} = \boldsymbol{\beta} \boldsymbol{x}_i$$

Maximum Likelihood Training

- Choose parameters (β_j's) that maximize probability (likelihood) of class labels (y_i's) given documents ($\boldsymbol{x}_i$'s)

$$L(\boldsymbol{\beta}) = p(\boldsymbol{\beta}|D) = (\prod_{i=1}^{n} \frac{1}{1+\exp(-\boldsymbol{\beta}^T \boldsymbol{x}_i y_i)})$$

- Tends to overfit
- Not defined if $d > n$
- Feature selection

Shrinkage Methods

- Avoid combinatorial challenge of feature selection
- L1 shrinkage: regularization + feature selection
- Expanding theoretical understanding
- Large scale
- Empirical performance

Ridge Logistic Regression

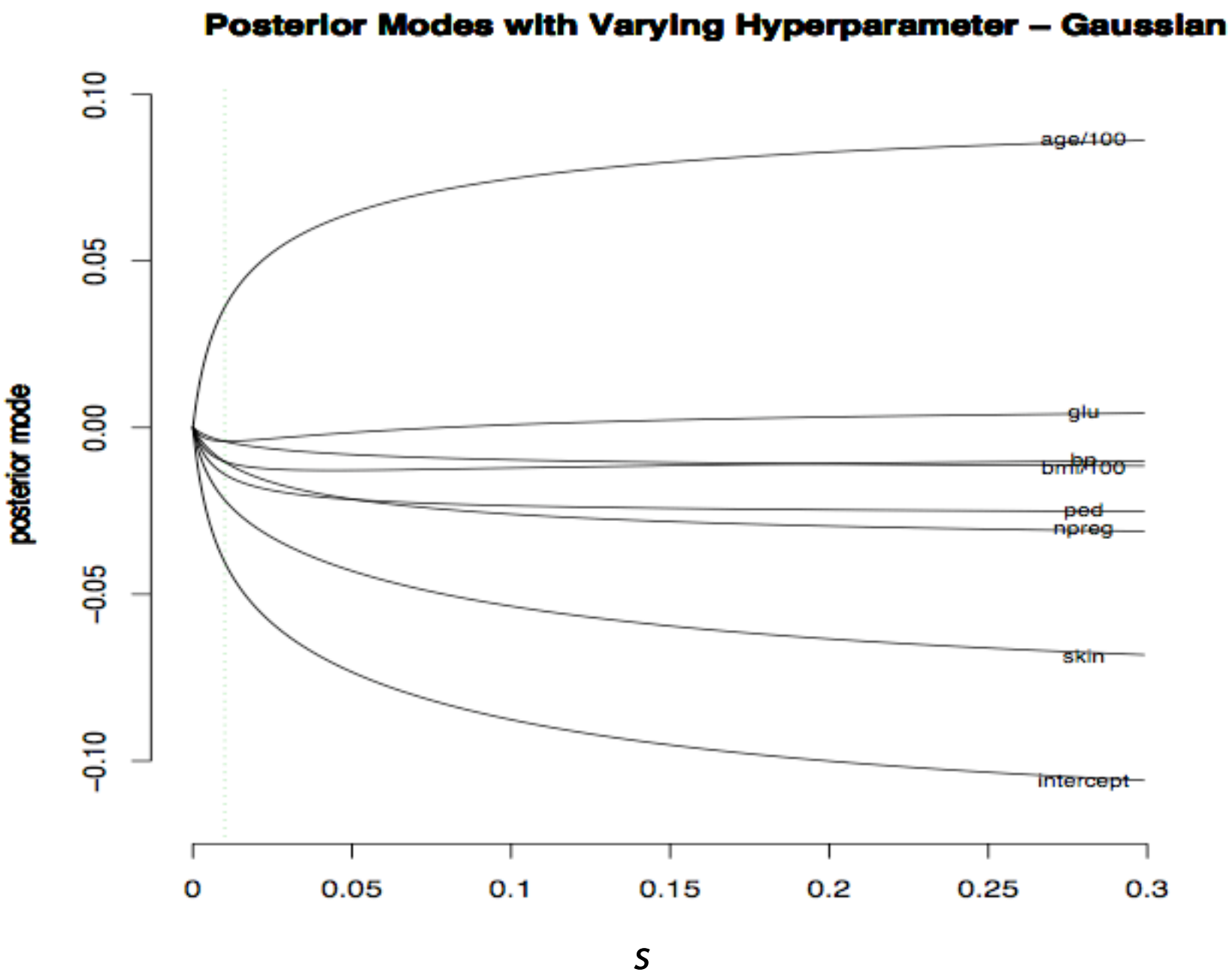
Maximum likelihood plus a constraint:

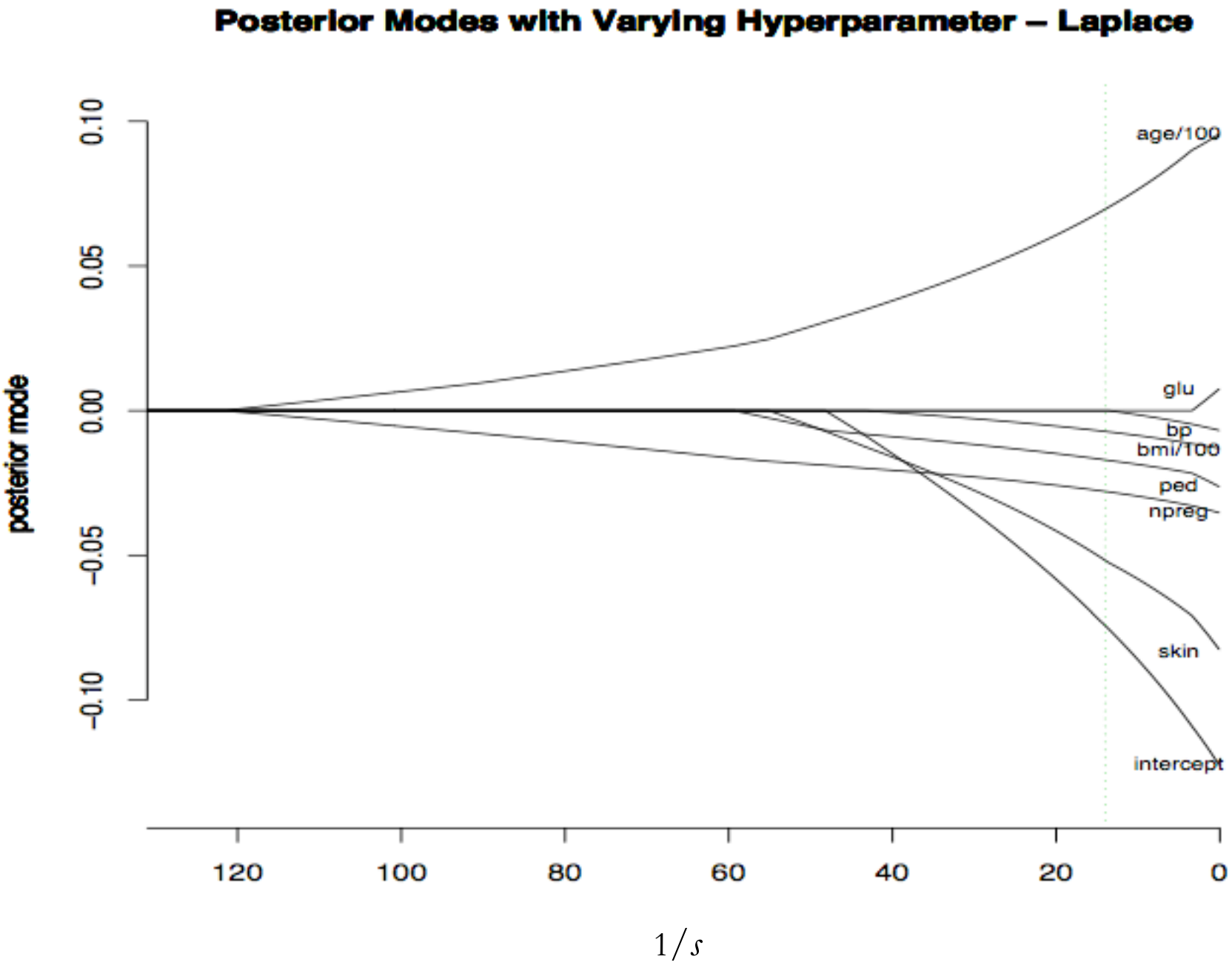
$$\sum_{j=1}^{p} \beta_j^2 \leq s$$

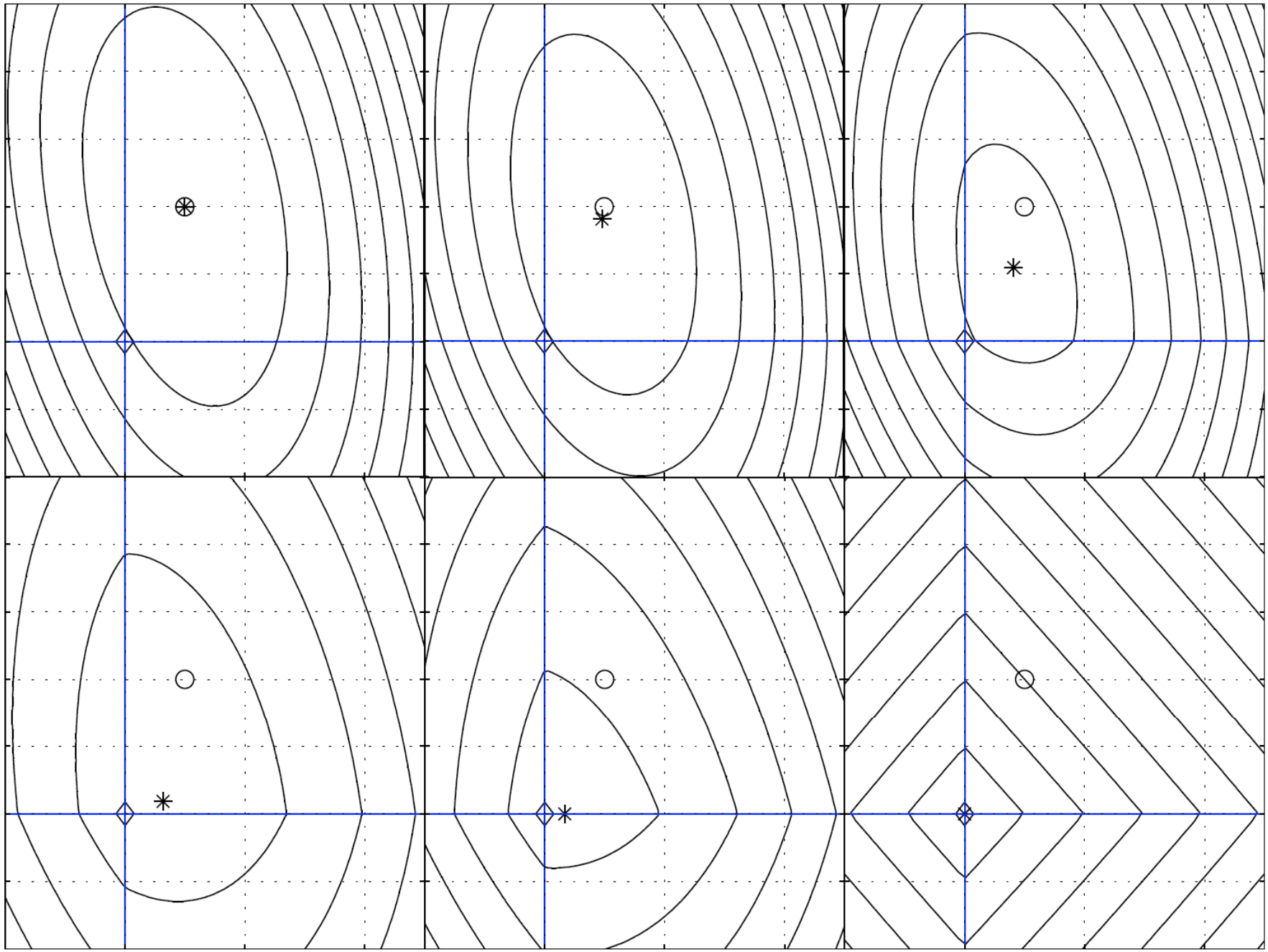
Lasso Logistic Regression

Maximum likelihood plus a constraint:

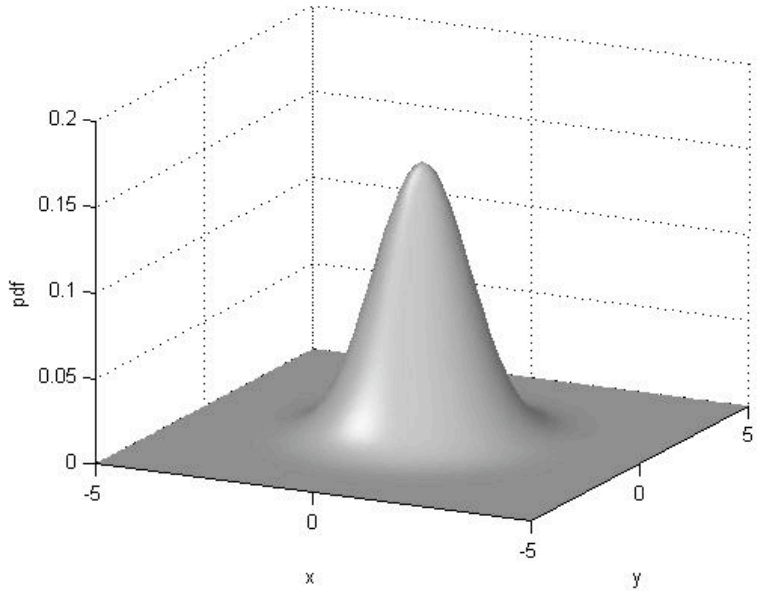
$$\sum_{j=1}^{p} \left| \beta_j \right| \leq s$$

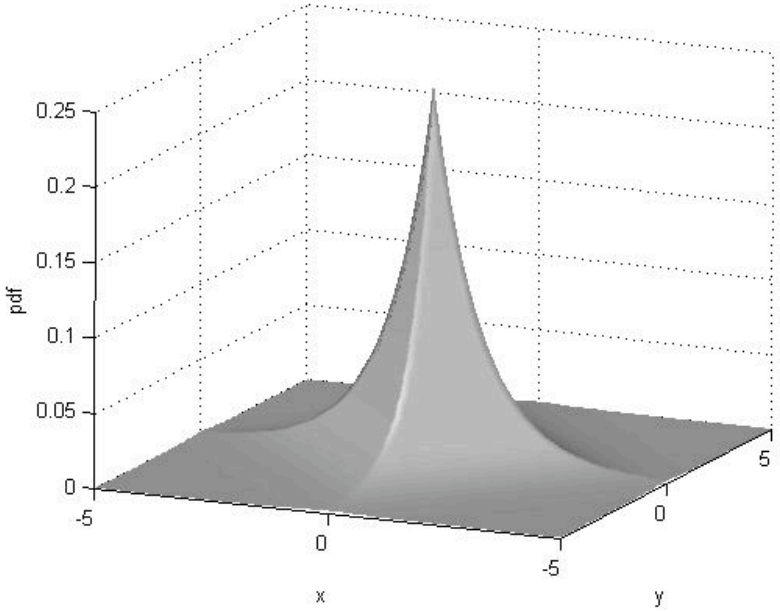

Posterior Modes with Varying Hyperparameter – Gaussian
posterior mode
0.10
0.05
0.00
−0.05
−0.10
age/100
glu
bp
bmi/100
ped
npreg
skin
intercept
0
0.05
0.1
0.15
0.2
0.25
0.3
s


Posterior Modes with Varying Hyperparameter – Laplace
posterior mode
0.10
0.05
0.00
−0.05
−0.10
age/100
glu
bp
bmi/100
ped
npreg
skin
intercept
120
100
80
60
40
20
0
1/s



Bayesian Perspective



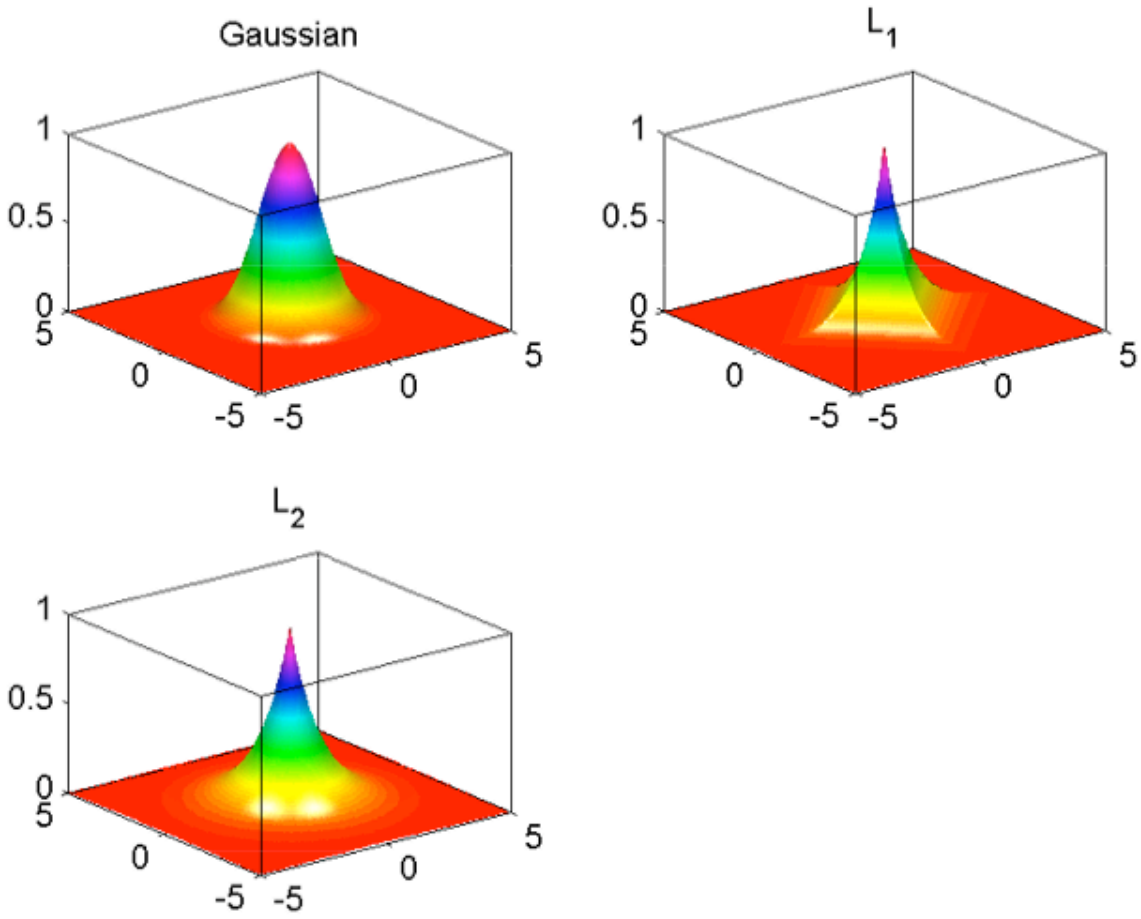


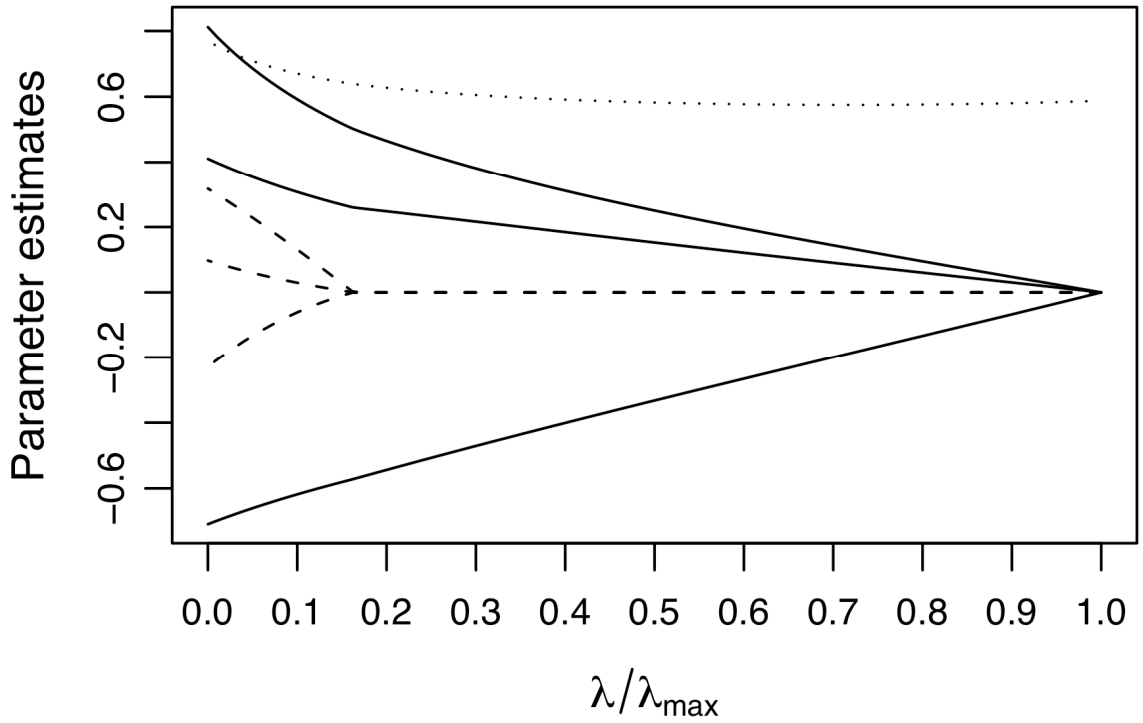
$$\beta_j \sim N(0, \tau^2)$$

$$\beta_j \sim N(0, \tau_j^2)$$

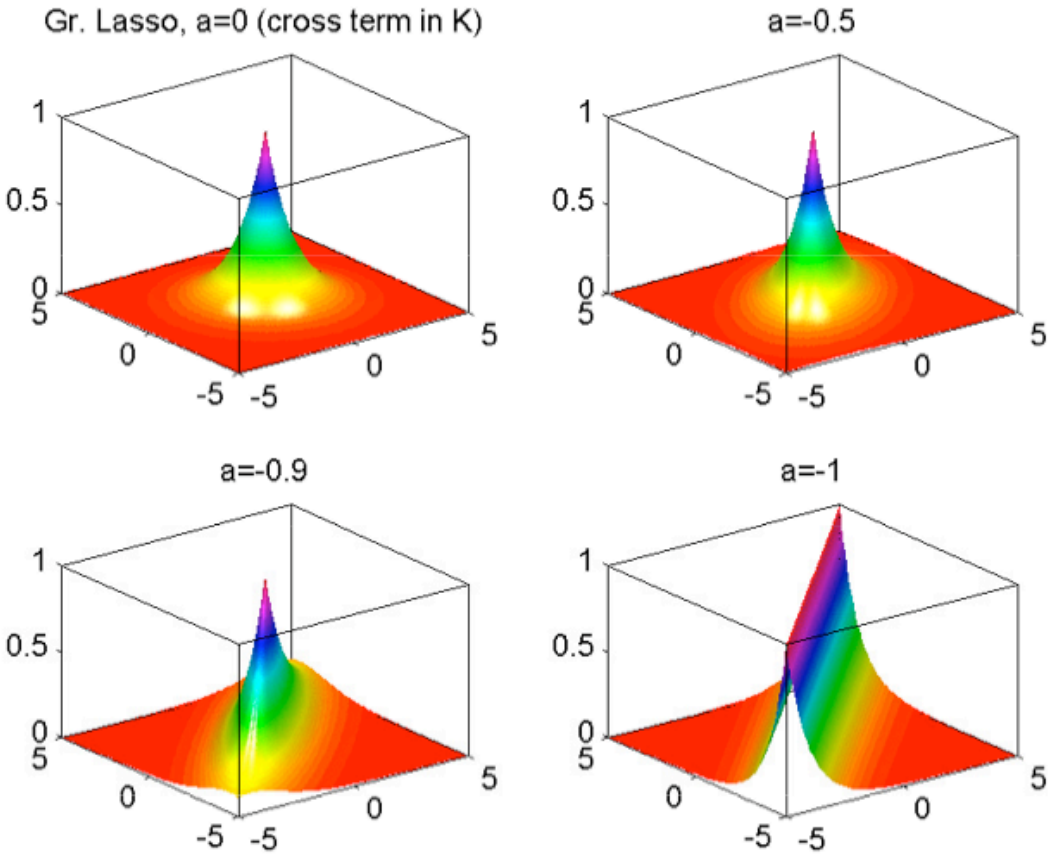
$$\tau_j^2 \sim \exp(\gamma)$$

Group Lasso




Parameter estimates
0.6
0.2
−0.2
−0.6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
λ/λmax

“soft fusion”



Balakrishnan and Madigan (2007)

“Consistency”

- Lasso not always consistent for variable selection
- SCAD (Fan and Li, 2001, JASA) consistent but non-convex
- relaxed lasso (Meinshausen and Buhlmann), adaptive lasso (Wang et al) have certain consistency results
- Zhao and Yu (2006) “irrepresentable condition”

Implementation

- Open source C++ implementation. Compiled versions for Linux, Windows, and Mac (soon)
- Binary and multiclass, hierarchical, informative priors
- Gauss-Seidel co-ordinate descent algorithm
- Fast? (parallel?)
- http://stat.rutgers.edu/~madigan/BBR

Aleks Jakulin's results

	log-loss / instance						
domain	BMR	DOT	NB	TAN	MAP	BKT	BK3
krkp	0.09	0.10	*-0.29*	0.19	0.06	0.11	**0.05**
monk2	0.65	0.64	*-0.65*	0.63	0.45	0.60	**0.45**
tic-tac-toe	0.09	0.08	*-0.55*	0.49	0.08	0.52	**0.07**
titanic	0.50	*-0.53*	0.52	0.48	0.48	0.48	**0.48**
lenses	0.61	0.72	2.44	*-2.99*	**0.34**	0.40	0.40
monk1	*-0.50*	0.49	0.50	0.09	**0.01**	0.08	0.02
mushroom	0.00	0.00	*-0.01*	**0.00**	0.00	0.00	0.00
shuttle	0.09	0.10	*-0.16*	**0.06**	0.07	0.07	0.07
soy-small*	0.27	*-0.31*	0.00	**0.00**	0.00	0.00	0.00
wine	0.10	0.09	**0.06**	0.29	0.19	0.11	0.11
yeast-class*	0.06	0.06	**0.01**	0.03	*-0.25*	0.12	0.12
anneal	0.07	**0.05**	0.07	*-0.17*	0.11	0.11	0.11
balance-scale	0.20	**0.17**	0.51	*-1.13*	0.51	0.51	0.51
lung-cancer*	1.11	**1.02**	5.41	*-6.92*	2.37	1.18	1.18
monk3	0.11	**0.11**	*-0.20*	0.11	0.11	0.11	0.11
post-op	0.67	**0.61**	0.93	1.78	0.79	0.67	0.67
promoters*	0.24	**0.23**	0.60	*-3.14*	0.59	0.52	0.52
adult	**0.28**	0.29	*-0.42*	0.33	0.30	0.30	0.30
audiology*	**1.04**	1.31	3.55	*-5.56*	2.24	2.21	2.21
australian	**0.33**	0.36	0.46	*-0.94*	0.41	0.37	0.37
breast-LJ	**0.55**	0.59	0.62	0.89	0.67	0.58	0.58
breast-wisc	**0.10**	0.12	0.21	0.23	0.21	0.16	0.16
bupa	**0.60**	0.60	0.62	0.60	0.62	0.61	0.61
car	**0.18**	0.18	*-0.32*	0.18	0.19	0.19	0.19
cmc	**0.91**	0.96	1.00	*-1.03*	0.93	0.92	0.92
crx	**0.33**	0.34	0.49	*-0.93*	0.37	0.35	0.35
ecoli	**0.45**	0.55	0.89	*-0.94*	0.85	0.81	0.81
german	**0.50**	0.51	0.54	*-1.04*	0.65	0.58	0.59
glass	**0.74**	0.78	1.25	*-1.76*	1.12	0.99	0.99
hayes-roth	**0.29**	0.35	0.46	*-1.18*	0.45	0.45	0.45
heart	**1.01**	1.03	1.25	*-1.53*	1.11	1.09	1.09
hepatitis	**0.36**	0.39	0.78	*-1.31*	0.48	0.39	0.39
horse-colic	**0.71**	0.71	1.67	*-5.97*	0.83	0.82	0.82
ionosphere	**0.19**	0.26	0.64	*-0.74*	0.39	0.30	0.30
iris	**0.16**	0.24	0.27	0.32	0.27	0.18	0.18
lymph	**0.50**	0.56	1.10	*-1.25*	0.98	0.79	0.79
o-ring	**0.66**	0.80	0.83	0.76	1.41	0.67	0.67
p-tumor*	**1.82**	1.93	3.17	*-4.76*	2.65	2.55	2.55
pima	**0.46**	0.48	0.50	0.49	0.51	0.48	0.48
segment	**0.13**	0.14	0.38	*-1.06*	0.17	0.17	0.17
soy-large*	**0.25**	0.46	0.57	0.47	*-0.68*	0.66	0.66
spam	**0.15**	0.16	*-0.53*	0.32	0.19	0.19	0.19
vehicle	**0.54**	0.56	*-1.78*	1.14	0.69	0.66	0.66
voting	**0.11**	0.13	*-0.60*	0.53	0.21	0.14	0.14
wdbc	**0.09**	0.10	0.26	*-0.29*	0.15	0.13	0.13
zoo*	**0.35**	*-0.47*	0.38	0.46	0.40	0.38	0.38
avg rank	2.13	2.87	5.62	5.60	4.74	3.68	3.36

1-of-K Sample Results: brittany-l

Feature Set	% errors	Number of Features
"Argamon" function words, raw tf	74.8	380
POS	75.1	44
1suff	64.2	121
1suff*POS	50.9	554
2suff	40.6	1849
2suff*POS	34.9	3655
3suff	28.7	8676
3suff*POS	27.9	12976
3suff+POS+3suff*POS+Argamon	27.6	22057
All words	23.9	52492

4.6 million parameters

89 authors with at least 50 postings. 10,076 training documents, 3,322 test documents.

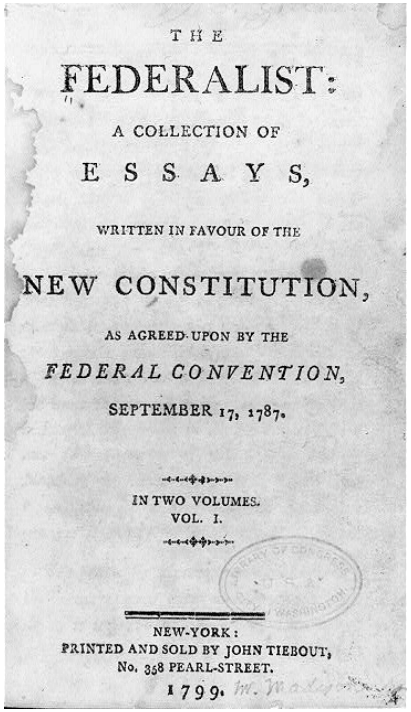
BMR-Laplace classification, default hyperparameter

Madigan et al. (2005)

The Federalist

- "The authorship of certain numbers of the 'Federalist' has fairly reached the dignity of a well-established historical controversy." (Henry Cabot Lodge, 1886)

- Historical evidence is muddled


THE
FEDERALIST:
A COLLECTION OF
ESSAYS,
WRITTEN IN FAVOUR OF THE
NEW CONSTITUTION,
AS AGREED UPON BY THE
FEDERAL CONVENTION,
SEPTEMBER 17, 1787.
IN TWO VOLUMES.
VOL. I.
NEW-YORK:
PRINTED AND SOLD BY JOHN TIEBOUT,
No. 358 PEARL-STREET.
1799.

Paper Number	Author
1	Hamilton
2-5	Jay
6-9	Hamilton
10	Madison
11-13	Hamilton
14	Madison
15-17	Hamilton
18-20	Joint: Hamilton and Madison
21-36	Hamilton
37-48	Madison
49-58	**Disputed**
59-61	Hamilton
62-63	**Disputed**
64	Jay
65-85	Hamilton

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 302 JUNE, 1963 *Volume 58*

INFERENCE IN AN AUTHORSHIP PROBLEM[1,2]

A comparative study of discrimination methods applied to the authorship of the disputed *Federalist* papers

FREDERICK MOSTELLER
Harvard University
and
Center for Advanced Study in the Behavioral Sciences
AND
DAVID L. WALLACE
University of Chicago

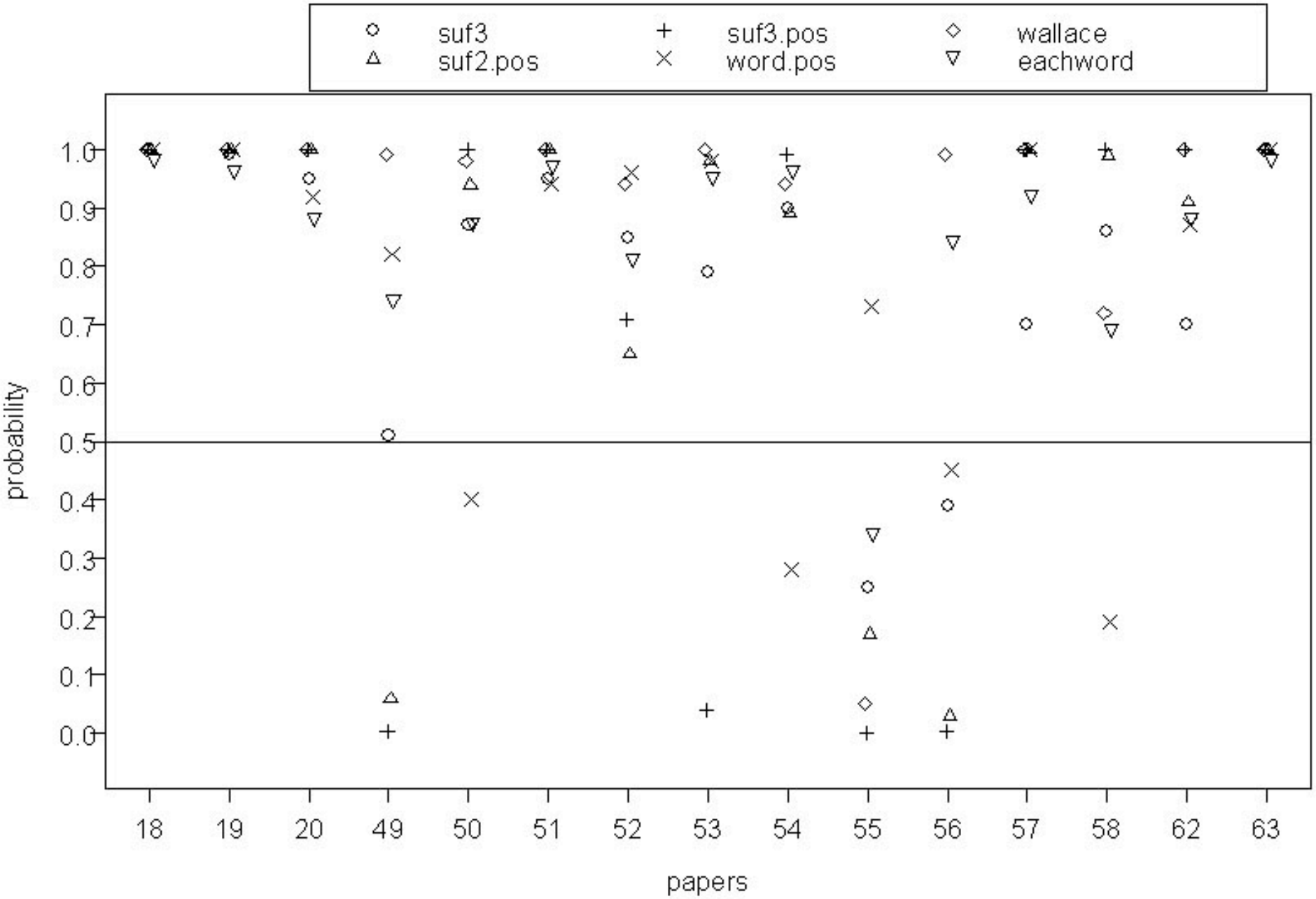
- Used function words with Naïve Bayes with Poisson and Negative Binomial model
- Out-of-sample predictive performance

F. Summing up

In summary, the following points are clear:

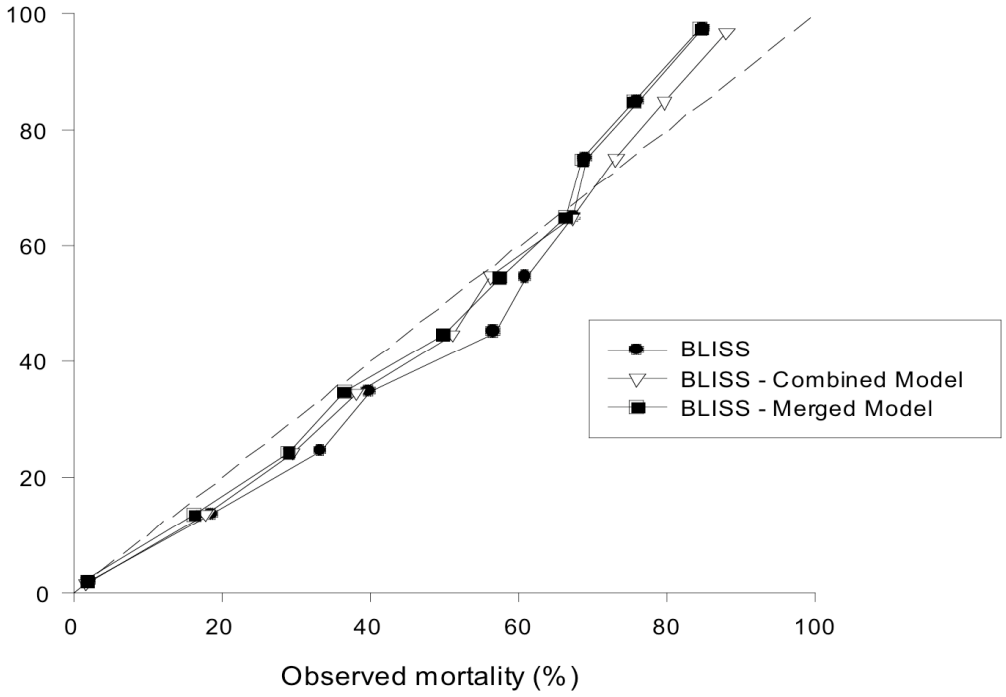
1) Madison is the principal author. These data make it possible to say far more than ever before that the odds are enormously high that Madison wrote the 12 disputed papers. Weakest support is given for No. 55. Support for Nos. 62 and 63, most in doubt by current historians, is tremendous.

Feature Set	**10-fold Error Rate**
Charcount	0.21
POS	0.19
Suffix2	0.12
Suffix3	0.09
Words	0.10
Charcount+POS	0.12
Suffix2+POS	0.08
Suffix3+POS	0.04
Words+POS	0.08
484 features	0.05
Wallace features	0.05
Words (>=2)	0.05
Each Word	0.05


suf3
suf2.pos
suf3.pos
word.pos
wallace
eachword
probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
18
19
20
49
50
51
52
53
54
55
56
57
58
62
63
papers

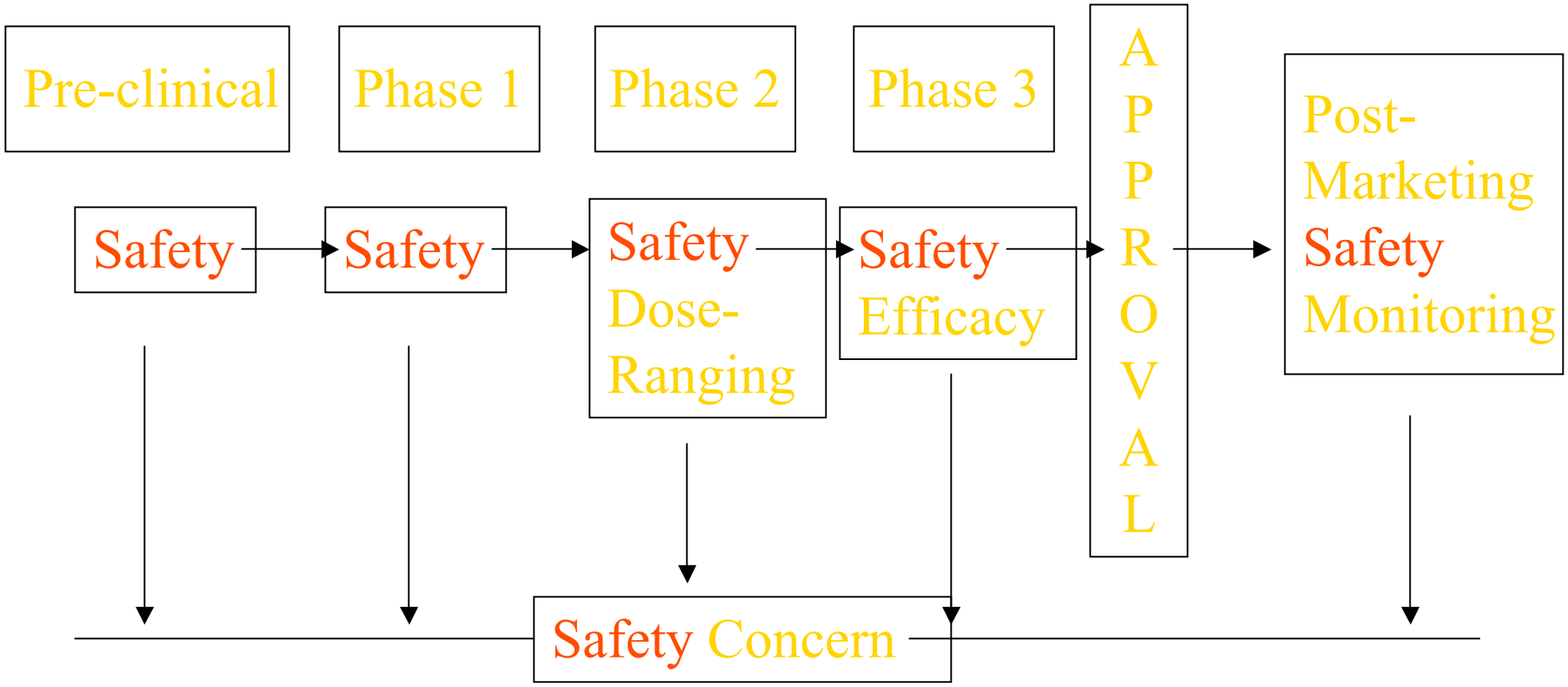
Risk Severity Score for Trauma

- Standard “ICISS” score poorly calibrated
- Lasso logistic regression with 2.5M predictors:



Burd and Madigan (2006)

Safety in Lifecycle of a Drug/Biologic product



Databases of Spontaneous ADRs

- FDA Adverse Event Reporting System (AERS)
  - Online 1997 – replace the SRS
  - Over 250,000 ADRs reports annually
  - 15,000 drugs - 16,000 ADRs
- CDC/FDA Vaccine Adverse Events (VAERS)
  - Initiated in 1990
  - 12,000 reports per year
  - 50 vaccines and 700 adverse events
- Other SRS
  - WHO - international pharmacovigilance program

U.S. Department of Health and Human Services

MEDWATCH

The FDA Safety Information and Adverse Event Reporting Program

For VOLUNTARY reporting of adverse events, product problems and product use errors

Page ____ of ____

Form Approved: OMB No. 0910-0291, Expires: 10/31/08
See OMB statement on reverse.

FDA USE ONLY

Triage unit sequence #

PLEASE TYPE OR USE BLACK INK

A. PATIENT INFORMATION

1. **Patient Identifier** In confidence
2. **Age at Time of Event, or Date of Birth:**
3. **Sex** ☐ Female ☐ Male
4. **Weight** _____ lb or _____ kg

B. ADVERSE EVENT, PRODUCT PROBLEM OR ERROR

Check all that apply:

1. ☐ **Adverse Event** ☐ **Product Problem** *(e.g., defects/malfunctions)*
   ☐ **Product Use Error** ☐ **Problem with Different Manufacturer of Same Medicine**

2. **Outcomes Attributed to Adverse Event** *(Check all that apply)*
   ☐ Death: _____ *(mm/dd/yyyy)*
   ☐ Life-threatening
   ☐ Hospitalization - initial or prolonged
   ☐ Required Intervention to Prevent Permanent Impairment/Damage (Devices)
   ☐ Disability or Permanent Damage
   ☐ Congenital Anomaly/Birth Defect
   ☐ Other Serious (Important Medical Events)

3. **Date of Event** *(mm/dd/yyyy)*
4. **Date of this Report** *(mm/dd/yyyy)*

5. **Describe Event, Problem or Product Use Error**

6. **Relevant Tests/Laboratory Data, Including Dates**

7. **Other Relevant History, Including Preexisting Medical Conditions** *(e.g., allergies, race, pregnancy, smoking and alcohol use, liver/kidney problems, etc.)*

C. PRODUCT AVAILABILITY

Product Available for Evaluation? *(Do not send product to FDA)*

☐ Yes ☐ No ☐ Returned to Manufacturer on: _____ *(mm/dd/yyyy)*

D. SUSPECT PRODUCT(S)

1. **Name, Strength, Manufacturer** *(from product label)*
   #1
   #2

2. **Dose or Amount** | **Frequency** | **Route**
   #1
   #2

3. **Dates of Use** *(If unknown, give duration) from/to (or best estimate)*
   #1
   #2

4. **Diagnosis or Reason for Use** *(Indication)*
   #1
   #2

5. **Event Abated After Use Stopped or Dose Reduced?**
   #1 ☐ Yes ☐ No ☐ Doesn't Apply
   #2 ☐ Yes ☐ No ☐ Doesn't Apply

8. **Event Reappeared After Reintroduction?**
   #1 ☐ Yes ☐ No ☐ Doesn't Apply
   #2 ☐ Yes ☐ No ☐ Doesn't Apply

6. **Lot #**
   #1
   #2

7. **Expiration Date**
   #1
   #2

9. **NDC # or Unique ID**

E. SUSPECT MEDICAL DEVICE

1. **Brand Name**
2. **Common Device Name**
3. **Manufacturer Name, City and State**

4. **Model #**	**Lot #**	5. **Operator of Device** ☐ Health Professional ☐ Lay User/Patient ☐ Other:
Catalog #	**Expiration Date** *(mm/dd/yyyy)*	
Serial #	**Other #**	

6. **If Implanted, Give Date** *(mm/dd/yyyy)*
7. **If Explanted, Give Date** *(mm/dd/yyyy)*
8. **Is this a Single-use Device that was Reprocessed and Reused on a Patient?** ☐ Yes ☐ No
9. **If Yes to Item No. 8, Enter Name and Address of Reprocessor**

F. OTHER (CONCOMITANT) MEDICAL PRODUCTS

Product names and therapy dates *(exclude treatment of event)*

G. REPORTER *(See confidentiality section on back)*

1. **Name and Address**

Phone # | **E-mail**

2. **Health Professional?** ☐ Yes ☐ No
3. **Occupation**
4. **Also Reported to:** ☐ Manufacturer ☐ User Facility ☐ Distributor/Importer
5. **If you do NOT want your identity disclosed to the manufacturer, place an "X" in this box:** ☐

FORM FDA 3500 (10/05) **Submission of a report does not constitute an admission that medical personnel or the product caused or contributed to the event.**

Weakness of SRS Data

- Passive surveillance
  - Underreporting
- Lack of accurate "denominator", only "numerator"
  - "Numerator": No. of reports of suspected reaction
  - "Denominator": No. of doses of administered drug
- No certainty that a reported reaction was causal
- Missing, inaccurate or duplicated data

Existing Methods

- Multi-item Gamma Poisson Shrinker (MGPS)
  - US Food and Drug Administration (FDA)
- Bayesian Confidence Propagation Neural Network
  - WHO Uppsala Monitoring Centre (UMC)
- Proportional Reporting Ratio (PRR and aPRR)
  - UK Medicines Control Agency (MCA)
- Reporting Odds Ratios and Incidence Rate Ratios
  - Other national spontaneous reporting centers and drug safety research units

Existing Methods (Cont'd)

- Focus on 2X2 contingency table projections

	AE j = Yes	*AE j = No*	*Total*
Drug *i* = Yes	*a*=20	*b*=100	120
Drug *i* = No	*c*=100	*d*=980	1080
Total	*120*	*1080*	1200

  - 15,000 drugs * 16,000 AEs = 240 million tables
  - Most $N_{ij} = 0$, even though $N..$ very large

The Different Measures

Measure of Association	*Formula*	*Probabilistic Interpretation*
RR Relative Risk*	$\frac{a * (a + b + c + d)}{(a + c) * (a + b)}$	$\frac{\Pr(ae \mid drug)}{\Pr(ae)}$
PRR Proportional Reporting Ratio	$\frac{a / (a + b)}{c / (c + d)}$	$\frac{\Pr(ae \mid drug)}{\Pr(ae \mid \neg drug)}$
ROR Reporting Odds Ratio	$\frac{a / c}{b / d}$	$\frac{\Pr(ae \mid drug) / \Pr(\neg ae \mid drug)}{\Pr(ae \mid \neg drug) / \Pr(\neg ae \mid drug)}$
Information Component	$Log_2 \frac{a * (a + b + c + d)}{(a + c) * (a + b)}$	$\log_2 \frac{\Pr(ae \mid drug)}{\Pr(ae)}$

Relative Reporting Ratio

$$(RR_{ij}=N_{ij}/E_{ij})$$

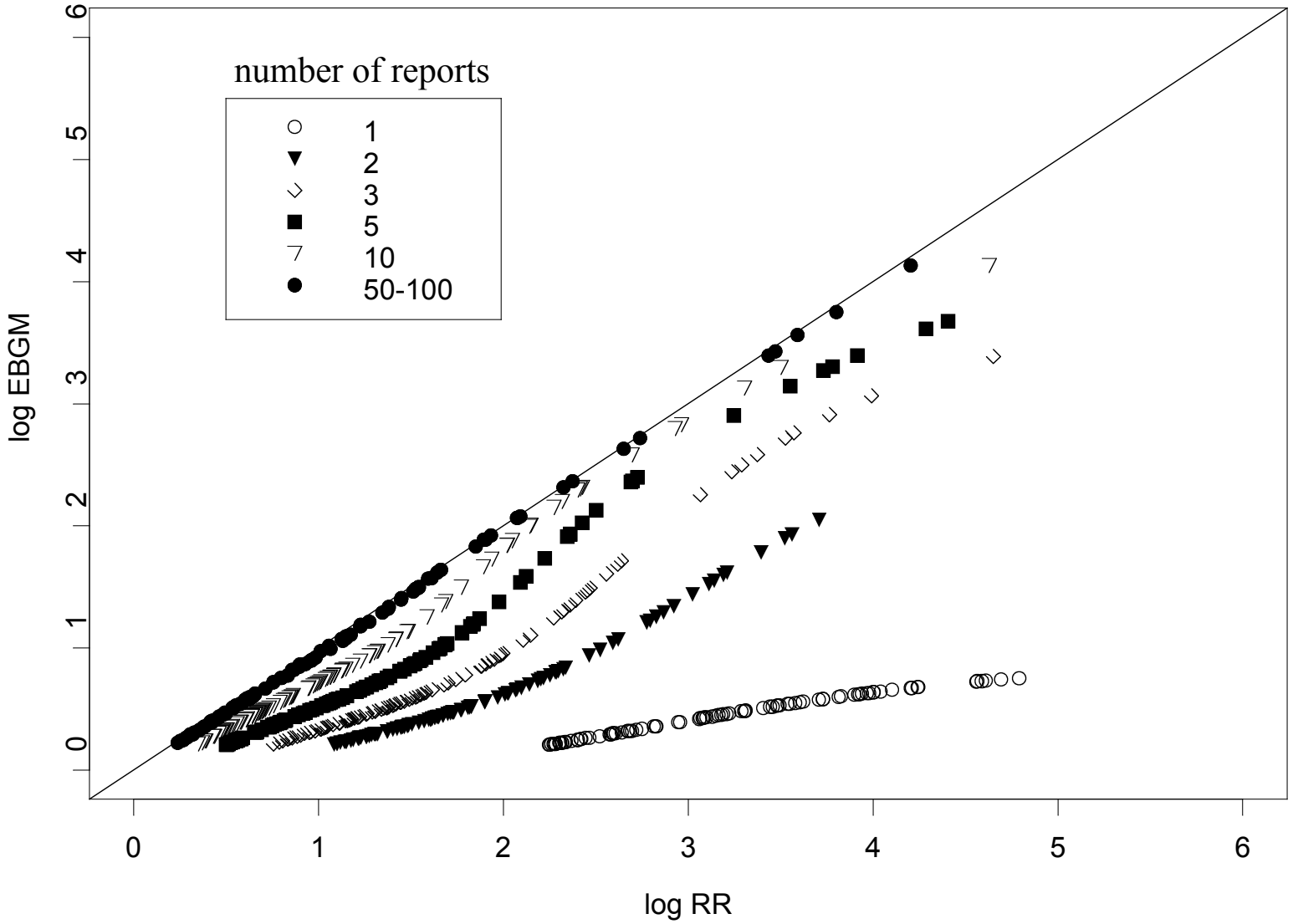
- Advantages
  - Simple
  - Easy to interpret
- Disadvantages
  - Extreme sampling variability when baseline and observed frequencies are small

    (N=1, E=0.01 v.s. N=100, E=1)
  - GPS provides a shrinkage estimate of RR that addresses this concern.

$$E_{ij}=N_{ij}*N../N_{i.}N_{.j}$$

	AE_j	Not AE_j	
$Drug_i$	N_{ij}		$N_{i.}$
Not $Drug_i$			
	$N_{.j}$		$N..$

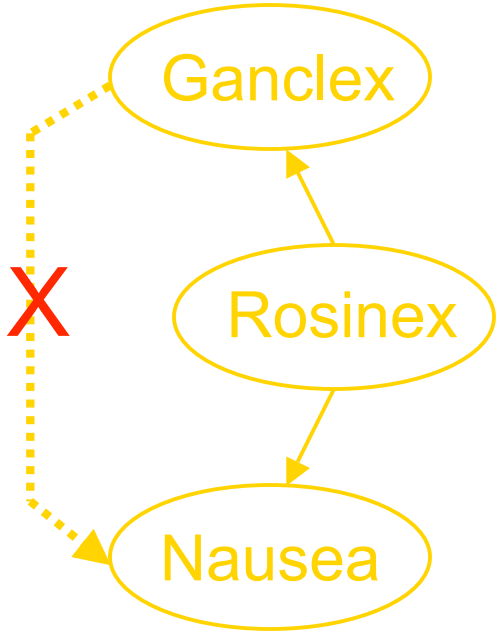
GPS SHRINKAGE – AERS DATA

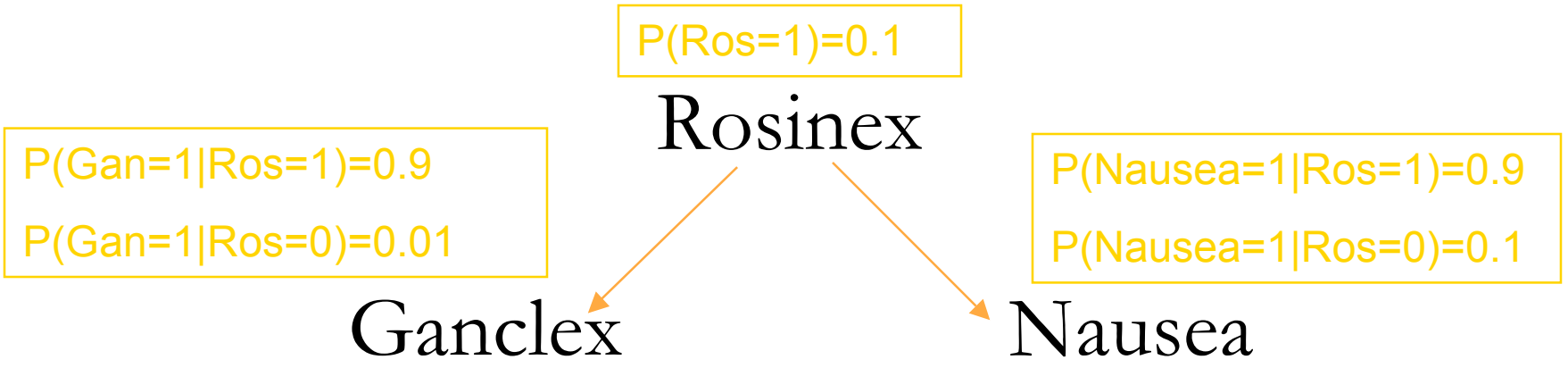


Confounding

- Contingency table analysis ignores effects of drug-drug association on drug-AE association

	Rosinex		No Rosinex		Total	
	Nausea	No Nausea	Nausea	No Nausea	Nausea	No Nausea
Ganclex	81	9	1	9	82	18
No Ganclex	9	1	90	810	99	811
RR	1		1		4.58	





		Nausea vs. Ganclex		Nausea vs. Rosinex	
		Value	Rank	Value	Rank
N		1673	2	1826	1
Bayesian Logistic Regression	Laplace-CV	0.0	9127	4.0	7
GPS EBGM		2.8	73	3.0	68
Observed RR		2.8	744	3.0	681

Logistic Regression

- log [$P/(1-P)$] = intercept + $\sum$ (each drug effect)
  - P = Pr (report with these drugs will have the AE)
- 15,000 logistic regressions with n≈3million
- 15,000 main effects
- millions of pairwise interactions???

Current Work

- Model associations between *groups* of drugs and *groups* of adverse events
- Bayesian generative approach applicable
- Sketch:
  - assign every drug to a latent group
  - assign every AE to a latent group
  - for each set of drugs and set of AE's, generate a report with probability defined by latent group memberships
- Major computational challenges
- Blei, Feinberg, Ghahramani, Roweis, etc.

- BR has 5 APs, site dimension: 225 ft X 144 ft
  - 259 blue (corridor) data points taken earlier

Hierarchical Model