High Dimensional Bayesian Classifiers

David Madigan
Columbia University
stat.columbia.edu/~madigan

Columbia's expansion faces substantial community opposition.

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

- Benzécri, 2005

Logistic Regression

- Linear model for log odds of category membership:

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

Maximum Likelihood Training

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

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

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

Shrinkage/Regularization/Bayes

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

Ridge Logistic Regression

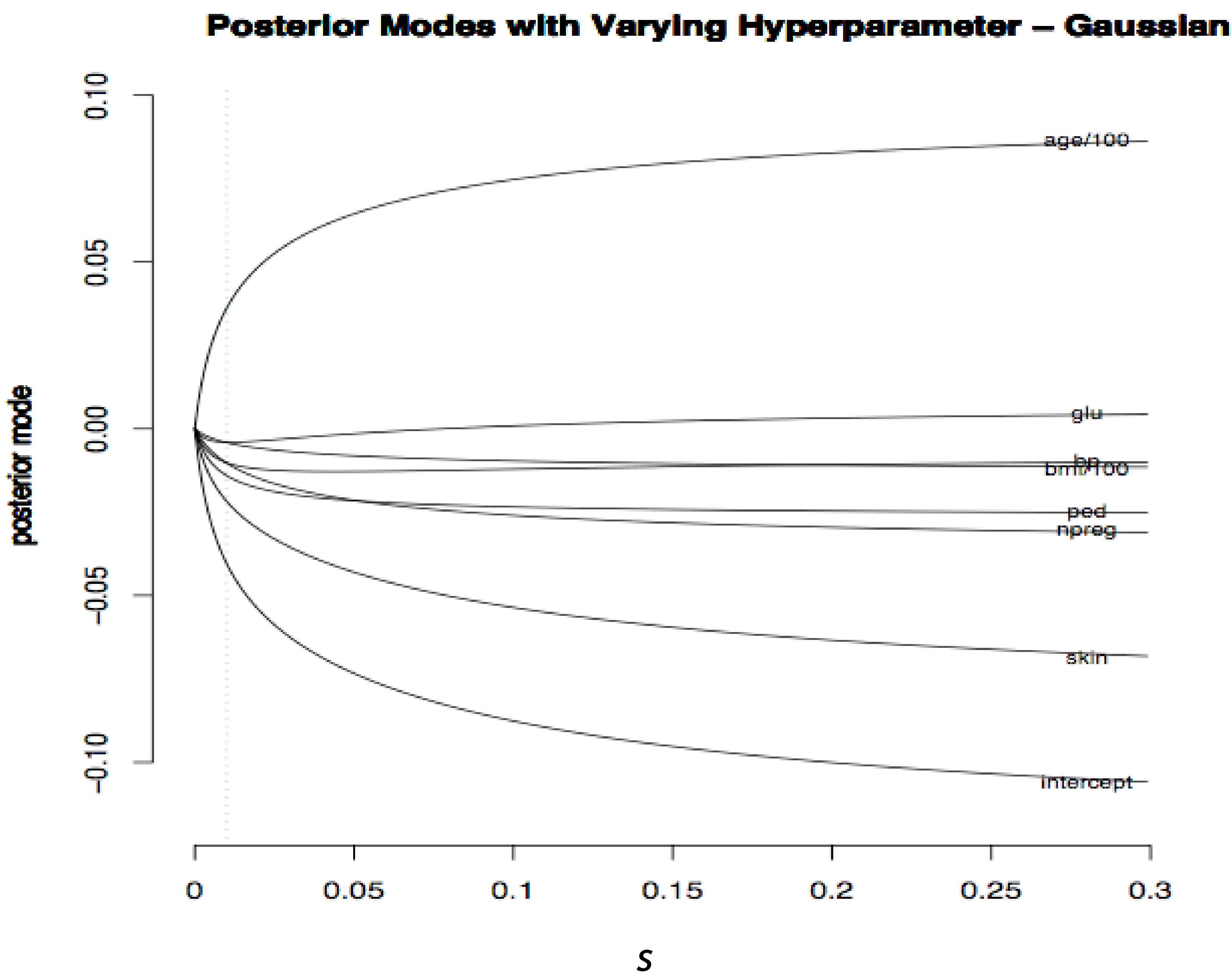
Maximum likelihood plus a constraint:

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

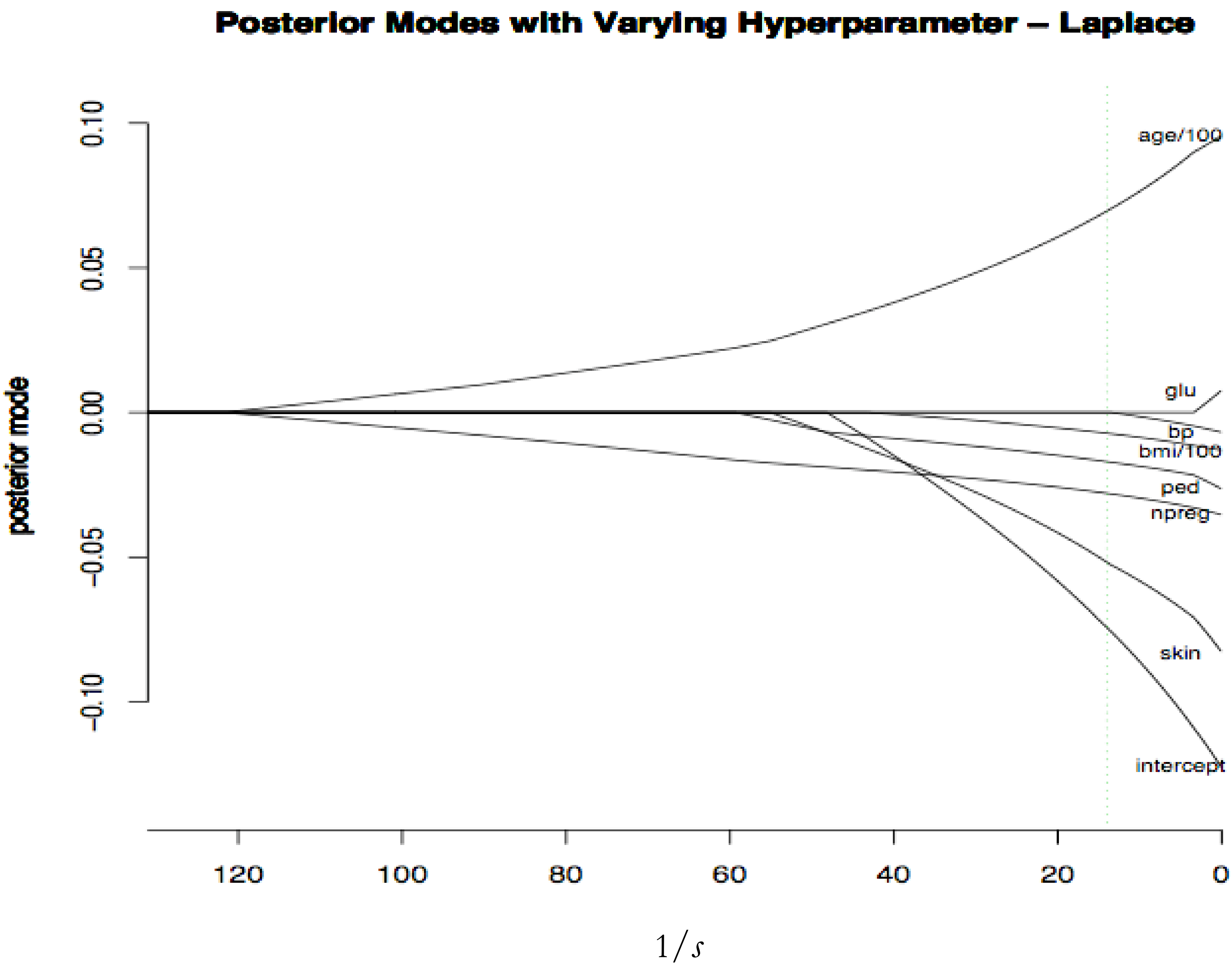
Lasso Logistic Regression

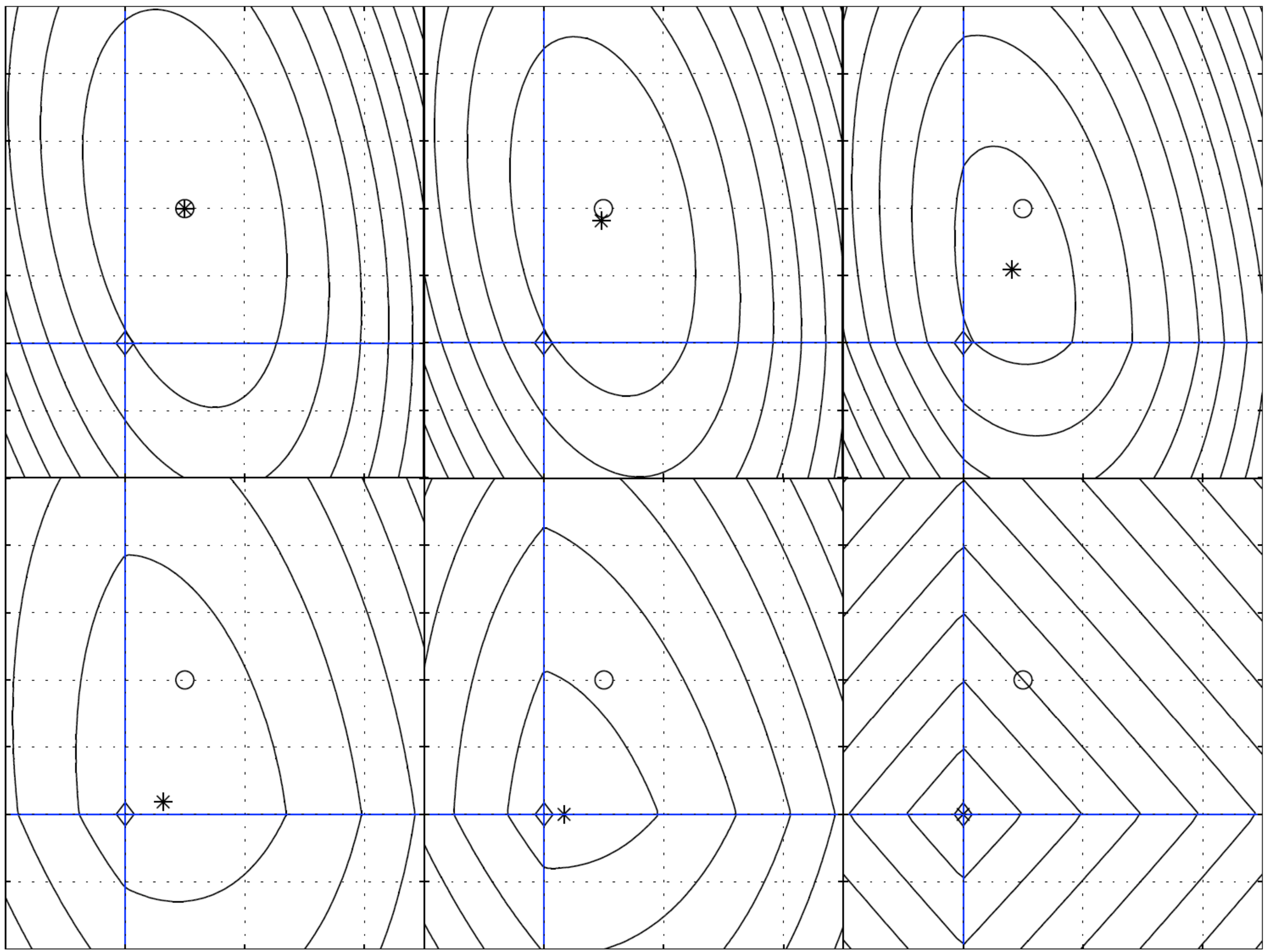
Maximum likelihood plus a constraint:

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

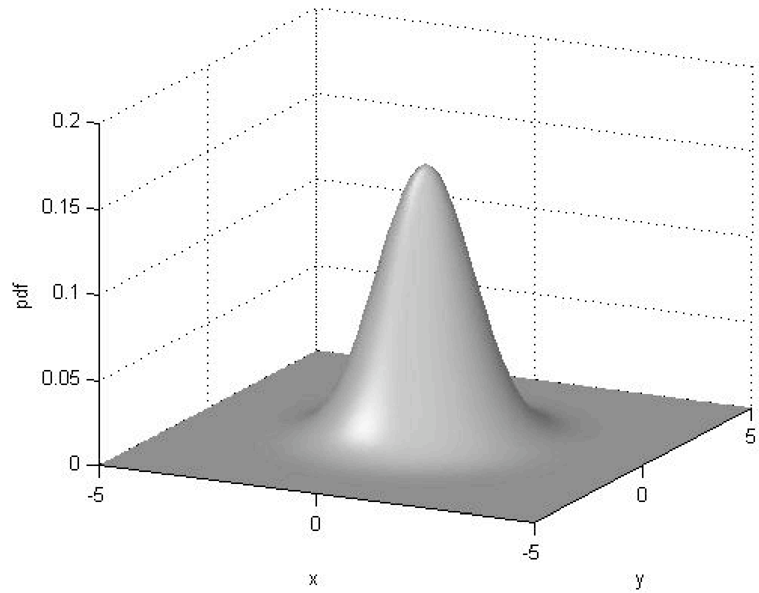


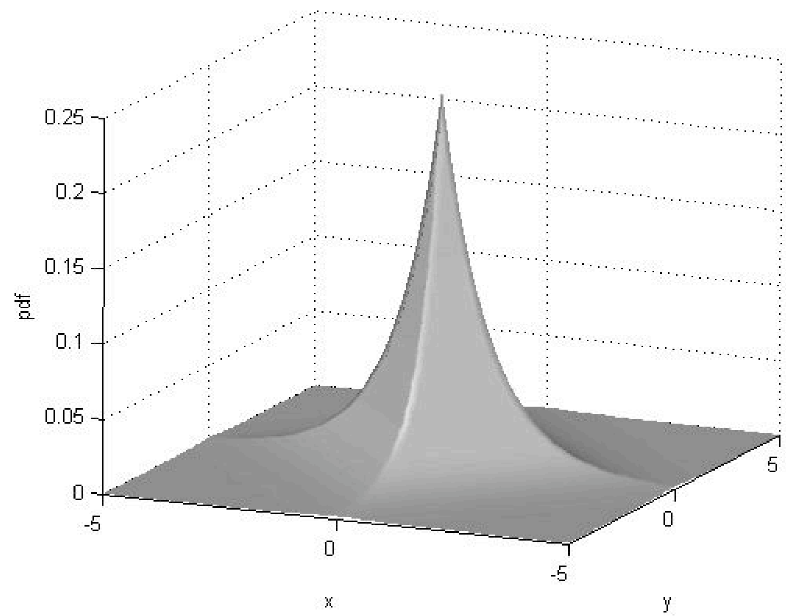
Posterior Modes with Varying Hyperparameter – Gaussian
posterior mode
0.10
0.05
0.00
−0.05
−0.10
age/100
glu
bp
bmi/100
ped
npreg
skin
intercept
0
0.05
0.1
0.15
0.2
0.25
0.3
s


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





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

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

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

Implementation: BXR

- Highly-optimized open source C++ implementation
- Compiled versions for Linux, Windows, and Mac
- Binary and multiclass, hierarchical, informative priors
- L1 and L2 regularization
- Gauss-Seidel co-ordinate descent algorithm
- Fast? (parallel?)
- http://www.bayesianregression.org

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.

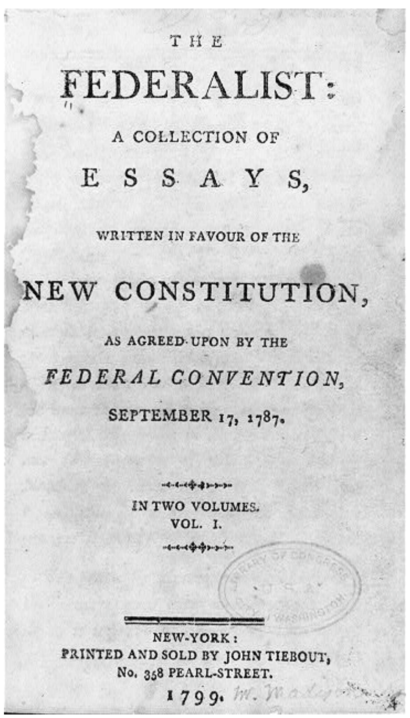
BMR-Laplace classification, default hyperparameter

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

- "High" dimensional Bayesian classification
- Used function words with Naïve Bayes with Poisson and Negative Binomial model
- Out-of-sample predictive performance

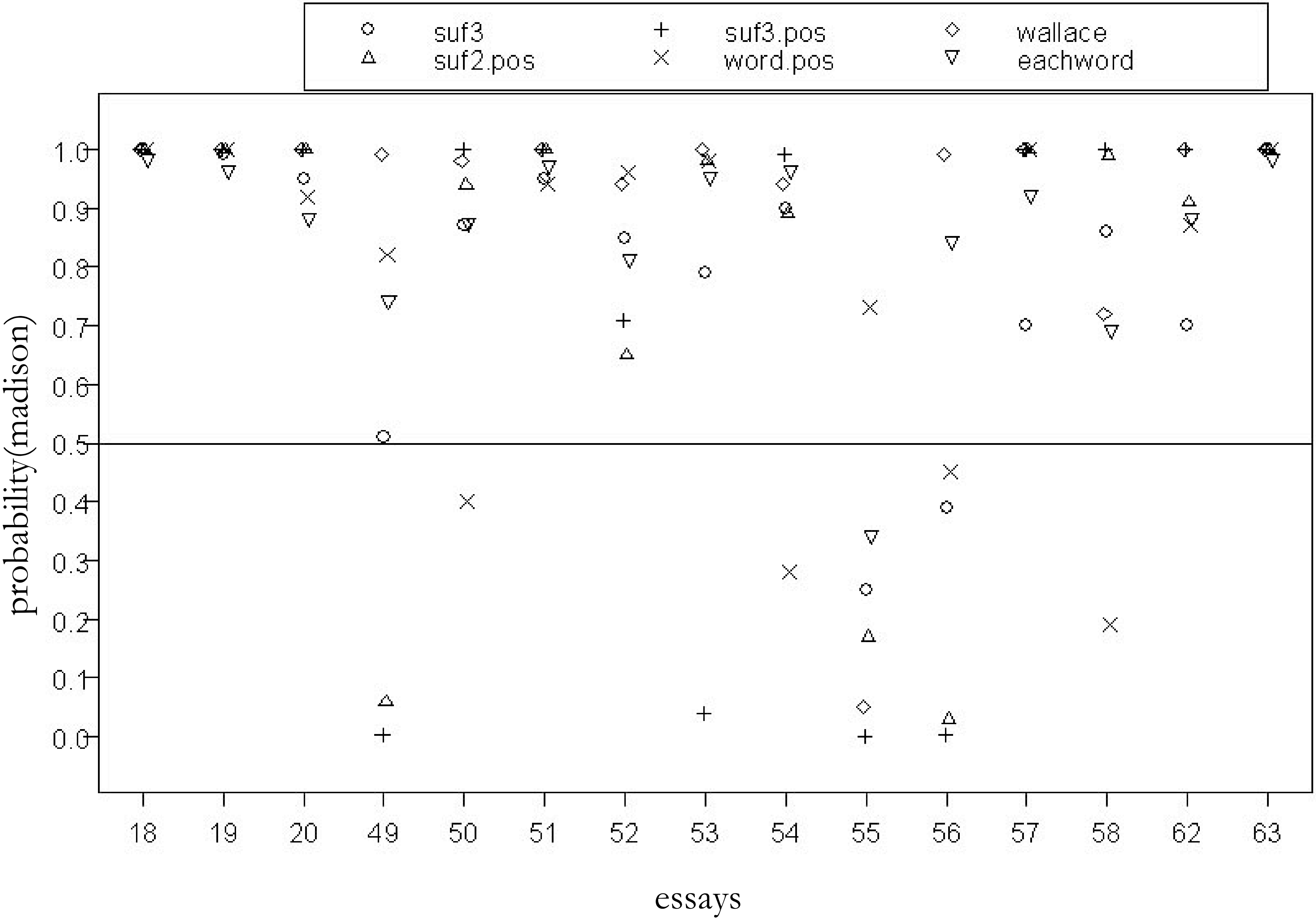
F. Summing up

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.

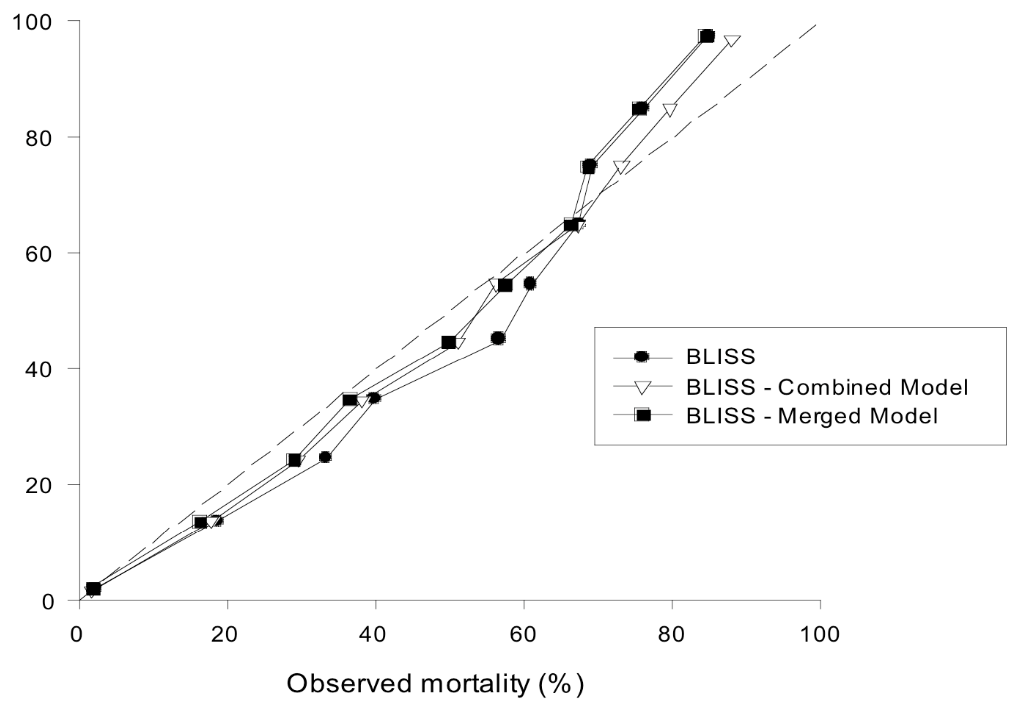
Feature Set	10-fold Error Rate
Charcount	0.21
POS	0.19
Suffix2	0.12
Suffix3	0.09
Words	0.10
Charcount+POS	0.12
Suffix2+POS	0.08
Suffix3+POS	0.04
Words+POS	0.08
484 features	0.05
Wallace features	0.05
Words (>=2)	0.05
Each Word	0.05

best


suf3
suf3.pos
wallace
suf2.pos
word.pos
eachword
probability(madison)
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
essays

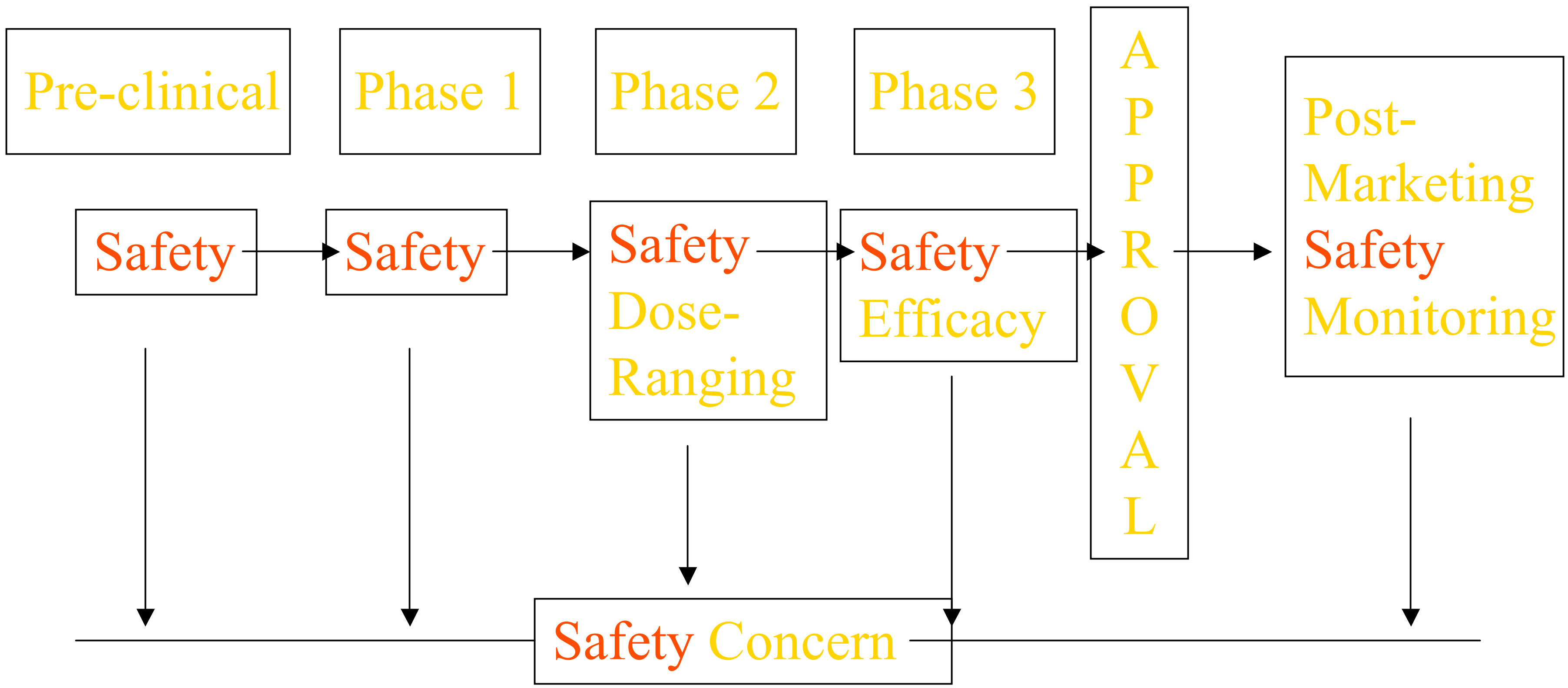
Risk Severity Score for Trauma

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



Burd and Madigan (2007)

Safety in Lifecycle of a Drug/Biologic product



U.S. Department of Health and Human Services

MEDWATCH

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

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

FDA USE ONLY

Triage unit sequence #

The FDA Safety Information and Adverse Event Reporting Program

Page ____ of ____

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)*
   - ☐ Disability or Permanent Damage
   - ☐ Life-threatening
   - ☐ Congenital Anomaly/Birth Defect
   - ☐ Hospitalization - initial or prolonged
   - ☐ Other Serious (Important Medical Events)
   - ☐ Required Intervention to Prevent Permanent Impairment/Damage (Devices)
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)*

PLEASE TYPE OR USE BLACK INK

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
6. **Lot #**
   #1
   #2
7. **Expiration Date**
   #1
   #2
8. **Event Reappeared After Reintroduction?**
   #1 ☐ Yes ☐ No ☐ Doesn't Apply
   #2 ☐ Yes ☐ No ☐ Doesn't Apply
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 #**
   Catalog # **Expiration Date** *(mm/dd/yyyy)*
   Serial # **Other #**
5. **Operator of Device**
   ☐ Health Professional
   ☐ Lay User/Patient
   ☐ 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:** ☐

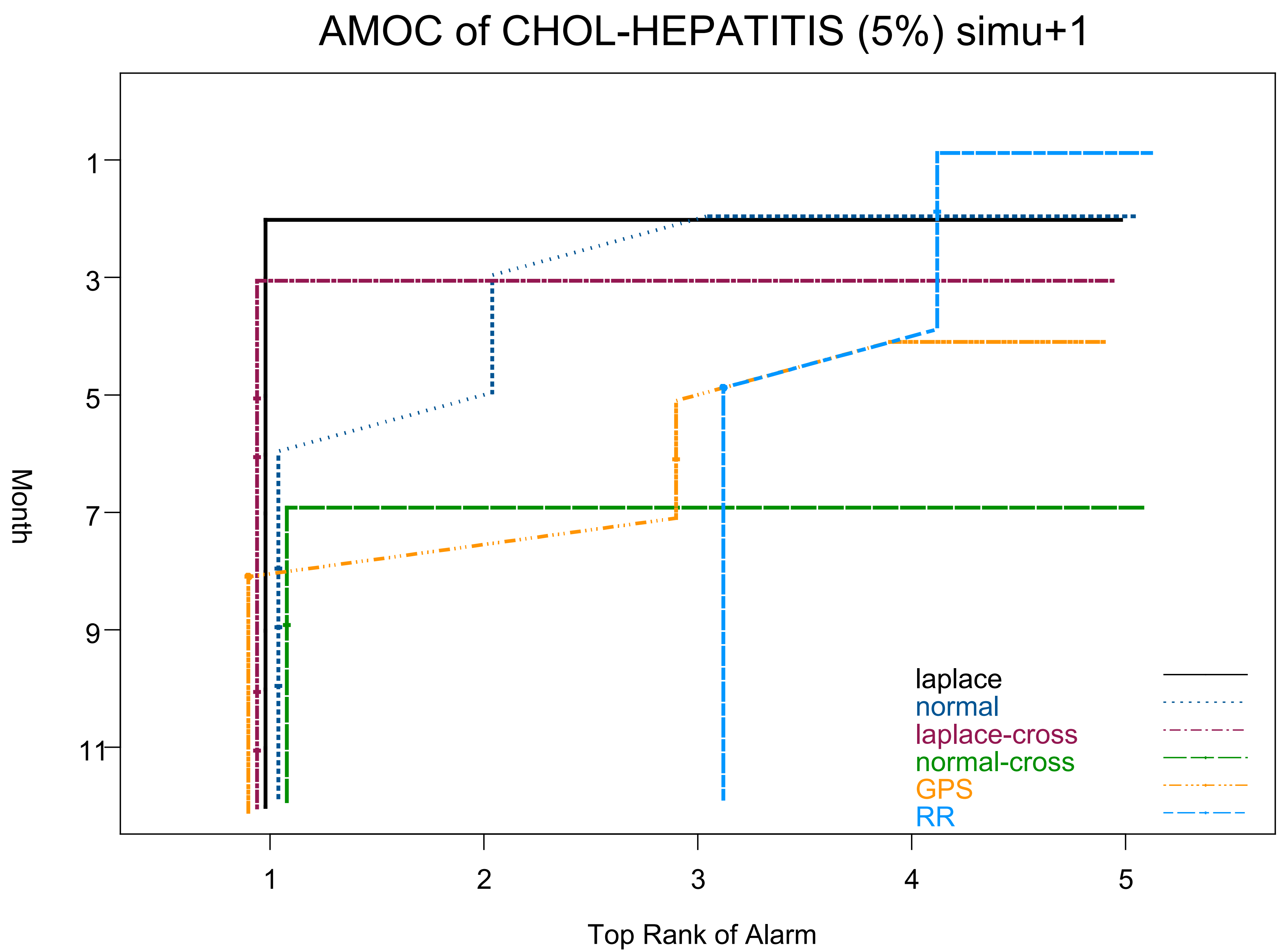
 Submission of a report does not constitute an admission that medical personnel or the product caused or contributed to the event.

Monitoring Spontaneous Drug Safety Reports

- Most reports contain several drugs and several AEs
- FDA, vendors, PhRMA, 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
  - Shrinkage methods better than e.g. chi square tests
  - "Innocent bystander" (i.e., confounding)
  - Regress each AE on all drugs
  - Regress all AE's on all drugs



AMOC of CHOL-HEPATITIS (5%) simu+1
Month
1
3
5
7
9
11
Top Rank of Alarm
1
2
3
4
5
laplace
normal
laplace-cross
normal-cross
GPS
RR

Consistency

- lasso consistently estimates the regression function (Greenshtein and Ritov, 2004)
- Lasso not always consistent for variable selection
- SCAD (Fan and Li, 2001, JASA) consistent but non-convex
- Zhao and Yu (2006) “irrepresentable condition”
- relaxed lasso (Meinshausen and Buhlmann), adaptive lasso (Wang et al) have certain consistency results
- Zou (2006, JASA) adaptive lasso --> BXR

High-Dimensional Bayes? Engineered Priors

(ModApte; training=100 documents)

	Macro F1	ROC
Laplace	37.2	76.2
Laplace & DK-based variance	65.3	87.1
Laplace & DK-based mode	72.0	93.5

Dayanik et al. (2006)

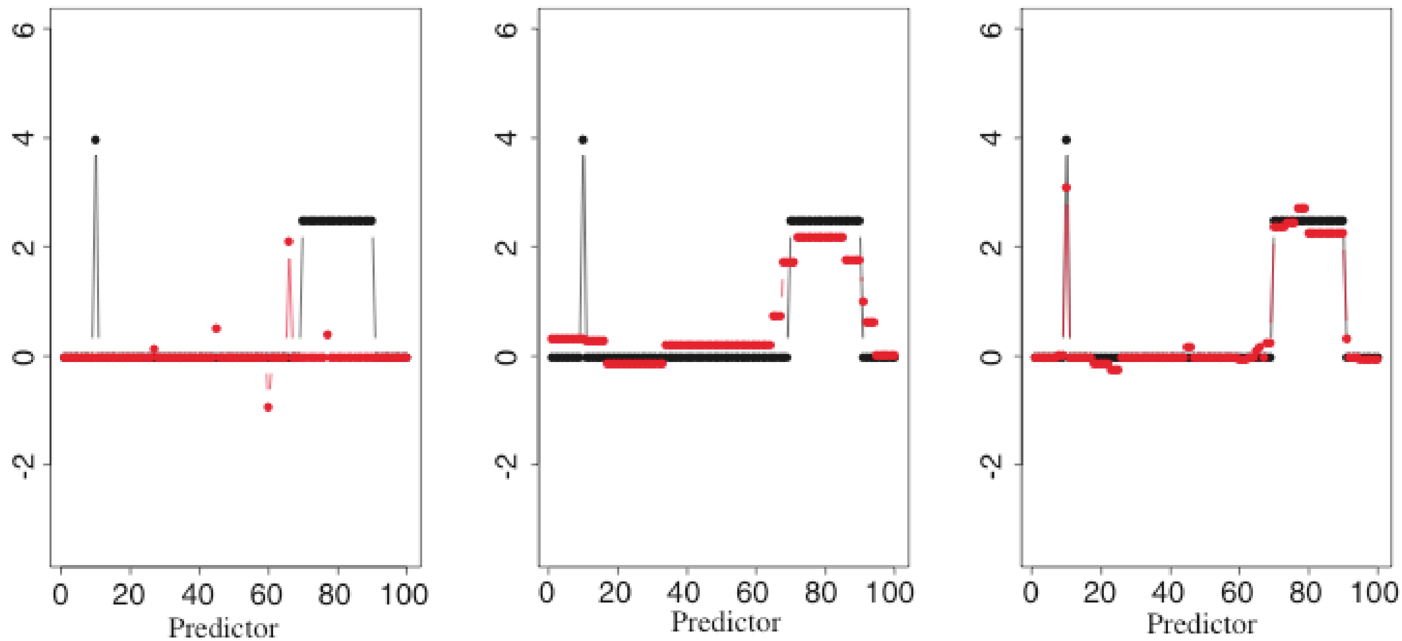
Fused Lasso

- If there are many correlated features, lasso gives non-zero weight to only one of them
- Maybe correlated features (e.g. time-ordered) should have similar coefficients?

$$\hat{\beta} = \arg\min \left\{ \sum_i \left(y_i - \sum_j x_{ij} \beta_j \right)^2 \right\}$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leqslant s_1 \text{ and } \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leqslant s_2$$

Tibshirani et al. (2005)



lasso-only   fusion-only   lasso+fusion

Group Lasso

- Suppose you represent a categorical predictor with indicator variables
- Might want the set of indicators to be in or out

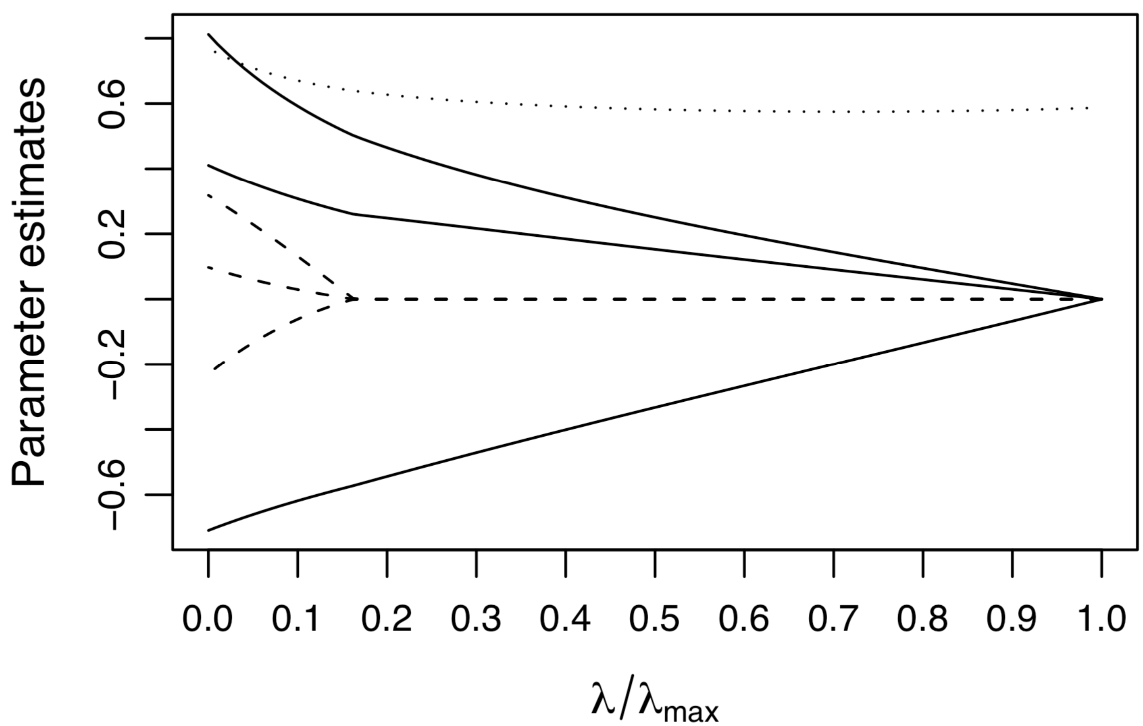
regular lasso:

$$\widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - X\boldsymbol{\beta}\|_2^2 + \lambda \sum_{i=1}^{p} |\beta_i|$$

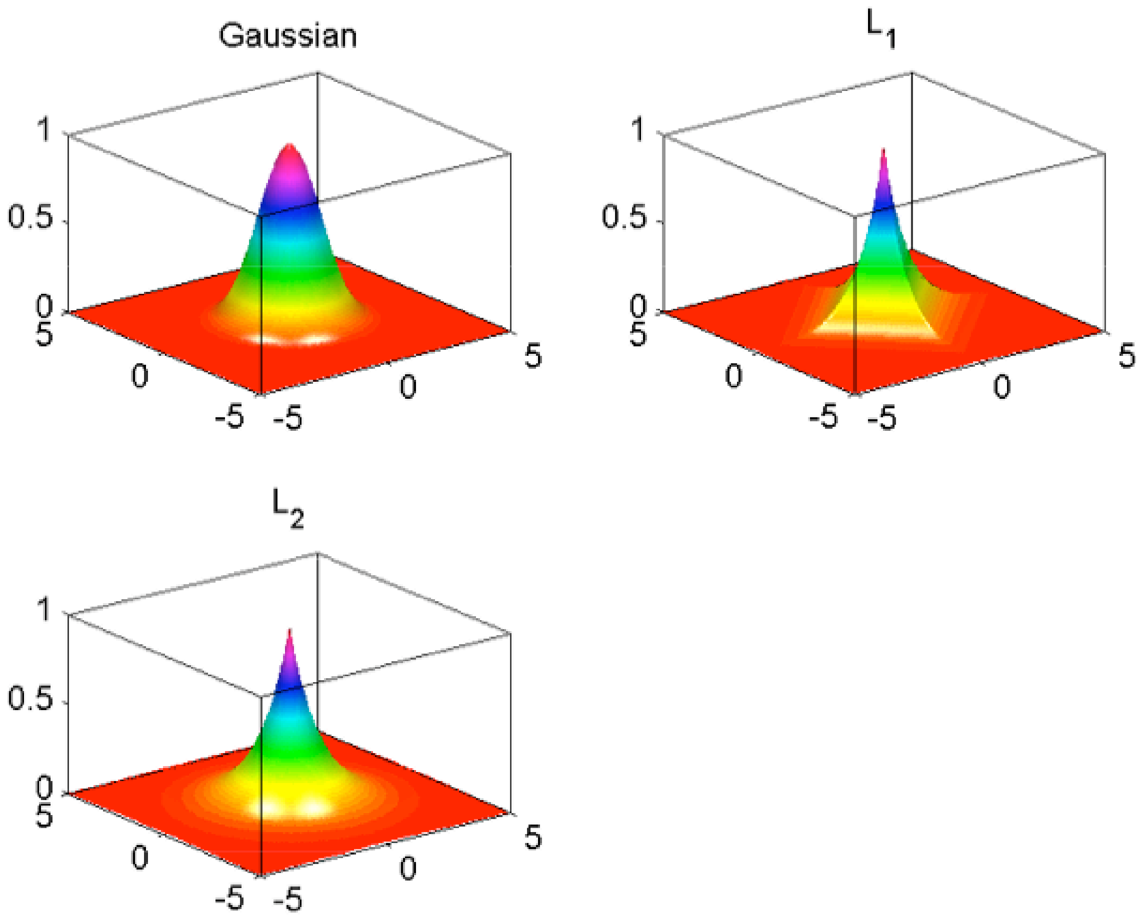
group lasso:

$$\widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \|Y - X\boldsymbol{\beta}\|_2^2 + \lambda \sum_{g=1}^{G} \|\boldsymbol{\beta}_{\mathcal{I}_g}\|_2$$

Yuan and Lin (2006)



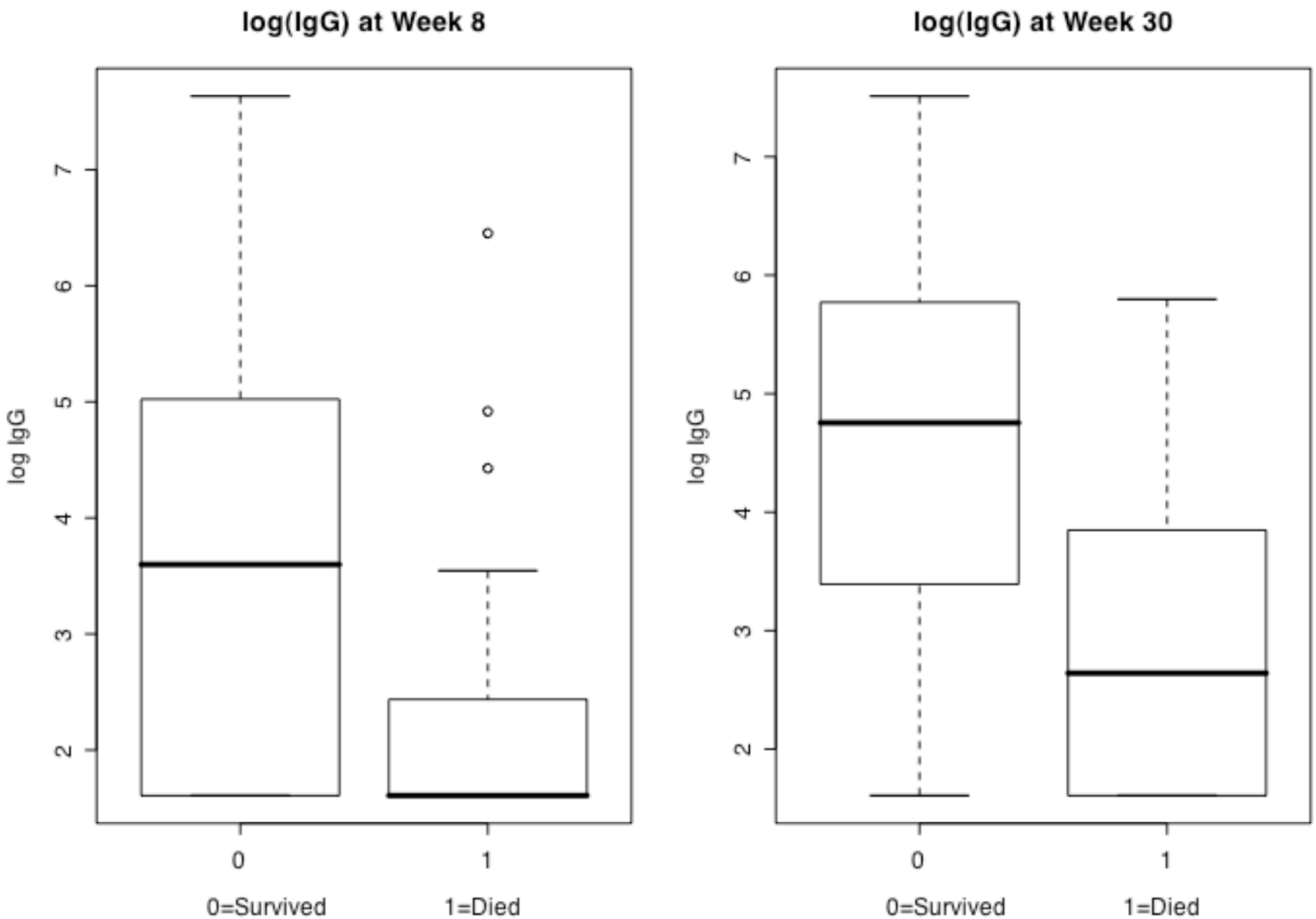
Parameter estimates
0.6
0.2
−0.2
−0.6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
λ/λ_{max}

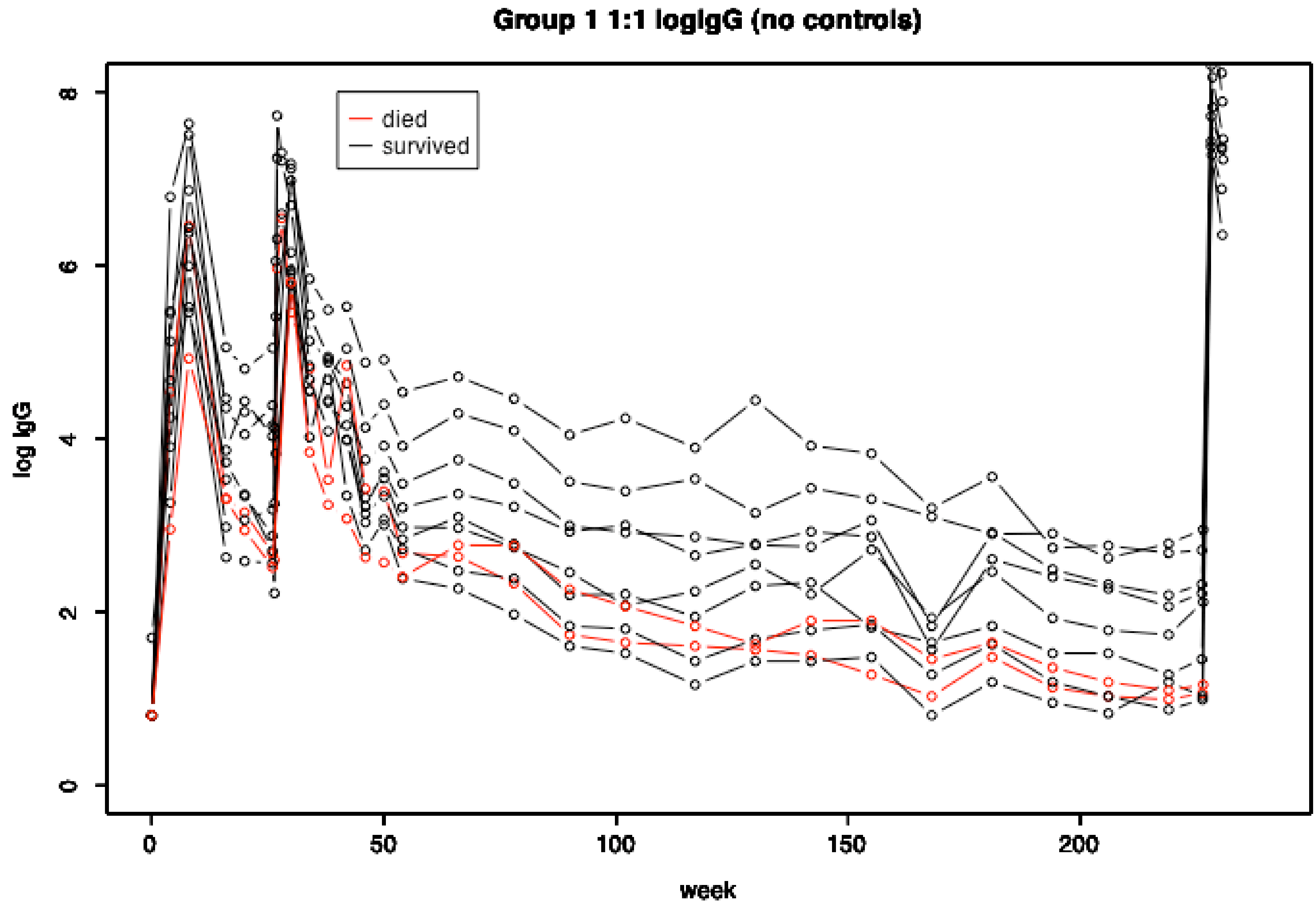

Gaussian
L1
L2
1
0.5
0
5
0
-5
-5
0
5

Anthrax Vaccine Study in Macaques

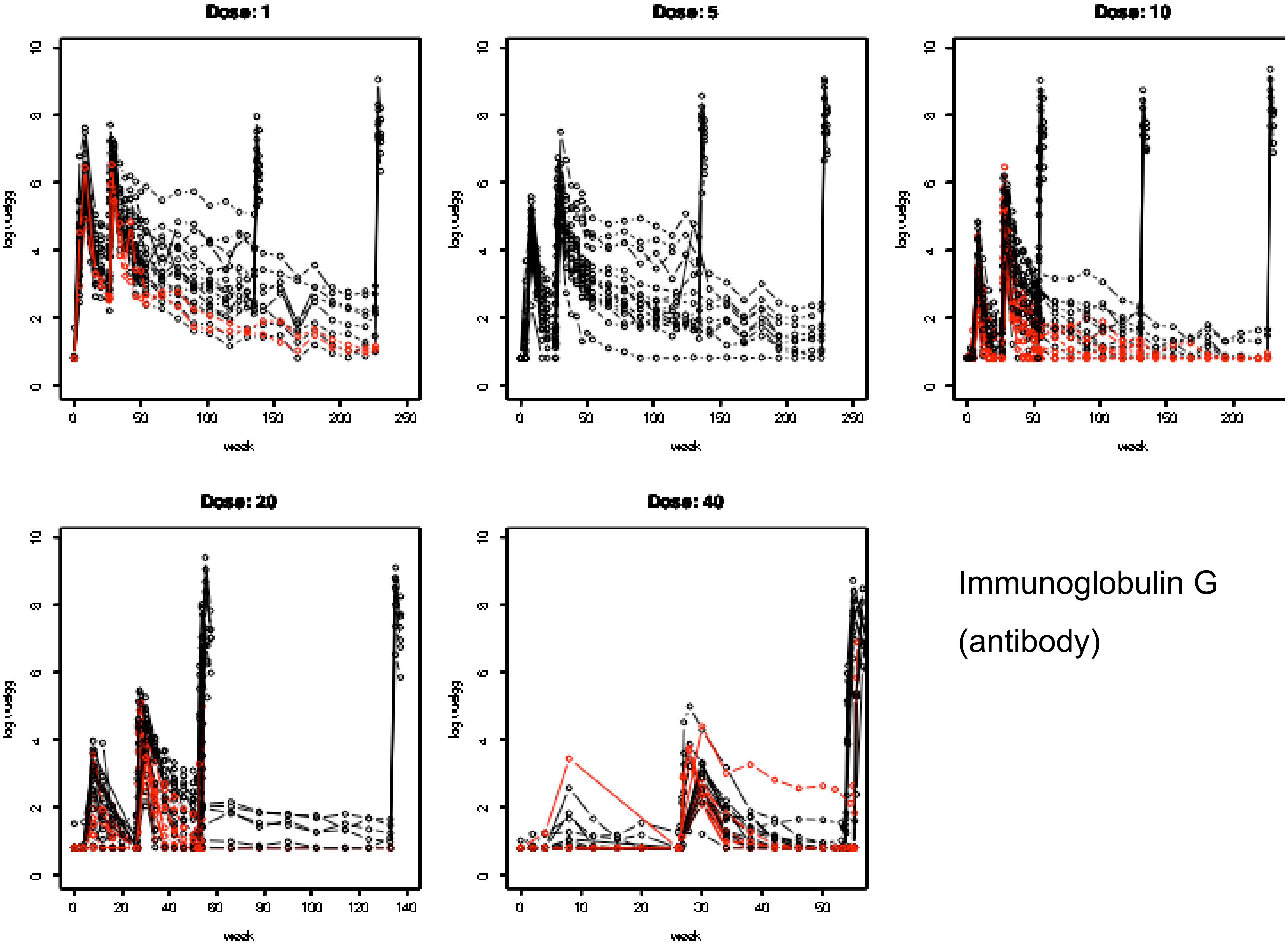
- Vaccinate macaques with varying doses; subsequently "challenge" with anthrax spores
- Are measurable aspects of the state of the immune system predictive of survival?
- Immunoglobulin G (IgG) expected to be important
- Problem: hundreds of different assay timepoints but fewer than one hundred macaques

Vaccine Dilution	Count	Outcome	
		Died	Death Rate
1:1	20	2	10%
1:5	17	0	0%
1:10	29	9	31%
1:20	28	10	36%
1:40	20	7	35%
control	23	16	70%
Total	137	44	32%


log(IgG) at Week 8
log(IgG) at Week 30
log IgG
2
3
4
5
6
7
0
1
0=Survived
1=Died

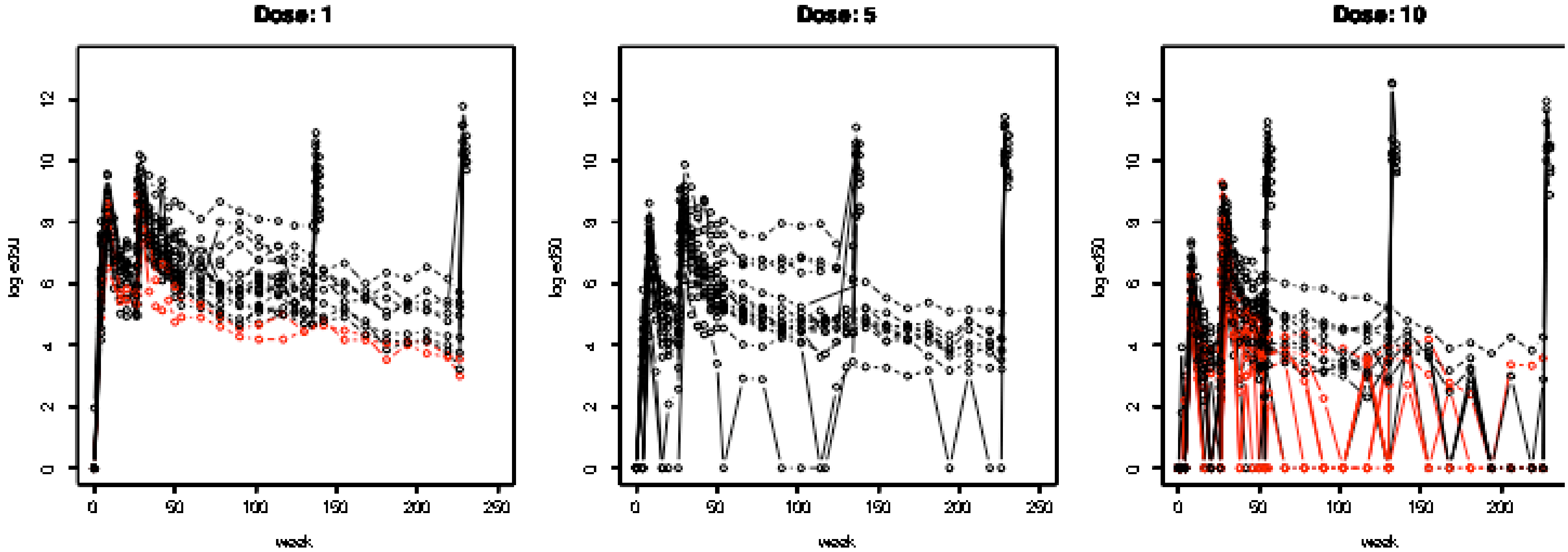


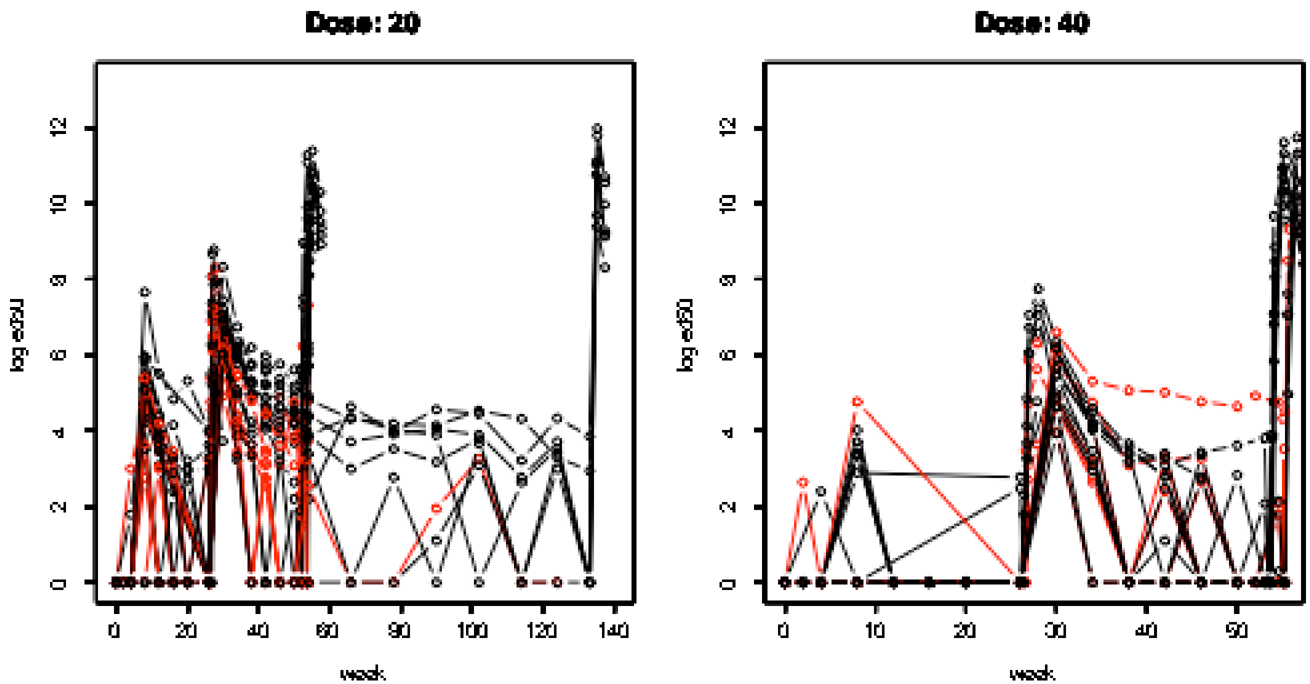
Group 1 1:1 logIgG (no controls)
died
survived
log IgG
0
2
4
6
8
0
50
100
150
200
week



Immunoglobulin G

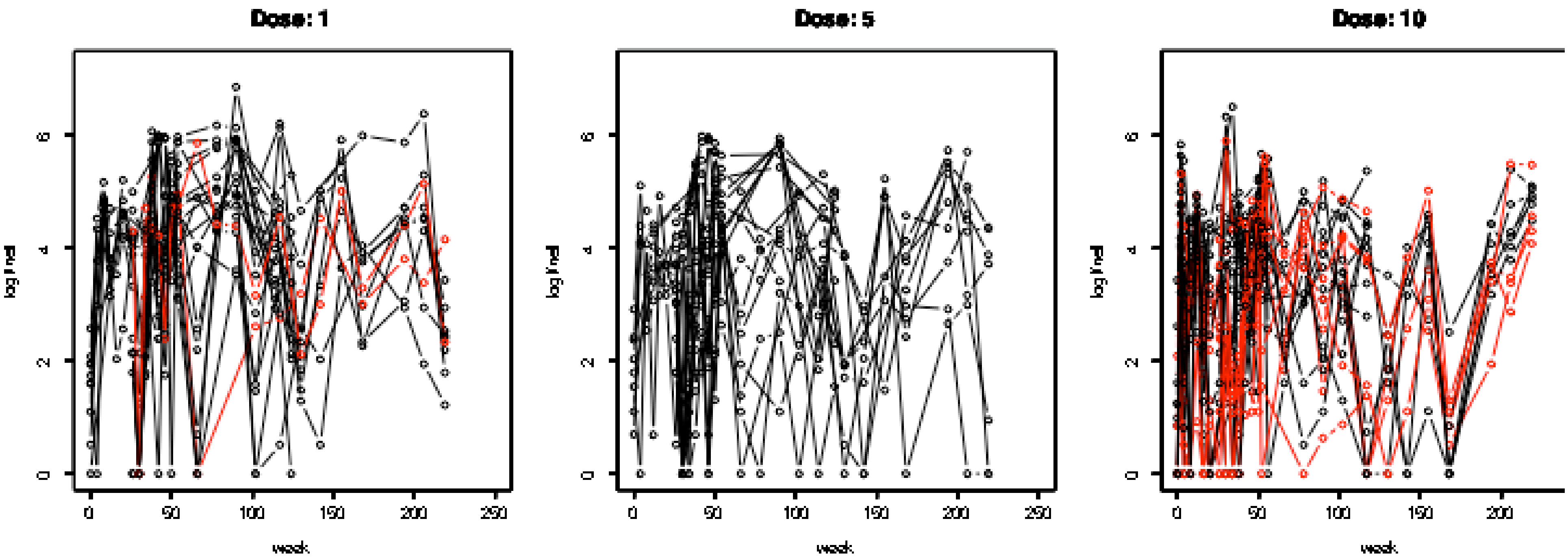
(antibody)

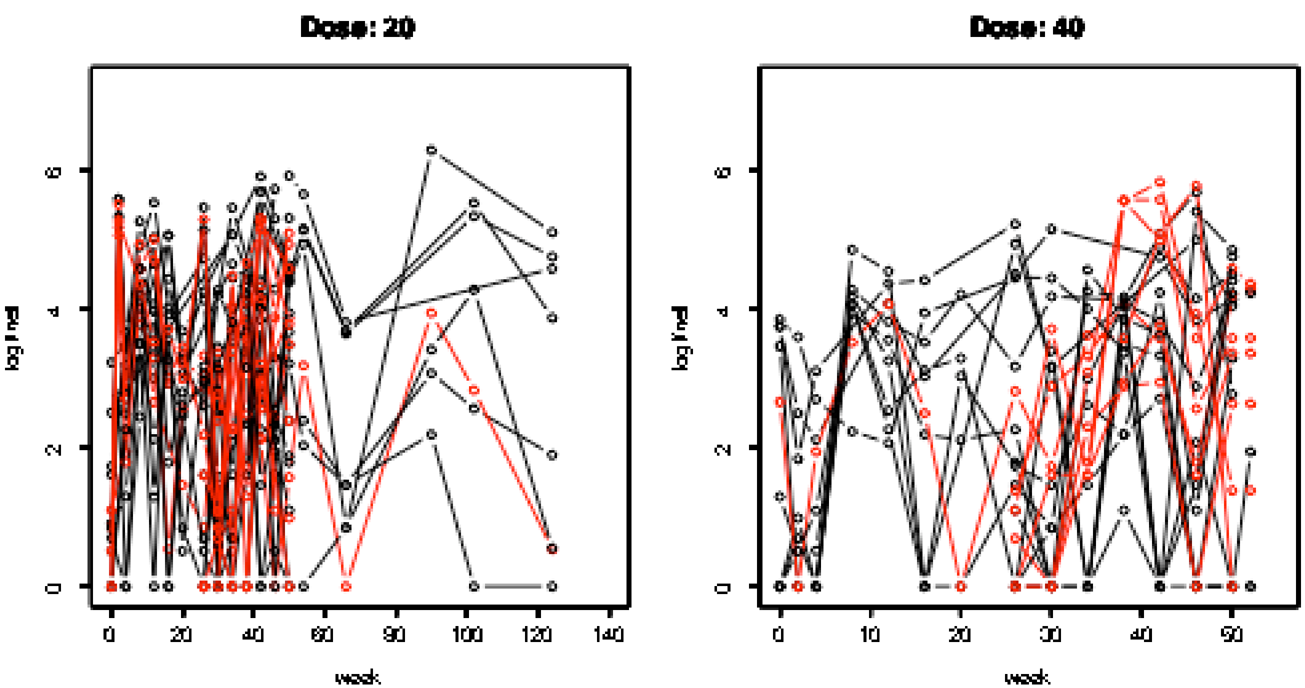




TNA

(toxin-neutralizing antibody)





IFNeli

(interferon - proteins produced by the immune system)

L1 Logistic Regression

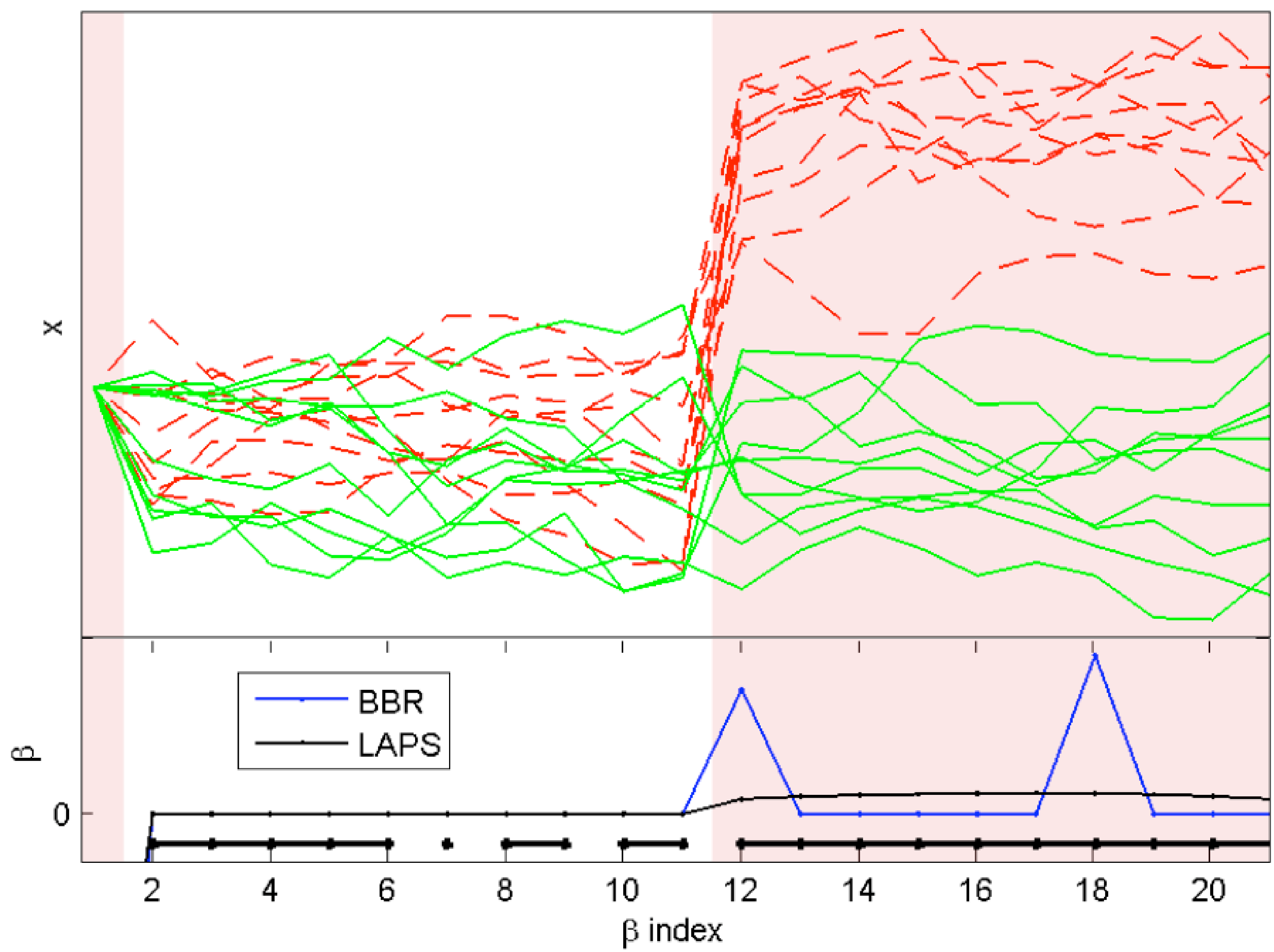
-imputation

-common weeks only (0,4,8,26,30,38,42,46,50)

-no interactions

IGG_38	ED50_30	SI_8
IFNeli_8	ED50_38	ED50_42
IFNeli_26	IL4/IFNeli_0	

group+fusion combined?

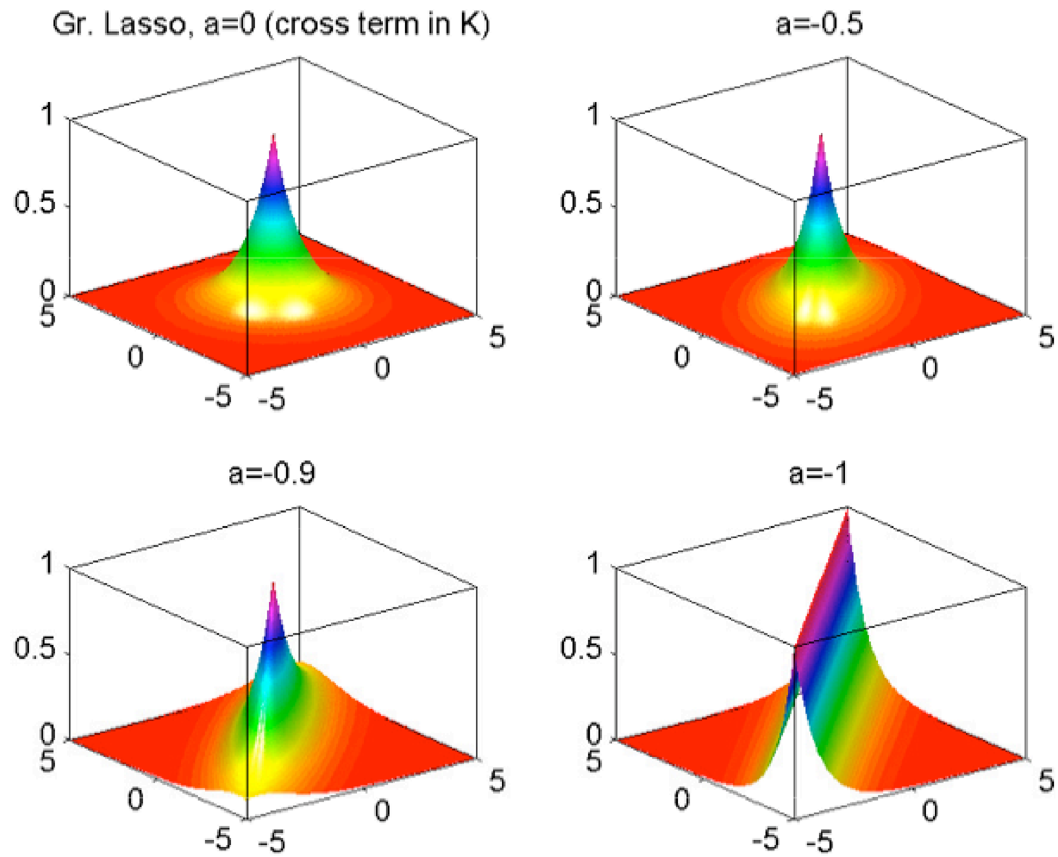

x
β
BBR
LAPS
0
2
4
6
8
10
12
14
16
18
20
β index

Group Lasso, Non-Identity

$$\frac{1}{2}\left\|Y - \sum_{j=1}^{J} X_j\beta_j\right\|^2 + \lambda \sum_{j=1}^{J} ||\beta_j||_{K_j}$$

$$\|\eta\|_K = (\eta' K \eta)^{1/2}$$

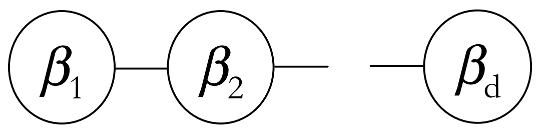
- multivariate power exponential prior
- KKT conditions lead to an efficient and straightforward block coordinate descent algorithm, similar to Tseng and Yun (2006).



“soft fusion”

LAPS: Lasso with Attribute Partition Search

- Group lasso
- Non-diagonal K to incorporate, e.g., serial dependence
- Within group have:



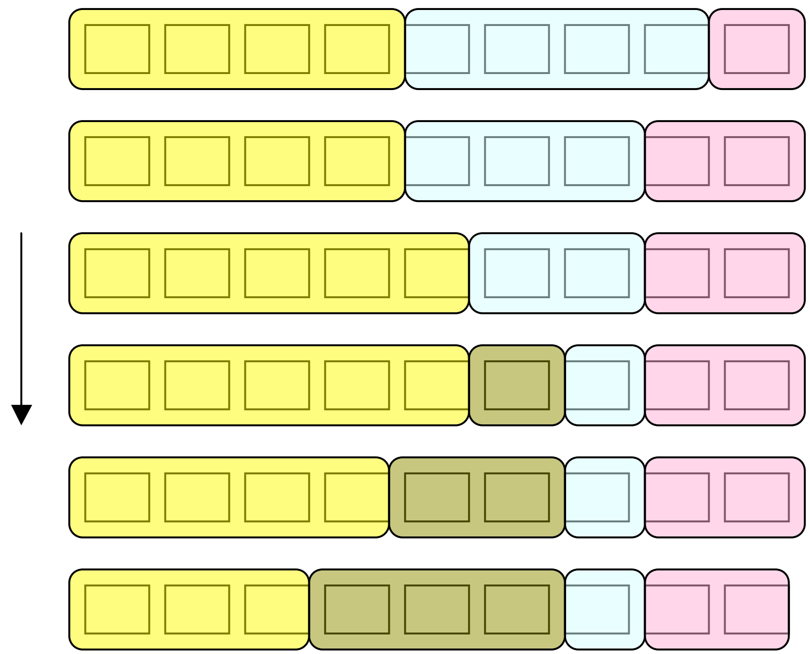
(block diagonal K)

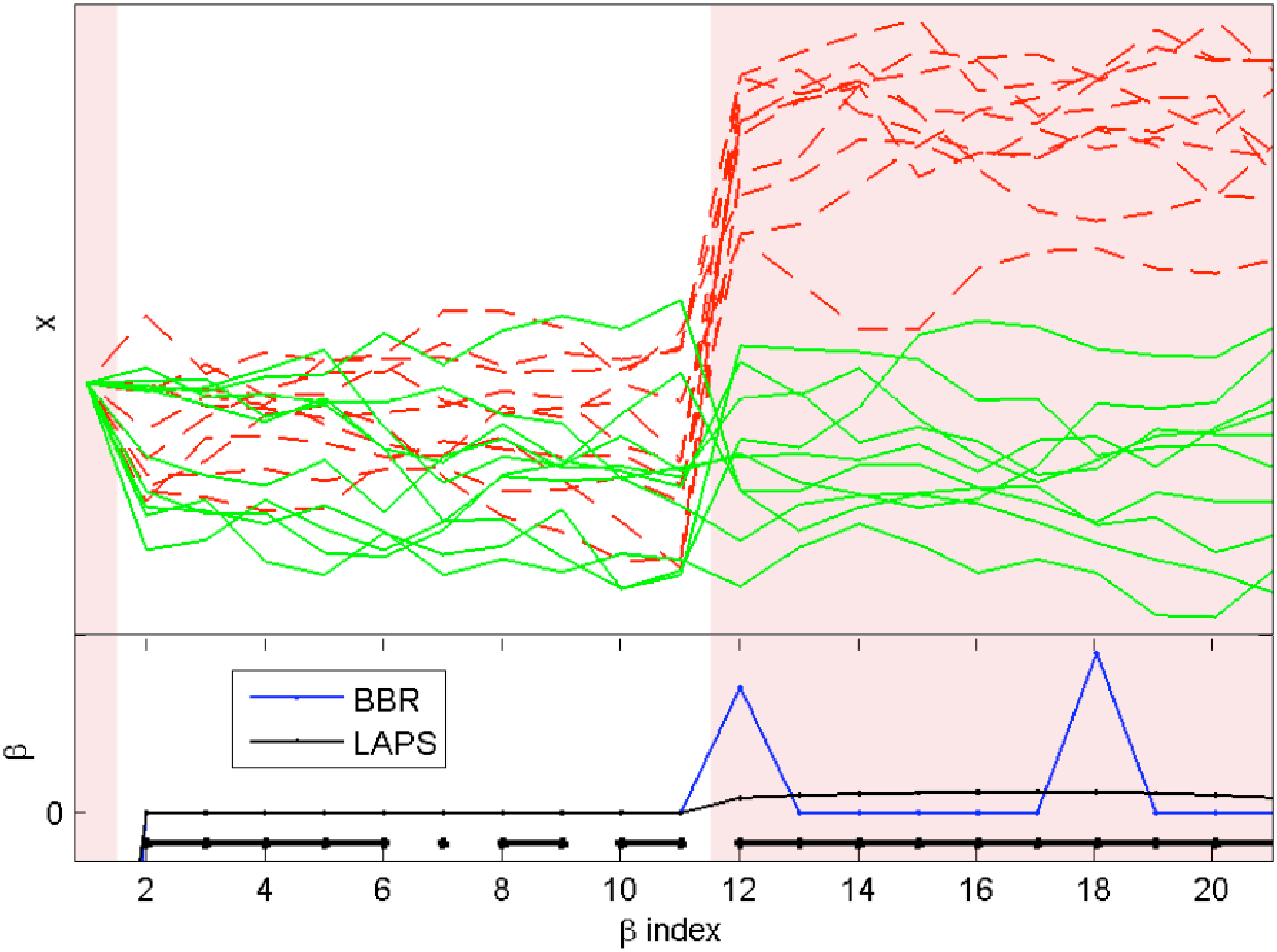
- Search for partitions that maximize a model score/average over partitions

Balakrishnan and Madigan (2007)

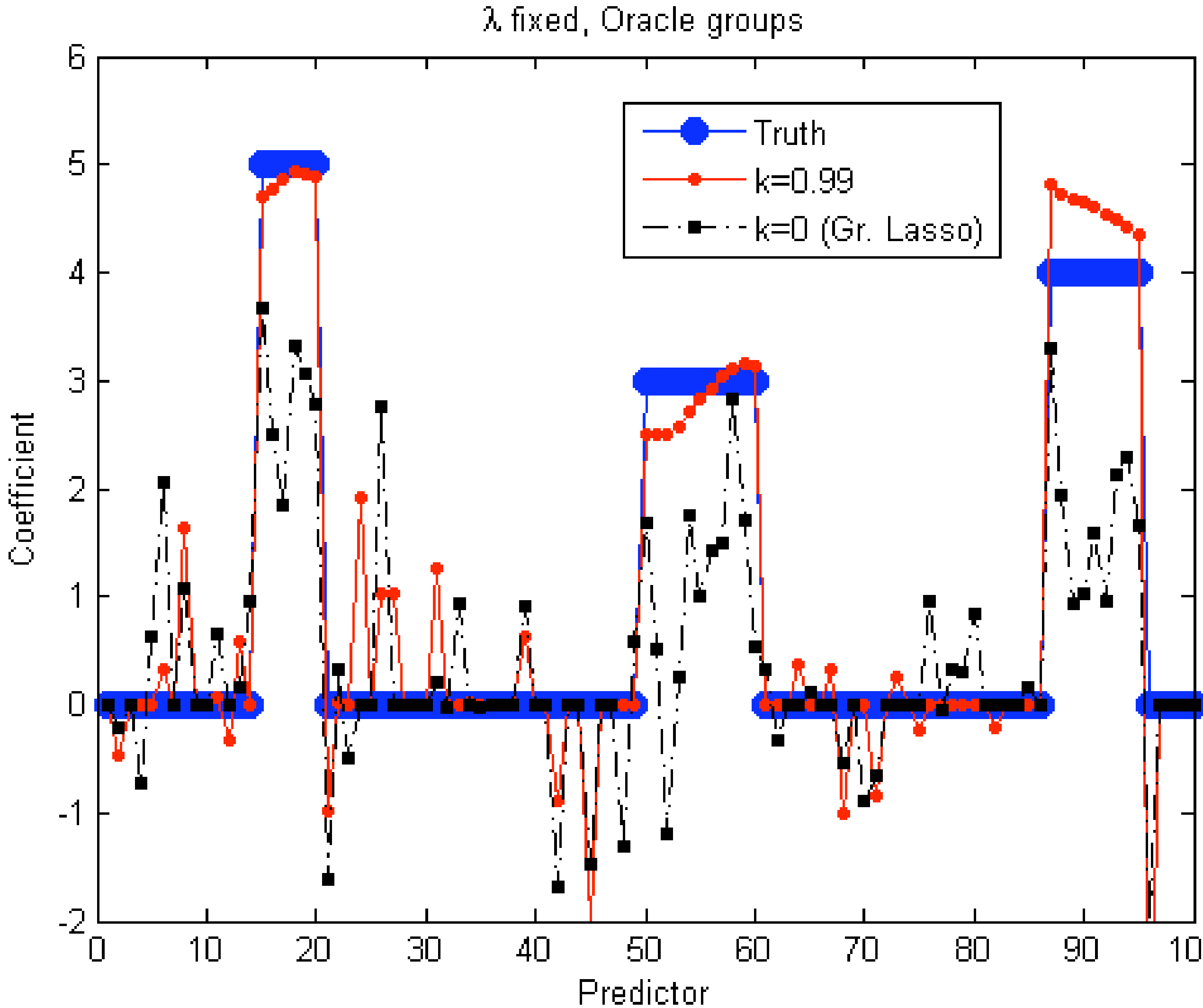
LAPS: Lasso with Attribute Partition Search

- Currently use a BIC-like score and/or test accuracy
- Hill-climbing vs. MCMC/BMA
- Uniform prior on partition space
- Consonni & Veronese (1995)
- Nonparametric



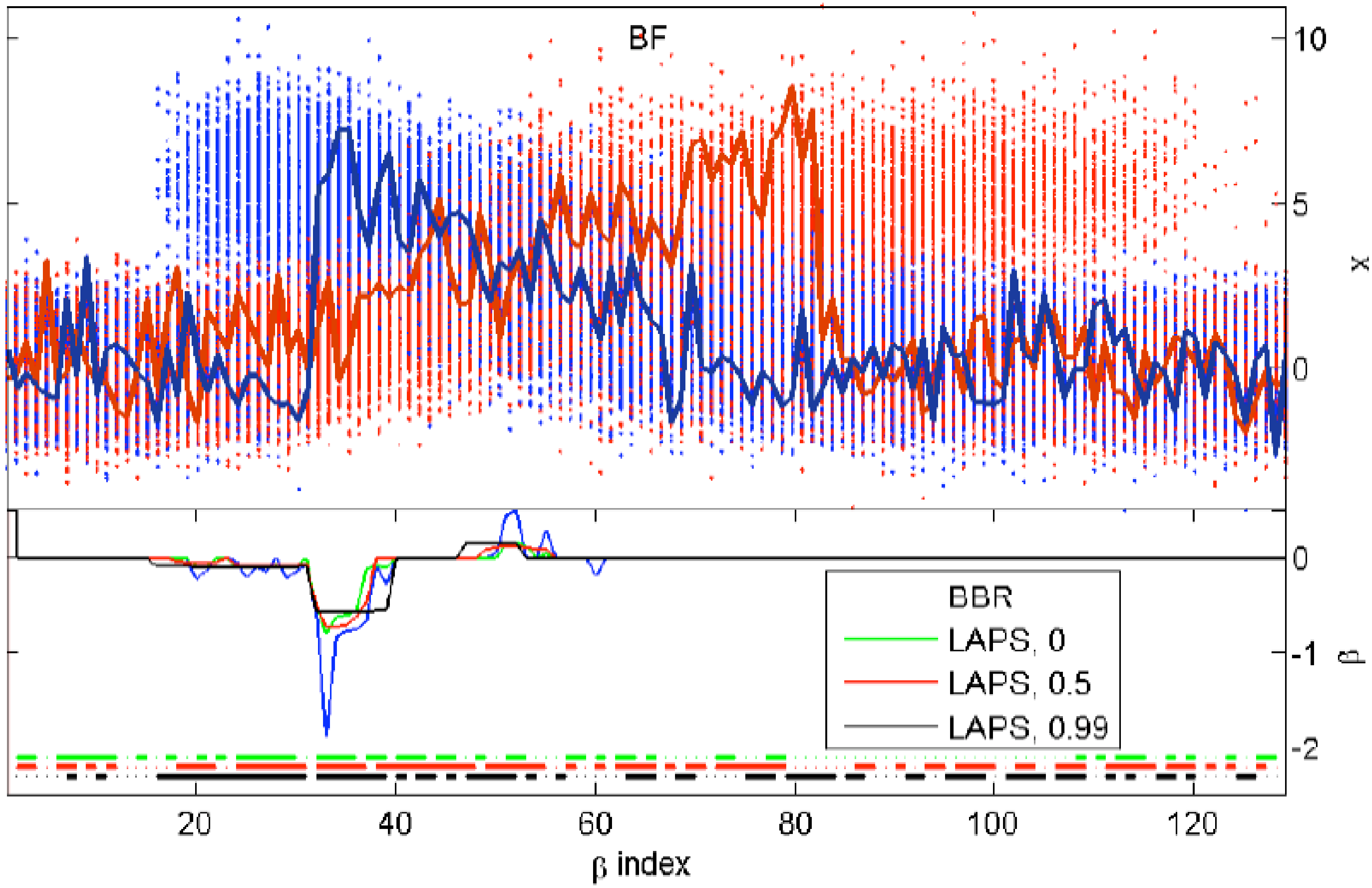


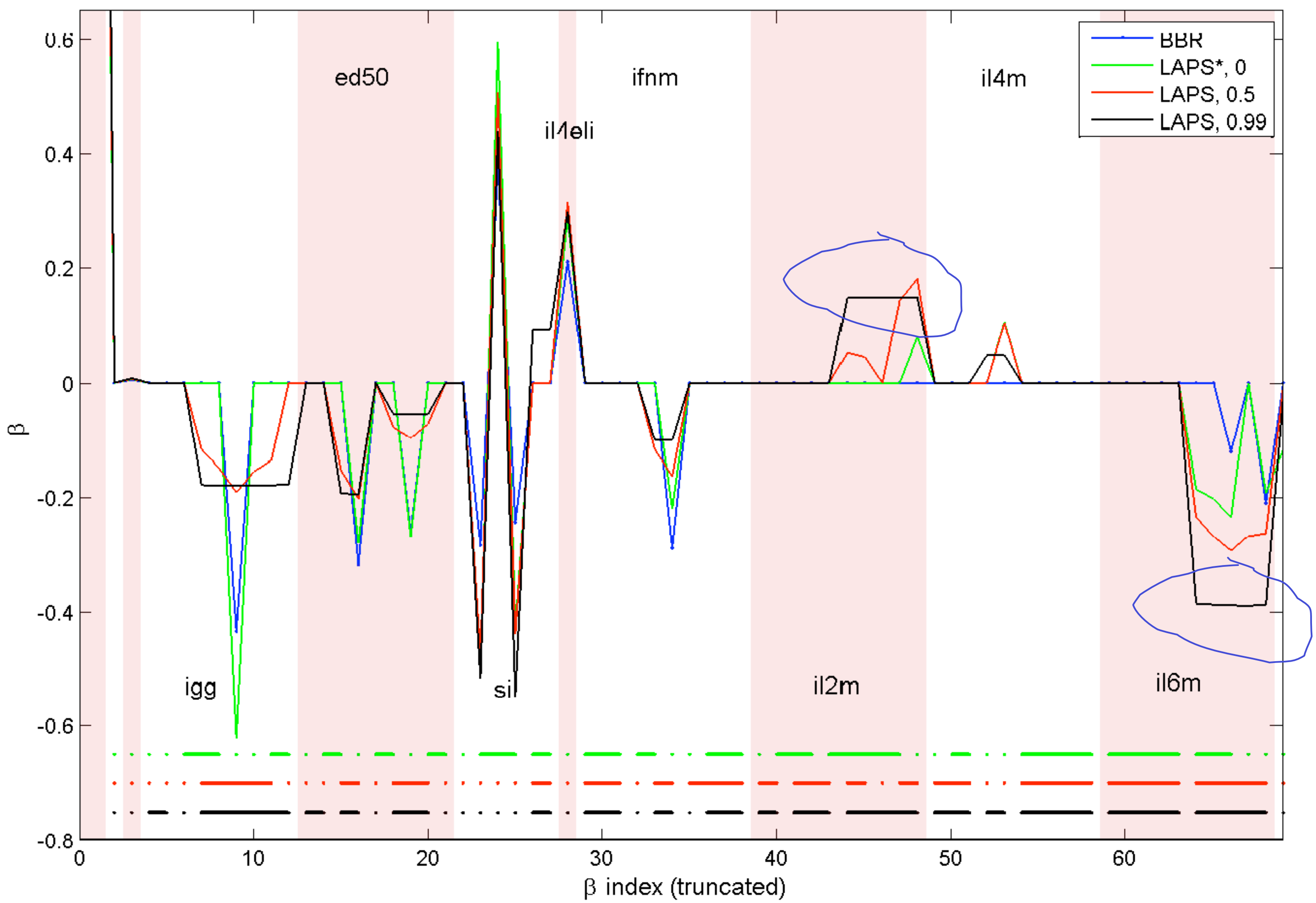
x
BBR
LAPS
β
0
2
4
6
8
10
12
14
16
18
20
β index



λ fixed, Oracle groups
Truth
k=0.99
k=0 (Gr. Lasso)
Coefficient
Predictor
6
5
4
3
2
1
0
-1
-2
0
10
20
30
40
50
60
70
80
90
100

LAPS: Bell-Cylinder example



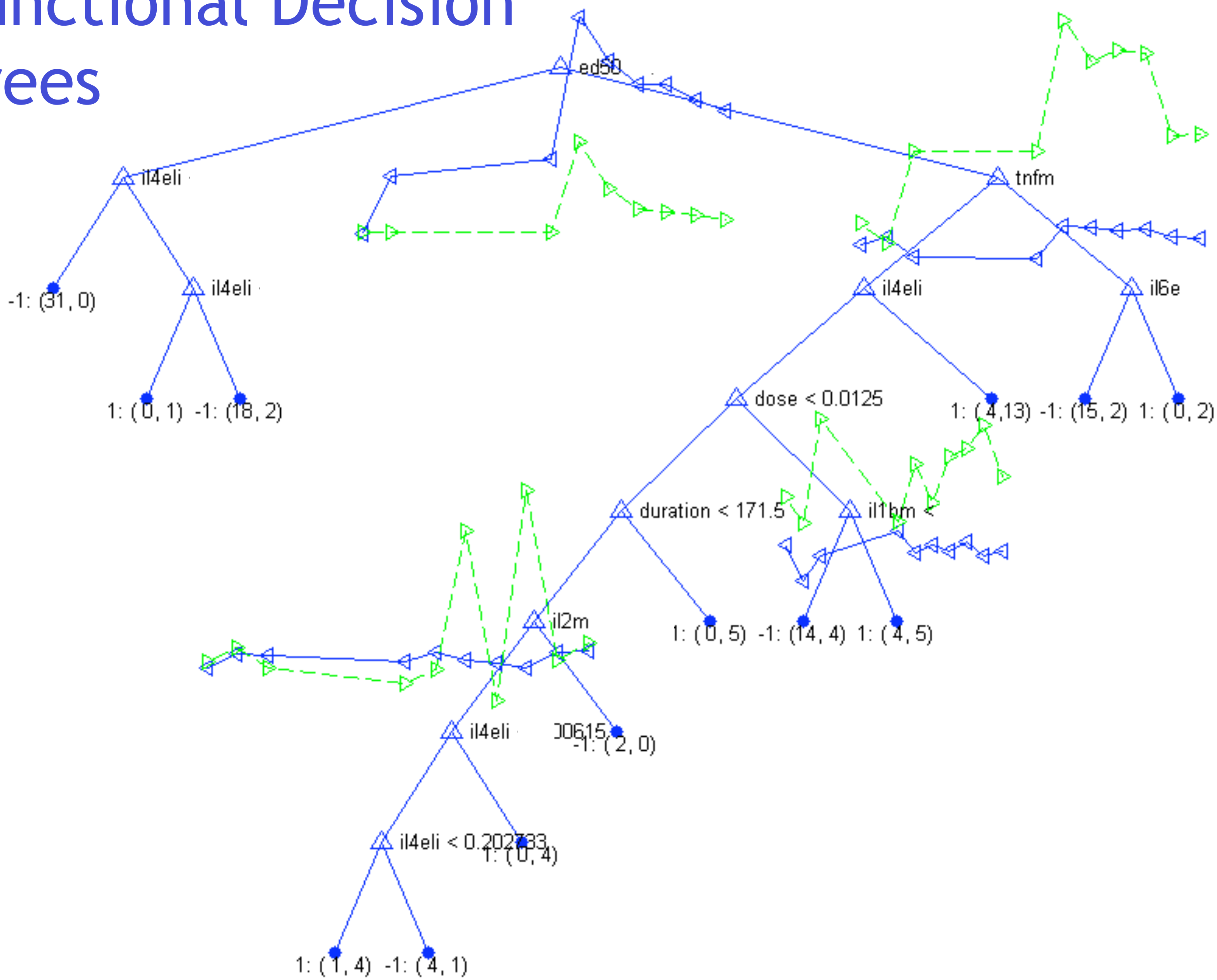


BBR
LAPS*, 0
LAPS, 0.5
LAPS, 0.99
ed50
ifnm
il4m
il4eli
igg
si
il2m
il6m
β
β index (truncated)
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
0
10
20
30
40
50
60

Predictive performance—estimated error rates[9]

Data	Lasso		LAPS		
	% Err	V^*	% Err	V^*	k^*
SM1	**25.43**	0.45	27.52	0.28	0
SM2	**30.83**	0.15	34.38	0.54	0.99
SM3	35.98	0.15	**30.62**	0.37	0.99
LG1	22.31	0.15	**22.09**	0.54	0.74
LG2	21.14	0.5	**21.09**	0.63	0
LG3	21.86	0.35	21.68	0.19	0.99
BF	0.1887 $\pm$ 0.6	200	0.1887 $\pm$ 0.6	0.45	0
NHP	30.81 $\pm$ 11.97	0.2	**28.02** $\pm$ 10.27	0.46	0

Functional Decision Trees



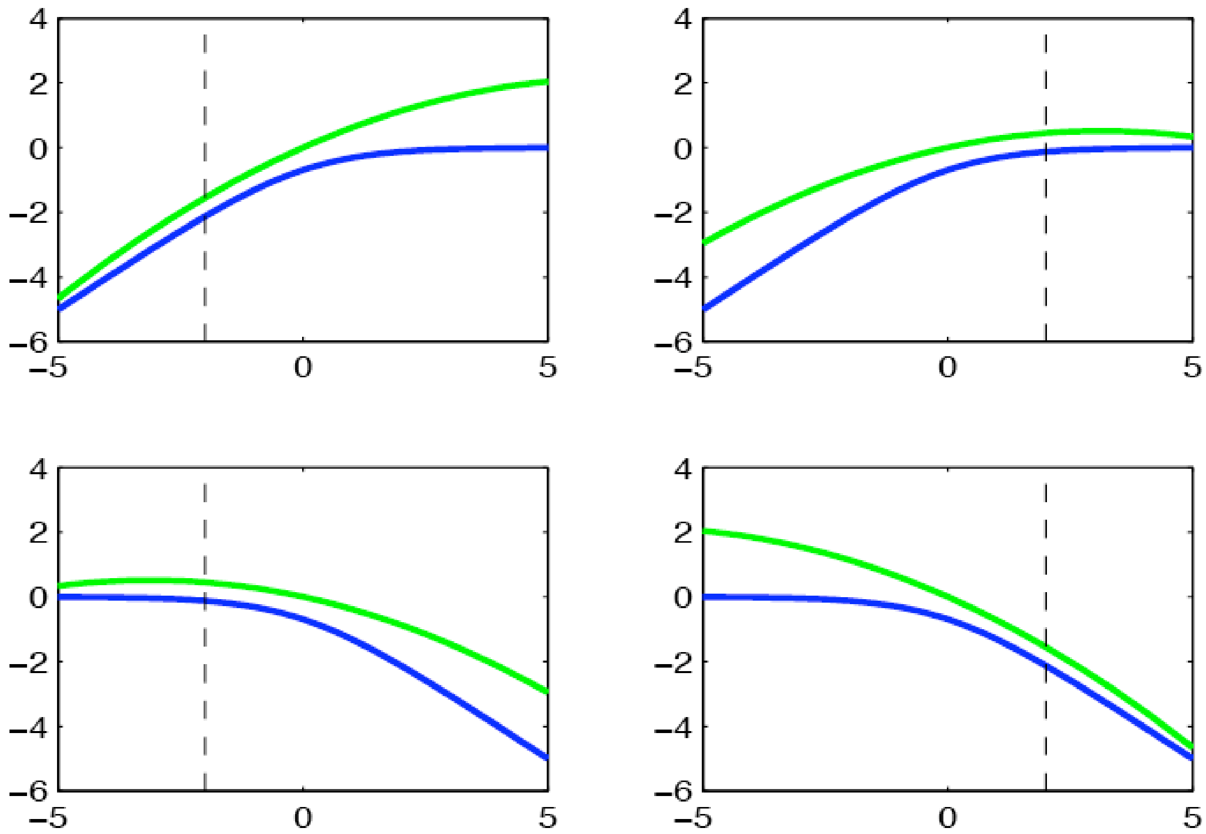
Balakrishnan and Madigan (2006)

Computational Landscape

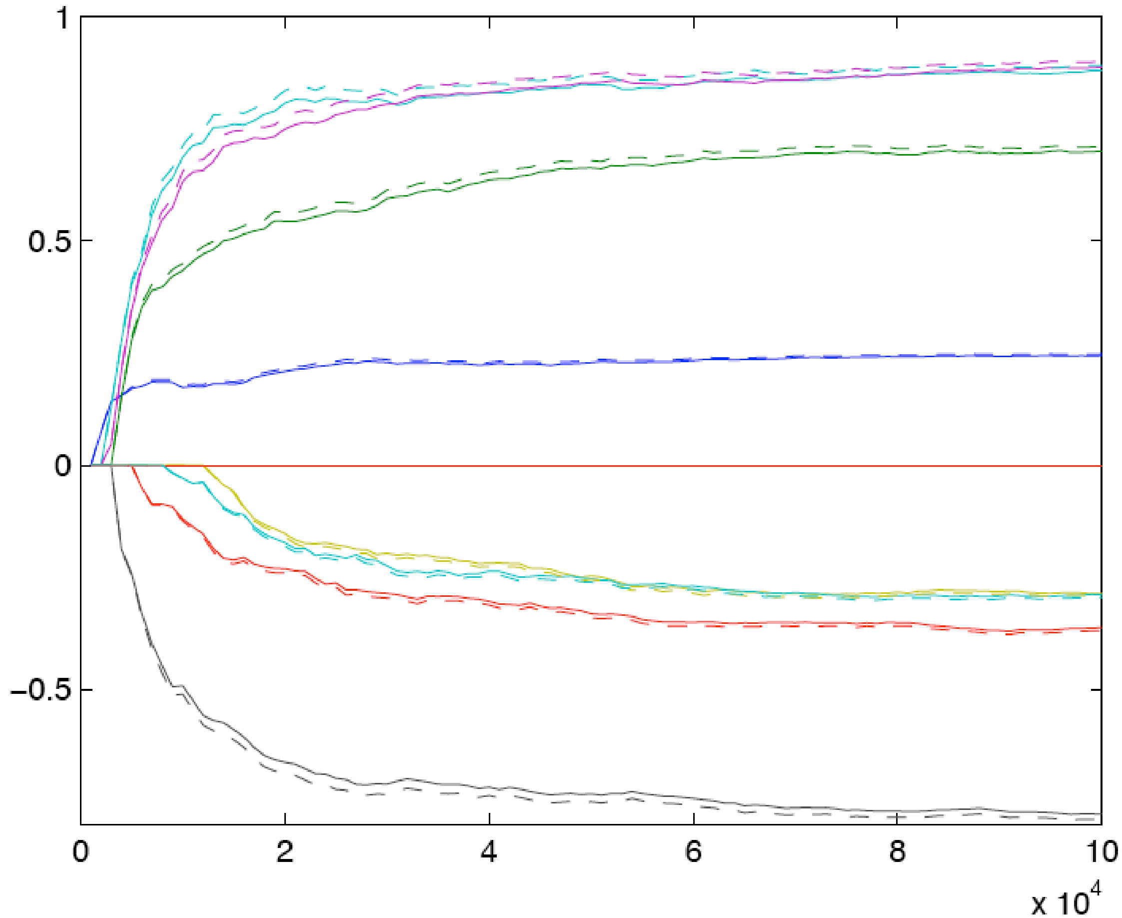
	Full Bayes	MAP Bayes
	$p(Y_{t+1}=1\|\bar{Y}_t) = \int p(Y_{t+1}\|\beta)p(\beta\|\bar{Y}_t)d\beta$	$p(Y_{t+1}=1\|\bar{Y}_t) \approx p(Y_{t+1}\|\hat{\beta}(\bar{Y}_t))$
Batch	Variational (Jordan & Jaakola) MCMC	Gauss-Seidel (BXR) Interior Point (Boyd)
Online	online variational Sequential MC (Chopin, 2002; Ridgeway & Madigan, 2003)	Online EM, Quasi-Bayes (Titterington, 1984; Smith & Makov, 1978)

Quadratic Approximation for Log-Likelihood Terms

$$\log\left(y_i\Phi(\boldsymbol{\beta}^T\mathbf{x}_i) + (1-y_i)(1-\Phi(\boldsymbol{\beta}^T\mathbf{x}_i))\right) \approx a_i(\boldsymbol{\beta}^T\mathbf{x}_i)^2 + b_i(\boldsymbol{\beta}^T\mathbf{x}_i) + c_i$$



Excellent Performance with Small *d*

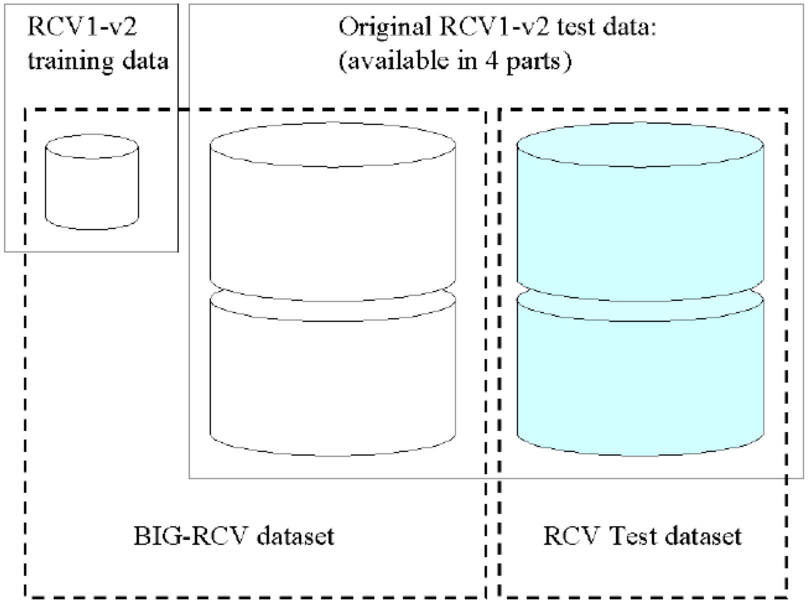


Big-*d*

- Multi-pass, limited memory algorithm
- Highly scaleable
- Example: RCV-1, n=420K, d=288K

Balakrishnan and Madigan (2008)

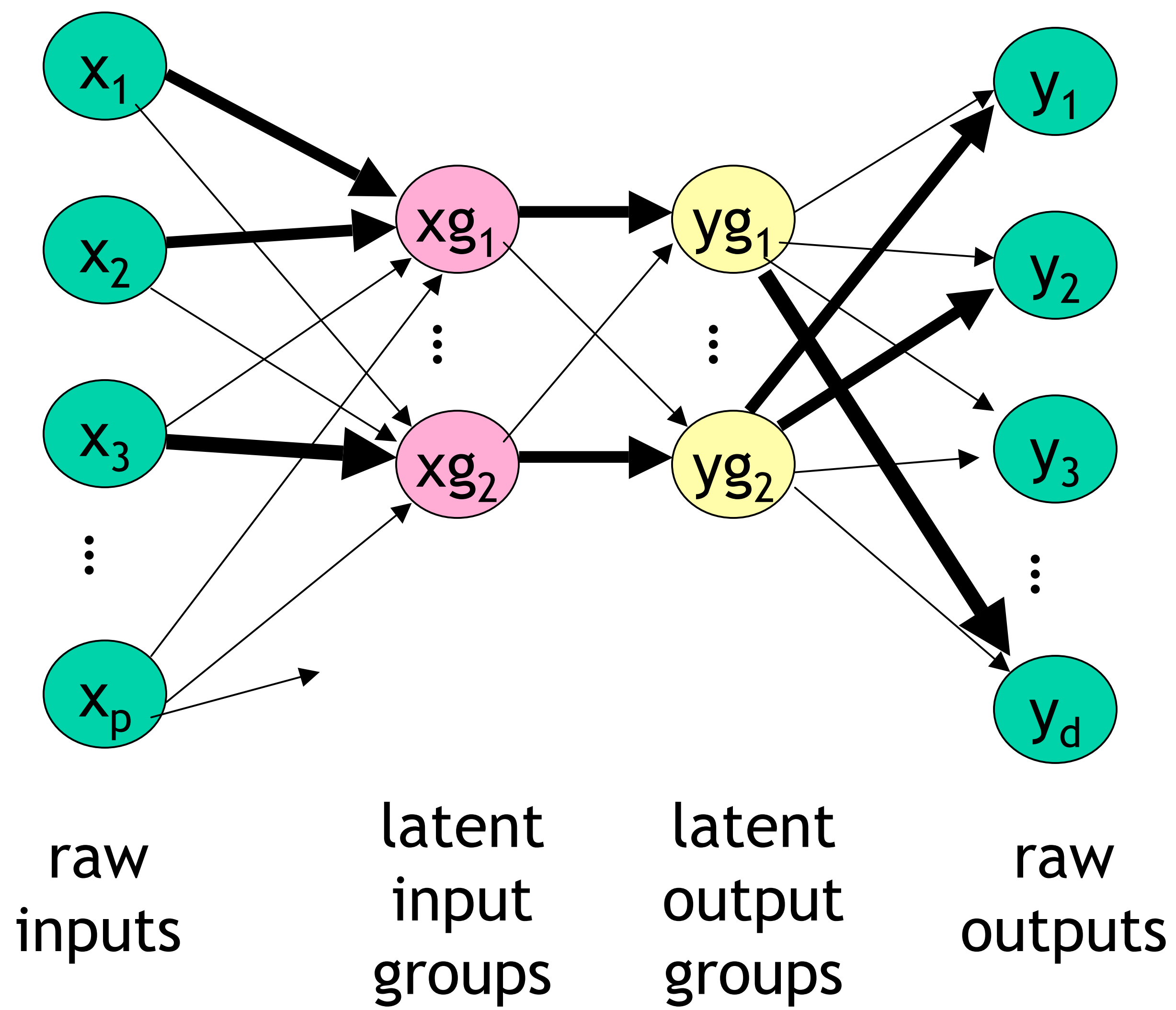
RCV-1 Results



	"Optimized" β trained on RCV1-v2 training data		"Naive" β trained on BIG-RCV	
	Relevant	Not Relevant	Relevant	Not Relevant
Retr.	38,821	7,415	40,655	6,017
Not R.	16,368	319,994	14,534	321,392
Pr.	**83.96%**		**87.11%**	
Re.	**70.34%**		**73.67%**	
	$d = 47,236,\ t = 23,149$		$d = 288,062,\ t = 421,816$	

Back to drug safety…

- Real question: which classes of drugs cause which groups of adverse events
- Example: COX-2 inhibitors cause cardiovascular thrombotic events



- idea: groups of x's (e.g. drugs) cause groups of y's (e.g. adverse events)
- all nodes binary; logistic regression for each node given parents
- need prior on number of hidden units, etc.

Latent Space Model

$$\operatorname{logit}(\Pr(Y_{i,j} = 1 \mid Z, X, \beta)) = \sum_{k=1}^{p} \beta_k X_{k,i,j} - \left\| Z_i - Z_j \right\|$$

Hoff, Raftery, and Handcock (2002), Krivitsky, Handcock, Raftery, and Hoff (2007)

- generalize to two classes of actors and groups bigger than two:

$$\operatorname{logit}\left[\Pr(X_I, Y_J \mid Z)\right] = \sum_{\substack{i: X_i \in X_I \\ j: X_j \in X_J}} \left\{ \sum_{k=1}^{K} \beta_k X_{k,i,j} - \left\| Z_{Xi} - Z_{Yj} \right\| \right\}$$

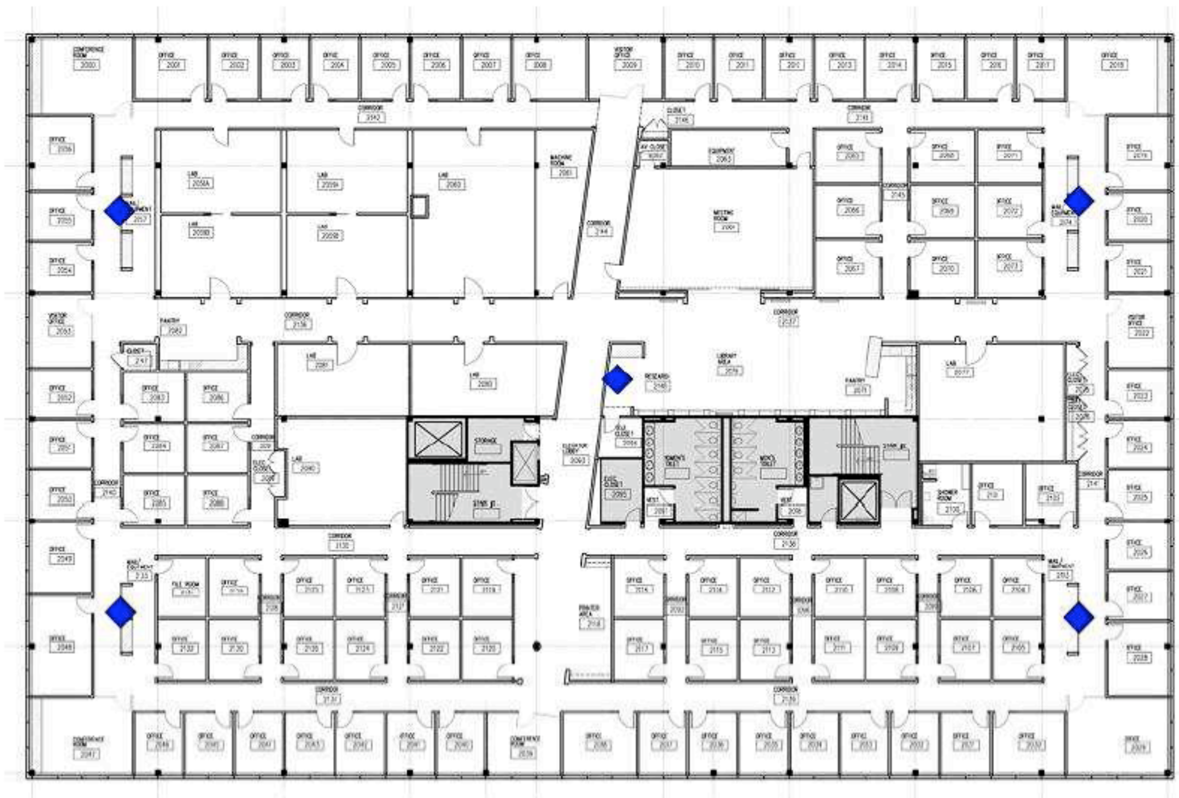
Gormley and Murphy (2006)

Final Comments

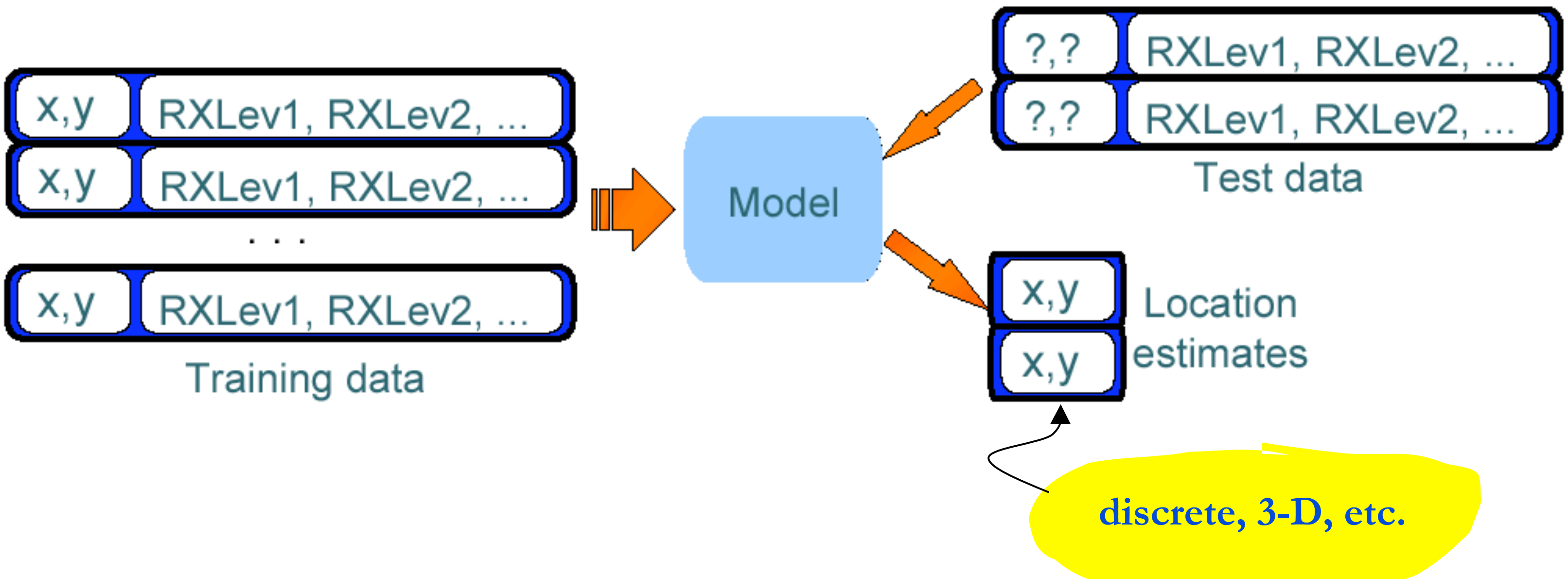
- Predictive modeling with 10^5-10^7 predictor variables is feasible and sometimes useful
- Google builds ad placement models with 10^8 predictor variables
- Computation is a central problem in Statistics

The Problem

- **Estimate the physical location of a wireless terminal/user in an enterprise**
  - Radio wireless communication network, specifically, 802.11-based

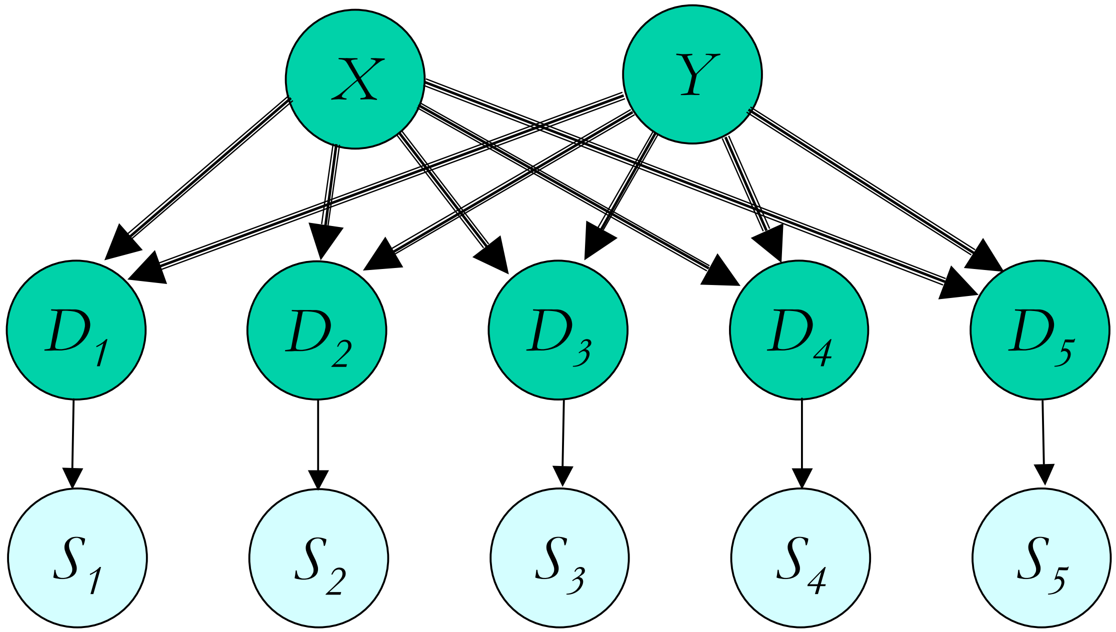


Prior Work



- Take signal strength measures at many points in the site and do a closest match to these points in signal strength vector space. [e.g. Microsoft's RADAR system]
- Take signal strength measures at many points in the site and build a multivariate regression model to predict location (e.g., Tirri's group in Finland)
- Some work has utilized wall thickness and materials

Bayesian Graphical Model Approach



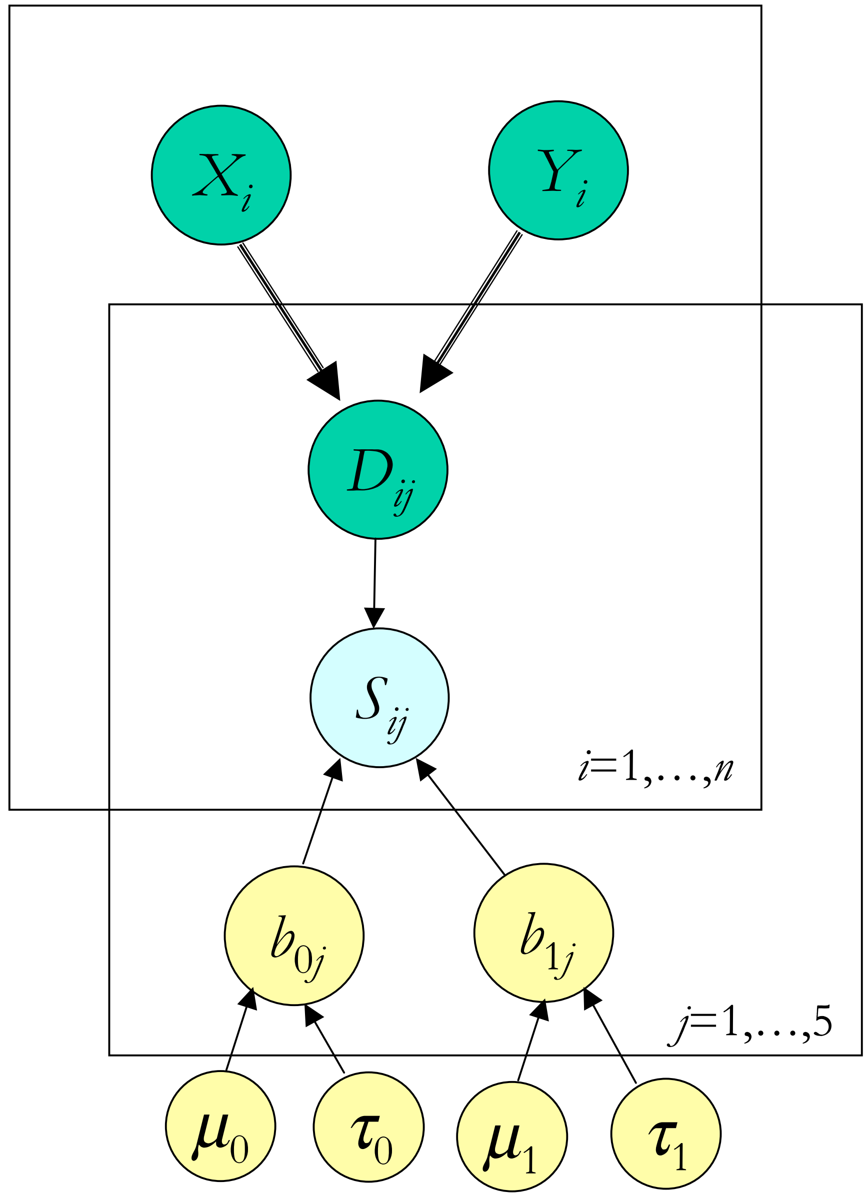
$$X, Y \sim unif$$

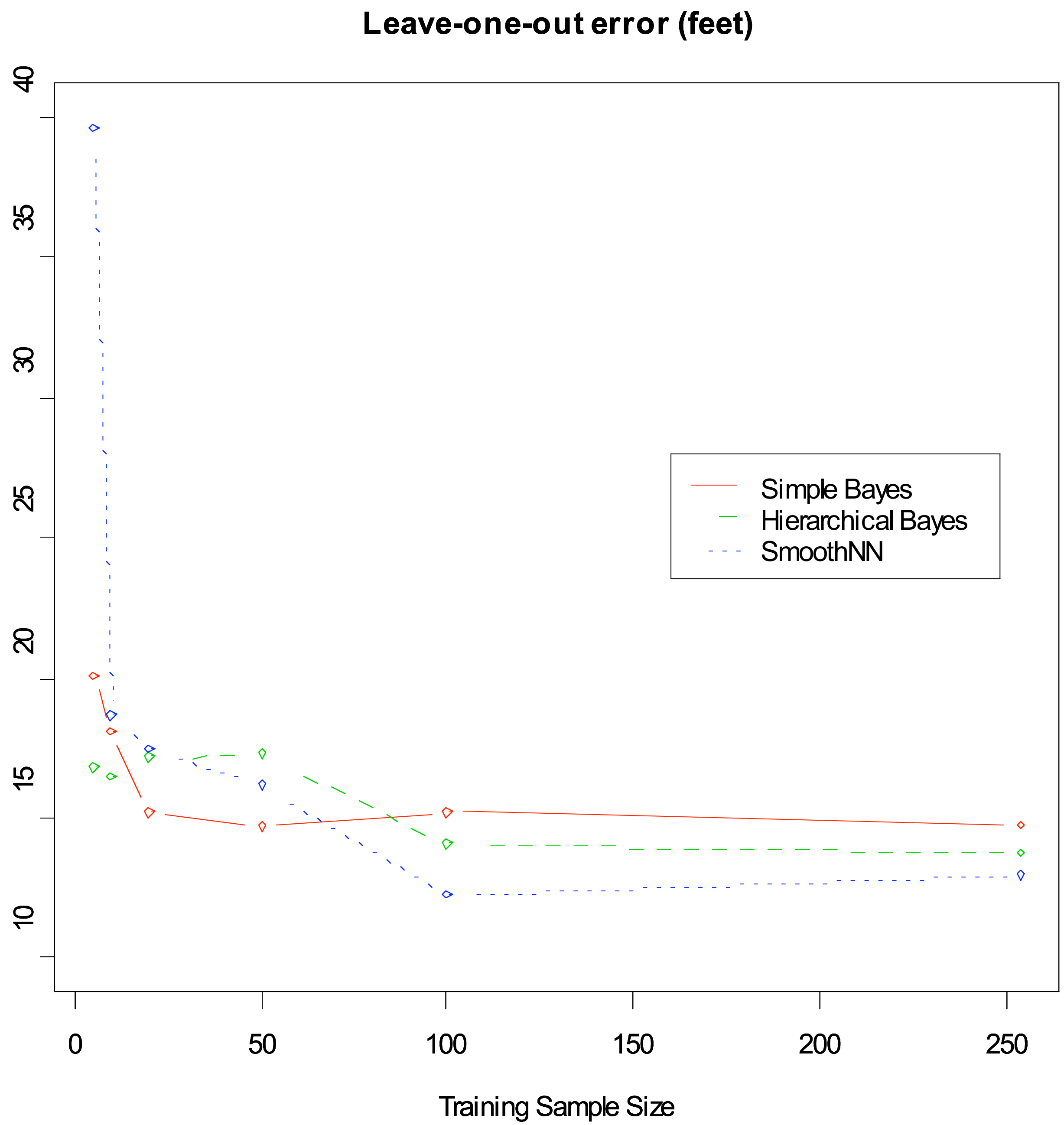
$D_i(X,Y) =$ distance to the ith access point

$$S_i \sim N(b_{i0} + b_{i1}\log D_i, \sigma_i^2),\ i = 1,\ldots,5$$

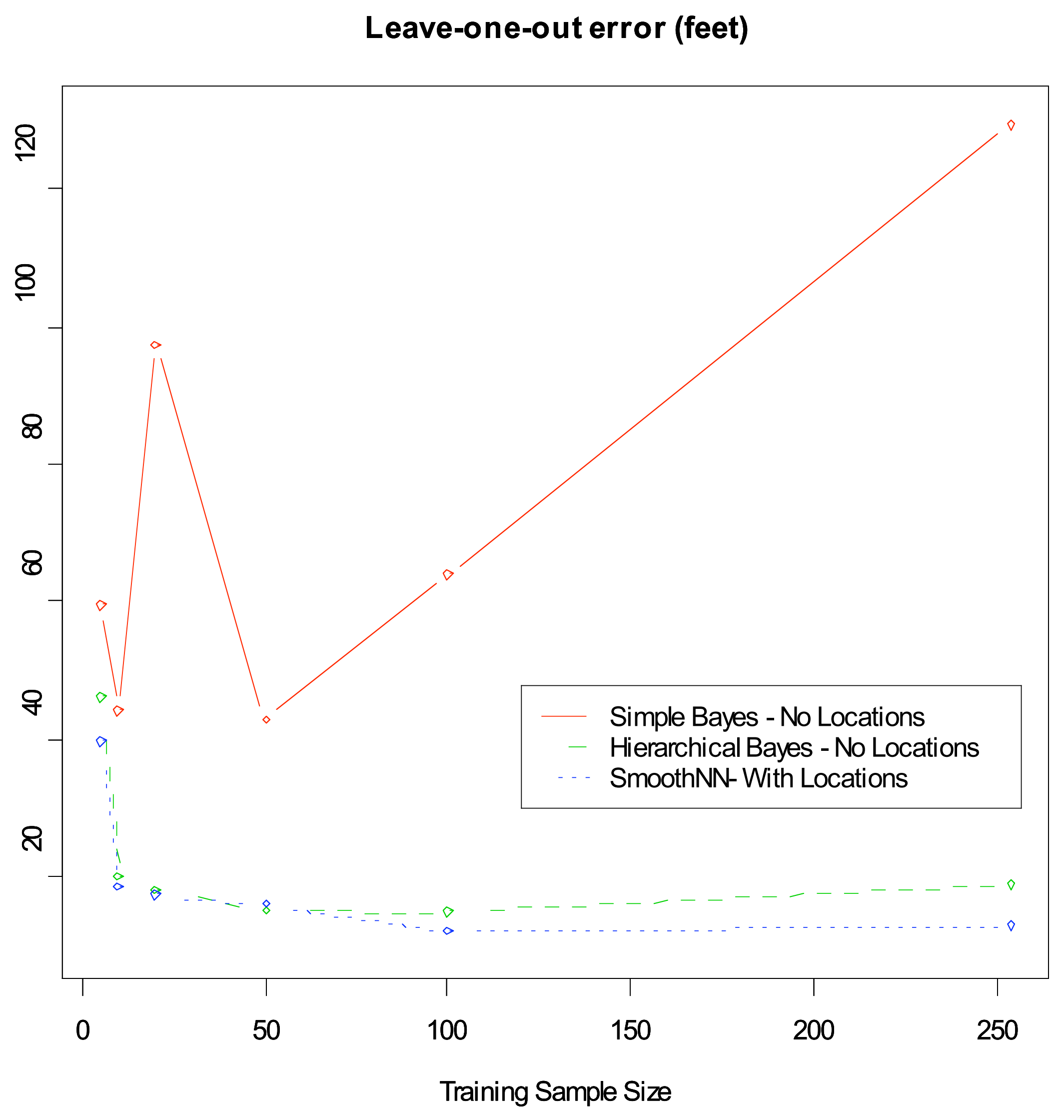
average

Hierarchical Model



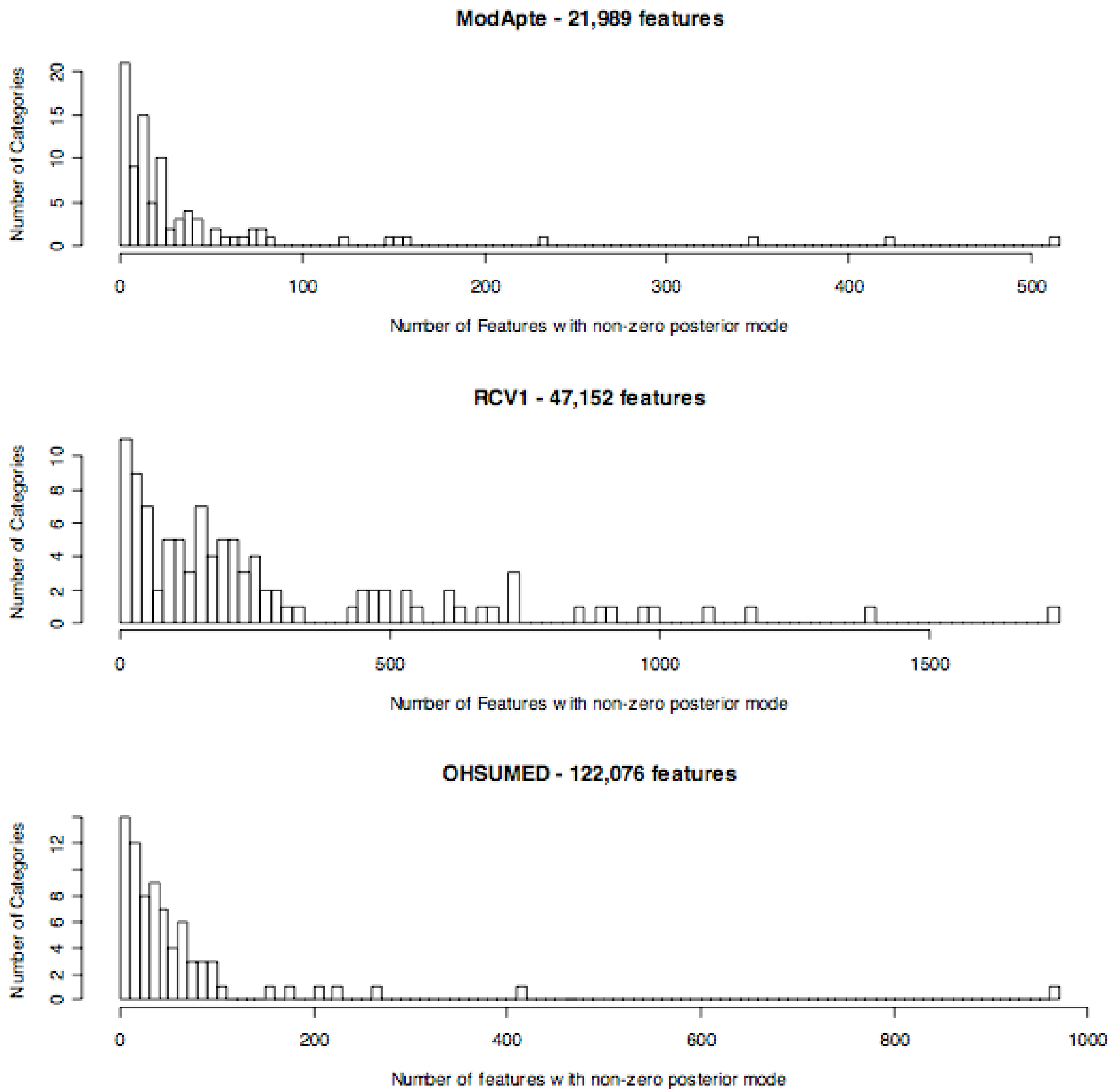

Leave-one-out error (feet)
Simple Bayes
Hierarchical Bayes
SmoothNN
10
15
20
25
30
35
40
0
50
100
150
200
250
Training Sample Size

What if we had no locations in the training data?



Future Work

- Rigorous derivation of BIC and df
- Prior on partitions
- Better search strategies for partition space
- Out of sample predictive accuracy
- LAPS C++ implementation
- Fully Bayesian alternative



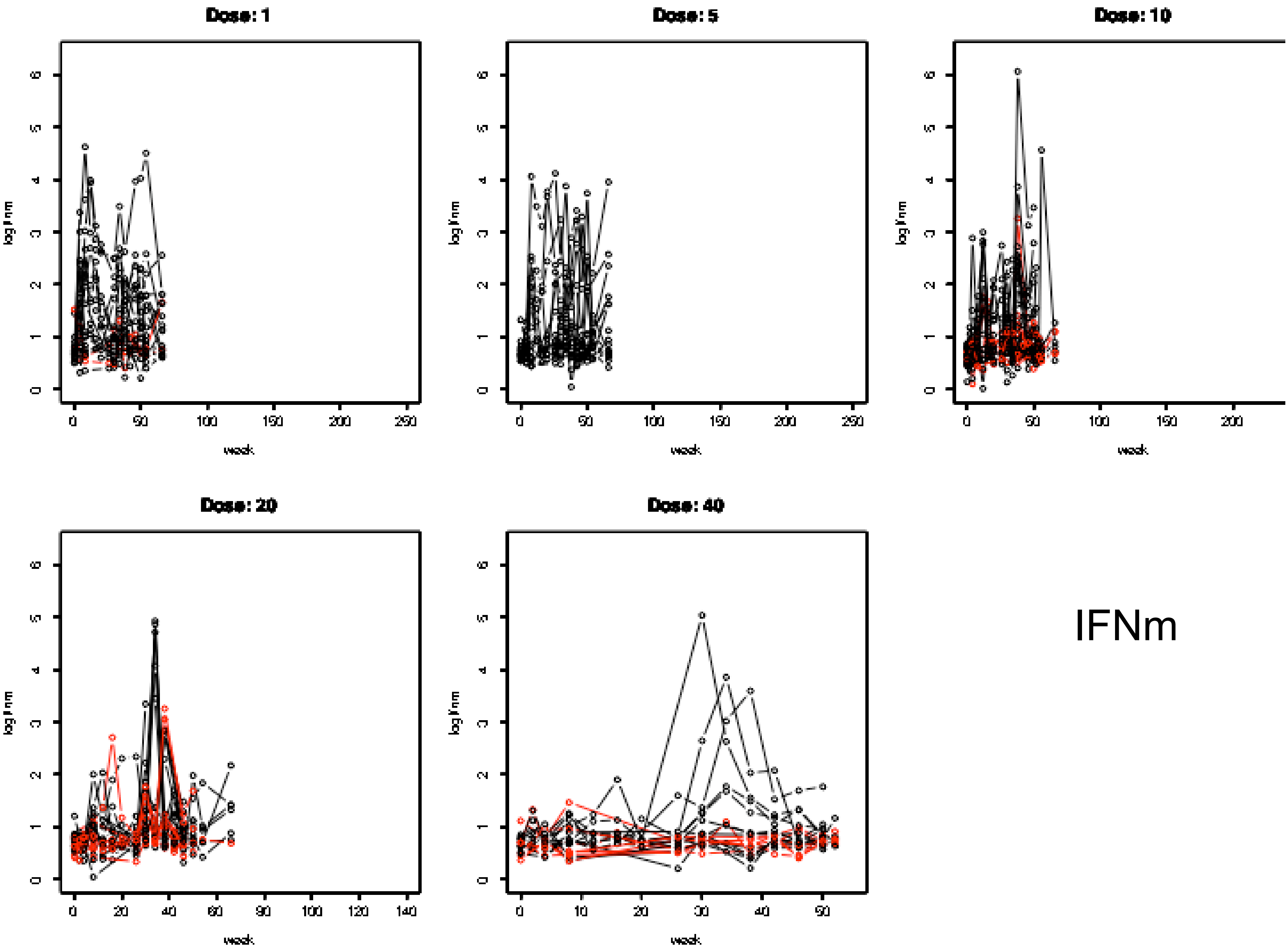
Genkin et al. (2004)

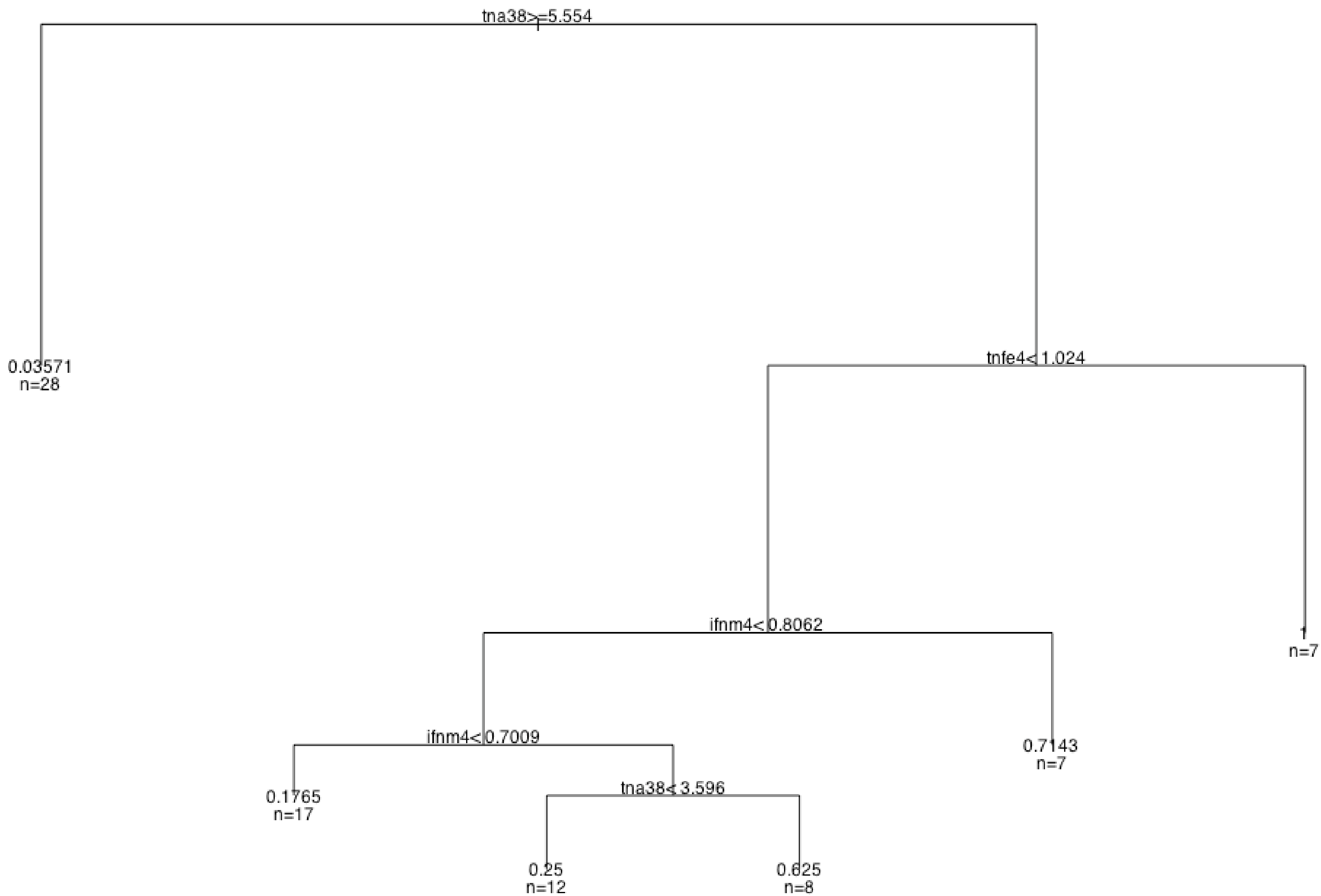
ModApte: Bayesian Perspective Can Help

(training: 100 random samples)

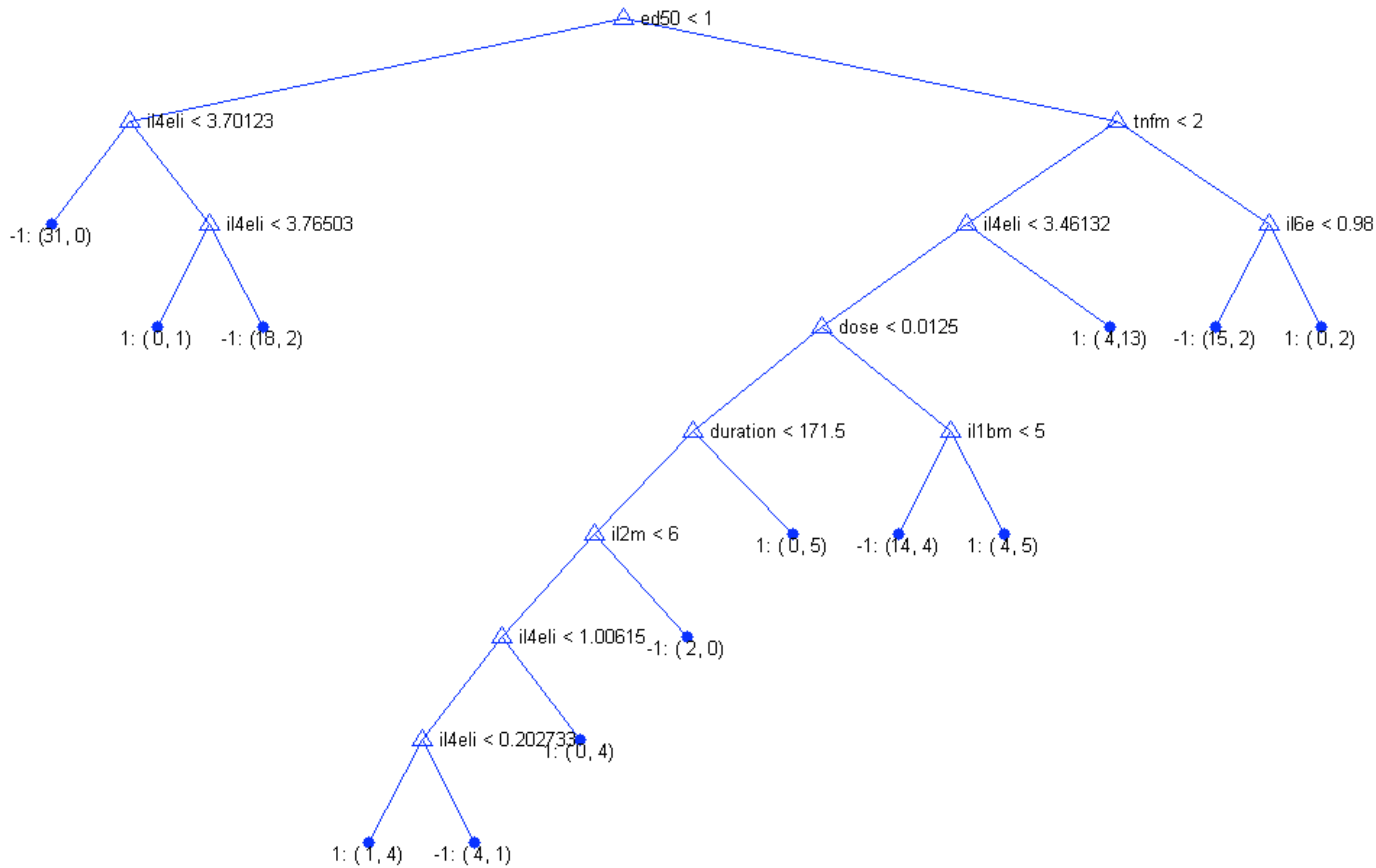
	Macro F1	ROC
Laplace	37.2	76.2
Laplace & DK-based variance	65.3	87.1
Laplace & DK-based mode	72.0	93.5

Dayanik et al. (2006)

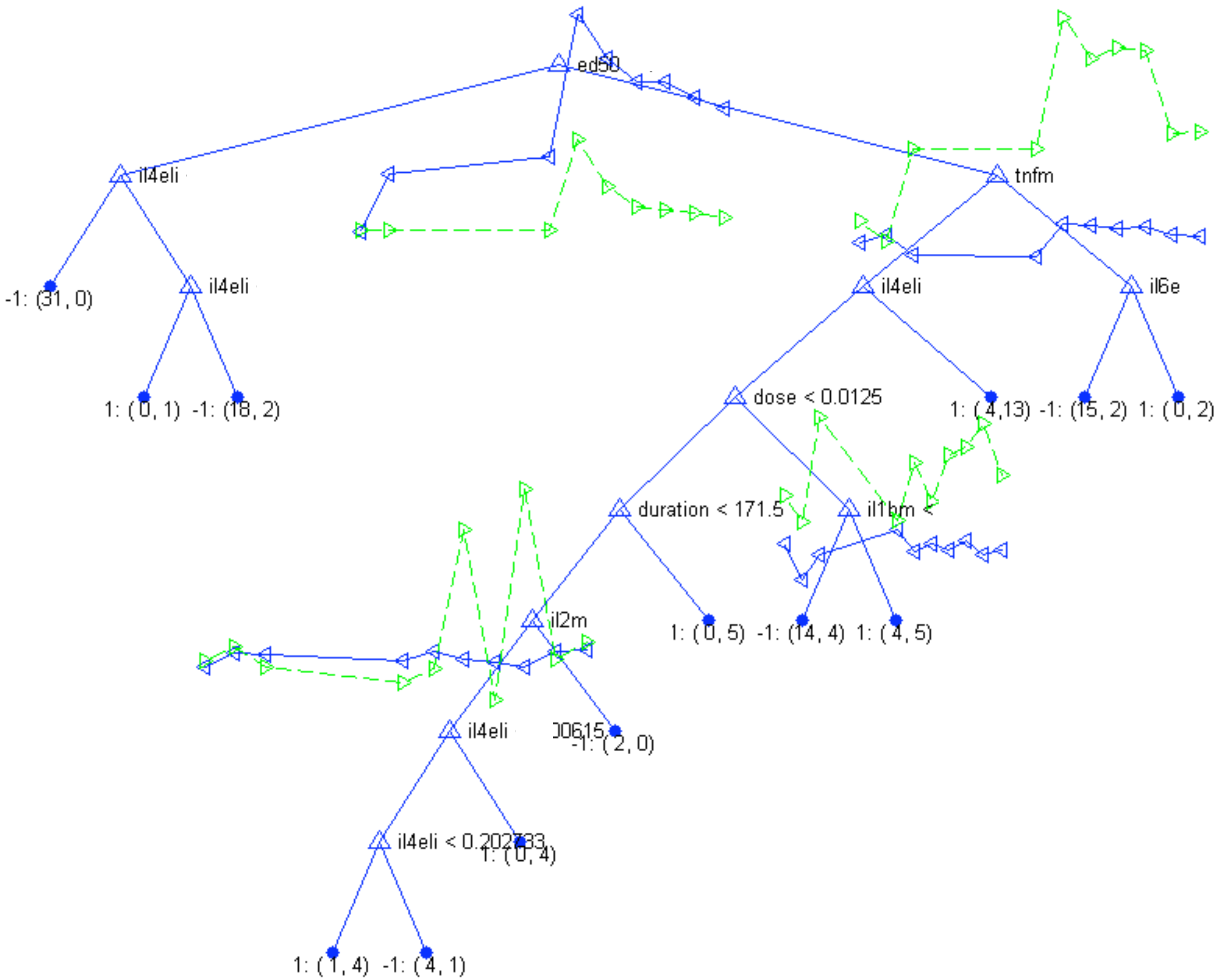

Dose: 1
Dose: 5
Dose: 10
Dose: 20
Dose: 40
log ifnm
week
IFNm


tna38>=5.554
0.03571
n=28
tnfe4< 1.024
ifnm4< 0.8062
1
n=7
ifnm4< 0.7009
0.7143
n=7
0.1765
n=17
tna38< 3.596
0.25
n=12
0.625
n=8

Groups 1-3


ed50 < 1
il4eli < 3.70123
tnfm < 2
-1: (31, 0)
il4eli < 3.76503
il4eli < 3.46132
il6e < 0.98
1: (0, 1)
-1: (18, 2)
dose < 0.0125
1: (4,13)
-1: (15, 2)
1: (0, 2)
duration < 171.5
il1bm < 5
il2m < 6
1: (0, 5)
-1: (14, 4)
1: (4, 5)
il4eli < 1.00615
-1: (2, 0)
il4eli < 0.202733
1: (0, 4)
1: (1, 4)
-1: (4, 1)

“



Balakrishnan and Madigan (2006)

L1 Logistic Regression

-imputation

-common weeks only (0,4,8,26,30,38,42,46,50)

-no interactions

IGG_38	-0.16 (0.17)
ED50_30	-0.11 (0.14)
SI_8	-0.09 (0.30)
IFNeli_8	-0.07 (0.24)
ED50_38	-0.03 (0.35)
ED50_42	-0.03 (0.36)
IFNeli_26	-0.02 (0.26)
IL4/IFNeli_0	+0.04 (0.36)

```
bbrtrain -p 1 -s --autosearch --accurate commonBBR.txt commonBBR.mod
```

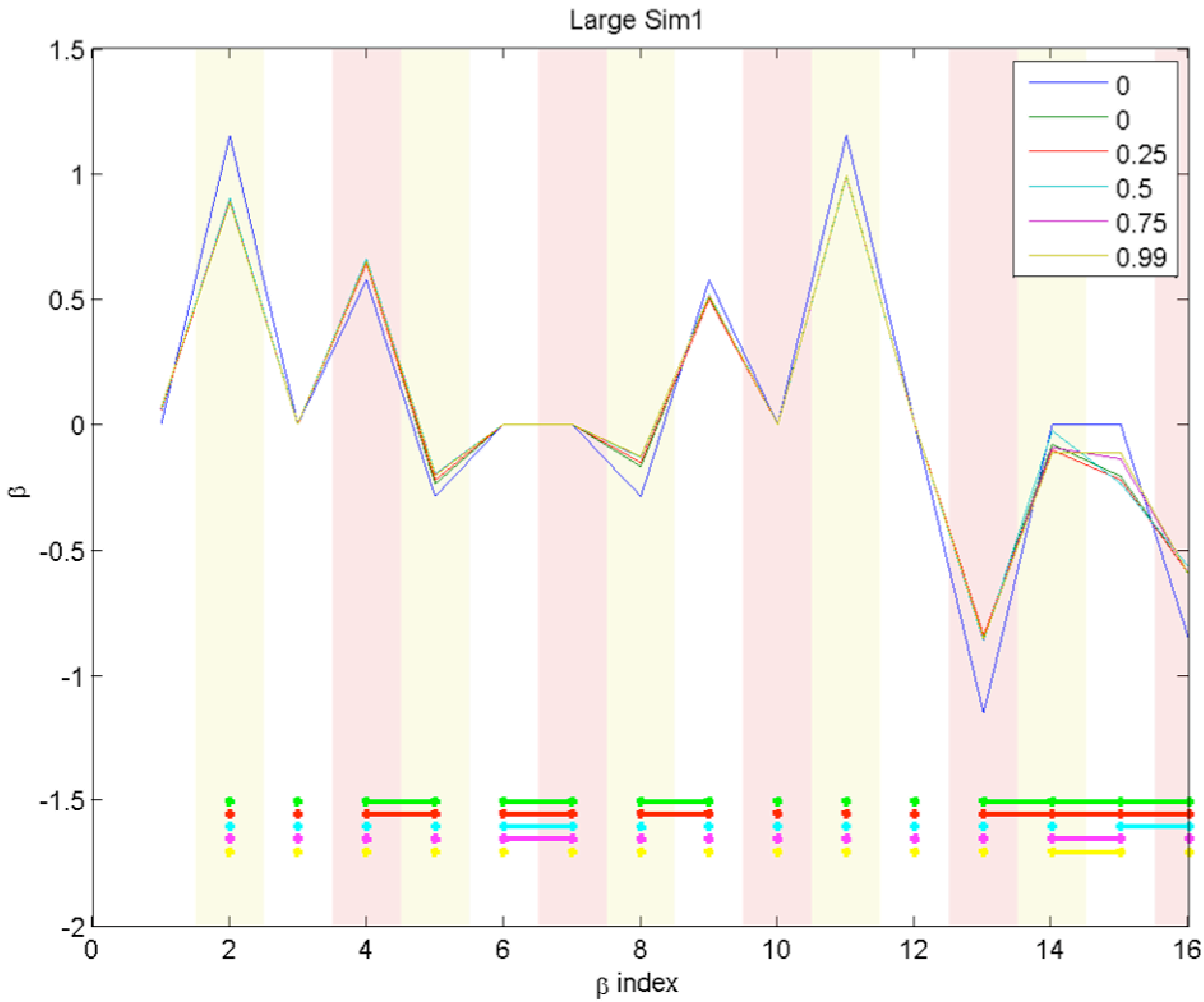
LAPS Simulation Study

X ~ N(0,1)^15 (iid, uncorrelated attributes)
Beta = one of three conditions (corresponding to Sim1, Sim2 and Sim3)

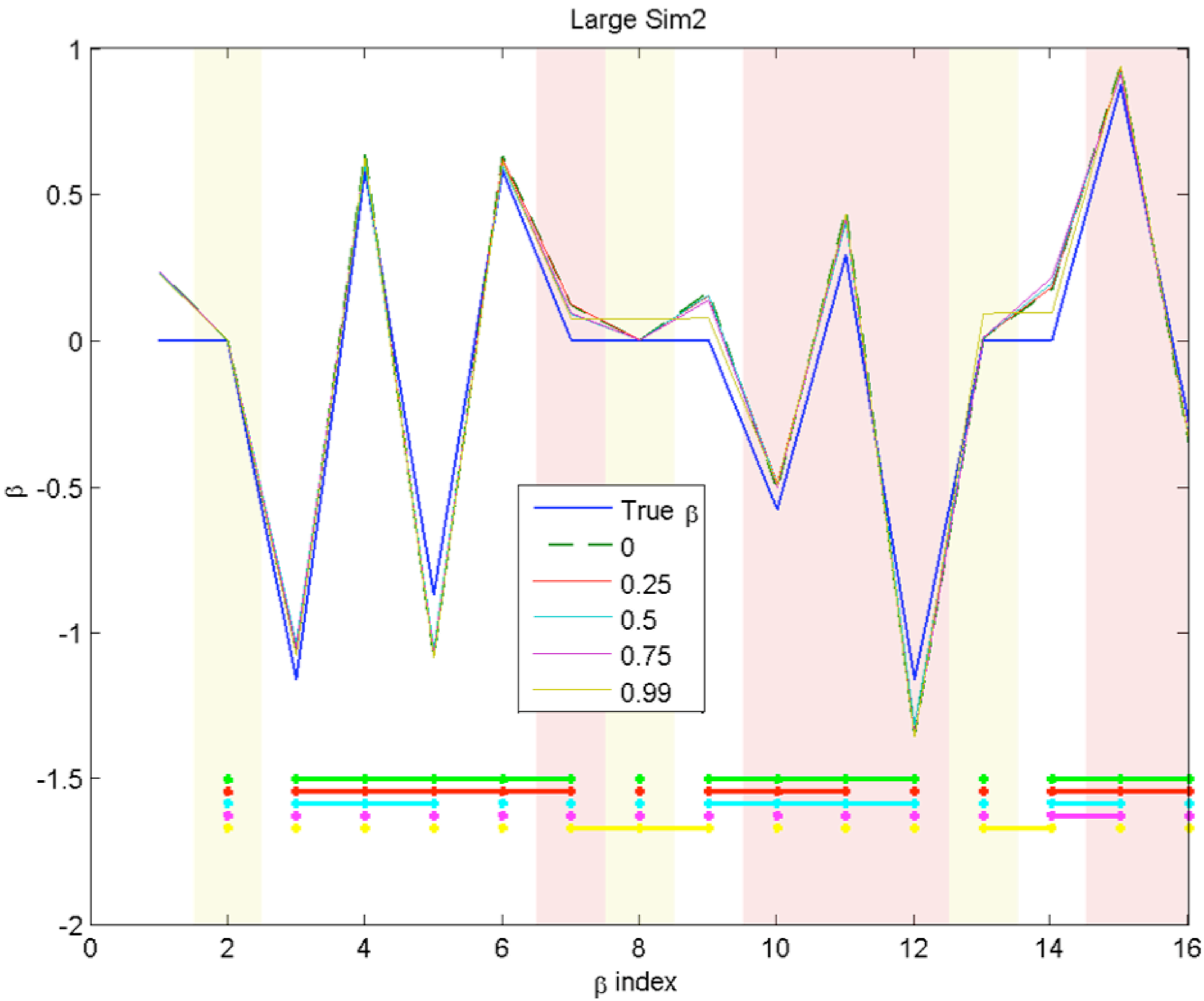
Small (or SM) => small sample = 50 observations
Large (or LG) => large sample = 500 observations

True betas (used to simulate data)
Adjusted so that Bayes error (on a large dataset) ~=0.20

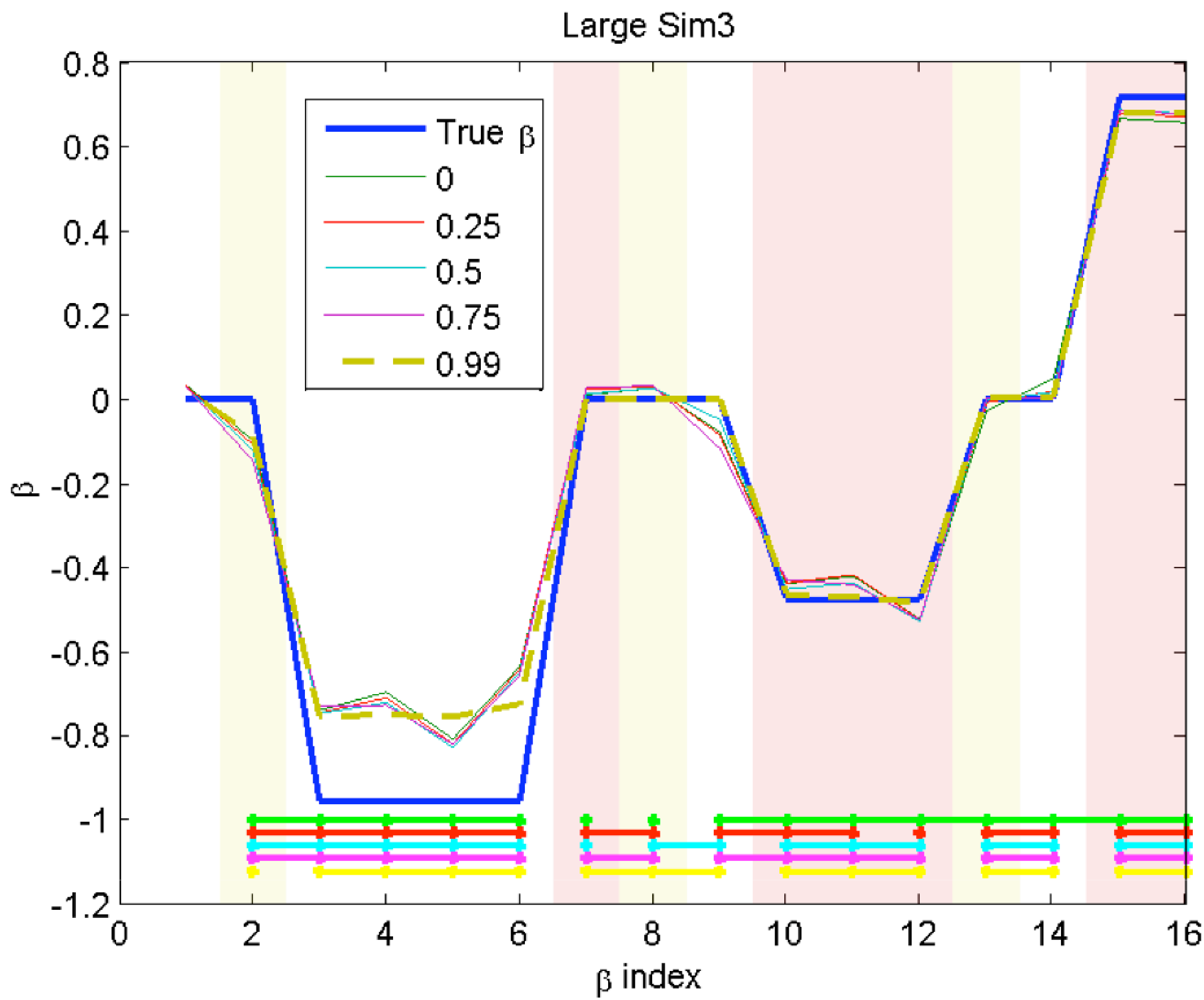
SIM1 (favors BBR)	SIM2 (fv GR. Lasso, kij=0)	SIM3 (fv Fused Gr Lasso, kij->1)
1.1500	0	0
0	-1.1609	-0.9540
0.5750	0.5804	-0.9540
-0.2875	-0.8706	-0.9540
0	0.5804	-0.9540
0	0	0
-0.2875	0	0
0.5750	0	0
0	-0.5804	-0.4770
1.1500	0.2902	-0.4770
0	-1.1609	-0.4770
-1.1500	0	0
0	0	0
0	0.8706	0.7155
-0.8625	-0.2902	0.7155



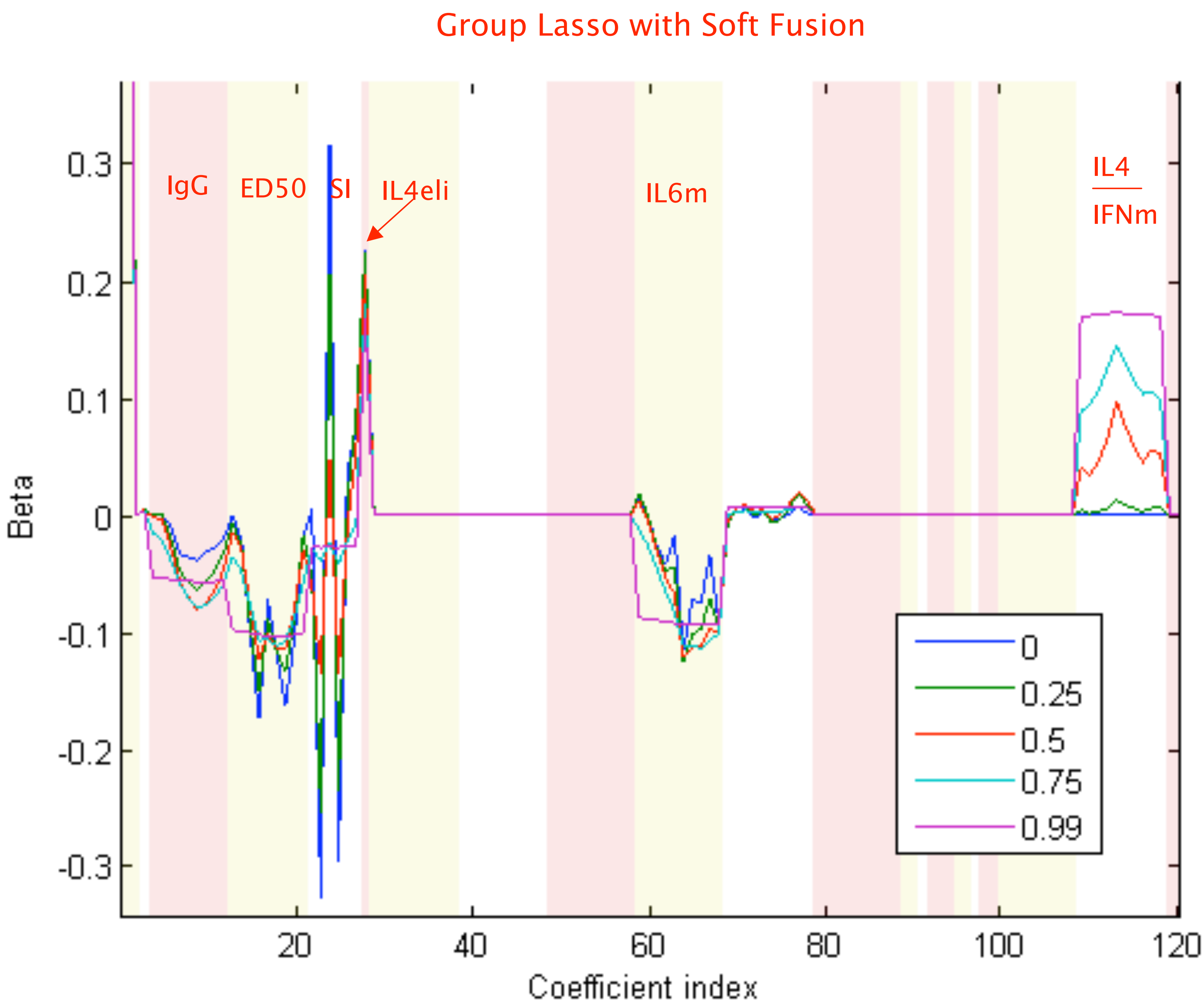
Large Sim1
β
β index
0
0
0.25
0.5
0.75
0.99
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
2
4
6
8
10
12
14
16

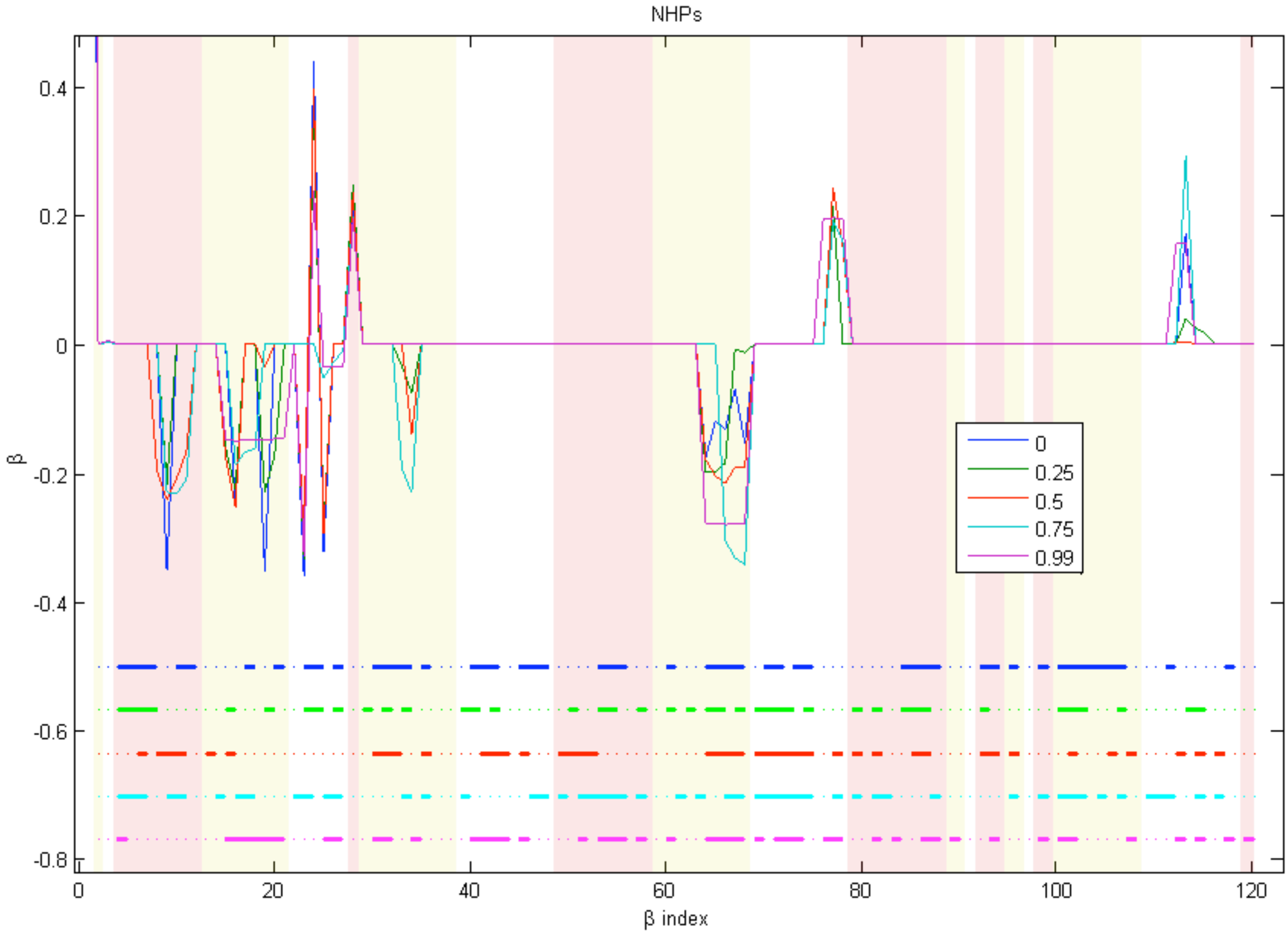


Large Sim2
β
β index
True β
0
0.25
0.5
0.75
0.99
1
0.5
0
-0.5
-1
-1.5
-2
0
2
4
6
8
10
12
14
16

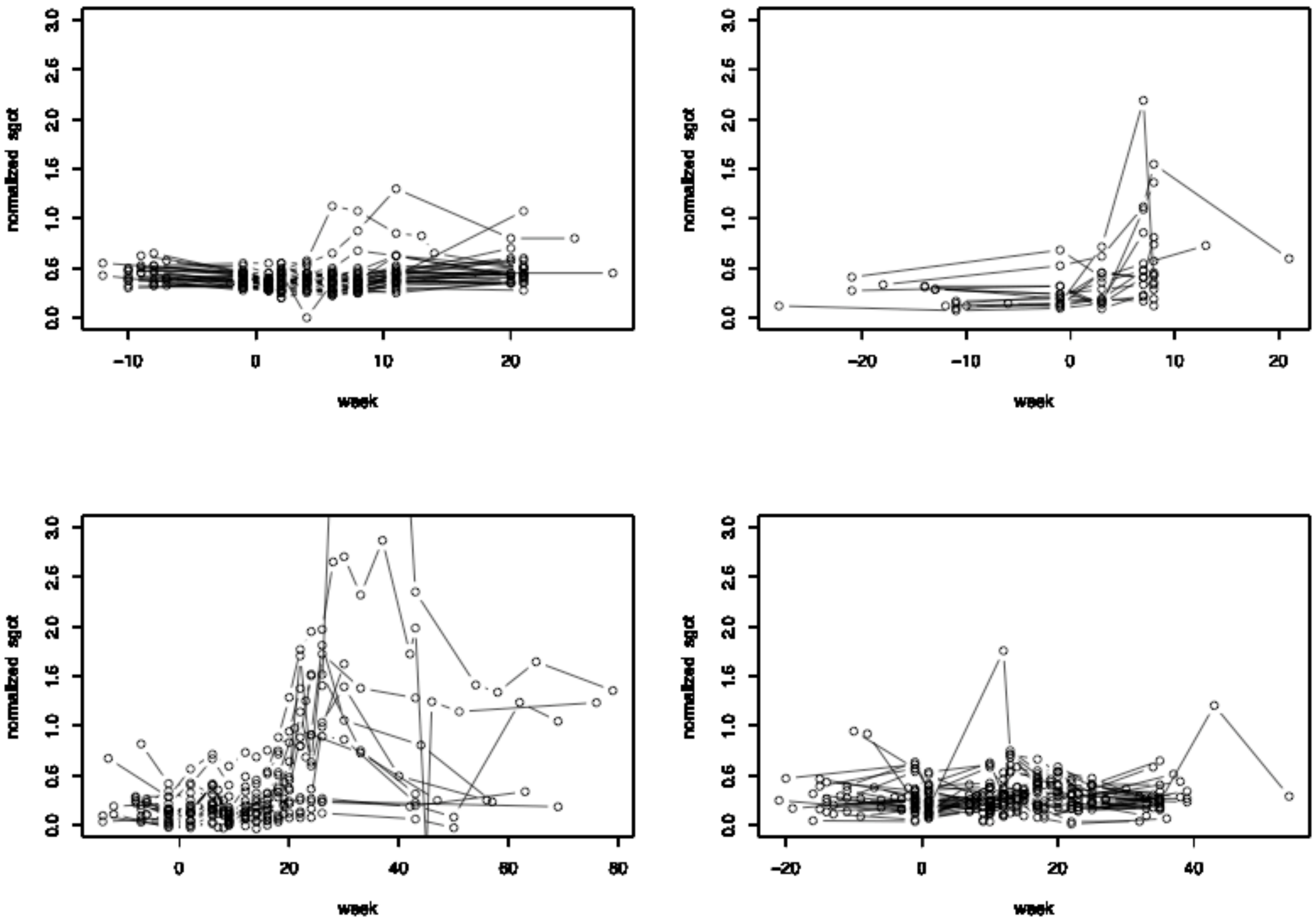


Large Sim3
True β
0
0.25
0.5
0.75
0.99
β
β index
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
0
2
4
6
8
10
12
14
16


Group Lasso with Soft Fusion
IgG
ED50
SI
IL4eli
IL6m
IL4
IFNm
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
Beta
20
40
60
80
100
120
Coefficient index
0
0.25
0.5
0.75
0.99


NHPs
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
β
0
20
40
60
80
100
120
β index
0
0.25
0.5
0.75
0.99

Drug Safety: Early Detection of Toxicity



Future Work

- Rigorous derivation of BIC and df
- Prior on partitions
- Better search strategies for partition space
- Out of sample predictive accuracy
- LAPS C++ implementation

Domain Knowledge in Text Classification

- Certain words are positively or negatively associated with category
- **Domain Knowledge:** textual descriptions for categories
- Prior mean quantifies the strength of positive or negative association
- Prior variance quantifies our confidence in the domain knowledge

An Example Model
(category "grain")

Word	Beta		Word	Beta
corn	29.78		formal	-1.15
wheat	20.56		holder	-1.43
rice	11.33		hungarian	-6.15
sindt	10.56		rubber	-7.12
madagascar	6.83		special	-7.25
import	6.79		…	…
grain	6.77		beet	-13.24
contract	3.08		rockwood	-13.61

Using Domain Knowledge (DK)

- Give domain words higher mean or variance
- **Two methods:** For each DK term *t* and category *Q*, and manually chosen *C*,
  - First method sets **DK-based variance**:

$$\text{variance}(t, Q) = C \cdot \text{significance}(t, Q) \cdot \sigma^2$$

  - Second method sets **DK-based mode**:

$$\text{mode}(t, Q) = C \cdot \text{significance}(t, Q) \cdot \sigma$$

  Here σ^2 is variance for all other words chosen by 5-fold CV on training data

- Used TFxIDF weighting on the prior knoweldge documents to compute *significance*(*t*, *Q*)

Experiments

- Data sets
  1) TREC 2004 Genomics data:
     - **Categories:** 32 MeSH categories under "Cells" hierarchy
     - **Documents:** 3742 training and 4175 test
     - **Prior Knowledge:** MeSH category descriptions

  2) ModApte subset of Reuters-21578
     - **Categories:** 10 most frequent categories
     - **Documents:** 9603 training and 3299 test
     - **Prior Knowledge:** keywords selected by hand (Wu & Srihari, 2004)

- Big (all training examples) and small size training data
- Limited, biased data often the case in applications

MeSH Prior Knowledge Example

- **MeSH Heading:** Neurons
- **Scope Note:** The basic cellular units of nervous tissue. Each neuron consists of a body, an axon, and dendrites. Their purpose is to receive, conduct, and transmit impulses in the nervous system.
- **Entry Term:** Nerve Cells
- **See Also:** Neural Conduction

MeSH Results (Big training data)

	Macro F1	ROC
Laplace	50.2	88.7
Laplace & DK-based variance	53.7	89.2
Laplace & DK-based mode	52.8	89.4

MeSH Results

(training: 500 random examples)

	Macro F1	ROC
Laplace	35.1	78.3
Laplace & DK-based variance	49.7	83.8
Laplace & DK-based mode	44.4	84.2

MeSH Results

(training: 5 positive and 5 random examples for each category)

	Macro F1	ROC
Laplace	29.3	65.9
Laplace & DK-based variance	43.7	77.6
Laplace & DK-based mode	35.8	83.3

Prior Knowledge for ModApte

Category	Prior Knowledge
earn	cents cts net profit quarter qtr revenue rev share shr
acq	acquire acquisition company merger stake
money-fx	bank currency dollar money
grain	agriculture corn crop grain wheat usda
crude	barrel crude oil opec petroleum
trade	deficit import surplus tariff trade
interest	bank money lend rate
wheat	wheat
ship	port ship tanker vessel warship
corn	corn

ModApte Results

(training: 100 random samples)

	Macro F1	ROC
Laplace	37.2	76.2
Laplace & DK-based variance	65.3	87.1
Laplace & DK-based mode	72.0	93.5

ModApte Results

(training: 5 positive + 5 random samples for each category)

	Macro F1	ROC
Laplace	42.7	77.8
Laplace & DK-based variance	63.8	88.1
Laplace & DK-based mode	66.5	94.4