Bayesian Performance Comparison of Text Classifiers

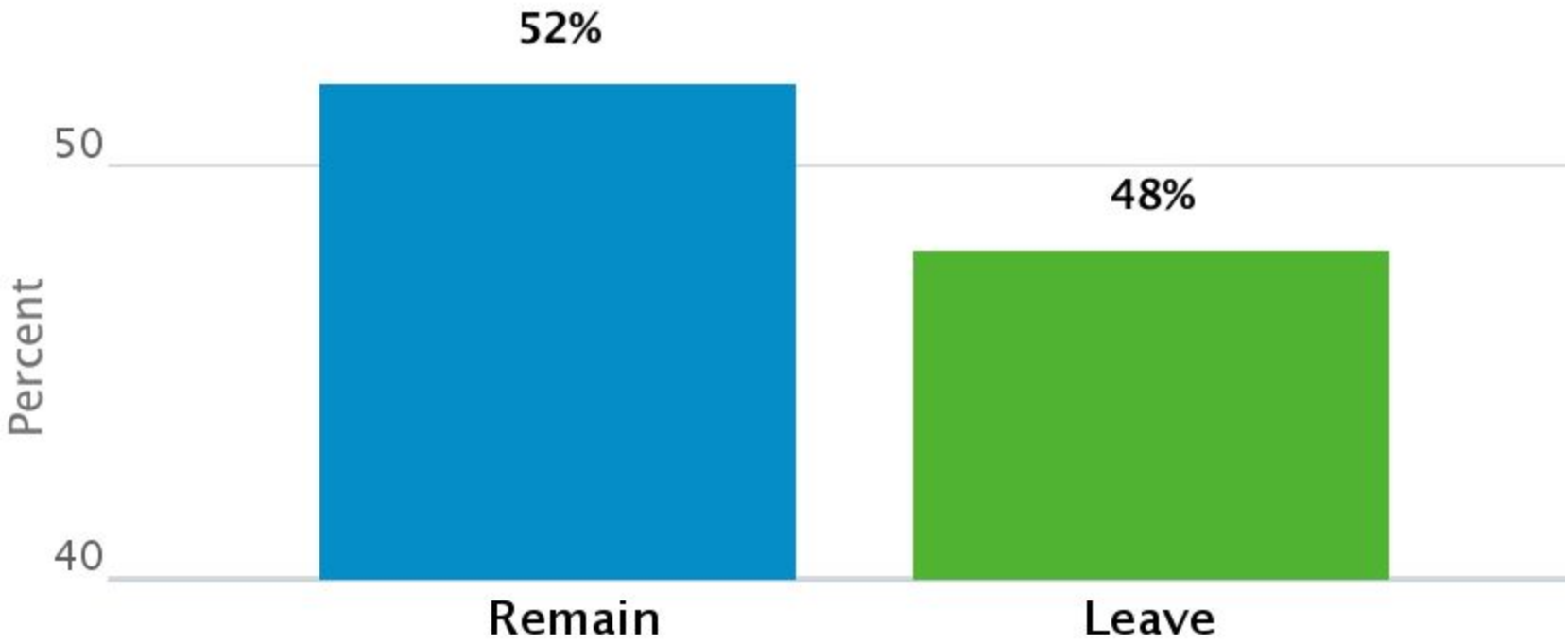
Dell Zhang

(joint work with Jun Wang, Emine Yilmaz, Xiaoling Wang, and Yuxin Zhou)


BREXIT
EU

Referendum Vote Intention Poll of Polls

Latest average of six polls from 16/06/16 to 22/06/16



Source data at www.WhatUKThinks.org/EU run by NatCen Social Research

EU REFERENDUM

Results

UK votes to **LEAVE** the EU

Leave
51.9%
17,410,742 VOTES

Remain
48.1%
16,141,241 VOTES

0 results left to declare

LEAVE UK votes to **LEAVE** the EU

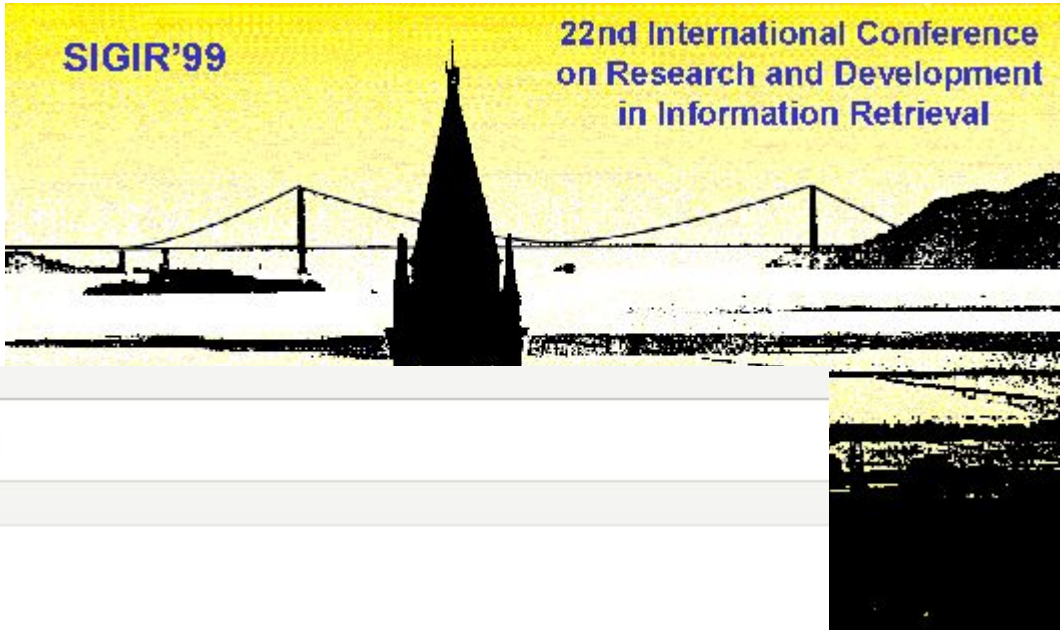
Electorate	46,501,241	
Turnout		72.2%
Rejected ballots		26,033

How results are calculated

Problem

If one classifier A got a higher score than the other classifier B on a test set of documents, how sure can we be that A is really better than B?

William Sealy Gosset
(“Student”)



Google Yang SIGIR 1999

Scholar

A re-examination of text categorization methods

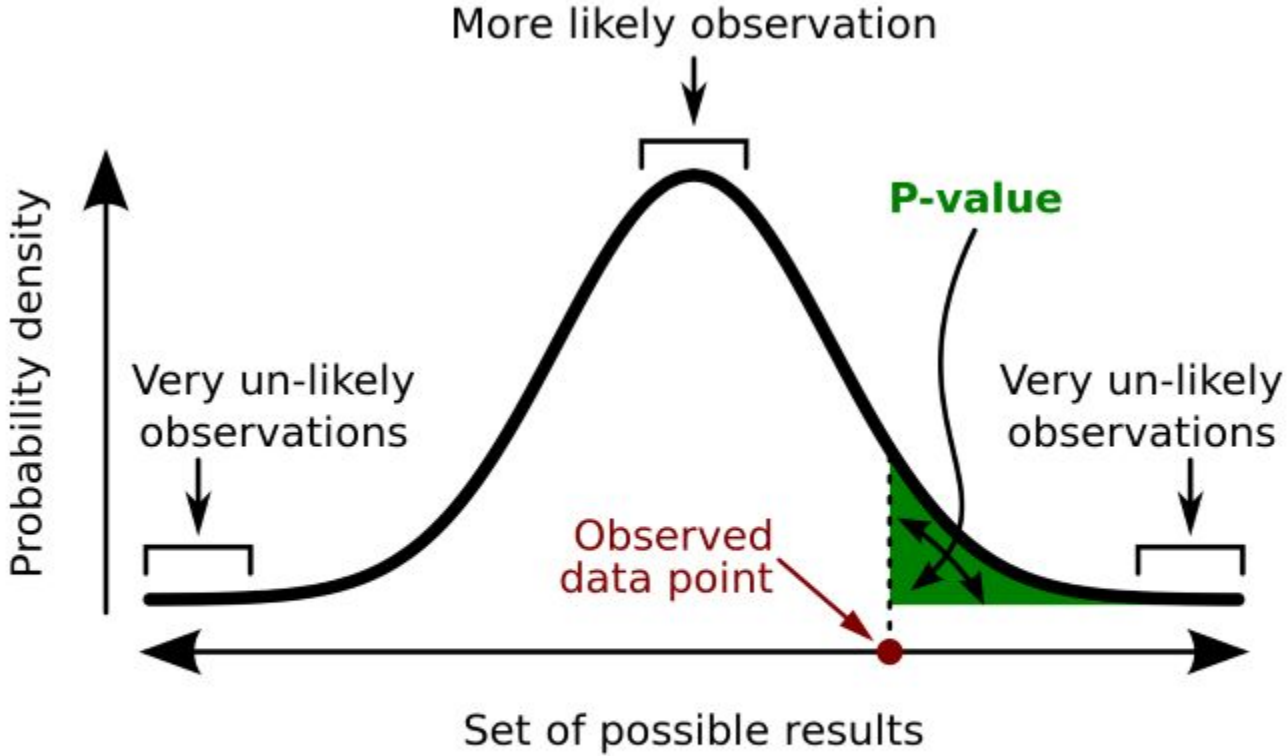
Y **Yang**, X Liu - … of the 22nd annual international ACM **SIGIR** …, **1999** - dl.acm.org

Abstract This paper reports a controlled study with statistical signicance tests on ve text categorization methods: the Support Vector Machines (SVM), a k-Nearest Neighbor (kNN) classi er, a neural network (NNet) approach, the Linear Leastsquares Fit (LLSF) mapping ...

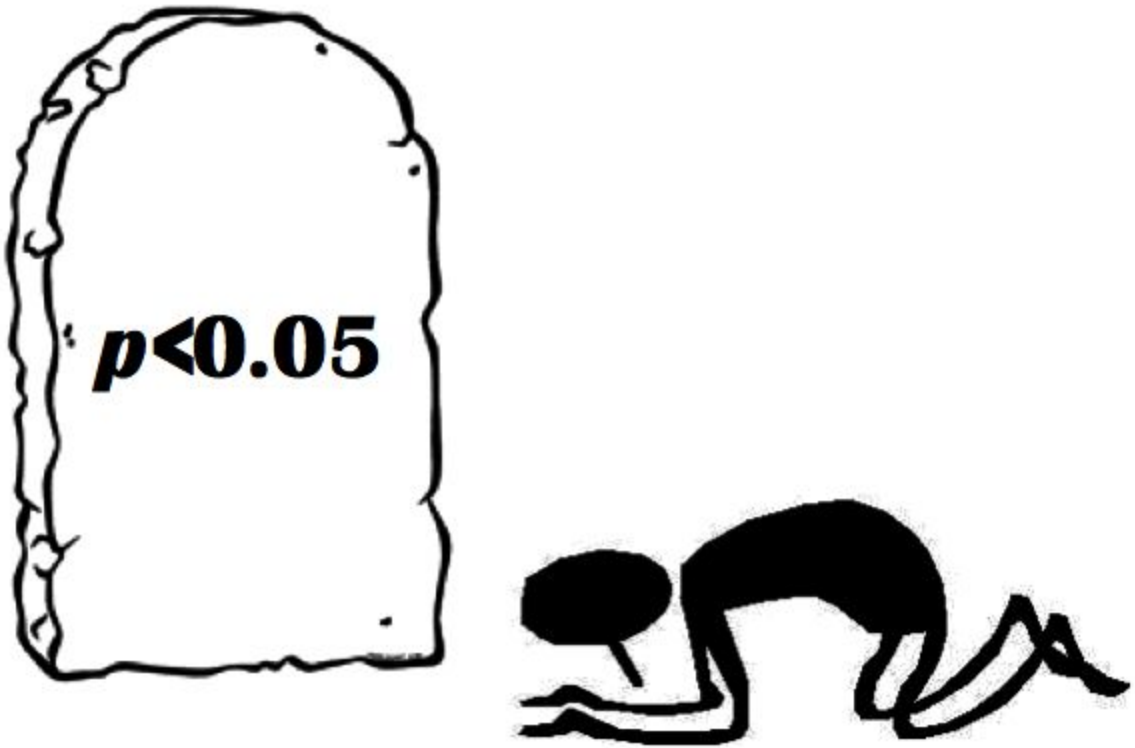
Cited by 3245 Related articles All 26 versions Save More

NHST [Yang1999]

- sign test
  - Micro sign test (s-test)
  - Macro sign test (S-test)
- t test
  - Comparing proportions (p-test)
  - Macro t test (T-test)



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.


$p<0.05$

nature International weekly journal of science

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

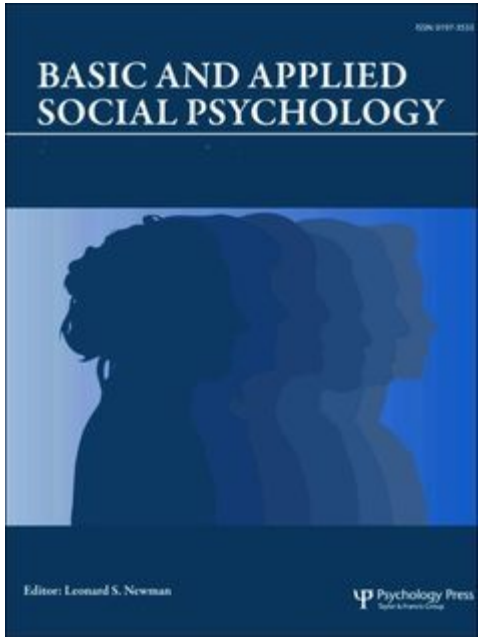
Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015

PDF | Rights & Permissions

A controversial statistical test has finally met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values because the statistics were too often used to support lower-quality research[1].



NHST: Deficiencies

- It cannot declare that two classifiers perform equally well
  - Failing to reject the null hypothesis does not mean that we can accept the null hypothesis.
- It cannot tell whether a non-zero performance difference really matters in practice
- It cannot deal with complex multivariate performance measures (F_1 etc.) on the document level
- …

Important:

Pr (observation | hypothesis) ≠ Pr (hypothesis | observation)

The probability of observing a result given that some hypothesis is true is *not equivalent* to the probability that a hypothesis is true given that some result has been observed.

Using the p-value as a “score” is committing an egregious logical error:
the transposed conditional fallacy.

Thomas Bayes

Journal of Experimental Psychology: General
2013, Vol. 142, No. 2, 573–603

0096-3445/13/$12.00 DOI: 10.1037/a0029146

Bayesian Estimation Supersedes the *t* Test

John K. Kruschke
Indiana University, Bloomington

Bayesian estimation for 2 groups provides complete distributions of credible values for the effect size, group means and their difference, standard deviations and their difference, and the normality of the data. The method handles outliers. The decision rule can accept the null value (unlike traditional *t* tests) when certainty in the estimate is high (unlike Bayesian model comparison using Bayes factors). The method also yields precise estimates of statistical power for various research goals. The software and programs are free and run on Macintosh, Windows, and Linux platforms.

Keywords: Bayesian statistics, effect size, robust estimation, Bayes factor, confidence interval

A probabilistic interpretation of precision, recall and F-score, with implication for evaluation

Authors:

Cyril Goutte, Eric Gaussier

Abstract :

We address the problem of 1/assessing the confidence of the standard point estimates, precision, recall and F-score, and 2/ comparing the results, in terms of precision, recall and F-score, obtained using two different methods. To do so, we use a probabilistic setting which allows us to obtain posterior distributions on these performance indicators, rather than point estimates. This framework is applied to the case where different methods are run on different datasets from the same source, as well as the standard situation where competing results are obtained on the same data.

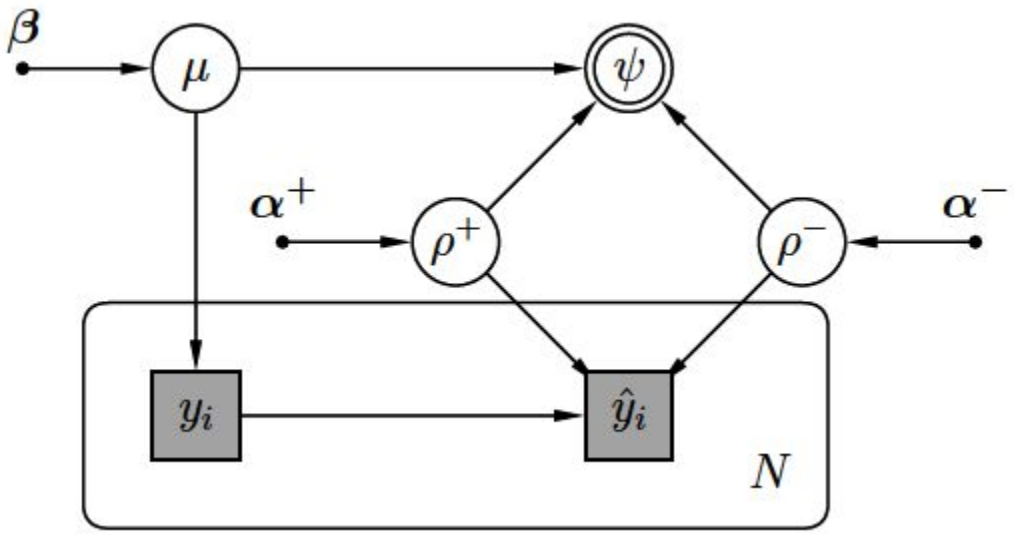
Citation :

ECIR 27th European Conference on Information Retrieval,Santiago de Compostela, Spain

Year :

2005

y_i		$\hat{y}_i$	
$+$	μ	1 0	ρ^+ $1-\rho^+$
$-$	$1-\mu$	1 0	ρ^- $1-\rho^-$



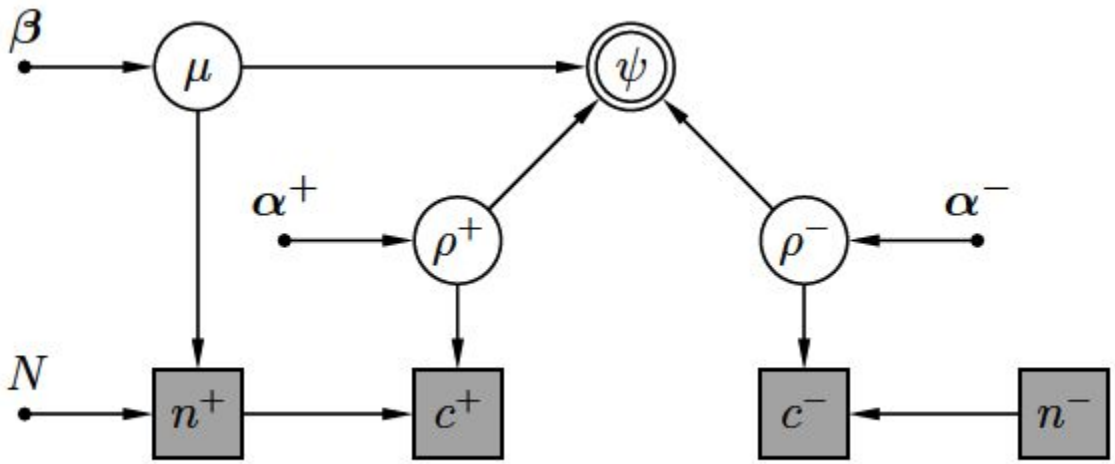
$$\mu \sim \text{Beta}(\boldsymbol{\beta})$$

$$y_i \sim \text{Bern}(\mu) \text{ for } i = 1, \ldots, N$$

$$\rho^+ \sim \text{Beta}(\boldsymbol{\alpha}^+) \qquad \rho^- \sim \text{Beta}(\boldsymbol{\alpha}^-)$$

$$\hat{y}_i \sim \begin{cases} \text{Bern}(\rho^+) \text{ for } i = 1, \ldots, N & \text{if } y_i = + \\ \text{Bern}(\rho^-) \text{ for } i = 1, \ldots, N & \text{if } y_i = - \end{cases}$$

$$\psi = f(\mu, \rho^+, \rho^-)$$

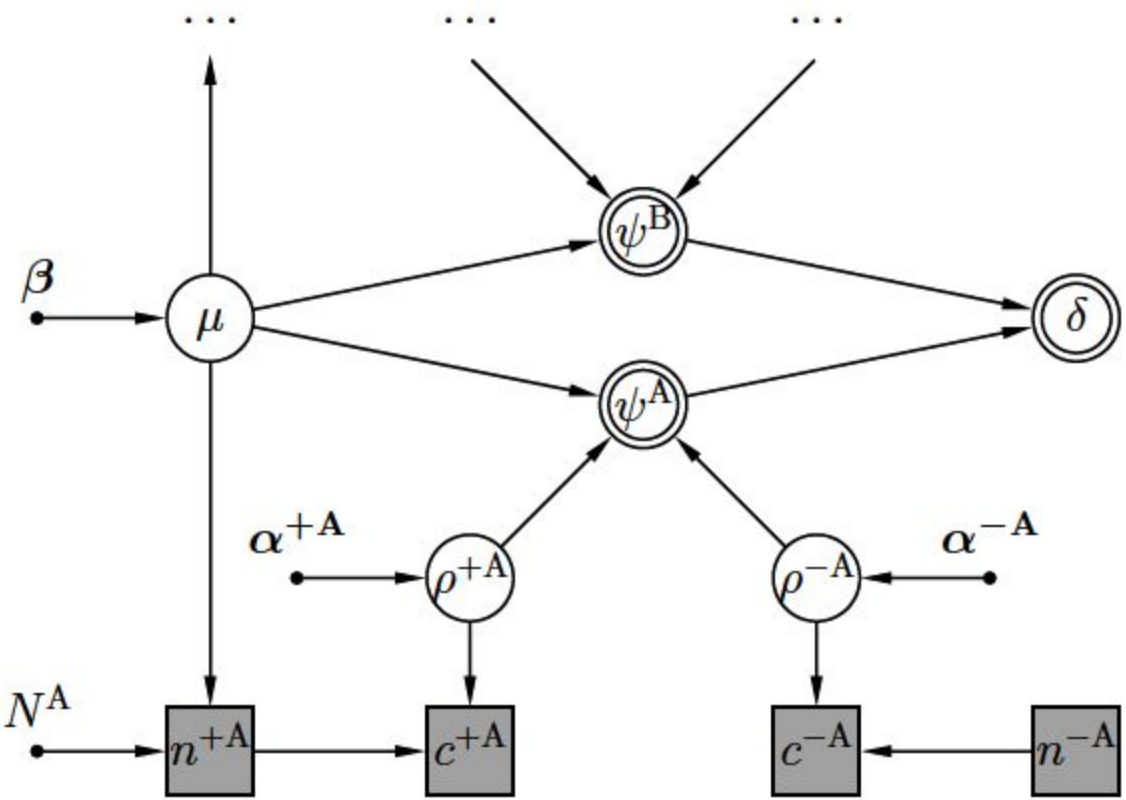


$$
\begin{aligned}
\mu &\sim \text{Beta}(\boldsymbol{\beta}) \\
n^+ &\sim \text{Bin}(N, \mu) \qquad & n^- &= N - n^+ \\
\rho^+ &\sim \text{Beta}(\boldsymbol{\alpha}^+) \qquad & \rho^- &\sim \text{Beta}(\boldsymbol{\alpha}^-) \\
c^+ &\sim \text{Bin}(n^+, \rho^+) \qquad & c^- &\sim \text{Bin}(n^-, \rho^-) \\
\psi &= f(\mu, \rho^+, \rho^-)
\end{aligned}
$$

$$tp = N\mu\rho^{+} \qquad fp = N(1-\mu)\rho^{-}$$
$$fn = N\mu(1-\rho^{+}) \qquad tn = N(1-\mu)(1-\rho^{-})$$

$$P = \frac{tp}{tp+fp} = \frac{\mu\rho^{+}}{\mu\rho^{+}+(1-\mu)\rho^{-}}$$
$$R = \frac{tp}{tp+fn} = \frac{\mu\rho^{+}}{\mu\rho^{+}+\mu(1-\rho^{+})} = \rho^{+}$$
$$F_1 = \frac{2PR}{P+R}$$

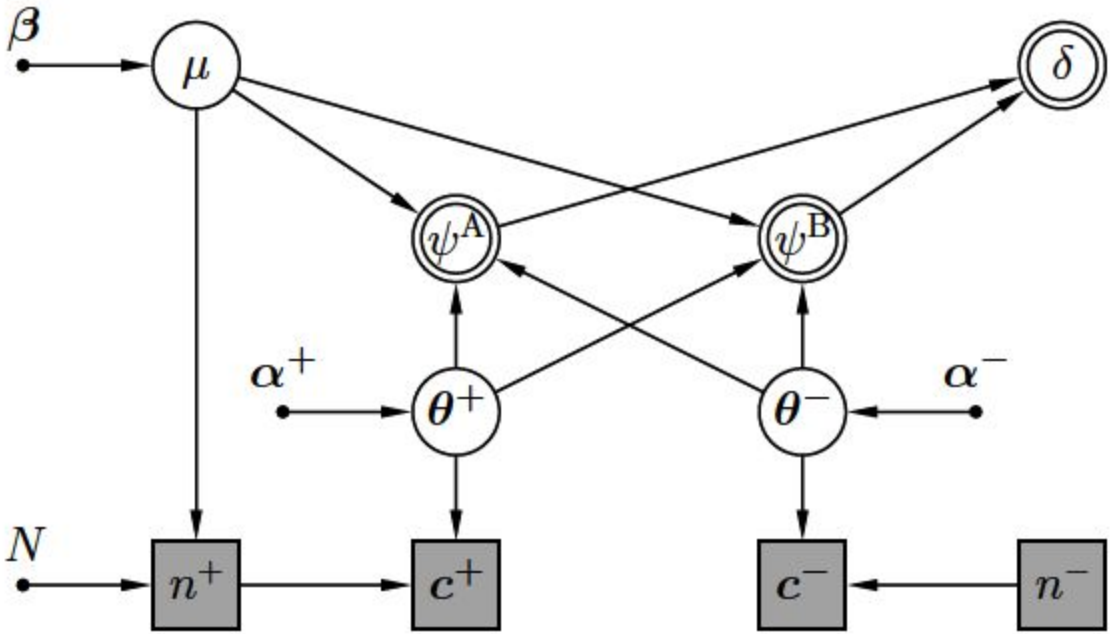
Unpaired Model



$$\delta = \psi^{\mathrm{A}} - \psi^{\mathrm{B}}$$

y_i		$\hat{y}_i^{\mathrm{A}}$	$\hat{y}_i^{\mathrm{B}}$	$\boldsymbol{o}_i$	
$+$	μ	1	1	(1,1)	$\theta^{+}_{(1,1)}$
		1	0	(1,0)	$\theta^{+}_{(1,0)}$
		0	1	(0,1)	$\theta^{+}_{(0,1)}$
		0	0	(0,0)	$\theta^{+}_{(0,0)}$
$-$	$1-\mu$	1	1	(1,1)	$\theta^{-}_{(1,1)}$
		1	0	(1,0)	$\theta^{-}_{(1,0)}$
		0	1	(0,1)	$\theta^{-}_{(0,1)}$
		0	0	(0,0)	$\theta^{-}_{(0,0)}$

Paired Model



$$\delta = \psi^{\mathrm{A}} - \psi^{\mathrm{B}}$$

Paired Model

$$
\begin{aligned}
\mu &\sim \mathrm{Beta}(\boldsymbol{\beta}) \\
n^+ &\sim \mathrm{Bin}(N, \mu) & n^- &= N - n^+ \\
\boldsymbol{\theta}^+ &\sim \mathrm{Dir}(\boldsymbol{\alpha}^+) & \boldsymbol{\theta}^- &\sim \mathrm{Dir}(\boldsymbol{\alpha}^-) \\
\boldsymbol{c}^+ &\sim \mathrm{Mult}(n^+, \boldsymbol{\theta}^+) & \boldsymbol{c}^- &\sim \mathrm{Mult}(n^-, \boldsymbol{\theta}^-) \\
\psi^{\mathrm{A}} &= f(\mu, \boldsymbol{\theta}^+, \boldsymbol{\theta}^-) & \psi^{\mathrm{B}} &= f'(\mu, \boldsymbol{\theta}^+, \boldsymbol{\theta}^-) \\
\delta &= \psi^{\mathrm{A}} - \psi^{\mathrm{B}}
\end{aligned}
$$

$$\rho^{+\mathrm{A}} = \theta^{+}_{(1,1)} + \theta^{+}_{(1,0)} \qquad \rho^{-\mathrm{A}} = \theta^{-}_{(1,1)} + \theta^{-}_{(1,0)}$$

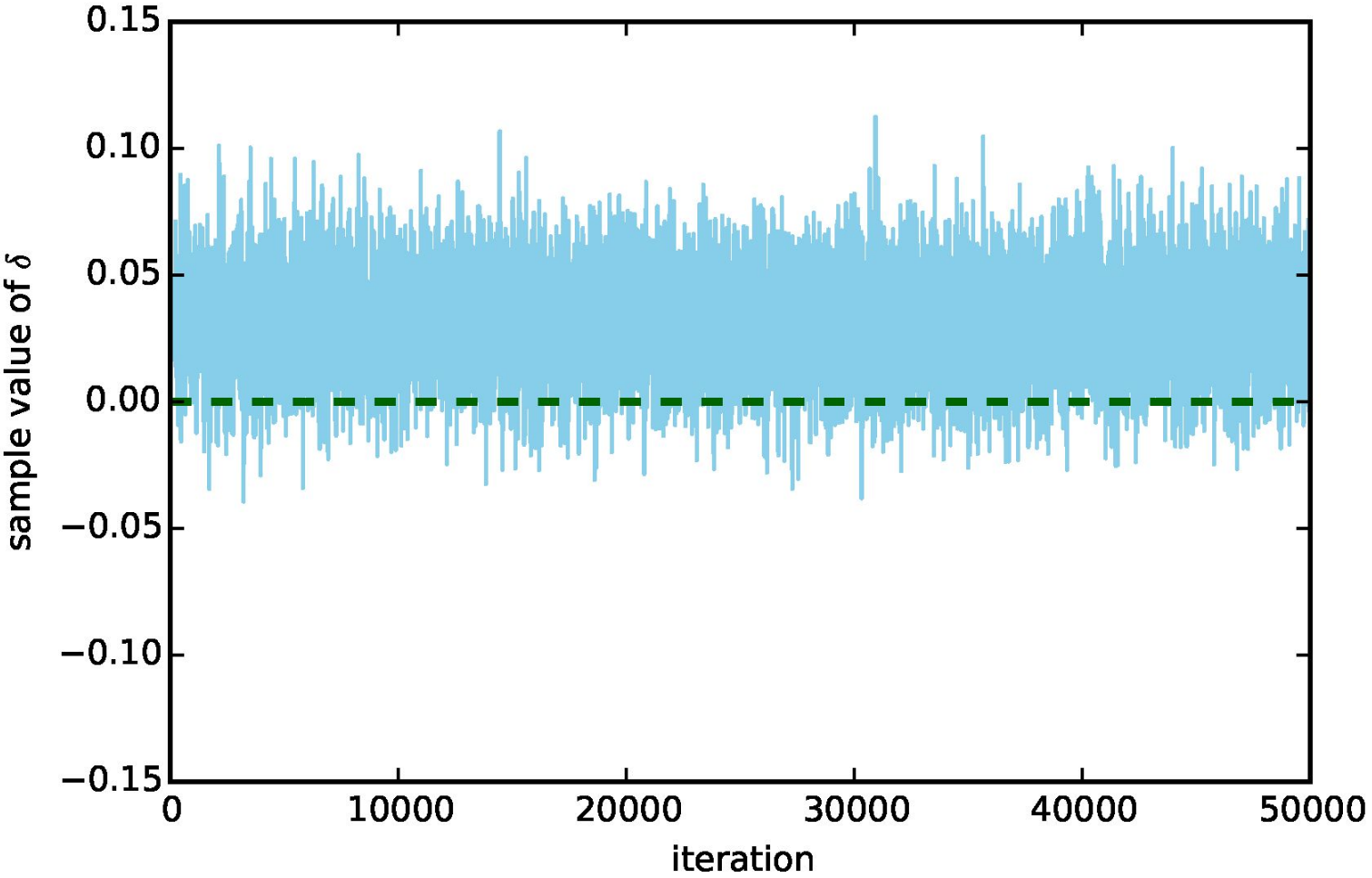
$$\rho^{+\mathrm{B}} = \theta^{+}_{(1,1)} + \theta^{+}_{(0,1)} \qquad \rho^{-\mathrm{B}} = \theta^{-}_{(1,1)} + \theta^{-}_{(0,1)}$$

$$tp = N\mu\rho^{+} \qquad fp = N(1-\mu)\rho^{-}$$
$$fn = N\mu(1-\rho^{+}) \qquad tn = N(1-\mu)(1-\rho^{-})$$

$$P = \frac{tp}{tp+fp} = \frac{\mu\rho^{+}}{\mu\rho^{+}+(1-\mu)\rho^{-}}$$
$$R = \frac{tp}{tp+fn} = \frac{\mu\rho^{+}}{\mu\rho^{+}+\mu(1-\rho^{+})} = \rho^{+}$$
$$F_1 = \frac{2PR}{P+R}$$

Markov Chain Monte Carlo (MCMC)

Metropolis-Hastings sampling


0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
sample value of δ
0
10000
20000
30000
40000
50000
iteration

Bayes Factor

$$\mathrm{BF} = \Pr[\mathcal{D}|\mathcal{M}_0] / \Pr[\mathcal{D}|\mathcal{M}_1]$$

Savage-Dickey Method

$$\mathrm{BF} = \Pr[\delta = 0|\mathcal{M}_1, \mathcal{D}] / \Pr[\delta = 0|\mathcal{M}_1]$$

(with Kernel Density Estimation)

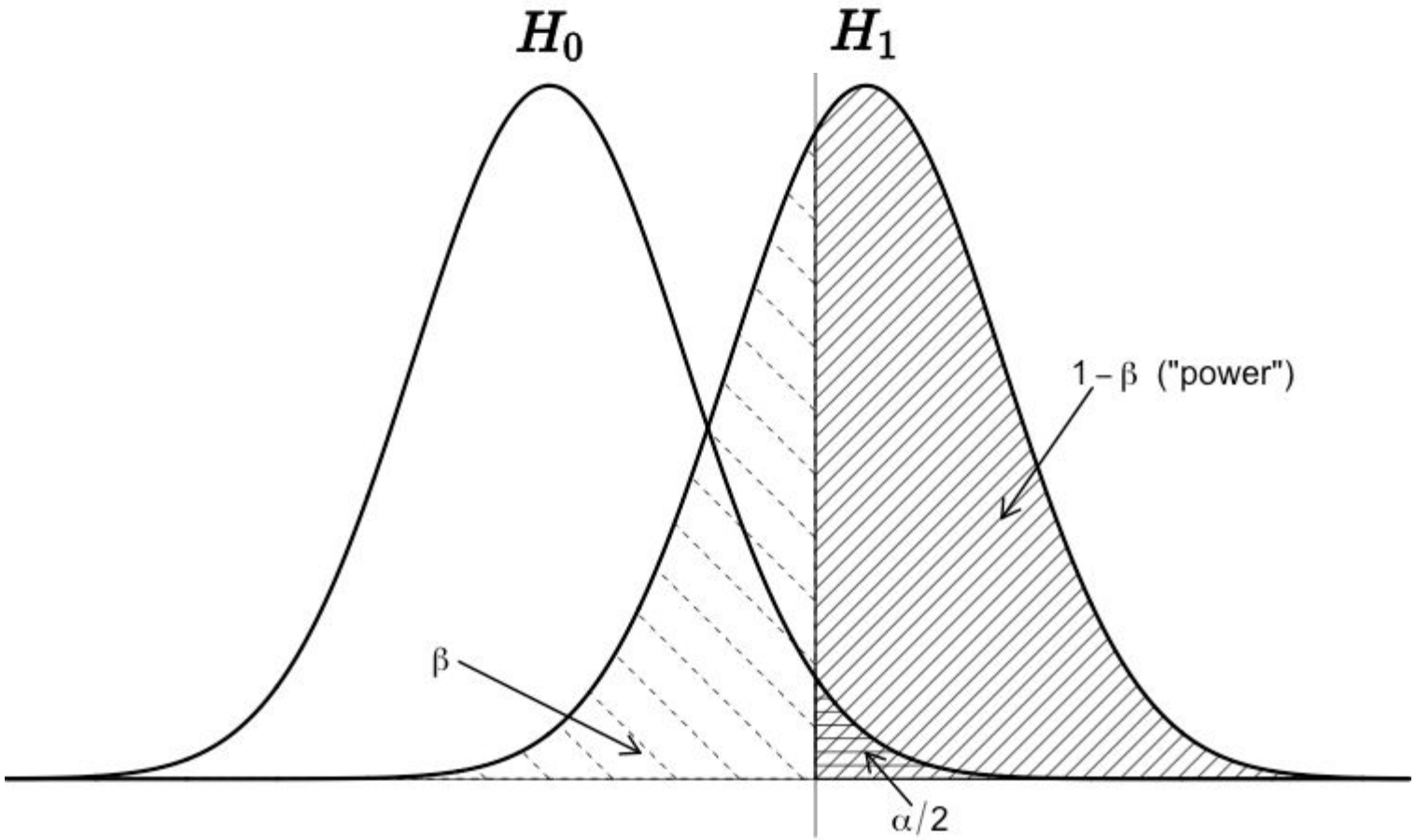
- BF > 3 : substantial evidence for M_0
  - The two classifiers perform equally well
- BF < 1/3: substantial evidence for M_1
  - One classier works better than the other

Bayesian Estimation

Comparing the 95% Highest Density Interval (HDI) of δ and the user-defined Region of Practical Equivalence (ROPE) of δ

- The HDI sits fully within the ROPE:
  - **A ≈ B**
- The HDI sits fully at the left/right side of the ROPE:
  - **A ≪ B** or **A ≫ B**
- The HDI sits mainly though not fully at the left/right side of the ROPE:
  - **A < B** or **A > B**

$$\text{power} = \mathbb{P}\big(\text{reject } H_0 \,\big|\, H_1 \text{ is true}\big)$$

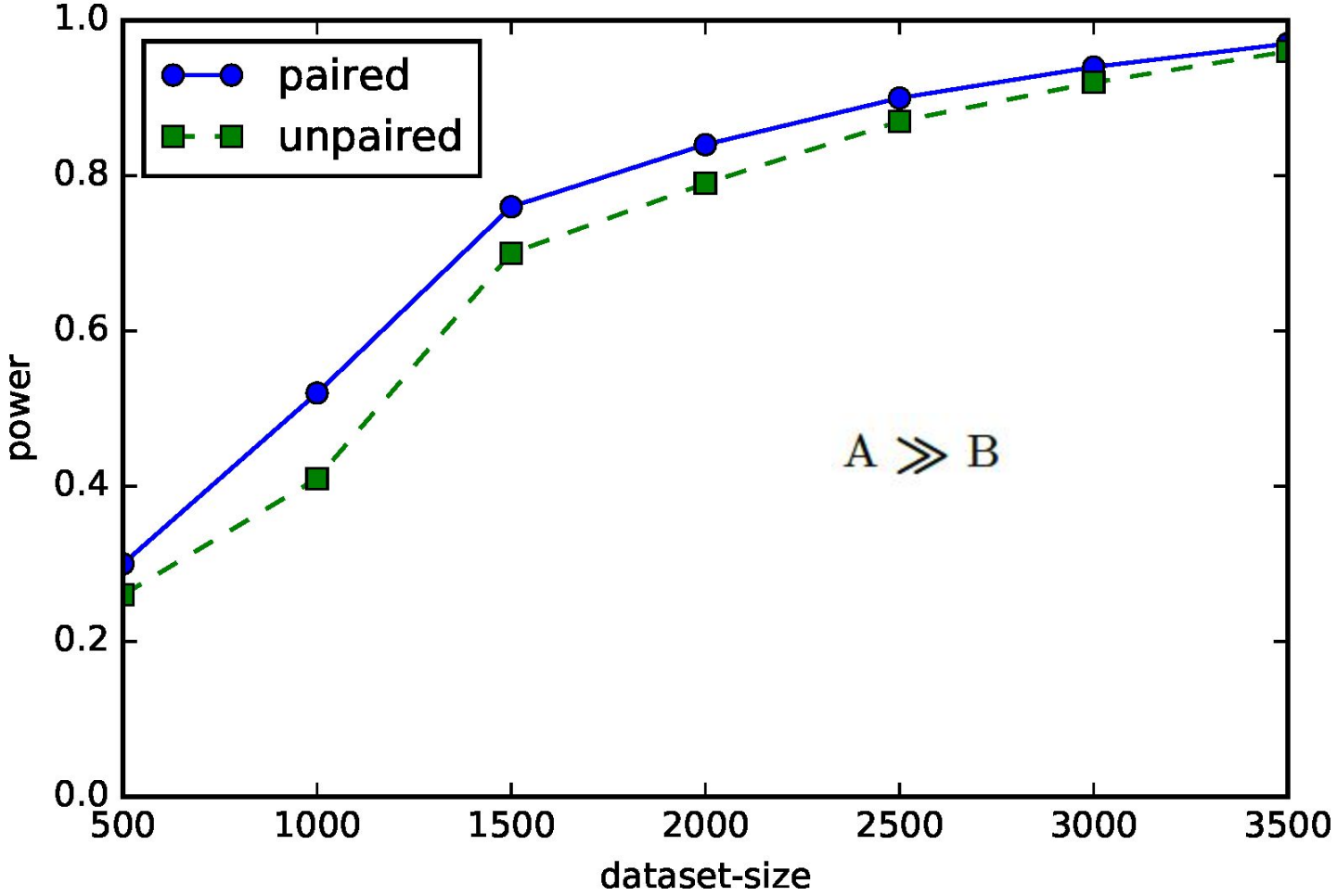


The scenario (a):

$$\Pr[+] = \mu \qquad = 0.5$$
$$\Pr[(1,1)|+] = \theta^{+}_{(1,1)} = 0.3,\ \Pr[(1,0)|+] = \theta^{+}_{(1,0)} = 0.3,$$
$$\Pr[(0,1)|+] = \theta^{+}_{(0,1)} = 0.2,\ \Pr[(0,0)|+] = \theta^{+}_{(0,0)} = 0.2,$$
$$\Pr[-] = 1 - \mu = 0.5$$
$$\Pr[(1,1)|-] = \theta^{-}_{(1,1)} = 0.2,\ \Pr[(1,0)|-] = \theta^{-}_{(1,0)} = 0.2,$$
$$\Pr[(0,1)|-] = \theta^{-}_{(0,1)} = 0.3,\ \Pr[(0,0)|-] = \theta^{-}_{(0,0)} = 0.3,$$

$$F_1^{\mathrm{A}} = 0.6 \qquad F_1^{\mathrm{B}} = 0.5$$

$$\mathrm{A} \gg \mathrm{B}$$


paired
unpaired
power
dataset-size
A ≫ B
0.0
0.2
0.4
0.6
0.8
1.0
500
1000
1500
2000
2500
3000
3500

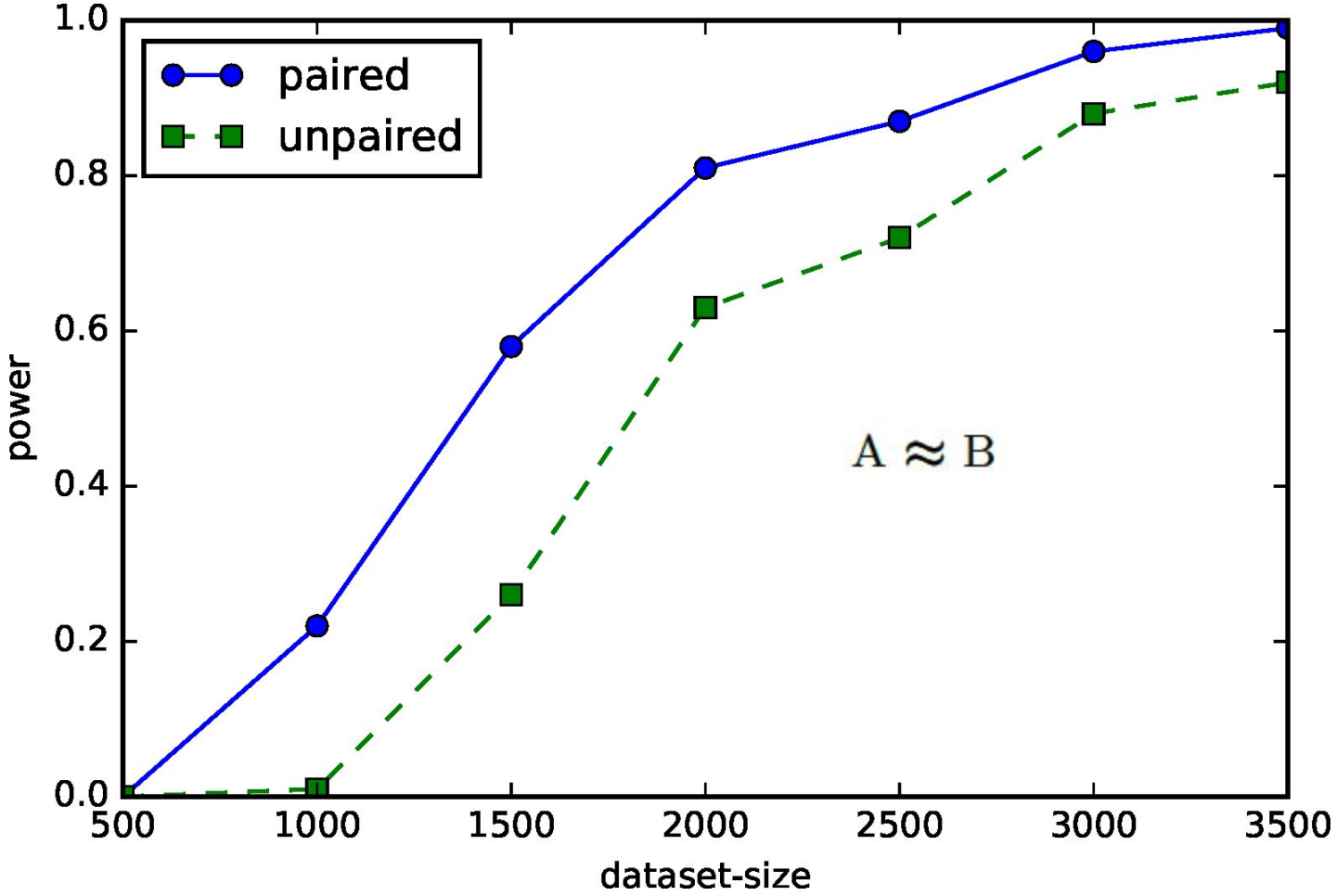
The scenario (b):

$$\Pr[+] = \mu \qquad = 0.5$$
$$\Pr[(1,1)|+] = \theta^{+}_{(1,1)} = 0.3,\ \Pr[(1,0)|+] = \theta^{+}_{(1,0)} = 0.2,$$
$$\Pr[(0,1)|+] = \theta^{+}_{(0,1)} = 0.2,\ \Pr[(0,0)|+] = \theta^{+}_{(0,0)} = 0.3,$$
$$\Pr[-] = 1 - \mu = 0.5$$
$$\Pr[(1,1)|-] = \theta^{-}_{(1,1)} = 0.3,\ \Pr[(1,0)|-] = \theta^{-}_{(1,0)} = 0.2,$$
$$\Pr[(0,1)|-] = \theta^{-}_{(0,1)} = 0.2,\ \Pr[(0,0)|-] = \theta^{-}_{(0,0)} = 0.3,$$

$$F_1^{\mathrm{A}} = 0.5 \qquad F_1^{\mathrm{B}} = 0.5$$

$$\mathrm{A} \approx \mathrm{B}$$

[NB] This goal is infeasible using the frequentist NHST.


paired
unpaired
power
dataset-size
A ≈ B
0.0
0.2
0.4
0.6
0.8
1.0
500
1000
1500
2000
2500
3000
3500

20newsgroups (hard version)

category	name	#train	#test	NB_{Bern}	NB_{Mult}	SVM_{L1}	SVM_{L2}
0	alt.atheism	480	319	0.398	0.480	0.453	0.480
1	comp.graphics	584	389	0.551	0.666	0.626	0.638
2	comp.os.ms-windows.misc	591	394	0.170	0.577	0.599	0.605
3	comp.sys.ibm.pc.hardware	590	392	0.545	0.641	0.591	0.611
4	comp.sys.mac.hardware	578	385	0.431	0.682	0.654	0.676
5	comp.windows.x	593	395	0.666	0.768	0.709	0.717
6	misc.forsale	585	390	0.645	0.781	0.716	0.768
7	rec.autos	594	396	0.634	0.725	0.561	0.695
8	rec.motorcycles	598	398	0.596	0.741	0.720	0.735
9	rec.sport.baseball	597	397	0.770	0.850	0.760	0.634
10	rec.sport.hockey	600	399	0.840	0.731	0.834	0.846
11	sci.crypt	595	396	0.632	0.725	0.737	0.742
12	sci.electronics	591	393	0.533	0.631	0.523	0.557
13	sci.med	594	396	0.681	0.796	0.716	0.736
14	sci.space	593	394	0.647	0.756	0.694	0.706
15	soc.religion.christian	599	398	0.655	0.682	0.656	0.691
16	talk.politics.guns	546	364	0.535	0.641	0.569	0.600
17	talk.politics.mideast	564	376	0.694	0.797	0.759	0.743
18	talk.politics.misc	465	310	0.398	0.487	0.457	0.472
19	talk.religion.misc	377	251	0.207	0.248	0.319	0.311

NB_{Bern} and NB_{Mult}

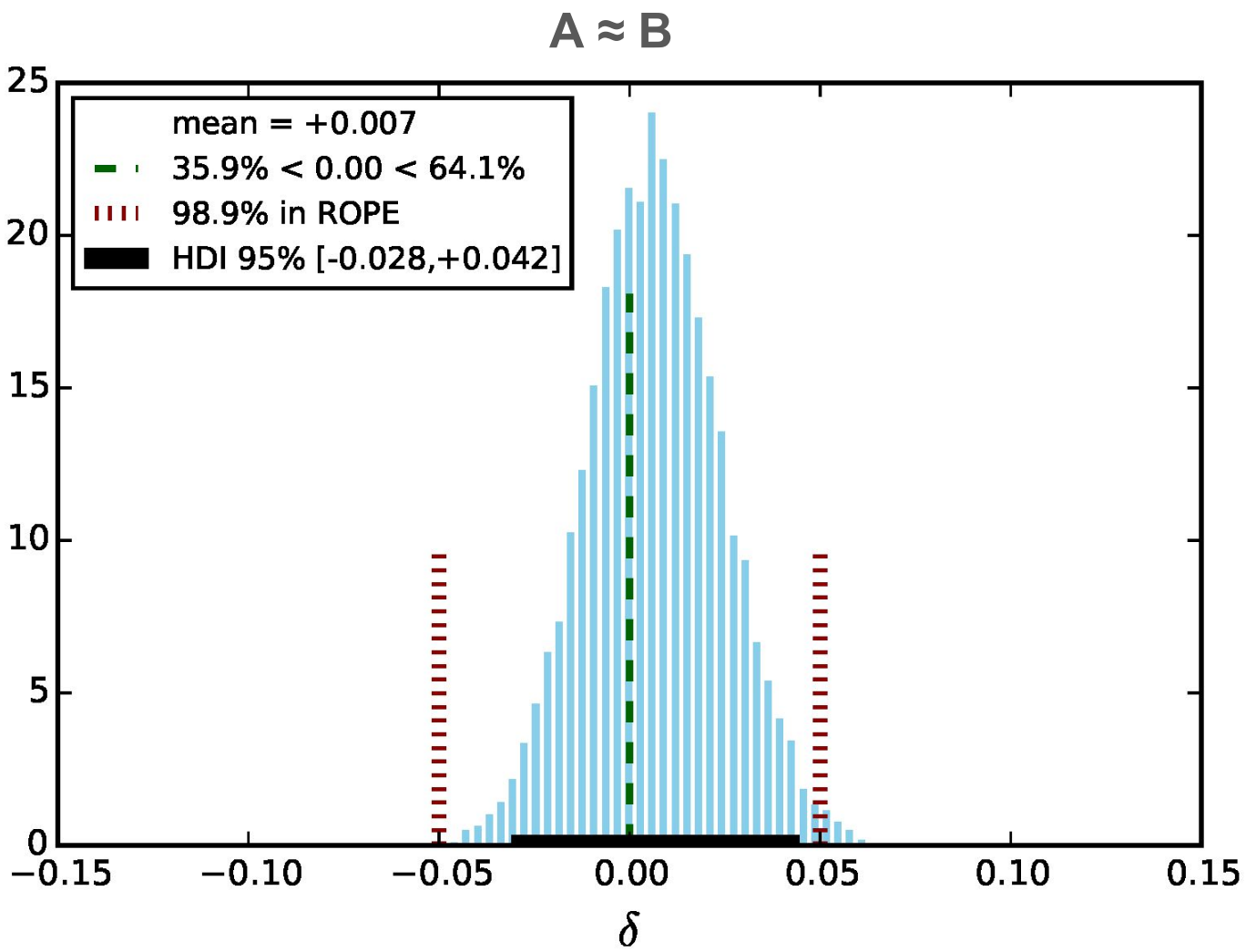
category	frequentist		Bayesian						
	sign-test	t-test	mean	std	BF_{SD}	LG pct	ROPE pct	HDI	**decision**
0	$\star$ 0.000	$\star$ 0.008	−0.081	0.021	$\star$ 0.003	100.0%<0<0.0%	6.6%	[−0.125, −0.041]	$<$
1	$\star$ 0.000	$\star$ 0.000	−0.114	0.017	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.148, −0.080]	$\ll$
2	$\star$ 0.000	$\star$ 0.000	−0.400	0.028	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.456, −0.345]	$\ll$
3	$\star$ 0.000	$\star$ 0.003	−0.095	0.016	$\star$ 0.000	100.0%<0<0.0%	0.2%	[−0.126, −0.062]	$\ll$
4	$\star$ 0.000	$\star$ 0.000	−0.249	0.019	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.286, −0.211]	$\ll$
5	$\star$ 0.000	$\star$ 0.000	−0.101	0.017	$\star$ 0.000	100.0%<0<0.0%	0.1%	[−0.135, −0.069]	$\ll$
6	$\star$ 0.000	$\star$ 0.000	−0.136	0.016	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.168, −0.105]	$\ll$
7	$\star$ 0.000	$\star$ 0.001	−0.092	0.016	$\star$ 0.000	100.0%<0<0.0%	0.4%	[−0.123, −0.061]	$\ll$
8	$\star$ 0.000	$\star$ 0.000	−0.144	0.017	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.178, −0.111]	$\ll$
9	$\star$ 0.000	$\star$ 0.001	−0.080	0.013	$\star$ 0.000	100.0%<0<0.0%	0.8%	[−0.105, −0.055]	$\ll$
10	$\star$ 0.000	$\star$ 0.000	+0.108	0.017	$\star$ 0.000	0.0%<0<100.0%	0.0%	[+0.074, +0.142]	$\gg$
11	$\star$ 0.000	$\star$ 0.002	−0.092	0.016	$\star$ 0.000	100.0%<0<0.0%	0.4%	[−0.125, −0.061]	$\ll$
12	$\star$ 0.000	$\star$ 0.002	−0.097	0.020	$\star$ 0.000	100.0%<0<0.0%	0.8%	[−0.137, −0.058]	$\ll$
13	$\star$ 0.000	$\star$ 0.000	−0.115	0.016	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.147, −0.084]	$\ll$
14	$\star$ 0.000	$\star$ 0.000	−0.109	0.016	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.139, −0.079]	$\ll$
15	0.064	0.178	−0.027	0.016	♯ 3.132	95.7%<0<4.3%	92.2%	[−0.059, +0.004]	$<$
16	$\star$ 0.024	0.151	−0.105	0.019	$\star$ 0.000	100.0%<0<0.0%	0.1%	[−0.141, −0.068]	$\ll$
17	$\star$ 0.000	$\star$ 0.000	−0.102	0.016	$\star$ 0.000	100.0%<0<0.0%	0.1%	[−0.134, −0.070]	$\ll$
18	$\star$ 0.000	$\star$ 0.034	−0.088	0.020	$\star$ 0.007	100.0%<0<0.0%	2.9%	[−0.128, −0.049]	$<$
19	$\star$ 0.000	$\star$ 0.011	−0.040	0.030	2.389	91.3%<0<8.7%	62.9%	[−0.097, +0.022]	$<$

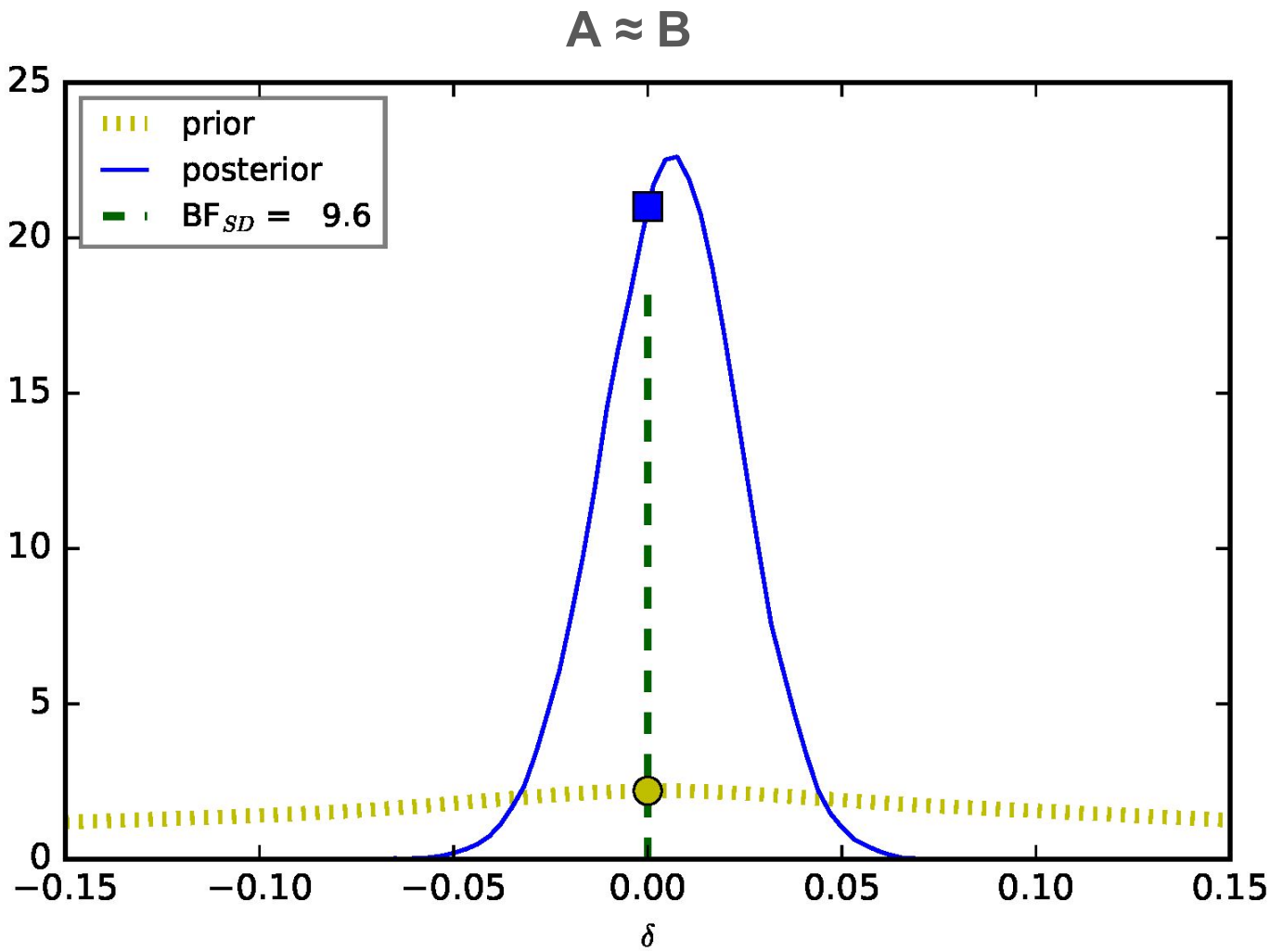
SVM_{L1} and SVM_{L2}

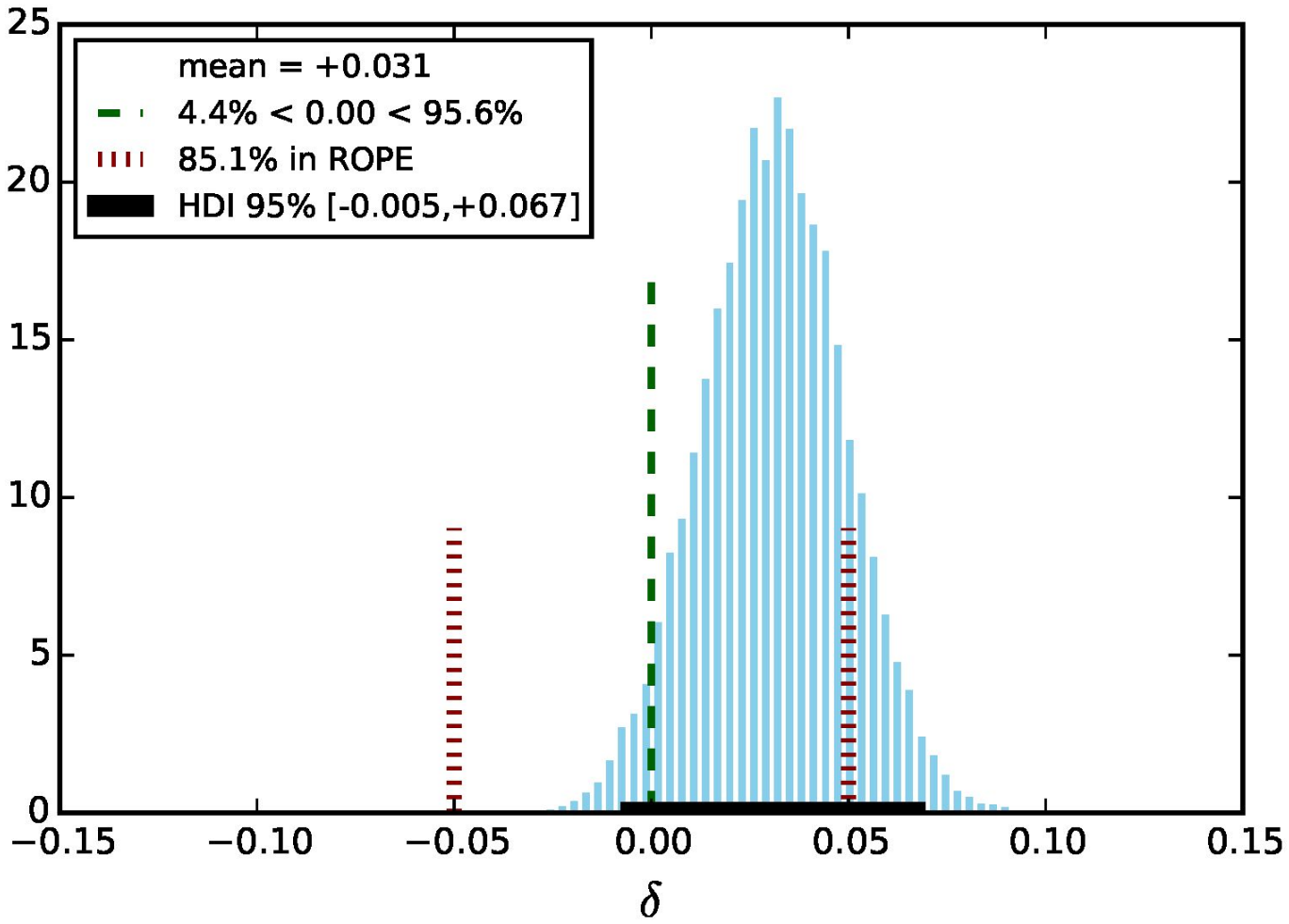
category	frequentist		Bayesian						
	sign-test	t-test	mean	std	BF_{SD}	LG pct	ROPE pct	HDI	**decision**
0	$\star$ 0.049	0.299	−0.027	0.018	♯ 3.399	93.6%<0<6.4%	89.9%	[−0.063, +0.008]	$<$
1	0.523	0.706	−0.011	0.013	♯ 9.427	80.6%<0<19.4%	99.9%	[−0.038, +0.014]	$\approx$
2	0.632	0.774	−0.007	0.014	♯ 12.058	68.3%<0<31.7%	99.8%	[−0.035, +0.020]	$\approx$
3	0.270	0.542	−0.020	0.014	♯ 4.593	92.3%<0<7.7%	98.5%	[−0.047, +0.007]	$\approx$
4	0.247	0.524	−0.022	0.014	♯ 3.781	94.8%<0<5.2%	97.4%	[−0.049, +0.005]	$\approx$
5	1.000	0.960	−0.009	0.014	♯ 12.206	73.2%<0<26.8%	99.8%	[−0.035, +0.019]	$\approx$
6	$\star$ 0.000	$\star$ 0.031	−0.053	0.013	$\star$ 0.011	100.0%<0<0.0%	41.9%	[−0.078, −0.029]	$<$
7	$\star$ 0.000	$\star$ 0.000	−0.133	0.017	$\star$ 0.000	100.0%<0<0.0%	0.0%	[−0.166, −0.098]	$\ll$
8	0.221	0.457	−0.015	0.014	♯ 8.165	84.4%<0<15.6%	99.3%	[−0.042, +0.014]	$\approx$
9	$\star$ 0.000	$\star$ 0.000	+0.126	0.017	$\star$ 0.000	0.0%<0<100.0%	0.0%	[+0.094, +0.160]	$\gg$
10	0.515	0.657	−0.013	0.011	♯ 9.261	87.7%<0<12.3%	99.9%	[−0.035, +0.009]	$\approx$
11	0.771	0.837	−0.005	0.013	♯ 12.937	64.0%<0<36.0%	100.0%	[−0.030, +0.021]	$\approx$
12	0.061	0.298	−0.033	0.016	1.209	98.2%<0<1.8%	84.7%	[−0.066, −0.003]	$<$
13	0.264	0.463	−0.019	0.015	♯ 5.513	90.2%<0<9.8%	97.9%	[−0.050, +0.009]	$\approx$
14	0.733	0.814	−0.011	0.014	♯ 10.671	79.3%<0<20.7%	99.7%	[−0.036, +0.018]	$\approx$
15	0.192	0.448	−0.035	0.013	$\star$ 0.330	99.6%<0<0.4%	86.9%	[−0.062, −0.011]	$<$
16	0.065	0.355	−0.030	0.014	1.225	98.7%<0<1.3%	92.1%	[−0.057, −0.003]	$<$
17	0.105	0.333	+0.014	0.014	♯ 6.589	15.7%<0<84.3%	99.3%	[−0.015, +0.041]	$\approx$
18	$\star$ 0.014	0.192	−0.014	0.016	♯ 8.065	82.0%<0<18.0%	98.7%	[−0.045, +0.017]	$\approx$
19	$\star$ 0.033	0.265	+0.008	0.022	♯ 7.488	36.5%<0<63.5%	96.9%	[−0.035, +0.051]	$>$

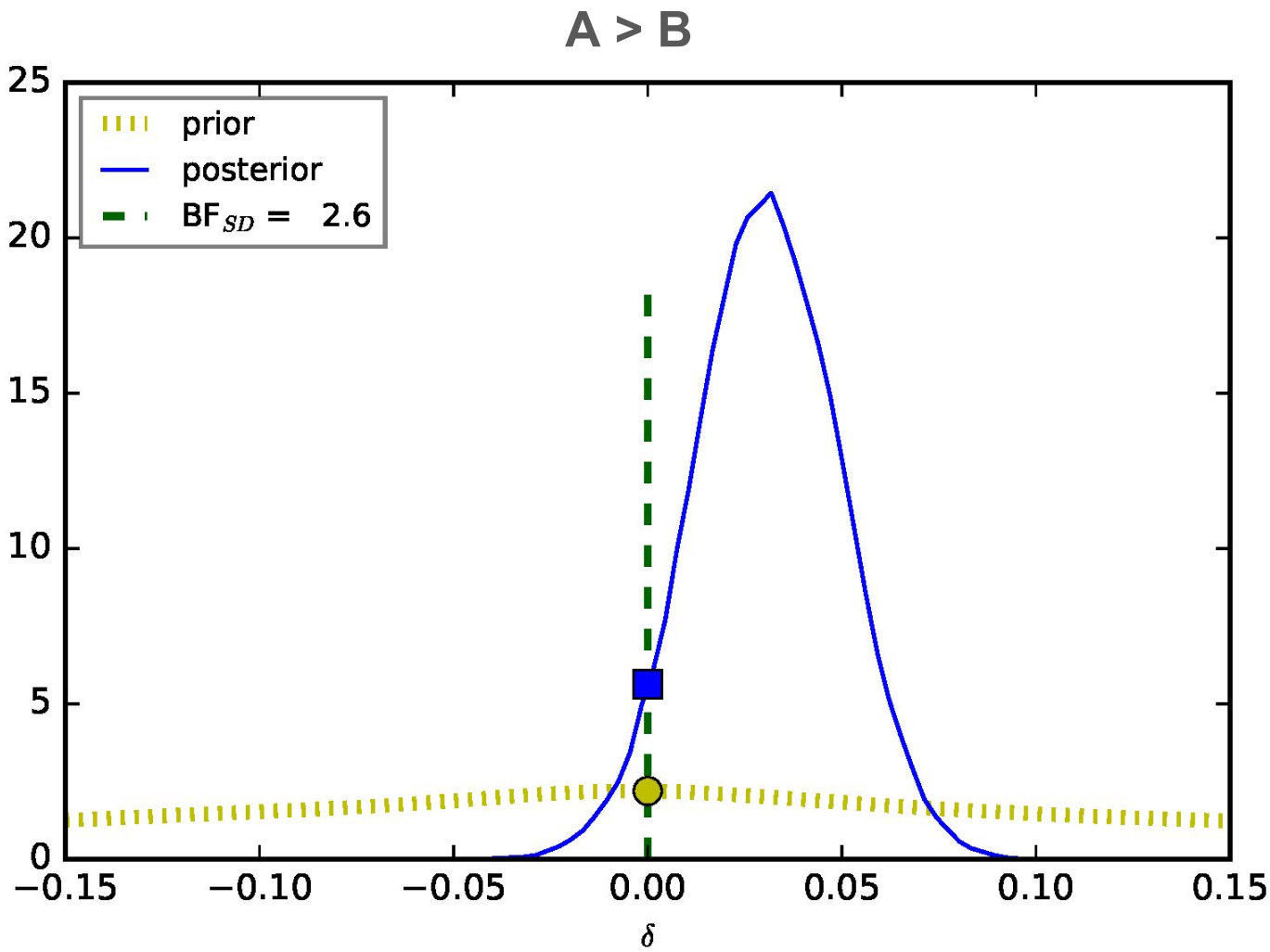
NB_{Mult} and SVM_{L2}

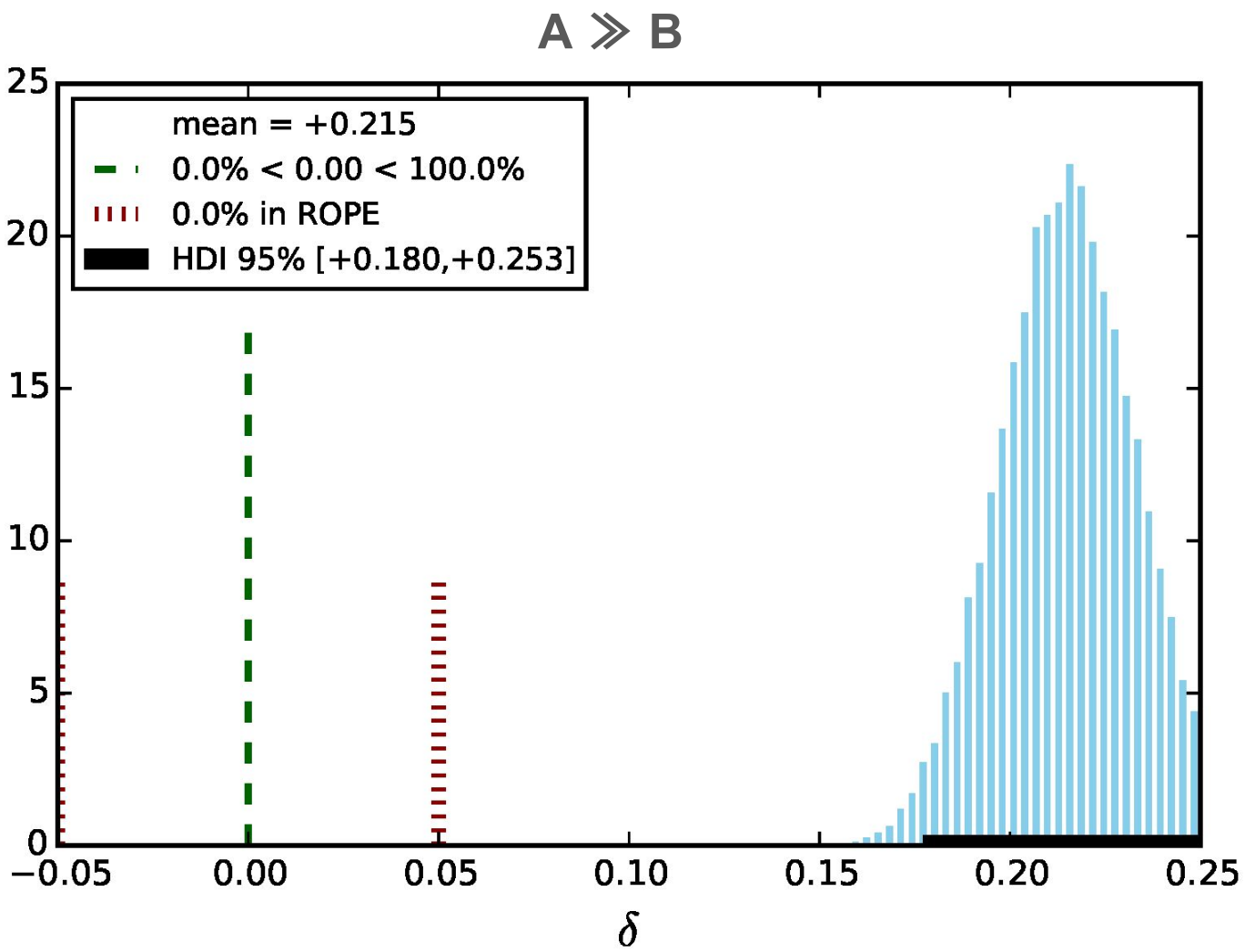
category	frequentist		Bayesian						
	sign-test	t-test	mean	std	BF_{SD}	LG pct	ROPE pct	HDI	**decision**
0	0.314	0.536	−0.001	0.022	♯ 8.542	50.5%<0<49.5%	97.7%	[−0.043, +0.042]	≈
1	0.386	0.521	+0.028	0.018	♯ 3.128	6.2%<0<93.8%	89.0%	[−0.007, +0.063]	>
2	0.056	0.205	−0.028	0.021	♯ 3.739	91.2%<0<8.8%	84.5%	[−0.069, +0.013]	<
3	1.000	1.000	+0.030	0.018	2.966	5.1%<0<94.9%	86.8%	[−0.006, +0.066]	>
4	0.840	0.853	+0.006	0.018	♯ 9.533	37.6%<0<62.4%	99.0%	[−0.031, +0.040]	≈
5	⋆ 0.017	0.089	+0.051	0.017	⋆ 0.075	0.2%<0<99.8%	48.0%	[+0.017, +0.083]	>
6	0.057	0.188	+0.013	0.016	♯ 8.805	20.6%<0<79.4%	99.1%	[−0.017, +0.046]	≈
7	0.180	0.312	+0.031	0.018	2.566	4.4%<0<95.6%	85.1%	[−0.005, +0.067]	>
8	1.000	1.000	+0.007	0.018	♯ 9.551	35.9%<0<64.1%	98.9%	[−0.028, +0.042]	≈
9	⋆ 0.000	⋆ 0.000	+0.215	0.019	⋆ 0.000	0.0%<0<100.0%	0.0%	[+0.180, +0.253]	≫
10	⋆ 0.000	⋆ 0.000	−0.115	0.017	⋆ 0.000	100.0%<0<0.0%	0.0%	[−0.149, −0.082]	≪
11	⋆ 0.024	0.102	−0.016	0.017	♯ 7.016	84.0%<0<16.0%	98.1%	[−0.048, +0.017]	≈
12	⋆ 0.000	⋆ 0.001	+0.073	0.020	⋆ 0.008	0.0%<0<100.0%	12.4%	[+0.034, +0.113]	>
13	⋆ 0.000	⋆ 0.009	+0.060	0.016	⋆ 0.004	0.0%<0<100.0%	26.1%	[+0.029, +0.090]	>
14	⋆ 0.038	0.129	+0.050	0.017	⋆ 0.194	0.2%<0<99.8%	50.7%	[+0.016, +0.082]	>
15	⋆ 0.009	0.064	−0.009	0.014	♯ 10.473	74.6%<0<25.4%	99.8%	[−0.037, +0.021]	≈
16	0.213	0.430	+0.041	0.017	0.650	0.8%<0<99.2%	71.1%	[+0.007, +0.074]	>
17	⋆ 0.010	0.070	+0.053	0.016	⋆ 0.058	0.0%<0<100.0%	43.2%	[+0.023, +0.085]	>
18	0.119	0.349	+0.015	0.019	♯ 6.927	22.2%<0<77.8%	96.5%	[−0.023, +0.053]	>
19	⋆ 0.021	0.179	−0.061	0.031	0.825	97.6%<0<2.4%	36.0%	[−0.119, +0.001]	<

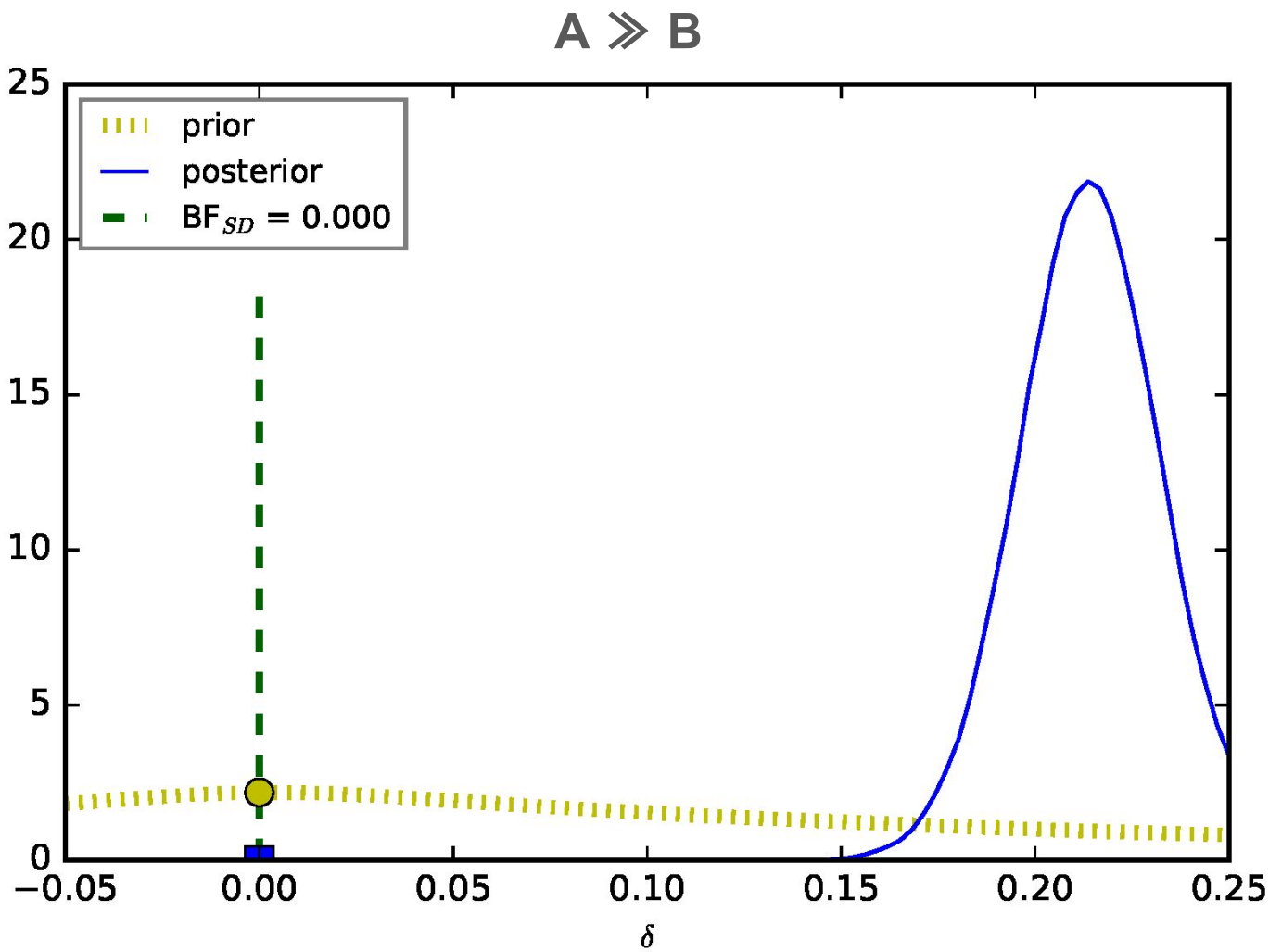

A ≈ B
mean = +0.007
35.9% < 0.00 < 64.1%
98.9% in ROPE
HDI 95% [-0.028,+0.042]
25
20
15
10
5
0
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
δ


A ≈ B
prior
posterior
BF$_{SD}$ = 9.6
25
20
15
10
5
0
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
δ


A > B
mean = +0.031
4.4% < 0.00 < 95.6%
85.1% in ROPE
HDI 95% [-0.005,+0.067]
25
20
15
10
5
0
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
δ

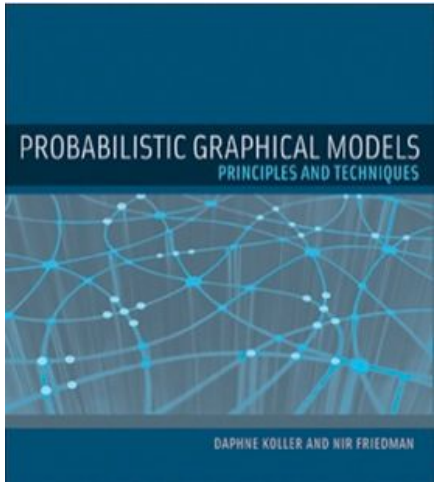

A > B
prior
posterior
BF$_{SD}$ = 2.6
25
20
15
10
5
0
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
δ


A ≫ B
mean = +0.215
0.0% < 0.00 < 100.0%
0.0% in ROPE
HDI 95% [+0.180,+0.253]
25
20
15
10
5
0
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
δ


A ≫ B
prior
posterior
BF$_{SD}$ = 0.000
25
20
15
10
5
0
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
δ

Extensions

- Multiple Classes
  - micro-averaging and macro-averaging
  - trade-off between statistical power and computational tractability
  - Bayesian hierarchical model: see our ICDM-2015 paper
- Other Performance Measures
  - F_β, AUC, …
- Other Tasks
  - search, recommendation, advertising, …



Performance Comparison: Bayesian vs Frequentist

- Main Advantage: Richer Information
  - A ≈ B
  - ROPE
  - F_1, …
- Main Disadvantage: Slower Speed
  - but still perfectly acceptable (< 20 sec)

This repository | Search

Pull requests **Issues** **Gist**

dell-zhang / **bperf**

Code | Issues 0 | Pull requests 0 | Wiki | Pulse | Graphs

Python code for Bayesian performance comparison of classifiers. — Edit

8 commits | **1** branch

Branch: **master** | **New pull request**

dell-zhang Update README.md

GitHub

Take Home Message

- Forget about *p*-values
- Report the HDI & ROPE instead

Thanks for Listening

to My Awesome Presentation


KEEP
CALM
AND
PLEASE DON'T ASK
HARD QUESTIONS