

Approximate inference for vector parameters

Nancy Reid

April 20, 2011

with Don Fraser, Anthony Davison, Nicola Sartori



Models and inference

Motivation

Directional tests

Contingency table

Continuous example

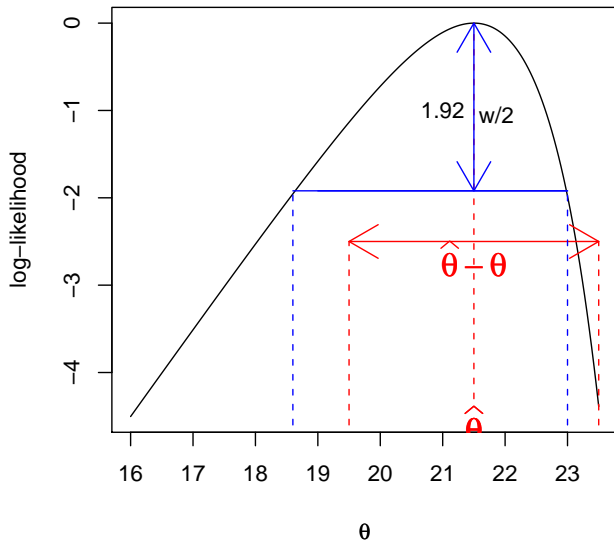
Tangent exponential model

Conclusion

Parametric models and likelihood

- ▶ model $f(y; \theta)$, $\theta \in \mathbb{R}^d$
- ▶ data $y = (y_1, \dots, y_n)$ independent observations
- ▶ log-likelihood function $\ell(\theta; y) = \log f(y; \theta)$
- ▶ parameter of interest $\theta = (\psi, \lambda)$, $\psi \in \mathbb{R}^{d_0}$
- ▶ max. likelihood estimate $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$; $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$
- ▶ log-likelihood ratio $w(\psi) = 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi)\}$
 $= 2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}$
- ▶ profile log-likelihood function $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$

log-likelihood function



Inference using likelihood ratio

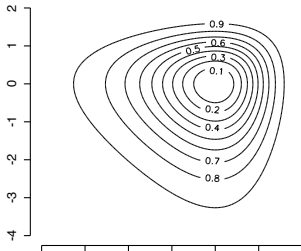
- ▶ $\theta = (\psi, \lambda)$, $\psi \in \mathbb{R}^{d_0}$ $H_0 : \psi = \psi_0$
- ▶ likelihood ratio test:

$$w(\psi_0) = 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi_0, \hat{\lambda}_{\psi_0})\} \sim \chi_{d_0}^2$$

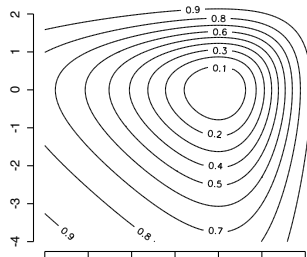
- ▶ Bartlett correction:

$$\tilde{w}(\psi_0) = \frac{w(\psi_0)}{1 + B(\psi_0)/n} \sim \chi_{d_0}^2 \{1 + O(n^{-2})\}$$

(a)



(c)



Targetted inference: single parameters

▶ $\theta = (\psi, \lambda), \lambda \in \mathbb{R}^{d-1}$

▶ likelihood ratio test $w(\psi) = 2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\} \sim \chi_1^2$

▶ likelihood root

$$r(\psi) = \text{sign}(\hat{\psi} - \psi) \sqrt{w(\psi)} \sim N(0, 1) \quad O_p(n^{-1/2})$$

▶ modified likelihood root

$$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)} \log \left\{ \frac{Q(\psi)}{r(\psi)} \right\} \quad O_p(n^{-3/2}) \text{ or } O_p(n^{-1})$$

▶

$$Q(\psi) = \underbrace{\frac{\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)}{\hat{\sigma}_\chi}}_{\text{scalar lined up with } \psi} \underbrace{\frac{|j_{\lambda\lambda}(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}}_{\text{correction re } \lambda} \quad j_{\lambda\lambda}(\theta) = -\partial^2 \ell(\theta) / \partial \lambda \partial \lambda^T$$

▶ $r^*(\psi)$ gives the ‘right’ inferential summary re ψ

▶ and is well approximated by standard normal

Right inferential summary

- Example: linear regression with non-normal errors¹

$$y_i = x_i' \beta + \sigma \epsilon_i; \quad \beta_{7 \times 1}, i = 1, \dots, 32$$

	Normal errors		t_4 errors	
	Est (SE)	z	Est (SE)	z
Constant	-13.26 (3.140)	-4.22	-11.86 (3.70)	-3.21
date	0.212 (0.043)	4.91	0.196 (0.049)	4.02
log(cap)	0.723 (0.119)	6.09	0.682 (0.129)	5.31
NE	0.249 (0.074)	3.36	0.239 (0.080)	2.97
CT	0.140 (0.060)	2.32	0.143 (0.063)	2.26
log(N)	-0.088 (0.042)	-2.11	-0.072 (0.048)	-1.51
PT	-0.226 (0.114)	-1.99	-0.265 (0.110)	-2.42

95% confidence interval for β_5
 (-0.173, -0.002), $\epsilon \sim \text{Normal}$

$r^* : (-0.161, -0.026), \epsilon \sim t_4$

¹BDR, Ch.5.2

... right inferential summary

Likelihood for discrete data 9

Table 1. Lung cancer deaths in British male physicians (Frome, 1983). The table gives man-years at risk/number of cases of lung cancer, T/y , cross-classified by years of smoking, taken to be age minus 20 years, and number of cigarettes smoked per day.

Years of smoking t	Daily cigarette consumption x						
	Nonsmokers	1-9	10-14	15-19	20-24	25-34	35+
15-19	10366/1	3121	3577	4317	5683	3042	670
20-24	8162	2937	3286/1	4214	6385/1	4050/1	1166
25-29	5969	2288	2546/1	3185	5483/1	4290/4	1482
30-34	4496	2015	2219/2	2560/4	4687/6	4268/9	1580/4
35-39	3512	1648/1	1826	1893	3646/5	3529/9	1336/6
40-44	2201	1310/2	1386/1	1334/2	2411/12	2424/11	924/10
45-49	1421	927	988/2	849/2	1567/9	1409/10	556/7
50-54	1121	710/3	684/4	470/2	857/7	663/5	255/4
55-59	826/2	606	449/3	280/5	416/7	284/3	104/1

$$E_{\theta}(Y | t, x, T) = T e^{\lambda_1} t^{\lambda_2} \{1 + e^{\lambda_3} x^{\psi}\}$$

T man-yrs. at risk x # cigarettes t Years smoking $\psi = \psi_0 = 1$

... inferential summary

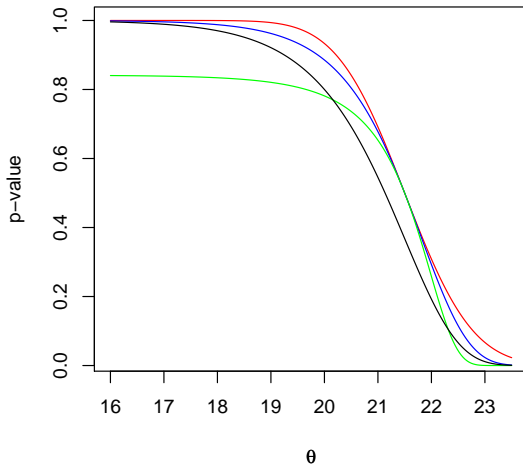
- ▶ $E_{\theta}(Y | t, x, T) = T e^{\lambda_1} t^{\lambda_2} \{1 + e^{\lambda_3} x^{\psi}\}$
- ▶ $H_0 : \psi = 1 \implies$ linear increase in death rate with 'dose'

log-likelihood root $r = 1.506$ $p = 0.066$

modified likelihood root $r^* = 1.491$ $p = 0.068$

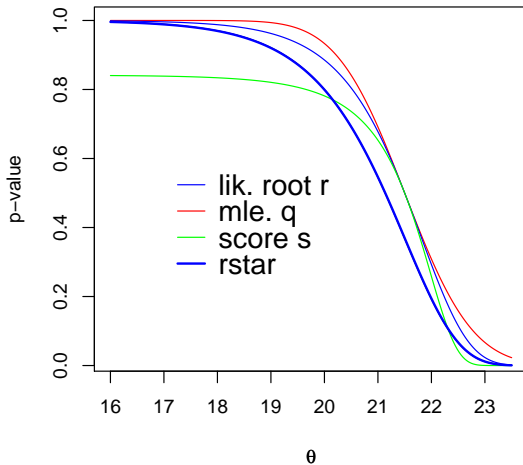
Accurate approximation

Pvalue functions

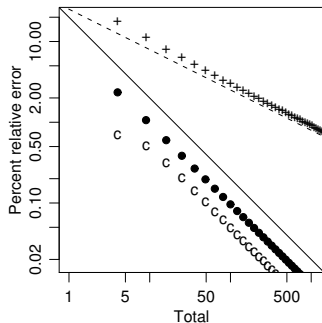
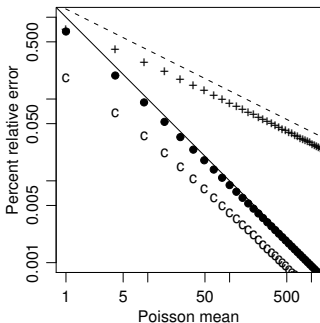


Accurate approximation

Pvalue functions



Accurate approximation



DFR, 2006

scalar ψ

vector ψ

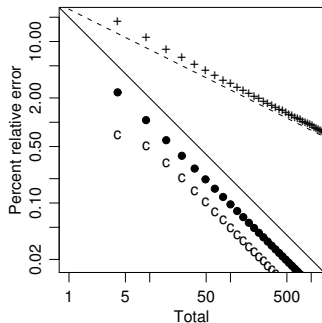
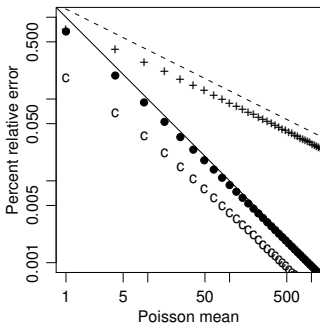
continuous response $r^* : O(n^{-3/2})$

directional: $O(n^{-1})$

discrete response $r^* : O(n^{-1})$

directional: $O(n^{-1})$

Accurate approximation



DFR, 2006

scalar ψ

vector ψ

continuous response

$r^* : O(n^{-3/2})$

directional: $O(n^{-1})$

discrete response

$r^* : O(n^{-1})$

directional: $O(n^{-1})$

Directional tests

- ▶ given a vector of ‘departures’ from ψ_0
e.g. $(\hat{\psi}_1 - \psi_{01}, \dots, \hat{\psi}_{d_0} - \psi_{0d_0})$
- ▶ compute a directional departure based on the magnitude of the vector, conditional on its direction
- ▶ m.l.e. not parameterization invariant; use instead a departure measure based on score variable \mathbf{s}
- ▶ **Propose: directed departure on profile sample space \mathcal{S}_{ψ_0}**
- ▶ all sample points that give the same estimate for the nuisance parameter $\hat{\lambda}_{\psi}$

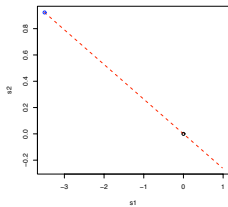
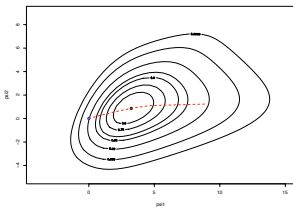


$$\mathcal{S}_{\psi} = \{\mathbf{s} : \ell_{\lambda}(\hat{\theta}_{\psi}^0; \mathbf{s}) = 0\} = \{\mathbf{s} : \hat{\theta}_{\psi} = \hat{\theta}_{\psi}^0\}$$

a surface of dimension d_0 , passing through data point y^0

... directional tests

- ▶ Directed departure on \mathcal{S}_{ψ_0} $\lambda = \hat{\lambda}_{\psi_0}$
- ▶ observed value s^0 , corresponding to m.l.e. $\hat{\theta}^0$
- ▶ expected value s_0 under H_0 , corresponding to constrained m.l.e. $\hat{\theta} = \hat{\theta}_{\psi_0}^0$ $s_0 = -\ell_{\theta}(\hat{\theta}_{\psi_0})$
- ▶ Distribution of magnitude of $|s - s_0|$; given the direction $(s - s_0)/|s - s_0|$

 \mathcal{S}_{ψ_0} line through s_0, s^0

Directional p -value

- ▶ line $\mathbf{s}(t)$ from hypothesis, \mathbf{s}_0 , to data, $\mathbf{s}^0 : \mathbf{s}_0 + t(\mathbf{s}^0 - \mathbf{s}_0)$
- ▶ $f(\mathbf{s}; \psi_0)$: distribution of \mathbf{s} under H_0 .
- ▶ Observed value \mathbf{s}^0 ($t = 1$)
- ▶ Expected value \mathbf{s}_0 ($t = 0$)
- ▶ along the line $\mathbf{s}(t)$ we have

$$f(\mathbf{s}; \psi_0) d\mathbf{s} = f\{\mathbf{s}(t); \psi_0\} dt = f\{\mathbf{s}_0 + t(\mathbf{s}^0 - \mathbf{s}_0); \psi_0\} dt .$$

- ▶ **directional p -value:**

$$p(\psi_0) = \frac{\int_1^\infty t^{d_0-1} f\{\mathbf{s}(t); \psi_0\} dt}{\int_0^\infty t^{d_0-1} f\{\mathbf{s}(t); \psi_0\} dt}$$

- ▶ one-dimensional integrals computed numerically

Example: 2×3 contingency table

- ▶ activity amongst psychiatric patients ²

	Affective disorders	Schizophrenics	Neurotics
Retarded	12	13	5
Not retarded	18	17	25

- ▶ model: log-linear $y \sim \text{Poisson}$, $\log\{E(y)\} = X\theta \in \mathbb{R}^6$
- ▶ nuisance parameter $\lambda \in \mathbb{R}^4$ main effects
- ▶ parameter of interest (ψ_1, ψ_2) interaction
- ▶ $H_0 : \psi = \psi_0 = (0, 0)$ independence
- ▶ line $s(t)$ from hypothesis, s_0 , to data, $s^0 : s_0 + t(s^0 - s_0)$
- ▶ Here $s \equiv y$ $\ell(\theta; y) = \exp \sum (y_i \log \theta_i - \theta_i)$

²Everitt, 1992 CH

... 2×3 contingency table

- ▶ expected frequencies under the null hypothesis $t = 0$

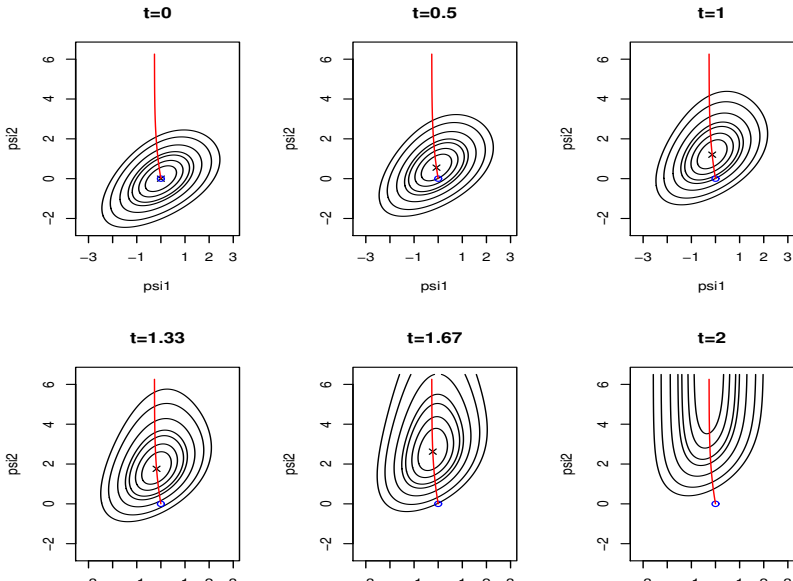
	Affective disorders	Schizophrenics	Neurotics
Retarded	10	10	10
Not retarded	20	20	20

- ▶ need to stop at $t = t_{\max} = 2$.
- ▶ the expected frequencies corresponding to $t_{\max} = 2$

	Affective disorders	Schizophrenics	Neurotics
Retarded	14	16	0
Not retarded	16	14	30

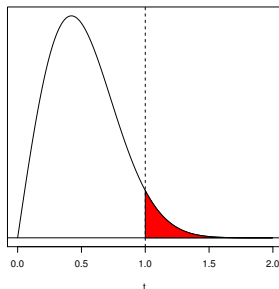
- ▶ All tables along the line $s(t)$ have the same margins.

Log-likelihood along the line $s(t)$



Directional p -value

- ▶ The directional p -value is equal to **0.050**



first order χ_2^2 approximation

$W(\psi_0)$ **0.047**

Skovgaard (2001 SJS) modified version

$W^*(\psi_0)$ **0.048**

simulated conditional

0.051

Another 2×3 table

- ▶ party identification by race³

	Democrat	Independent	Republican
Black	103	15	11
White	341	105	405

- ▶ H_0 : independence nested in saturated model
- ▶ first order likelihood ratio p -value: 2.43×10^{-20}
- ▶ directional p -value: 3.14×10^{-20} .

³Agresti, 2002 Wiley

... party identification example

- ▶ Expected ($t = 0$)

	Democrat	Independent	Republican
Black	58.44	15.80	54.76
White	385.56	104.20	361.24

- ▶ Observed ($t = 1$)

	Democrat	Independent	Republican
Black	103	15	11
White	341	105	405

- ▶ Boundary ($t_{\max} = 1.251$)

	Democrat	Independent	Republican
Black	114.20	14.80	0.00
White	329.80	105.20	416.00

Example: Infant survival data⁴

Age	Smoking	Gestation	Infant Survival	
			No	Yes
< 30	< 5	≤ 260	50	315
		> 260	24	4012
	5+	≤ 260	9	40
		> 260	6	459
30+	< 5	≤ 260	41	147
		> 260	14	1594
	5+	≤ 260	4	11
		> 260	1	124

⁴Agresti, 1992

...infant survival data

- ▶ response variables: length of gestation and infant survival
- ▶ explanatory variables: age of mother (A), number of cigarettes smoked per day (S)
- ▶ null model: all main effects and three first order interactions (IG, IA and SA) $\lambda \in \mathbb{R}^8$
- ▶ full model: two additional first order interaction parameters ψ : for IS and GA
- ▶ first order likelihood ratio p -value = 0.052.
- ▶ directional p -value = 0.056.

Example: suicide behavior data

		age	m	o	y
cause	sex				
drug	f	450	154	259	
	m	399	93	398	
gas	f	13	5	15	
	m	82	6	121	
gun	f	26	7	14	
	m	168	33	155	
hang	f	450	185	95	
	m	797	316	455	
jump	f	71	38	40	
	m	51	26	55	
other	f	60	10	38	
	m	82	14	124	

... suicide behavior data

- ▶ suicide classified by sex, age and cause⁵
- ▶ H_0 : independence, nested in saturated model
- ▶ nuisance parameter λ : intercept, main effects, first order interaction parameters $\in \mathbb{R}^{26}$
- ▶ saturated model: add parameter of interest $\psi \in \mathbb{R}^{10}$; second order interaction between the three variables.
- ▶ The first order likelihood ratio p -value is **0.136**.
- ▶ The directional p -value is **0.141**.

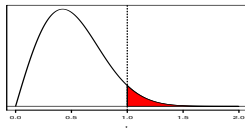
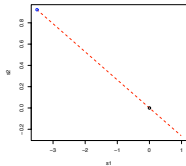
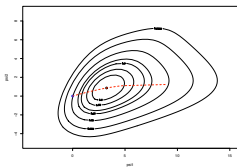
⁵Everitt, 1992 CH

Score variable?

- ▶ exponential family model

$$f(y; \theta) = \exp\{\theta' s(y) - c(\theta) - d(y)\}$$

- ▶ score s carries all the information about θ
- ▶ $\theta = (\psi, \lambda)$, $\psi \in \mathbb{R}^{d_0}$
- ▶ $f(s_1 | s_2; \psi)$ free of λ
- ▶ available from saddlepoint approximation
- ▶ we use this on line $s(t) = s_0 + t(s^0 - s_0)$
- ▶ s_0 : hypothesis (e.g. independence table)
- ▶ s^0 : observed value (e.g. observed table)



Example: continuous data

- ▶ $Y_{ij} \sim N(\mu_i, \sigma_i^2)$, $j = 1, \dots, n_i; i = 1, \dots, g$
- ▶ $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_g^2$
- ▶ $\psi = \sigma_i^2 / \sigma_g^2$, $\lambda = (\sigma_g^2, \mu_1, \dots, \mu_g)$
- ▶ score $s = (n_1 \bar{y}_{1.}, \dots, n_g \bar{y}_{g.}, \Sigma y_{1j}^2, \dots, \Sigma y_{gj}^2)$
- ▶ maximum likelihood estimates: $v_i^2 = \Sigma_j (y_{ij} - \bar{y}_{i.})^2 / n_i$
- ▶ under the hypothesis: $\bar{v}^2 = \Sigma_{ij} (y_{ij} - \bar{y}_{i.})^2 / \Sigma n_i$
- ▶ observed $s^0 = 0$
- ▶ expected $s_0 = (0, \dots, 0, n_1(\bar{v}^2 - v_1^2), \dots, n_g(\bar{v}^2 - v_g^2))$
- ▶ need density along the line $f\{s_0 + t(s^0 - s_0); \psi_0\} dt$

... continuous data

- ▶ $Y_{ij} \sim N(\mu_i, \sigma_i^2)$, $j = 1, \dots, n_i; i = 1, \dots, g$
- ▶ $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_g^2$
- ▶ density along the line $f\{s_0 + t(s^0 - s_0); \psi_0\} dt = h(t)$, say
- ▶

$$h(t) \propto \prod_{i=1}^g \{tv_i^2 + (1-t)\bar{v}^2\}^{(n_i-3)/2}$$

▶

$$p(\psi_0) = \frac{\int_1^\infty t^{d_0-1} h(t) dt}{\int_0^\infty t^{d_0-1} h(t) dt}$$

... continuous data

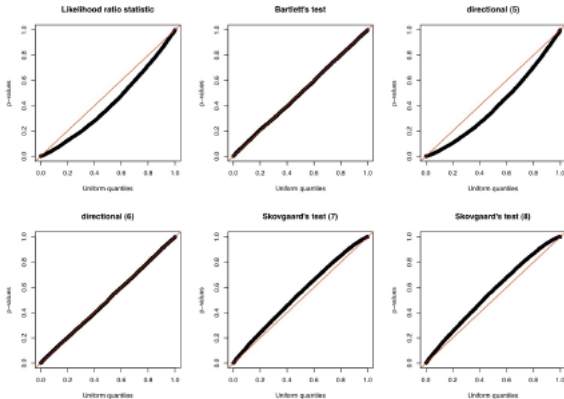


Figure 5: Simulations with $g = 10$ and $n_i = 10$, $i = 1, \dots, g$ and 10000 replications. Uniform quantile plot of p -values when generating data under the null hypothesis.

... continuous data

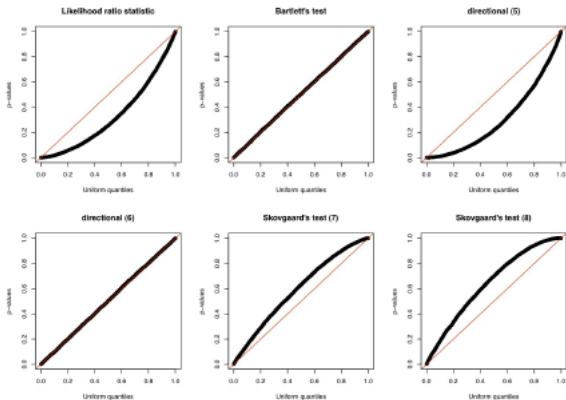


Figure 4: Simulations with $g = 10$ and $n_i = 5$, $i = 1, \dots, g$ and 10000 replications. Uniform quantile plot of p -values when generating data under the null hypothesis.

Tangent exponential model

- ▶ Every model $f(y; \theta)$ on \mathbb{R}^n can be approximated by an exponential family model:

$$f_{\text{TEM}}(s; \theta) ds = \exp\{\varphi(\theta)'s + \ell(\theta)\} h(s) ds \quad (1)$$

- ▶ s is a score variable on \mathbb{R}^d $s(y) = -\ell_{\varphi}(\hat{\theta}^0; y)$
- ▶ $\ell(\theta) = \ell(\theta; y^0)$ is the observed log-likelihood function
- ▶ $\varphi(\theta) = \varphi(\theta; y^0)$ is the canonical parameter $\in \mathbb{R}^d$
to be described
- ▶ has the same observed log likelihood function as the original model
- ▶ has same first derivative on the sample space, at y^0 , as the original model by definition of φ
- ▶ (1) approximates original model to $O(n^{-1})$

Exponential family models

- ▶ linear exponential family:

$$f(y; \theta) = \exp\{\varphi(\theta)'s(y) - c(\theta) - d(y)\}$$

- ▶ canonical parameter obtained as

$$\frac{\partial \ell(\theta; y)}{\partial s(y)} = \varphi(\theta)$$

- ▶ Example $N(\mu, \sigma^2)$:

$$\ell(\theta; y) = \frac{\mu}{\sigma^2} \sum y_i - \frac{1}{2\sigma^2} \sum y_i^2 - \frac{n\mu^2}{2\sigma^2} - n \log \sigma$$

- ▶ Example $Bin(n, p)$:

$$\ell(\theta; y) = \log \frac{p}{1-p} y + n \log(1-p)$$

Inference with TEM

$$f_{\text{TEM}}(\mathbf{s}; \theta) = \exp\{\varphi(\theta)' \mathbf{s} + \ell(\theta)\} h(\mathbf{s})$$

$$\varphi(\theta) = \varphi(\theta; \mathbf{y}^0), \quad \ell(\theta) = \ell(\theta; \mathbf{y}^0)$$

1. marginalize to eliminate nuisance parameter using tangent exponential model

- ▶ likelihood root $r^*(\psi) = r(\psi) + \frac{1}{r(\psi)} \log\left\{\frac{Q(\psi)}{r(\psi)}\right\}$

- ▶ $r(\psi) = \pm\sqrt{[2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}]}$ likelihood root

- ▶ $Q(\psi) = \frac{|\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi) \quad \varphi_\lambda(\hat{\theta}_\psi)|}{|\varphi_\theta(\hat{\theta})|} \frac{|j(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}$

2. saddlepoint approximation gives distribution approximation

- ▶ $r^* \sim N(0, 1) \quad O(n^{-3/2})$

Canonical parameter $\varphi(\theta)$

- ▶ if $f(y; \theta)$ is an exponential family, φ is sitting in the model
- ▶ if not
- ▶ if y is continuous, define

$$V = \left. \frac{dy}{d\theta} \right|_{y=y^0, \theta=\hat{\theta}^0} \quad y = (y_1, \dots, y_n)$$

- ▶ ??
- ▶ $z_i = z_i(y_i; \theta)$ with a fixed distribution, e.g. $(y_i - \mu)/\sigma$
- ▶ $V = - \left(\frac{\partial z}{\partial y} \right)^{-1} \frac{\partial z}{\partial \theta} \Big|_{y=y^0, \theta=\hat{\theta}^0} \quad n \times p$
- ▶

$$\varphi(\theta) = \varphi(\theta; y^0) = \left. \frac{\partial \ell(\theta; y)}{\partial V} \right|_{y=y^0} = \sum_{i=1}^n \frac{\partial \ell(\theta; y^0)}{\partial y_i} V_i$$

... canonical parameter $\varphi(\theta)$

- ▶ φ is a data-derivative of log-likelihood, $\ell; \nu(\theta; y^0)$
- ▶ doesn't work if the data is discrete
- ▶ $y \rightarrow s$ score variable

- ▶ $\frac{dy}{d\theta} \rightarrow \frac{dE(s; \theta)}{d\theta}$

DFR, 2006

- ▶
$$V_i = \left. \frac{\partial}{\partial \theta} E(s_i; \theta) \right|_{\theta = \hat{\theta}_0}$$

- ▶
$$\varphi(\theta) = \sum_{i=1}^n \left. \frac{\partial \ell(\theta; y^0)}{\partial s_i} \right|_{\theta = \hat{\theta}_0} V_i$$

Vector parameter inference

- ▶ use $\ell(\theta; y^0)$ to construct generalized likelihood ratio $w(\psi)$
- ▶ use $\varphi(\theta; y^0)$ to get a modified version
- ▶ combine with nuisance parameter adjustment
- ▶ saddlepoint approximation
- ▶ evaluation p -value conditionally

Summary

- ▶ exponential family, parameter of interest a single canonical parameter
- ▶ saddlepoint approximation
- ▶ tangent exponential model, with locally defined canonical parameter
- ▶ gives both inference summary and accurate approximation for arbitrary scalar parameter of interest
- ▶ extending to discrete data \longrightarrow requires new definition of canonical parameter

... summary

- ▶ exponential family, parameter of interest a vector of canonical parameters
- ▶ directional p -value based on saddlepoint approximation
- ▶ tangent exponential model, for either discrete or continuous data, gives joint density
- ▶ **eliminating nuisance parameter seems to be more difficult for non-canonical parameters** – wip
- ▶ hoping for insight from the structure of the solution – also wip
- ▶ Thanks!! Nicola, Don, Anthony

References

Fraser, D.A.S. and Massam, H. (1985). Conical tests. *Stat. Hefte*

Skovgaard, I.M. (1988). Saddlepoint expansions for directional test probabilities. *JRSS B*

Skovgaard, I.M. (2001). Likelihood asymptotics. *SJS*

Cheah, P.K., Fraser, D.A.S., Reid, N. (1994). Multiparameter testing in exponential models. *Biometrika*

Davison, A.C., Fraser, D.A.S., Reid, N. (2006). Improved likelihood inference for discrete data. *JRSS B*

Davison, A.C., Fraser, D.A.S., Reid, N., Sartori, N. (2011). On assessing vector valued parameters. in progress

`\beamertemplatenavigationsymbolsempty`