What is the chance that the match is a coincidence?

Richard Gill
Mathematical Institute, University Leiden
27 September 2013

What is the chance that the match is a coincidence?

Part 1: joint work with
Stefan Zohren (Oxford/Rio), Dragi Anevski (Lund)

Part 2: joint work with Helene van Eijck (Leiden)

What is the chance it's just a coincidence?

- DNA match
- Finger print match
- Handwriting match
- Locations and times of mobile phone calls
- ... and so on ...

$P(\text{coincidence}|H_{\text{defence}})$, or perhaps
$P(\text{coincidence}|H_{\text{defence}}) : P(\text{coincidence}|H_{\text{prosecution}})$

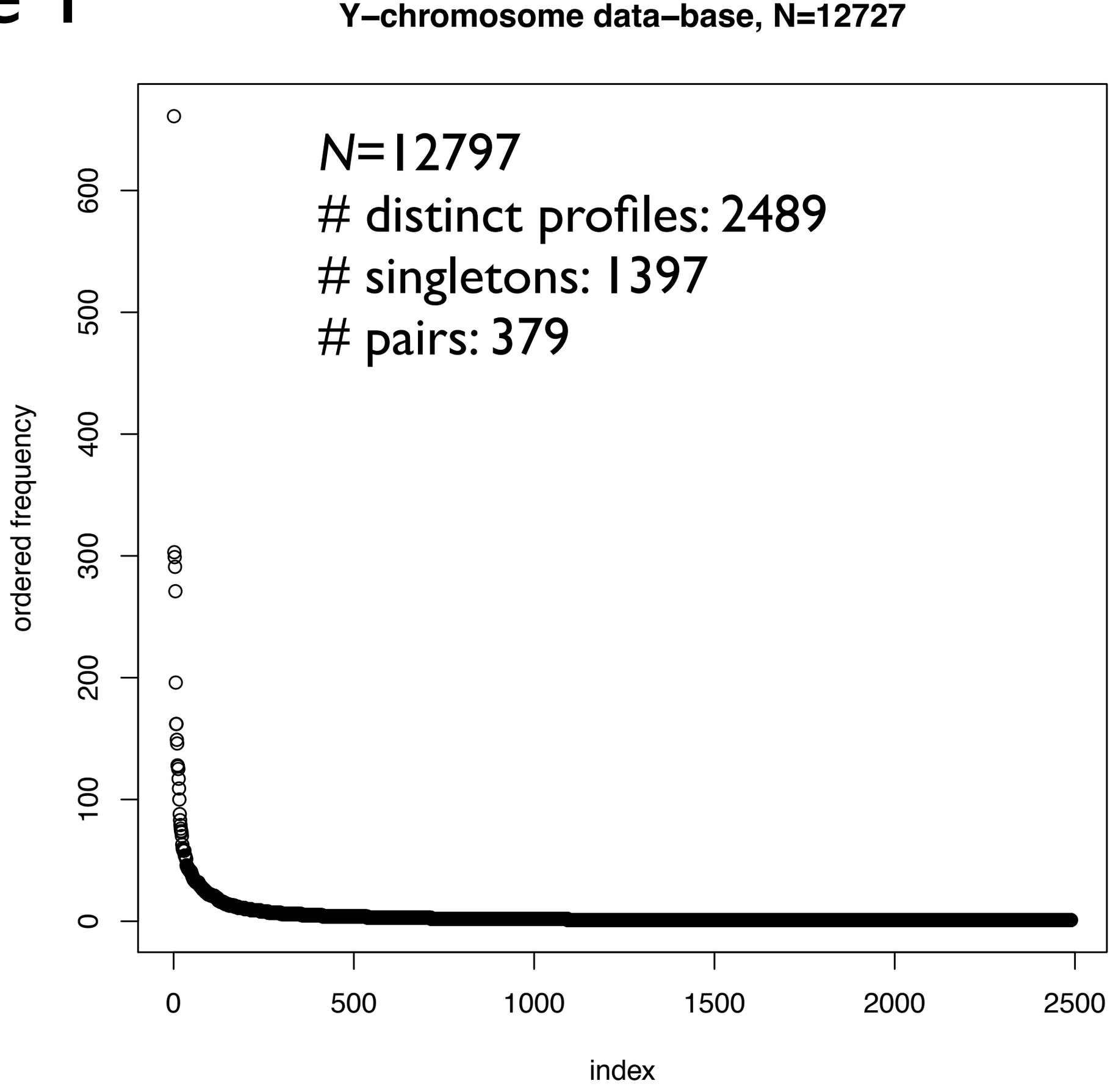
Example 1

- We have a data-base of Y-chromosome DNA profiles
  (pretend it's a random sample)
- We have a crime, and we have a suspect
- Profile of DNA found at crime scene matches DNA profile of suspect, doesn't occur in data-base

Example 2

- Mobile phone co-location
  - Phone 1 is anonymous, connected to crime
  - Phone 2 is not anonymous
  - Phones 1 and 2 seem to be in the same places at the same times

Example 1


Y–chromosome data–base, N=12727
N=12797
# distinct profiles: 2489
# singletons: 1397
# pairs: 379
ordered frequency
0
100
200
300
400
500
600
0
500
1000
1500
2000
2500
index

Defence position

- Defence: suspect and perpetrator (donor) are different.
- What is chance that if we pick two people from population, they have same, not yet observed, profile?
    - Good-type estimator: (2x379 / *N*) x 1/*N*

N=12797
distinct profiles: 2489
singletons: 1397
pairs: 379

Prosecution position

- Suspect and perpetrator are the same
- What is chance that if we pick one person from population, he'll have a not yet observed profile?
- Good-type estimator: 1397/*N*

N=12797
distinct profiles: 2489
singletons: 1397
pairs: 379

Likelihood ratio

- ((2x379 / *N*) x 1/*N*) : (1397/*N*)
  = 2 x 379 / 12797 x 1397
  ≈ 1 : 400 000

N=12797
# distinct profiles: 2489
# singletons: 1397
# pairs: 379

Conventional approaches:
1/12798 (prosecution)
2/12799 (defence)

ESTIMATED likelihood ratio

- ((2x379 / *N*) x 1/*N*) : (1397/*N*)
  = 2 x 379 / 12797 x 1397
  ≈ 1 : 400 000

N=12797
# distinct profiles: 2489
# singletons: 1397
# pairs: 379

How accurate is this? Should we care?

- Why we should care: new technologies give initially rather small new data-bases!
- e.g. mitochondrial DNA

- Proposal: estimate underlying distribution, estimate distribution of quantities like previous by plug-in
- The naive estimator of the underlying distribution might not be a good idea...
- The non-parametric maximum likelihood estimator is different and seems to be a lot better...

Problem

- Underlying data:
  X ~ multinomial (N;p)
  (X_1,X_2,...) ~ multinomial (N; p_1,p_2,...), $p_1 \geq p_2 \geq$ ...
- Observed data:
  Y = sort(X)
  (sort = *monotone ordering* = sort in decreasing order)
- Problem: estimate p = sort(p)
- Naive estimator: Y/N (sorted empirical)
- Missing data: map from observed data categories (ordered by observed relative frequency) to true categories (ordered by true probability)

Previous work

- Alon Orlitsky and collaborators introduce "*hi-profile estimator*" = NPMLE
- Compute with EM + Metropolis-Hastings (MH *within* EM)
- Outline proof of consistency ("incomplete" to put it kindly – yet in my opinion quite brilliant)
- Many "small data" examples

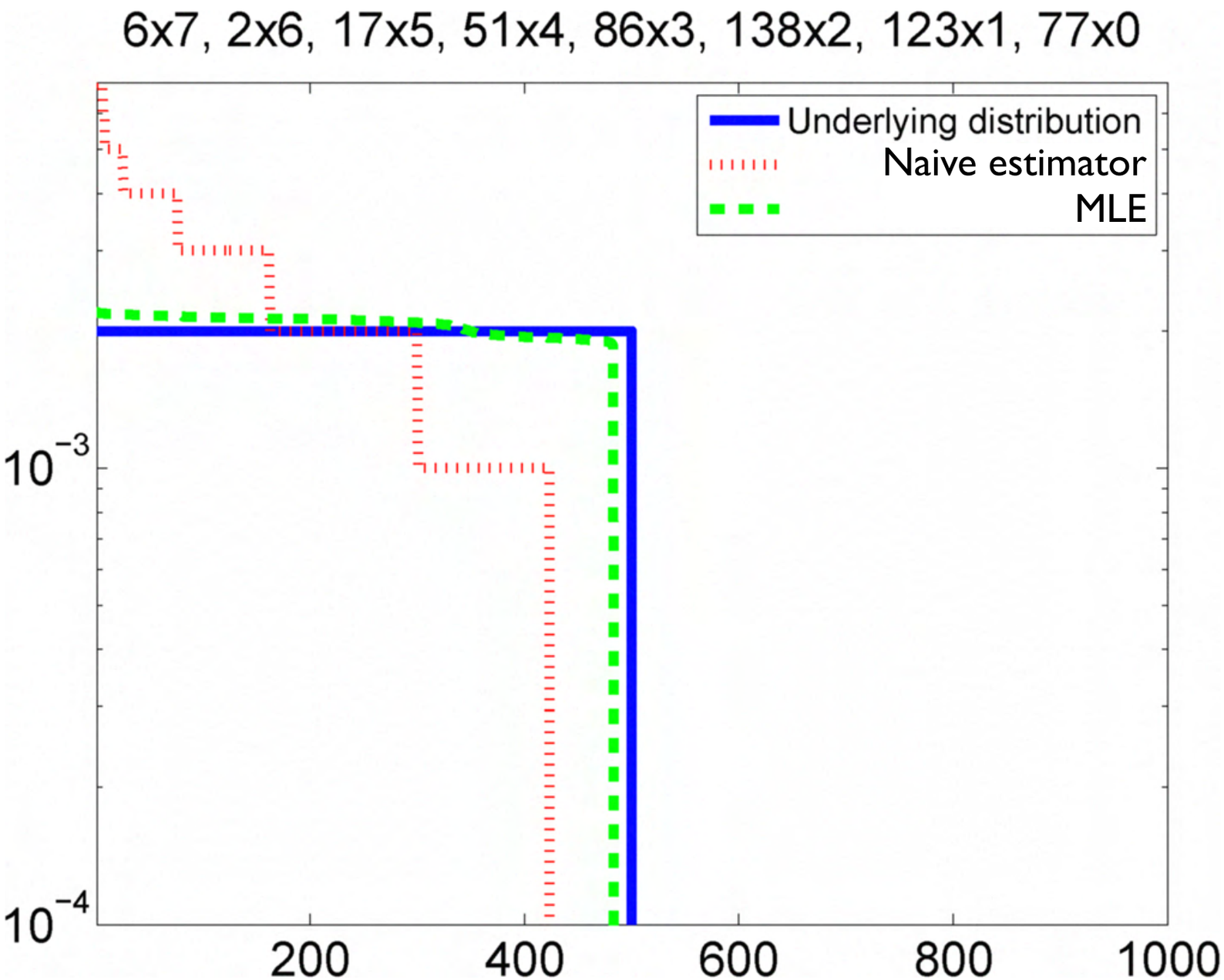
ISIT 2009, Seoul, Korea, June 28 - July 3, 2009

The Maximum Likelihood Probability of Unique-Singleton, Ternary, and Length-7 Patterns

Jayadev Acharya
ECE Department, UCSD
Email: jayadev@ucsd.edu

Alon Orlitsky
ECE & CSE Departments, UCSD
Email: alon@ucsd.edu

Shengjun Pan
CSE Department, UCSD
Email: s1pan@ucsd.edu



N=1000

The Maximum Likelihood Probability of Unique-Singleton, Ternary, and Length-7 Patterns

Jayadev Acharya
ECE Department, UCSD
Email: jayadev@ucsd.edu

Alon Orlitsky
ECE & CSE Departments, UCSD
Email: alon@ucsd.edu

Shengjun Pan
CSE Department, UCSD
Email: s1pan@ucsd.edu

Canonical $\overline{\psi}$	$\widehat{P}_{\overline{\psi}}$	Reference
1	any distribution	Trivial
11, 111, 111, ...	(1)	Trivial
12, 123, 1234, ...	$()$	Trivial
112, 1122, 1112, 11122, 111122	$(1/2, 1/2)$	[12]
11223, 112233, 1112233	$(1/3, 1/3, 1/3)$	[13]
111223, 1112223,	$(1/3, 1/3, 1/3)$	Corollary 5
1123, 1122334	$(1/5, 1/5, \ldots, 1/5)$	[12]
11234	$(1/8, 1/8, \ldots, 1/8)$	[13]
11123	(3/5)	[15]
11112	$(0.7887.., 0.2113..)$	[12]
111112	$(0.8322.., 0.1678..)$	[12]
111123	(2/3)	[15]
111234	(1/2)	[15]
112234	$(1/6, 1/6, \ldots, 1/6)$	[13]
112345	$(1/13, \ldots, 1/13)$	[13]
1111112	$(0.857.., 0.143..)$	[12]
1111122	$(2/3, 1/3)$	[12]
1112345	(3/7)	[15]
1111234	(4/7)	[15]
1111123	(5/7)	[15]
1111223	$\left(\frac{1}{\sqrt{7}}, \frac{\sqrt{7}-1}{2\sqrt{7}}, \frac{\sqrt{7}-1}{2\sqrt{7}}\right)$	Corollary 7
1123456	$(1/19, \ldots, 1/19)$	[13]
1112234	$(1/5, 1/5, \ldots, 1/5)$?	Conjectured

TABLE I
PML DISTRIBUTIONS OF ALL PATTERNS OF LENGTH ≤ 7

Anevski, Gill, Zohren

Estimating a probability mass function with unknown labels

Dragi Anevski, Richard Gill, Stefan Zohren,
Lund University, Leiden University, Oxford University

- Study *sieved* NPMLE estimator
- Proof of consistency
- Computation by SA-MH-EM
  (*stochastic approximation* for *interleaved* Metropolis-Hastings - EM)

SA-MH-EM

T= conditional expectation of *X*/*N* given observed data *Y*

- Re-sample *X* from law of *X* given *Y* under p, e.g. by Metropolis-Hastings
- $T_{new} = (1-\gamma)T_{old}+\gamma X/N$
- Replace p by mle of p based on *T* = isotonic regression of *T* ("pool adjacent violators")
- $\gamma = \gamma_k = k^{-1}$, *k* = iteration step

Kuhn and Lavielle, 2004 (it converges...)

Does this work?

- Almost! MLE often doesn't exist!
- Fix 1: *enlarge* parameter space
  - $\sum_i p_i \leq 1$ (before: $\sum_i p_i = 1$)
- Fix 2: *reduce* (sieve) enlarged parameter space
  - $p=(p_1,\ldots p_K)$, $K < \infty$
  - $p_K \geq \varepsilon > 0$

(Orlitsky et al. already use Fix 1)

Results

Estimating a probability mass function with unknown labels

Dragi Anevski, Richard Gill, Stefan Zohren,
Lund University, Leiden University, Oxford University

I: (not sieved) NPMLE

Theorem 1 *Let $\hat{\theta} = \hat{\theta}^{(n)}$ be the maximum likelihood estimator. Then for any $\delta > 0$*

$$P^{n,\theta}(||\hat{\theta} - \theta||_1 > \delta) \leq \frac{1}{\sqrt{3}n} e^{\pi\sqrt{\frac{2n}{3}} - n\frac{\epsilon^2}{2}}(1 + o(1)) \quad as \quad n \to \infty$$

where $\epsilon = \delta/(4r)$ and $r = r(\theta, \delta)$ such that $\sum_{i=r+1}^{\infty} \theta_i \leq \delta/4$.

Results I: (not sieved) NPMLE

Theorem 1 *Let* $\hat{\theta} = \hat{\theta}^{(n)}$ *be the maximum likelihood estimator. Then for any* $\delta > 0$

$$P^{n,\theta}(||\hat{\theta} - \theta||_1 > \delta) \leq \frac{1}{\sqrt{3n}} e^{\pi\sqrt{\frac{2n}{3}} - n\frac{\epsilon^2}{2}} (1 + o(1)) \quad as \quad n \to \infty$$

where $\epsilon = \delta/(4r)$ *and* $r = r(\theta, \delta)$ *such that* $\sum_{i=r+1}^{\infty} \theta_i \leq \delta/4$.

Theorem 2 *Let* $\Theta_\kappa = \{\theta : \theta_x = l(x)x^{-\kappa}\}$ *for* $\kappa > 1$ *fixed and with* l *some function slowly varying at infinity. Then, if* $\theta \in \Theta_\kappa$,

$$n^{1/4}||\hat{\theta}^{(n)} - \theta|| \overset{a.s.}{\to} 0$$

as $n \to \infty$.

Remark: naive estimator is root n consistent!

Results 2: sieved NPMLE

Theorem 3 *Let $\hat{P}_{SML}$ be the sieved ML estimator defined in* (13). *Then for any $\delta > 0$*

$$\mathbb{P}_P(||\hat{P}_{SML} - P||_1 > \delta) \leq \frac{1}{2\sqrt{3}n} e^{\pi\sqrt{\frac{2n}{3}}} (e^{-n(\epsilon+\frac{1}{n})^2/2} + e^{-n(\epsilon-\frac{1}{n})^2/2})(1 + o(1))$$

as $n \to \infty$, where $\epsilon = \delta/(4r)$ and $r = r(P, \delta)$ such that $\sum_{i=r+1}^{\infty} \theta_i \leq \delta/4$.

Theorem 4 *Let $\Theta_{\nu,\beta} = \{\theta : \theta_x = o(x^{\nu-1/2}e^{-\beta x^{\nu+1/2}})$ as $x \to \infty\}$ for $\nu > 0, \beta > 0$ fixed. Then, if $\theta \in \Theta_{\nu,\beta}$,*

$$n^{\alpha}||\hat{\theta}^{(n)}_{(s)} - \tilde{\theta}|| \overset{a.s.}{\to} 0$$

as $n \to \infty$, with $\alpha < 1/4$.

Remark: naive estimator is root n consistent!

Tools

- Kiefer-Dvoretsky-Wolfowitz

$$\Pr\left(\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|>\varepsilon\right)\leq 2e^{-2n\varepsilon^2}$$

- *Monotone ordering* is contraction mapping (wrt sup norm)
- Hardy-Ramanujan: number of partitions of *n* grows as $\frac{1}{4n\sqrt{3}}e^{\pi\sqrt{\frac{2n}{3}}}$

Prelimary steps

- By Kiefer-Wolfowitz, empirical relative frequencies are close to true probabilities, in sup norm, ordering known
- By contraction property, same is true after monotone reordering of empirical
    - 1. Naive estimator is close to the truth (with large probability)
    - 2. Naive estimator is far from any particular *distant* non-truth (with large probability)

Change of notation!

Key lemma

P, *Q* probability measures; *p*, *q* densities

- Find event *A* depending on *P* and δ
  - $P(A) \geq 1-\varepsilon$
  - $Q(A) \leq \varepsilon$ for all *Q*: $d(Q,P) \geq \delta$
  - Hence $P(\ p/q \geq 1) \geq 1-2\varepsilon$ if $d(Q,P) \geq \delta$

Application:
P, *Q* are probability distributions of data,
depending on parameters θ, ϕ respectively
A is event that *Y*/*n* is within δ of θ (sup norm)
d is L_1 distance between θ and ϕ

Proof outline

- By key lemma, probability MLE is any particular q distant from (true) p is very small
- By Hardy, there are not many q to consider
- Therefore probability MLE is far from p is small
- Must be very careful – sup norm on data, L_1 norm on parameter space, truncation of parameter vectors...

Conclusions

- We have consistency but not with the expected rate (but our proof is very crude)
- We did not yet study behaviour of functionals of estimated distribution
- More work needs to be done on computation (SA-MH-EM)

We still don't know if the whole thing is a good idea, either in theory or in practice – but at least we made a start

Example 2

- Colocation analysis of mobile phone call data records

Colocation analysis

- Prosecution alleges that members of small terrorist gang use several mobile phones, both “hidden” and “public”
- Mobile calls link one of hidden networks to crime
- **Colocation** of phones links hidden networks to one another and finally to public phones

Colocation analysis

- Two phones *colocate* if they are never used far apart in space close together in time
- NB cell phone records:
  - which cell towers
  - which phone called which phone
  - when

Colocation analysis

- continued on other slides

Conclusions

- "Forensic statistics" (statistics in crime investigation, statistics in crime prosecution) is doing statistics in the most alien environment imaginable
- The standard paradigmas don't work
- Big challenges for statisticians...