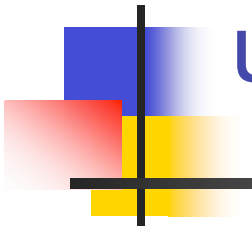
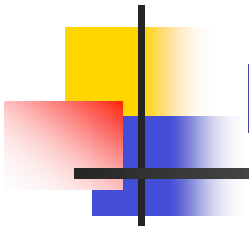


Coincidences and entropies of undersampled distributions

Ilya Nemenman
CCS-3

http://nsb-entropy.sourceforge.net



IT based data analysis

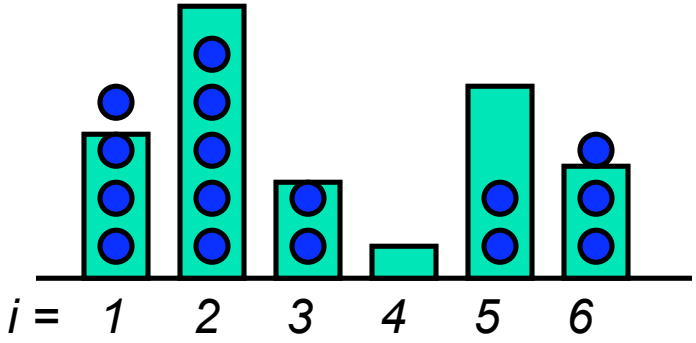
- Bioinformatics
  - Correlation analysis (network topology)
  - Overabundance of elements (TF binding sites search)
  - Channel capacity of signaling pathways
- Neurobiology
  - Channel capacity of neuronal pathways
  - Adaptation to maximize information transmission
- Structural biology
  - Free energy calculations for alternative protein confirmations
- Rare events search
- Others: Linguistics, Finance, TR…

Undersampling and entropy estimation

Maximum likelihood estimation:

$p_i, \quad i = 1 \ldots K$

(*K* - # of bins)



$$p_i^{ML} = \frac{n_i}{N}$$

(*N* - sample size)

$$S_{ML} = -\sum_i \frac{n_i}{N} \log \frac{n_i}{N}$$

!

$$\langle S_{ML} \rangle \leq -\sum_i \frac{\langle n_i \rangle}{N} \log \frac{\langle n_i \rangle}{N} = S$$

Undersampling and entropy estimation

$$\langle S_{ML} \rangle \leq -\sum_i \frac{\langle n_i \rangle}{N} \log \frac{\langle n_i \rangle}{N} = S$$

log K

$$\text{bias} \propto -\frac{2^S}{N} \gg (\text{variance})^{1/2} \propto \frac{1}{\sqrt{N}}$$

- Fluctuations underestimate entropies and overestimate mutual informations.
- Universal bias correction possible IFF $K<N$ (Grassberger 89-03, Antos and Kontoyiannins 02, Wyner and Foster 03, Batu et al. 02, Paninski 03, Panzeri and Treves 96, Strong et al. 98)

For $N<K$:

- Assumptions needed (won't work uniformly).
- Estimate entropies *without* estimating distributions.

Hope (Ma, 1981)

For uniform *K*-bin distribution the first coincidence occurs for

$$N_c \sim \sqrt{K} = \sqrt{2^S}$$

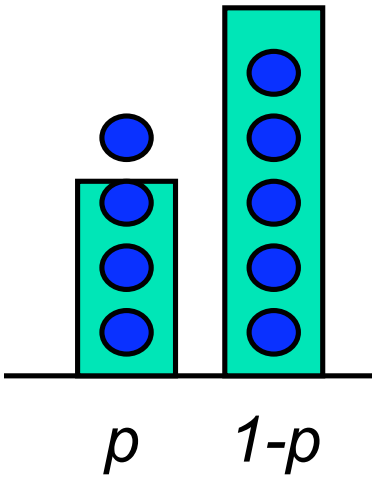
$$S \sim 2 \log N_c$$

Time of first coincidence

- Can make estimates for square-root-fewer samples!
- Can this be extended to nonuniform cases?

What is unknown?

Binomial distribution:



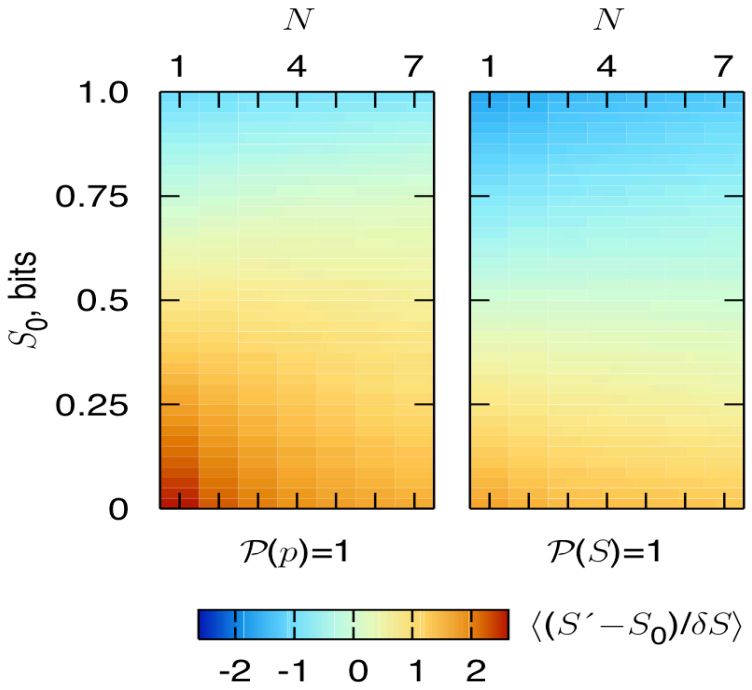
$$S = -p\log p - (1-p)\log(1-p)$$

Selection of wrong "unknown" biases estimation

Assume (Bayes)

uniform (no assumptions)

p S

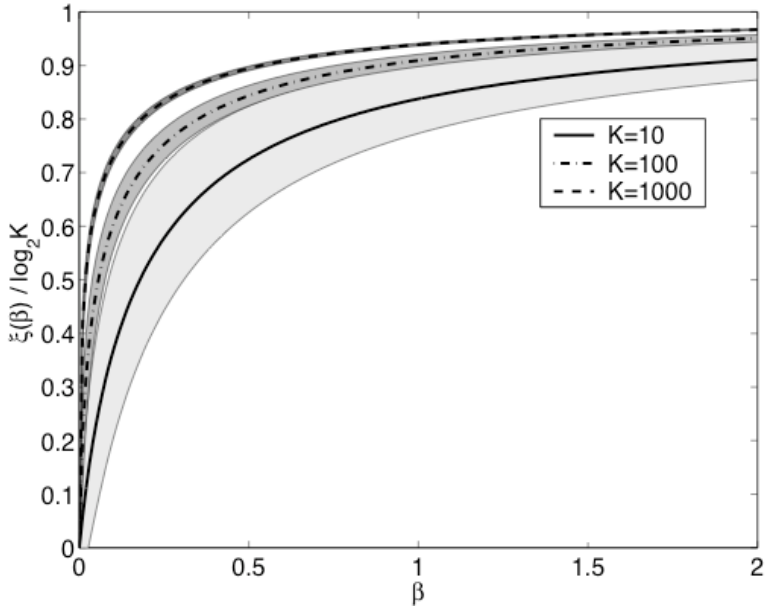


$$\varepsilon = \left\langle \frac{S_{est} - S_{true}}{\delta S_{est}} \right\rangle$$

One possible S-uniformization strategy

For Dirichlet pseudocount priors (uniform, ML, KT, etc.)

$$P_\beta(\{q_i\}) \propto \delta\left(1-\sum_{i=1}^{K} q_i\right) \prod_{i=1}^{K} q_i^{\beta-1}$$



The entropy is known a priori for *K*>>1.

Uniformize:

$$P_\beta(\{q_i\},\beta) = \frac{1}{Z}\,\delta\left(1-\sum_{i=1}^{K} q_i\right) \prod_{i=1}^{K} q_i^{\beta}\; \left.\frac{dS}{d\beta}\right|_{N=0} P(S|_{N=0})$$

- Infinite Dirchlet mixture.
- A delta-function sliding along the a priori entropy expectation -- producing (almost) uniform expectations.

Properties of the NSB estimator

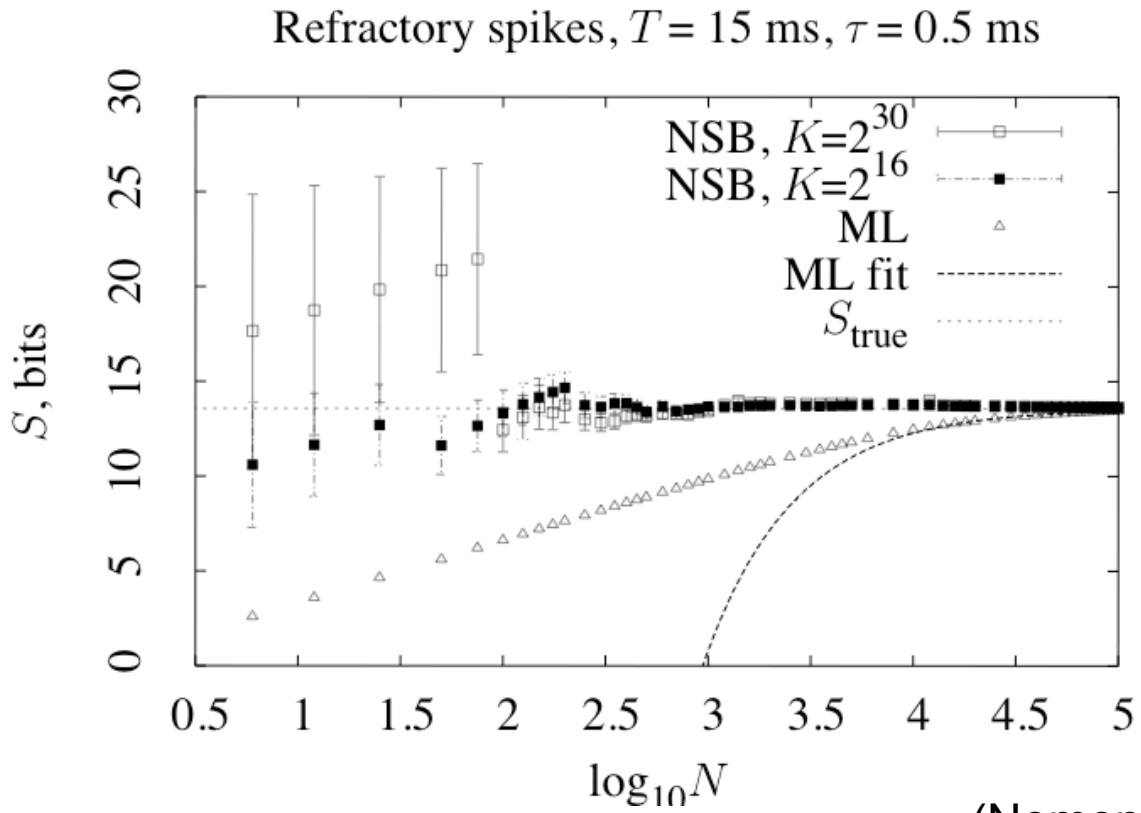
- Posterior variance scales as $N^{-0.5}$.
- Asymptotically consistent (is guaranteed correct for large N).
- Allows infinite # of bins.
- Little bias for light rank-order tail distributions.
- Is also Bayesian model selection (choosing the right mixture component).
- Has error bars!
- Counts coincidences and works in Ma regime (if works).

$$\left\langle \delta^2 S \right\rangle = \frac{1}{(\#\text{ of coincidences})} + \cdots$$

(Nemenman et al. 2002, 03, 04, 06)

Synthetic test

Refractory Poisson, rate 0.26 spikes/ms, refractory period 1.8 ms, T=15ms. discretization 0.5ms. true entropy 13.57 bits.



(Nemenman et al. 2004)