Estimating information quantities from biological data

Ilya Nemenman

Thanks to: William Bialek, Rob de Ruyter van Steveninck, Fariel Shafee

(UCSB, Princeton University, Indiana University)

http://arxiv.org/abs/physics/0306063
http://arxiv.org/abs/physics/0207009
http://arxiv.org/abs/physics/0108025
http://arxiv.org/abs/physics/0103088

Talk outline

Problem setup Why bother?

Talk outline

Problem setup Why bother?
Developing intuition Why hard?

Talk outline

Problem setup Why bother?
Developing intuition Why hard?
The method An idea, analysis, asymptotics.

Talk outline

Problem setup Why bother?
Developing intuition Why hard?
The method An idea, analysis, asymptotics.
Applications Synthetic and natural neural data.

Entropy and information

Assumptions-free measures of randomness and dependence.

Entropy and information

Assumptions-free measures of randomness and dependence.

$$S[x] = -\sum_x q(x) \log q(x)$$

Entropy and information

Assumptions-free measures of randomness and dependence.

$$S[x] = -\sum_x q(x)\,\log q(x)$$

$$I[x,y] = \sum_{x,y} q(x,y)\,\log\frac{q(x,y)}{q(x)q(y)} = S[x] - S[x|y]$$

$$I[x,y,z] = \sum_{x,y,z} q(x,y,z)\,\log\frac{q(x,y,z)}{q(x)q(y)q(z)}\;\ldots$$

Entropy and information

Assumptions-free measures of randomness and dependence.

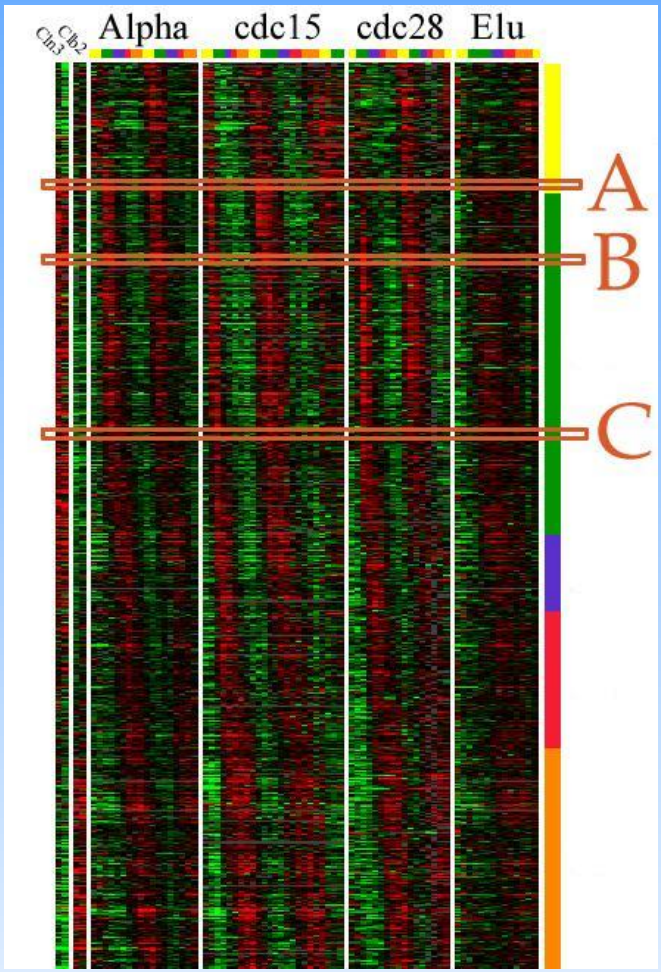
$$S[x] = -\sum_x q(x)\,\log q(x)$$

$$I[x,y] = \sum_{x,y} q(x,y)\,\log\frac{q(x,y)}{q(x)q(y)} = S[x] - S[x|y]$$

$$I[x,y,z] = \sum_{x,y,z} q(x,y,z)\,\log\frac{q(x,y,z)}{q(x)q(y)q(z)}\ \ldots$$

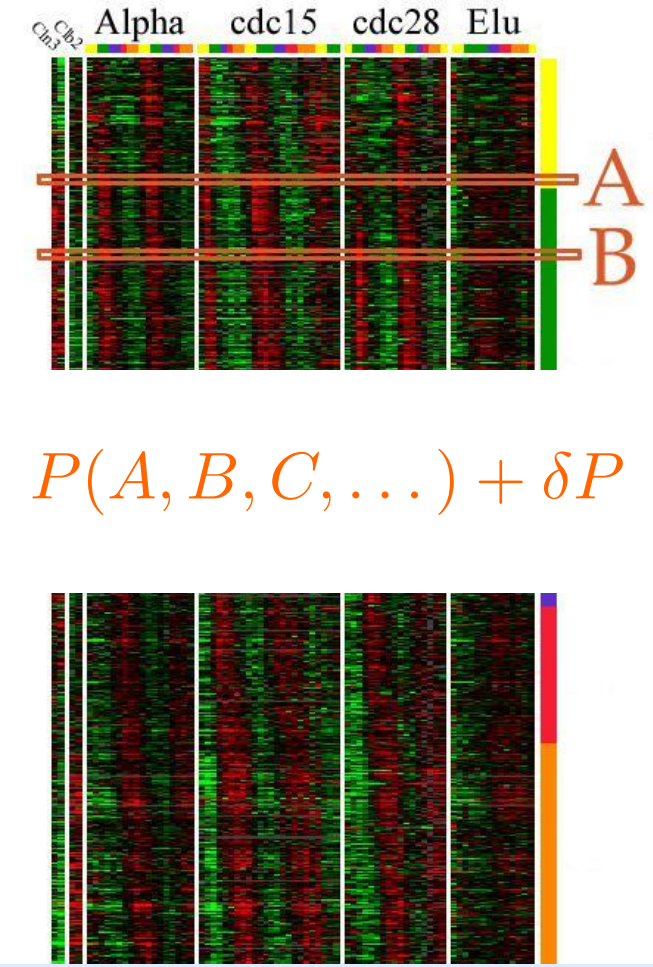
How can we estimate entropy (with error bars) from undersampled data?

A use: Inferring regulatory networks



(Spellman et al., 1998)

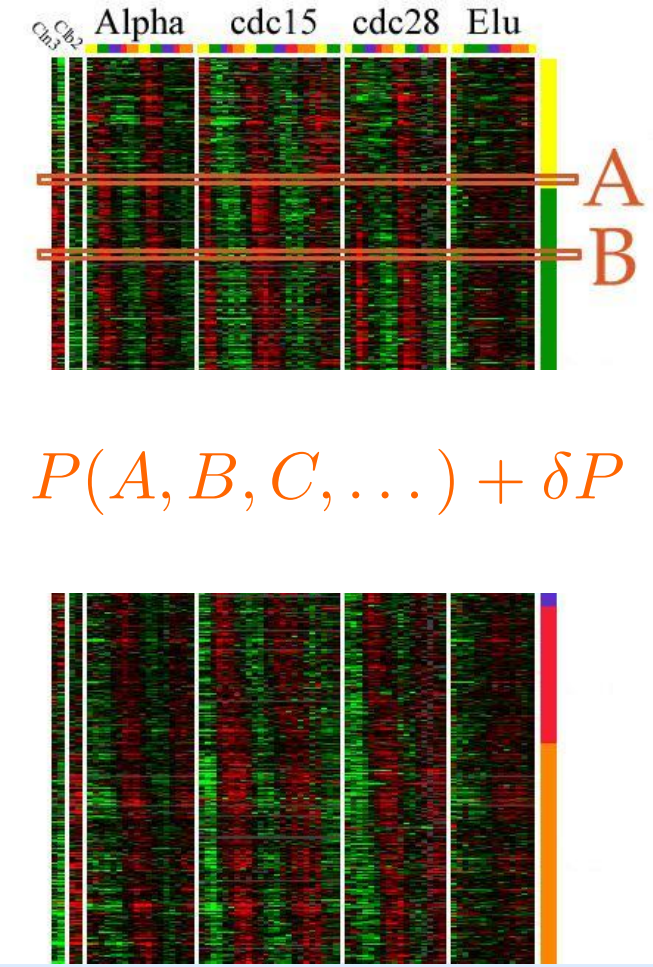
A use: Inferring regulatory networks



$$P(A, B, C, \ldots) + \delta P$$

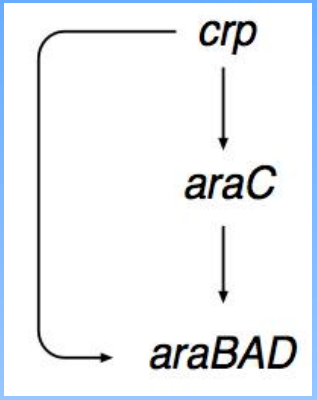
(Spellman et al., 1998)

A use: Inferring regulatory networks

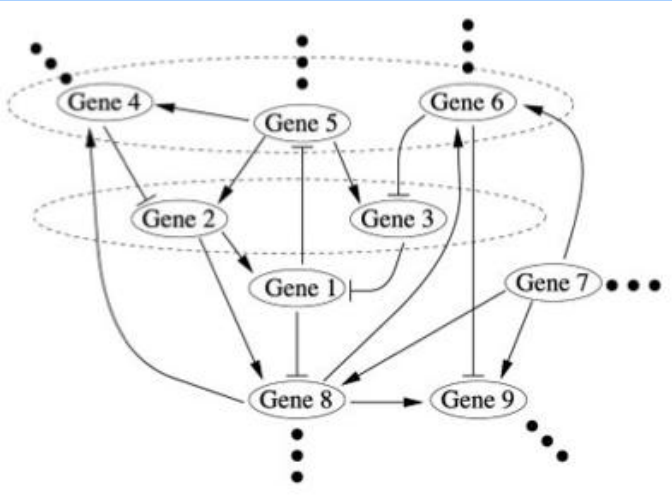


$$P(A, B, C, \ldots) + \delta P$$

(Spellman et al., 1998)

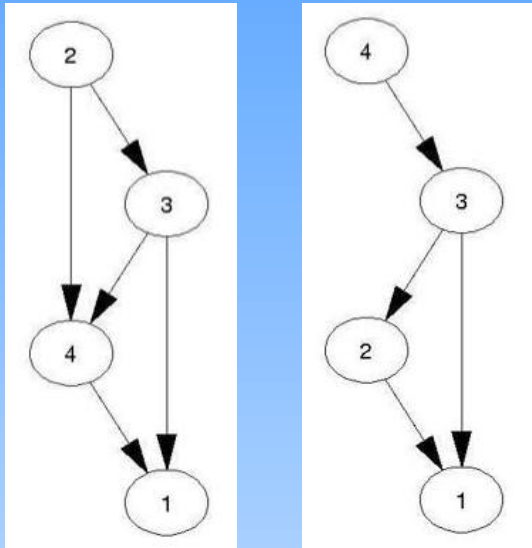


(Shen–Orr et al., 2002)



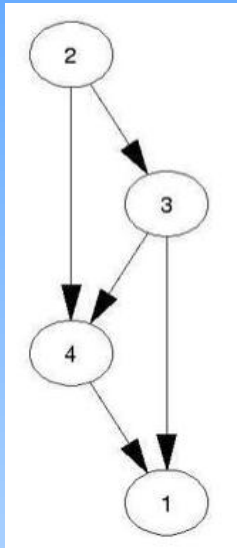
(Yeung et al., 2002)

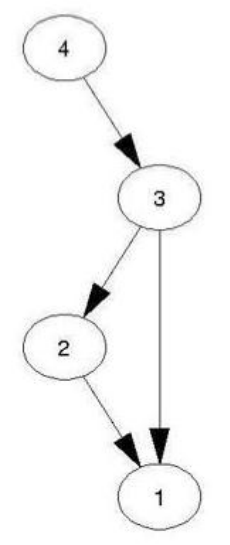
. . . but



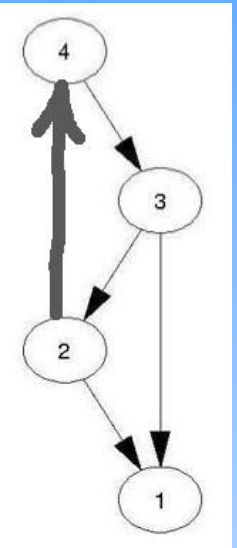
(Ziv et al., 2003)

. . . but

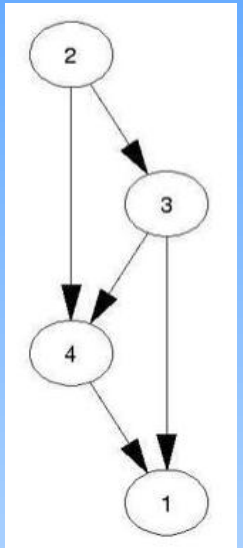


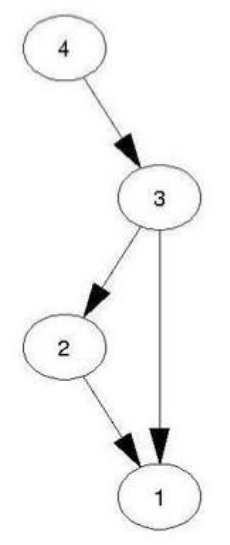


(Ziv et al., 2003)

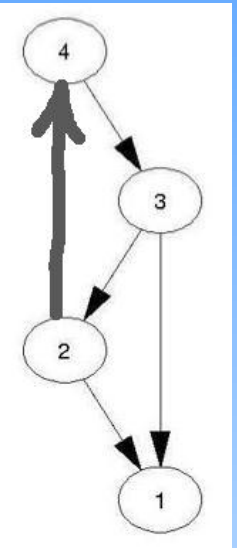


. . . but





(Ziv et al., 2003)



Does the data support the dependence between 2 and 4?

Solved by estimating various multiinformations (Nemenman, 2004).

A use: conserved elements search

$$\cdots x_{-1}x_0 \boxed{x_1 x_2 \cdots x_N} \underbrace{\cdots}_{D} \boxed{x_{N+D+1} x_{N+D+2} \cdots x_{N+D+M}} \cdots$$

x — aminoacid, nucleic acid, angles in the protein structure

A use: conserved elements search

$$\cdots x_{-1}x_0 \boxed{x_1 x_2 \cdots x_N} \underbrace{\cdots}_{D} \boxed{x_{N+D+1} x_{N+D+2} \cdots x_{N+D+M}} \cdots$$

x — aminoacid, nucleic acid, angles in the protein structure

Study predictability $S_D(M|N)$.

A use: conserved elements search

$$\cdots x_{-1}x_0 \boxed{x_1 x_2 \cdots x_N} \underbrace{\cdots}_{D} \boxed{x_{N+D+1} x_{N+D+2} \cdots x_{N+D+M}} \cdots$$

x — aminoacid, nucleic acid, angles in the protein structure

Study predictability $S_D(M|N)$.

- change in S indicates new region (coding–noncoding, helix–sheet,. . .)
- search for conserved sequences (motifs, new structural elements,. . .)
- protein length ~ 100, $N, M, D \sim 10$ — severe undersampling

A use: phylogeny and haplotyping

$$\begin{array}{ccc} & \overbrace{}^{N} & \\ \text{gccta} & \boxed{\text{accGt}} & \text{ggtccatatataaggaa} \\ \text{gccta} & \boxed{\text{accAt}} & \text{ggtccatatatatggac} \\ \text{accta} & \boxed{\text{accAt}} & \text{ggtcgatatataaggac} \end{array}$$

A use: phylogeny and haplotyping

	$\overbrace{\text{accGt}}^{N}$	
gccta	accGt	ggtccatatataaggaa
gccta	accAt	ggtccatatatatggac
accta	accAt	ggtcgatatataaggac

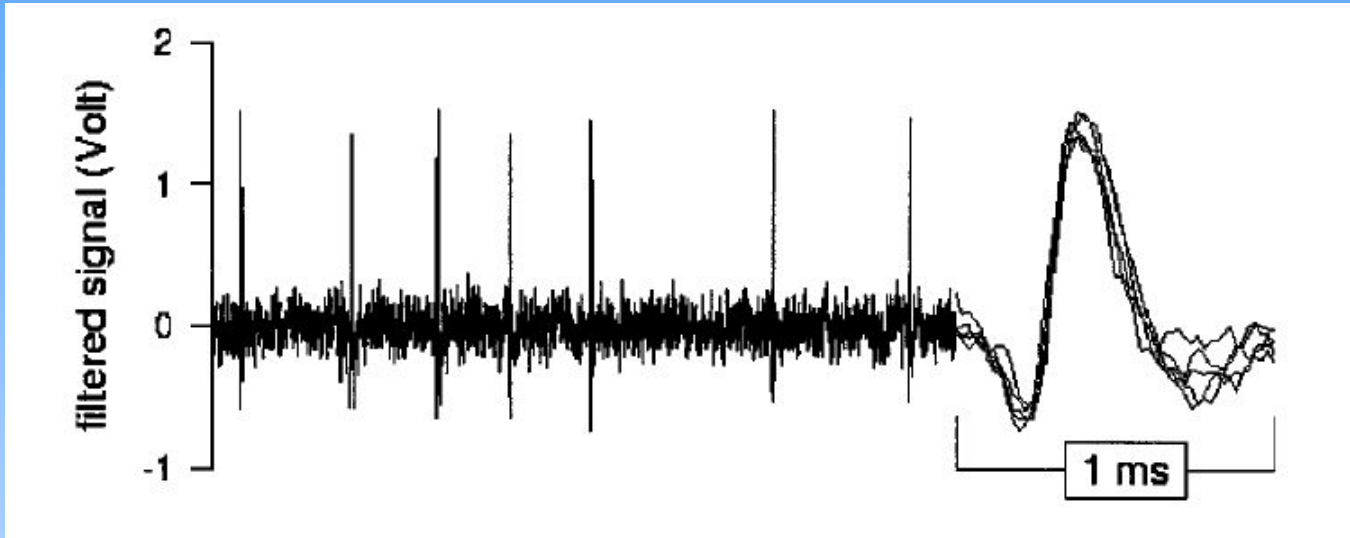
- length $10^6 \ldots 10^9$
- N up to 20
- < 100 repeats

Severe undersampling.

Other uses

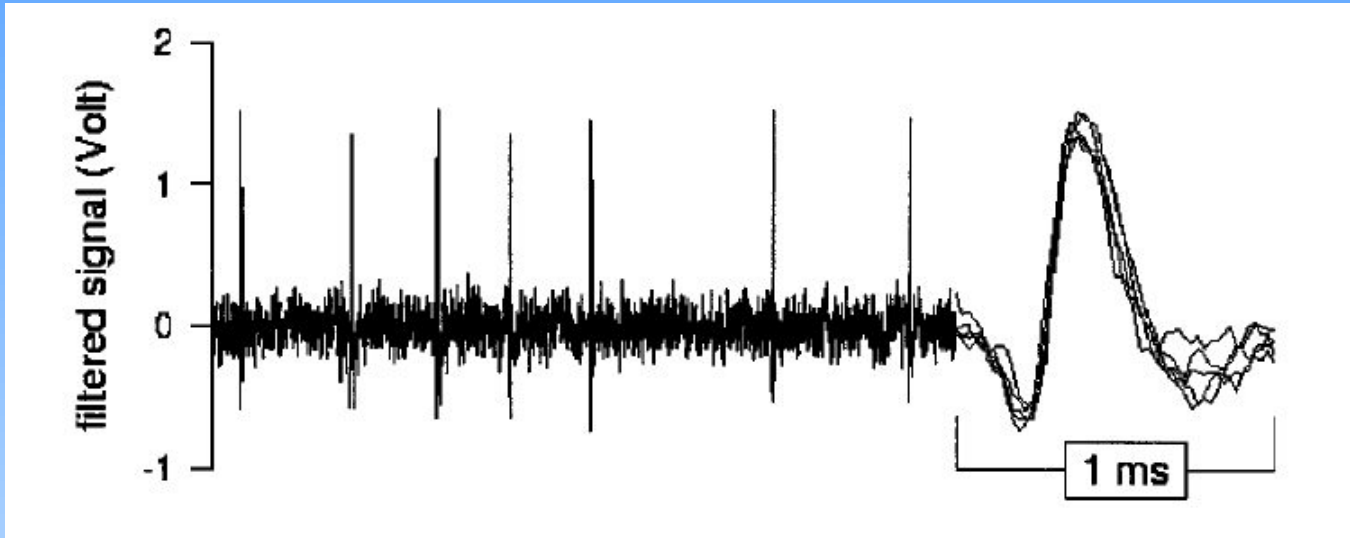
- information transmission in molecular cell signals
- cross–compression: comparative texts analysis (authorship of texts, similarity between languages,. . .)
- financial data and other prediction games
- dimensions of attractors in dynamical systems

Neurophysiological applications



(Strong et al., 1998)

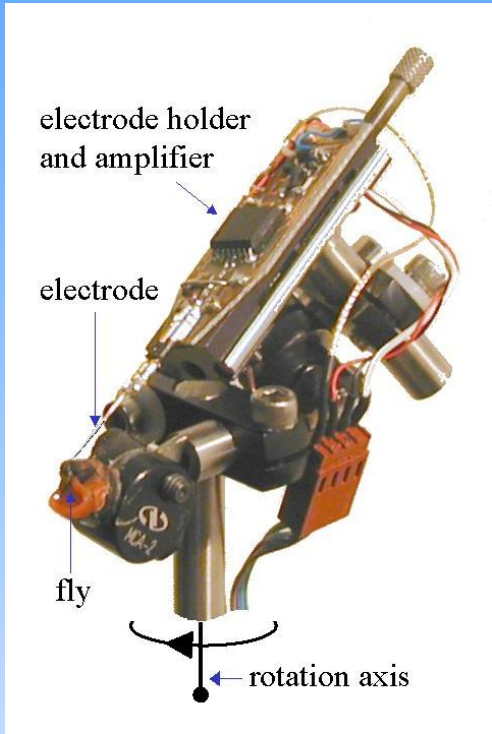
Neurophysiological applications



(Strong et al., 1998)

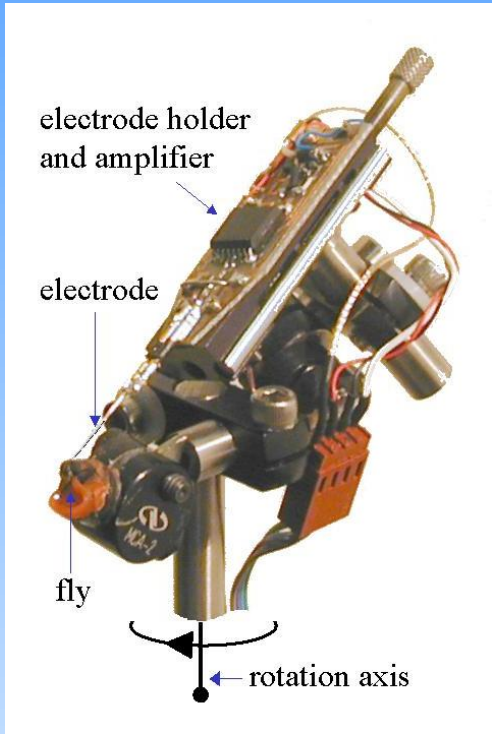
Neurons communicate by stereotypical pulses (spikes). Information is transmitted by spike rates and (possibly) precise positions of the spikes.

Experimental setup

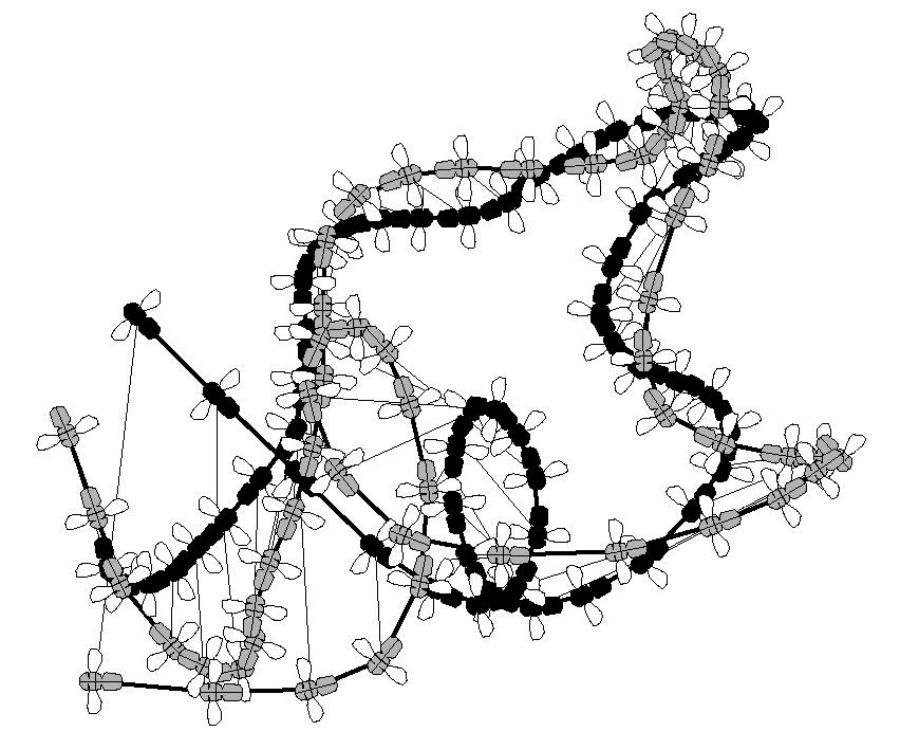


(Lewen, Bialek, and de Ruyter van Steveninck, 2001)

Experimental setup

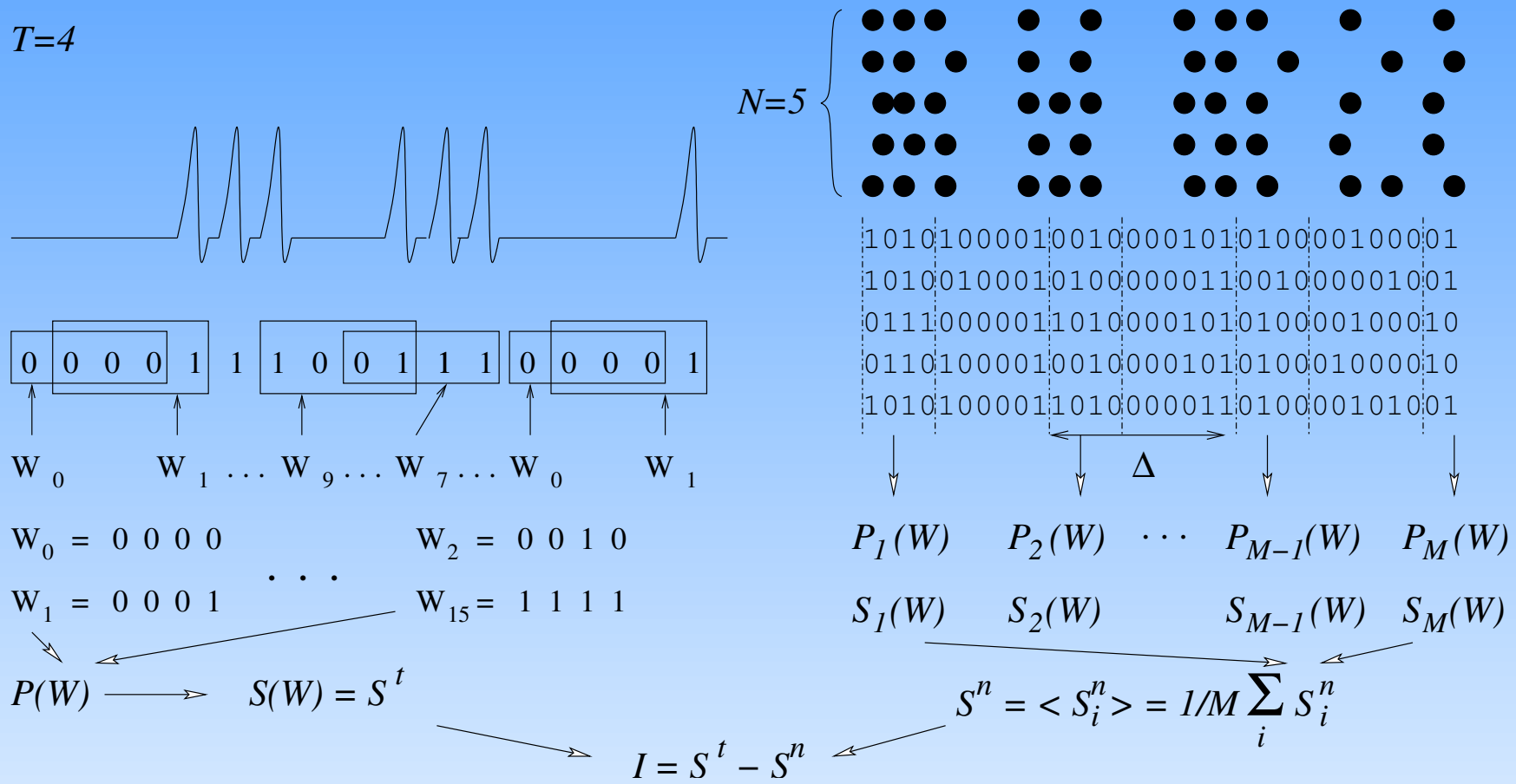


(Lewen, Bialek, and de Ruyter van Steveninck, 2001)



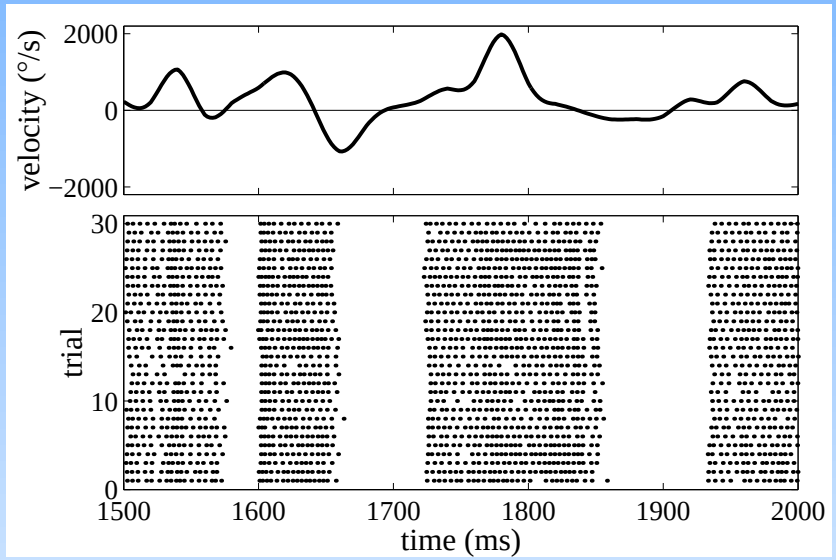
(Bialek and de Ruyter van Steveninck, 2002; Land and Collett, 1974)

Estimating information rate in spike trains



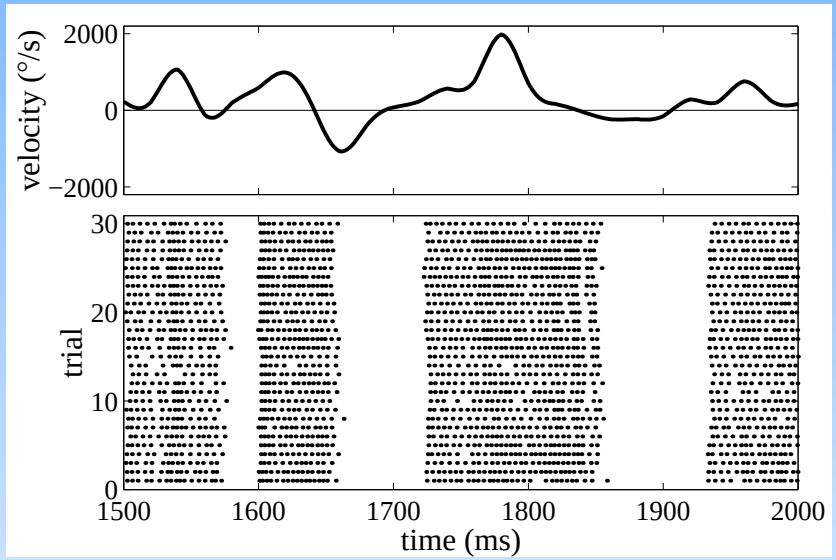
Recordings and problems

100–200 repeats of 5–10 s roller coasters rides



Recordings and problems

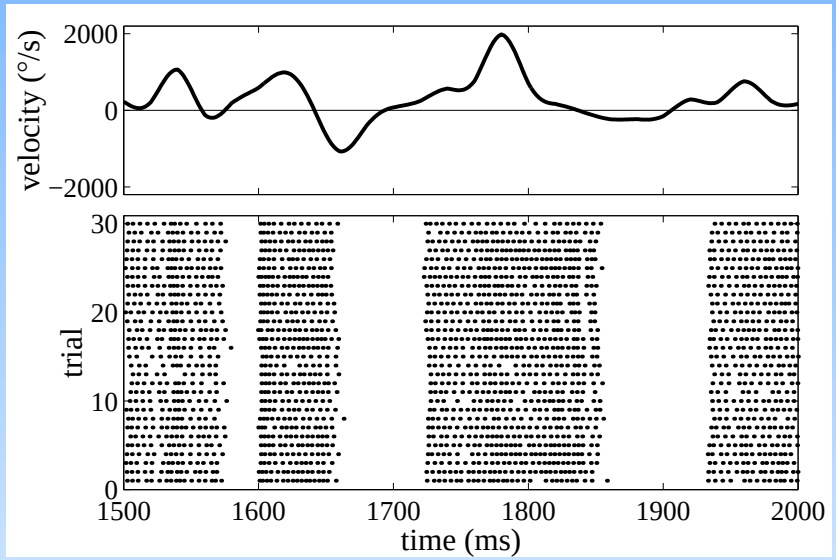
100–200 repeats of 5–10 s roller coasters rides



1. Need to take $T \to \infty$, $T > 30$ms for behavioral resolution.

Recordings and problems

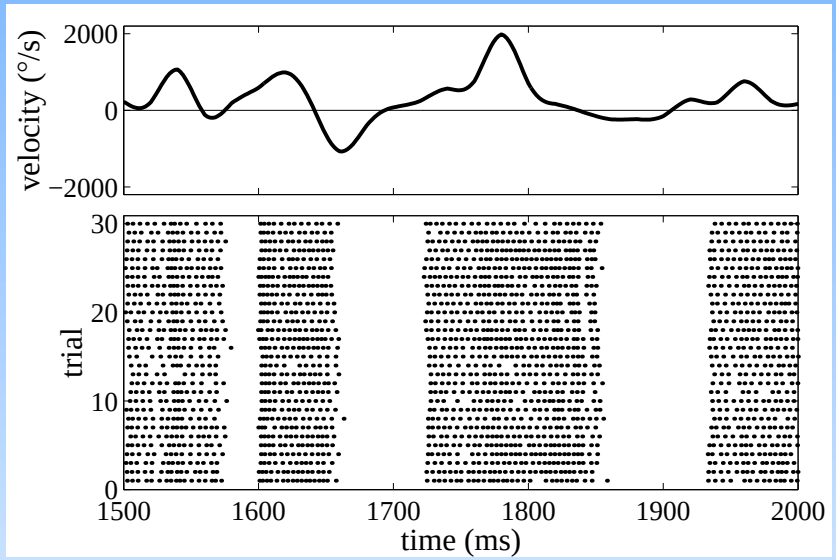
100–200 repeats of 5–10 s roller coasters rides



1. Need to take $T \to \infty$, $T > 30$ms for behavioral resolution.
2. Need to take $\tau \lesssim 1$ms.

Recordings and problems

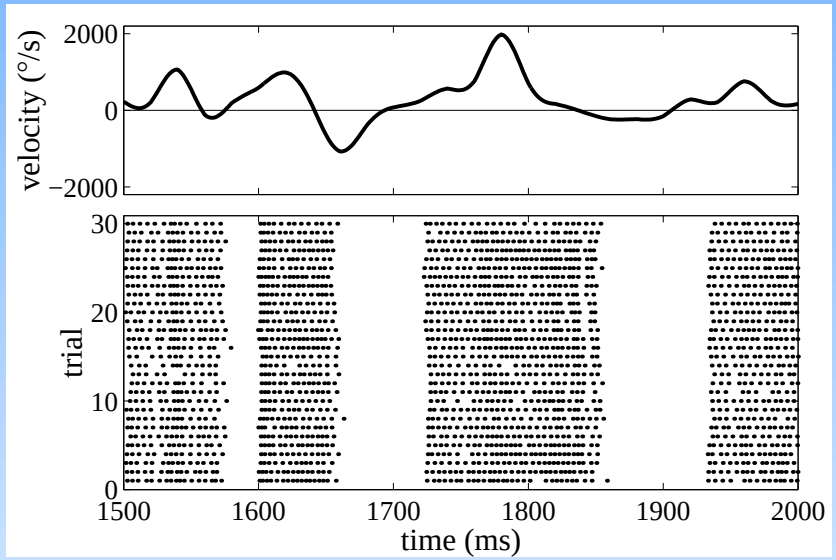
100–200 repeats of 5–10 s roller coasters rides



1. Need to take $T \to \infty$, $T > 30$ms for behavioral resolution.
2. Need to take $\tau \lesssim 1$ms.
3. Need to have $\Delta \approx 100$ms due to natural stimulus correlations.

Recordings and problems

100–200 repeats of 5–10 s roller coasters rides



1. Need to take $T \to \infty$, $T > 30$ms for behavioral resolution.
2. Need to take $\tau \lesssim 1$ms.
3. Need to have $\Delta \approx 100$ms due to natural stimulus correlations.

Need to estimate entropies of words of length ~ 40 from < 200 samples. Undersampled!

Why is this a difficult problem?

An asymptotically ($K/N \to 0$) easy problem.

But for $K \gg N$?

Why is this a difficult problem?

An asymptotically ($K/N \to 0$) easy problem.

But for $K \gg N$?

$$\lim_{p \to 0} \frac{p \log p}{p} = \infty$$

improbable events but large entropy

small errors in p but large errors in S

Why is this a difficult problem?

An asymptotically ($K/N \to 0$) easy problem.

But for $K \gg N$?

$$\lim_{p\to 0} \frac{p \log p}{p} = \infty$$

improbable events but large entropy

small errors in p but large errors in S

$$S_{\mathrm{ML}} \equiv -\hat{p}\log\hat{p} - (1-\hat{p})\log(1-\hat{p})$$ is convex

$$\Longrightarrow \quad E\,S_{\mathrm{ML}} < S(E\,\hat{p}) = S(p)$$

unknown negative bias, variance is much smaller

- no finite variance universally consistent unbiased entropy estimators for $N \ll K$, including string matching (Grassberger, 2003; Antos and Kontoyiannis, 2002; Wyner and Foster, 2003)
- no universally consistent multiplicative estimator (Batu et al., 2002)
- universal consistent entropy estimation is possible only for $K/N \to \mathrm{const}$, $K \to \infty$ (Paninski, 2003)
- bias–variance balanced estimators built for $K/N \to \mathrm{const}$, $K \to \infty$ (Paninski, 2003; Grassberger, 1989, 2003)

Correcting for bias

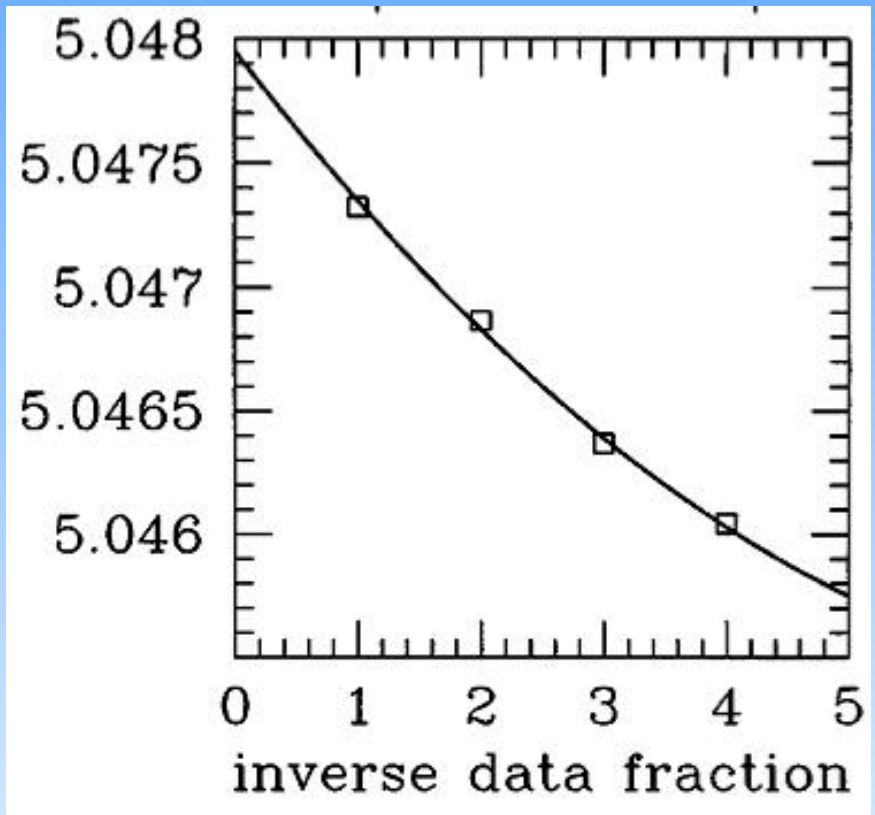
Correcting for bias as a power series in $2^S/N$:

- replica–averaging over samples (Panzeri and Treves, 1996)
- least bias + variance (Paninski, 2003; Grassberger, 2003)

Correcting for bias

Correcting for bias as a power series in $2^S/N$:

- replica–averaging over samples (Panzeri and Treves, 1996)
- least bias + variance (Paninski, 2003; Grassberger, 2003)
- empirical evaluation of bias (Strong et al., 1998); so far the best

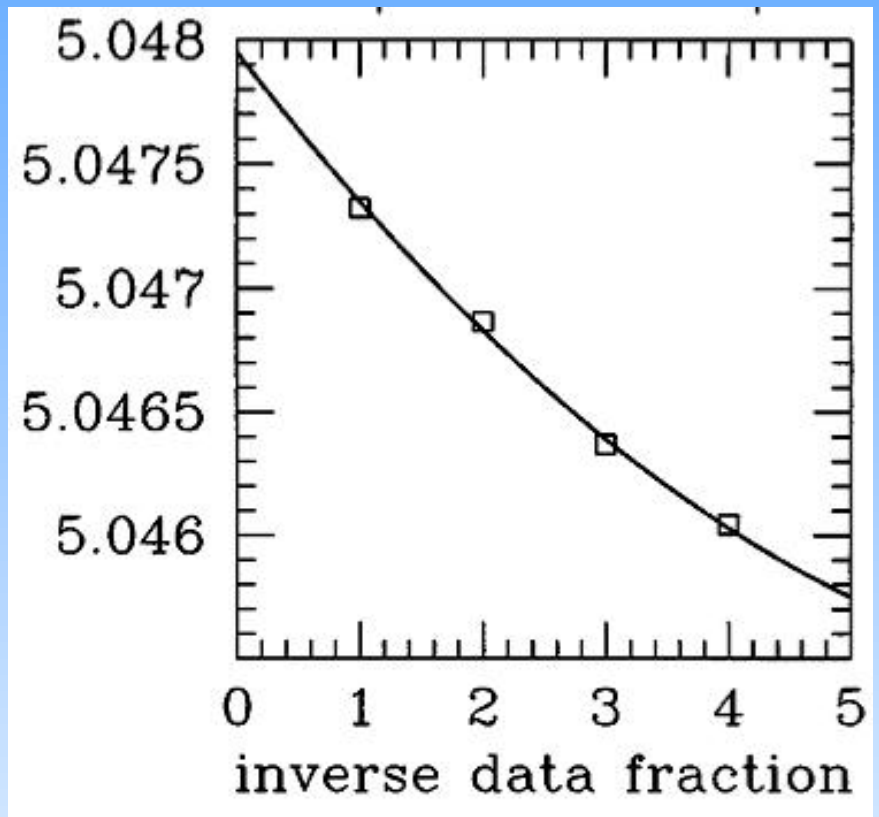


(Strong et al., 1998)

Correcting for bias

Correcting for bias as a power series in $2^S/N$:

- replica–averaging over samples (Panzeri and Treves, 1996)
- least bias + variance (Paninski, 2003; Grassberger, 2003)
- empirical evaluation of bias (Strong et al., 1998); so far the best
- ALL WORK FOR $2^S \ll N \ll K$



(Strong et al., 1998)

The hope

Ma's (1981) argument, the birthday problem.

For uniform K–bin distribution: for $N_c \sim \sqrt{K}$, probability of coincidences ~ 1.

$$S = \log K \approx \log N_c^2 = 2 \log N_c$$

Works in nonasymptotic regime $N \sim \sqrt{2^S}$. Better than it should!

$\delta S \sim 1$, but this is all we often need.

Extensions?

For Ma–type ideas to work for nonuniform cases

- forget universality, make assumptions about distributions
- do not learn distributions, learn entropies
- equate smoothness and long tails as high entropy (rapidly decaying Zipf plot)

Learning with nearly uniform priors

(ultra–local, Dirichlet priors)

$\{q_i\}$, $i = 1 \ldots K$:

$$\mathcal{P}_\beta(\{q_i\}) = \frac{1}{Z(\beta)} \delta\left(1 - \sum_{i=1}^{K} q_i\right) \prod_{i=1}^{K} q_i^{\beta-1}$$

Learning with nearly uniform priors

(ultra–local, Dirichlet priors)

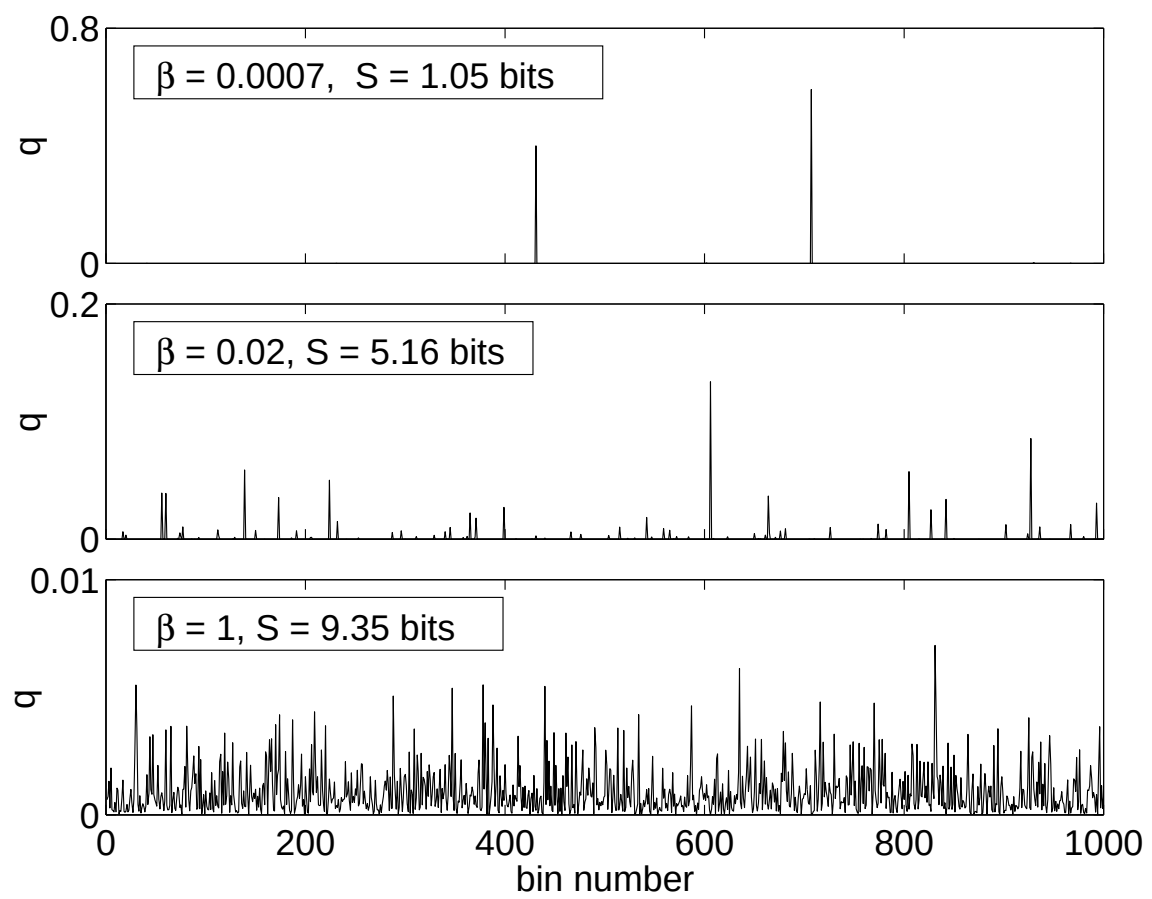
$\{q_i\}$, $i = 1 \ldots K$:

$$\mathcal{P}_\beta(\{q_i\}) = \frac{1}{Z(\beta)} \delta\left(1 - \sum_{i=1}^{K} q_i\right) \prod_{i=1}^{K} q_i^{\beta-1}$$

Some common choices:

Maximum likelihood	$\beta \to 0$
Laplace's successor rule	$\beta = 1$
Krichevsky–Trofimov (Jeffreys) estimator	$\beta = 1/2$
Schurmann–Grassberger estimator	$\beta = 1/K$

Typical distributions for $K = 1000, S \approx 9.97$



Typical rank–ordered plots

$$q_i \approx 1 - \left[\frac{\beta B(\beta, \kappa - \beta)(K-1)\, i}{K}\right]^{1/(\kappa-\beta)}, \quad i \ll K\,,$$

$$q_i \approx \left[\frac{\beta B(\beta, \kappa - \beta)(K-i+1)}{K}\right]^{1/\beta}, \quad K - i + 1 \ll K$$

Usually only the first regime is observed.

Gets to zero at finite i.

Faster decaying – too rough.

Slower decaying – too smooth.

Bayesian inference with Dirichlet priors

$$
\begin{aligned}
P_\beta(\{q_i\}|\{n_i\}) &= \frac{P(\{n_i\}|\{q_i\})\mathcal{P}_\beta(\{q_i\})}{P_\beta(\{n_i\})} \\
P(\{n_i\}|\{q_i\}) &= \prod_{i=1}^{K}(q_i)^{n_i} \\
\langle q_i \rangle_\beta &= \frac{n_i+\beta}{N+K\beta}
\end{aligned}
$$

Bayesian inference with Dirichlet priors

$$
\begin{aligned}
P_\beta(\{q_i\}|\{n_i\}) &= \frac{P(\{n_i\}|\{q_i\})\mathcal{P}_\beta(\{q_i\})}{P_\beta(\{n_i\})} \\
P(\{n_i\}|\{q_i\}) &= \prod_{i=1}^{K}(q_i)^{n_i} \\
\langle q_i \rangle_\beta &= \frac{n_i+\beta}{N+K\beta} \\
\langle S \rangle_\beta &= \text{known (Wolpert and Wolf, 1995)} \\
\langle \delta^2 S \rangle_\beta &= \text{known}
\end{aligned}
$$

Bayesian inference with Dirichlet priors

$$
\begin{aligned}
P_\beta(\{q_i\}|\{n_i\}) &= \frac{P(\{n_i\}|\{q_i\})\mathcal{P}_\beta(\{q_i\})}{P_\beta(\{n_i\})} \\
P(\{n_i\}|\{q_i\}) &= \prod_{i=1}^{K}(q_i)^{n_i} \\
\langle q_i \rangle_\beta &= \frac{n_i+\beta}{N+K\beta} \\
\langle S \rangle_\beta &= \text{known (Wolpert and Wolf, 1995)} \\
\langle \delta^2 S \rangle_\beta &= \text{known}
\end{aligned}
$$

Equal pseudocounts added to each bin.

Bayesian inference with Dirichlet priors

$$
\begin{aligned}
P_\beta(\{q_i\}|\{n_i\}) &= \frac{P(\{n_i\}|\{q_i\})\mathcal{P}_\beta(\{q_i\})}{P_\beta(\{n_i\})} \\
P(\{n_i\}|\{q_i\}) &= \prod_{i=1}^{K}(q_i)^{n_i} \\
\langle q_i \rangle_\beta &= \frac{n_i+\beta}{N+K\beta} \\
\langle S \rangle_\beta &= \text{known (Wolpert and Wolf, 1995)} \\
\langle \delta^2 S \rangle_\beta &= \text{known}
\end{aligned}
$$

Equal pseudocounts added to each bin.

Larger β means less sensitivity to data, thus more smoothing.

A problem: A priori entropy expectation

$$\mathcal{P}_\beta(S) = \int dq_1 dq_2 \cdots dq_K \, P_\beta(\{q_i\}) \, \delta \left[S + \sum_{i=1}^{K} q_i \log_2 q_i \right]$$

A problem: A priori entropy expectation

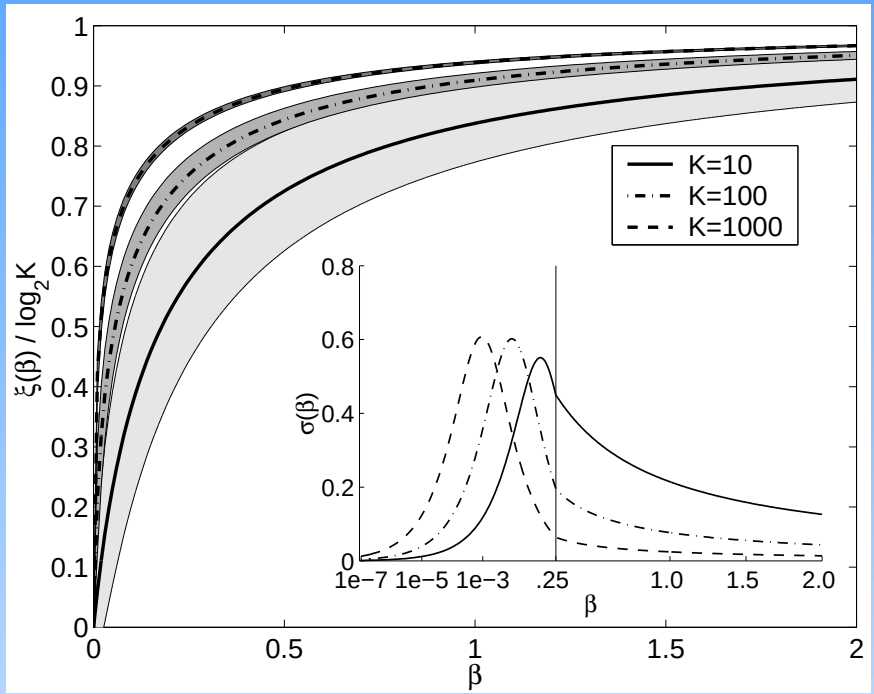
$$\mathcal{P}_\beta(S) = \int dq_1 dq_2 \cdots dq_K \, P_\beta(\{q_i\}) \, \delta \left[S + \sum_{i=1}^{K} q_i \log_2 q_i \right]$$

$$\xi(\beta) \equiv \langle \, S|_{N=0} \, \rangle_\beta = \psi_0(K\beta + 1) - \psi_0(\beta + 1)$$

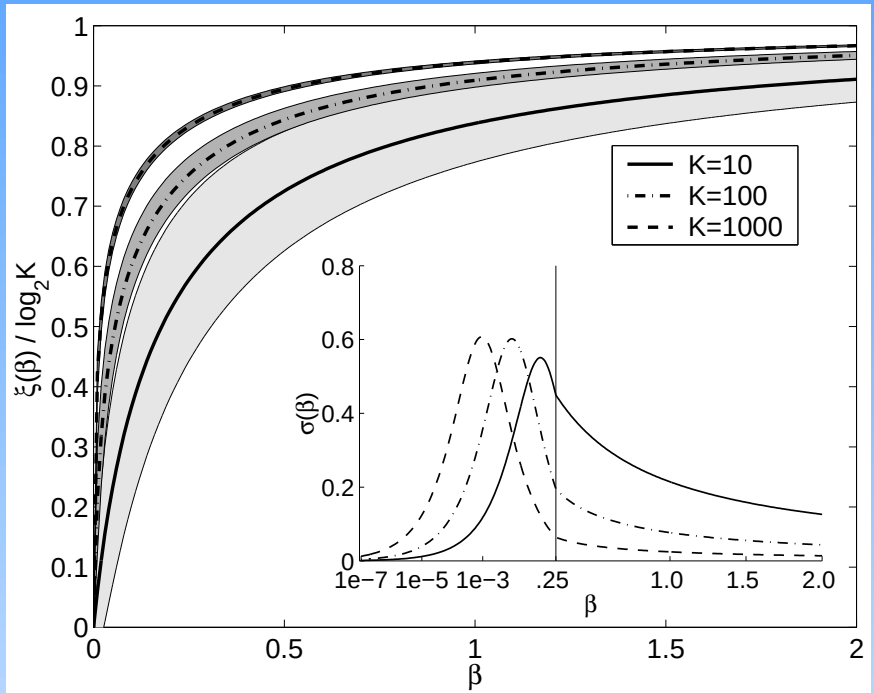
$$\sigma^2(\beta) \equiv \langle \, (\delta S)^2 \big|_{N=0} \rangle_\beta = \frac{\beta + 1}{K\beta + 1} \psi_1(\beta + 1) - \psi_1(K\beta + 1)$$

$\psi_m(x) = (d/dx)^{m+1} \log_2 \Gamma(x)$ –the polygamma function

The problem: Analysis

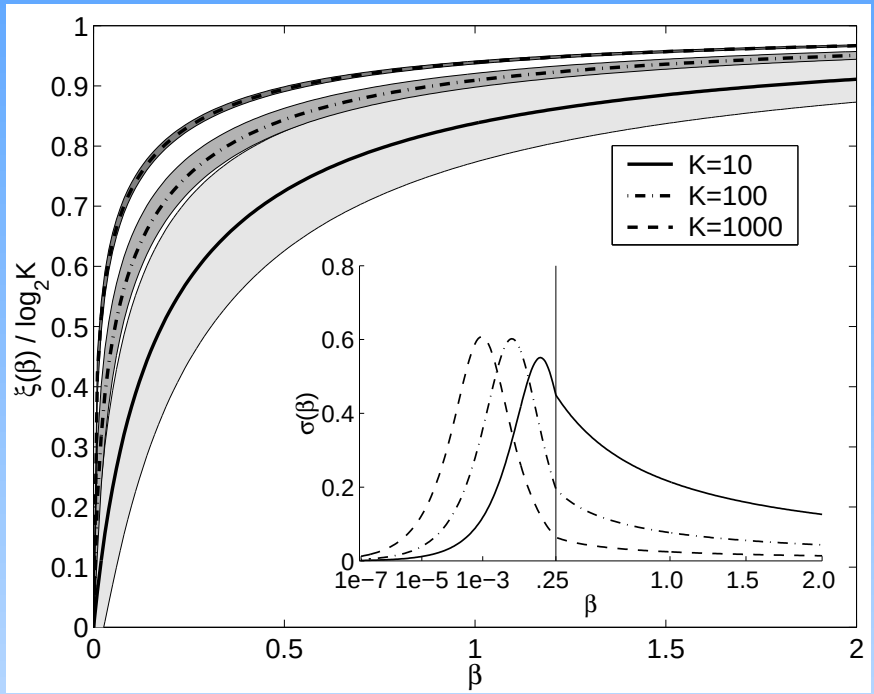


The problem: Analysis



- Because of the Jacobian of $\{q_i\} \to S$, a priori distribution of entropy is strongly peaked.
- *Narrow* peak: $\sigma(\beta) \propto 1/\sqrt{K\beta}$, $\max \sigma(\beta) = 0.61\,\text{bits}$.
- As β varies from 0 to ∞, the peak smoothly moves from 0 to $\log_2 K$. For $\beta \sim 1$, $\xi(\beta) = \log_2 K - O(K^0)$.

The problem: Analysis



- Because of the Jacobian of $\{q_i\} \to S$, a priori distribution of entropy is strongly peaked.
- *Narrow* peak: $\sigma(\beta) \propto 1/\sqrt{K\beta}$, $\max \sigma(\beta) = 0.61\,\text{bits}$.
- As β varies from 0 to ∞, the peak smoothly moves from 0 to $\log_2 K$. For $\beta \sim 1$, $\xi(\beta) = \log_2 K - O(K^0)$.

- No a priori way to specify β.
- Choosing β fixes allowed "shapes" of $\{q_i\}$, and defines the a priori expectation of entropy.
- Such expectation dominates data until $N \gg K\beta$.
- All common estimators are bad for learning entropies.

Removing the entropy bias at the source

Need such $\mathcal{P}(\{q_i\})$ that $\mathcal{P}(S[q_i])$ is (almost) uniform.

Removing the entropy bias at the source

Need such $\mathcal{P}(\{q_i\})$ that $\mathcal{P}(S[q_i])$ is (almost) uniform.

Our options:

1. $\mathcal{P}_\beta^{\text{flat}}(\{q_i\}) = \frac{\mathcal{P}_\beta(\{q_i\})}{\mathcal{P}_\beta(S[q_i])}$.

Removing the entropy bias at the source

Need such $\mathcal{P}(\{q_i\})$ that $\mathcal{P}(S[q_i])$ is (almost) uniform.

Our options:

1. $\mathcal{P}_\beta^{\text{flat}}(\{q_i\}) = \frac{\mathcal{P}_\beta(\{q_i\})}{\mathcal{P}_\beta(S[q_i])}$. Difficult.

Removing the entropy bias at the source

Need such $\mathcal{P}(\{q_i\})$ that $\mathcal{P}(S[q_i])$ is (almost) uniform.

Our options:

1. $\mathcal{P}_\beta^{\text{flat}}(\{q_i\}) = \frac{\mathcal{P}_\beta(\{q_i\})}{\mathcal{P}_\beta(S[q_i])}$. Difficult.
2. $\mathcal{P}(S) \sim 1 = \int \delta(S - \xi) d\xi$.

Removing the entropy bias at the source

Need such $\mathcal{P}(\{q_i\})$ that $\mathcal{P}(S[q_i])$ is (almost) uniform.

Our options:

1. $\mathcal{P}_\beta^{\text{flat}}(\{q_i\}) = \frac{\mathcal{P}_\beta(\{q_i\})}{\mathcal{P}_\beta(S[q_i])}$. Difficult.
2. $\mathcal{P}(S) \sim 1 = \int \delta(S - \xi) d\xi$. Easy: $\mathcal{P}_\beta(S)$ is almost a δ-function!

Solution

Average over β — infinite Dirichlet mixtures.

$$\mathcal{P}(\{q_i\};\beta) = \frac{1}{Z}\delta\left(1-\sum_{i=1}^{K} q_i\right)\prod_{i=1}^{K} q_i^{\beta-1} \qquad \frac{d\xi(\beta)}{d\beta} \qquad \mathcal{P}(\xi(\beta))$$

Solution

Average over β — infinite Dirichlet mixtures.

$\beta \longrightarrow \xi$ Jacobian entropy prior

$$\mathcal{P}(\{q_i\};\beta) = \frac{1}{Z}\,\delta\left(1-\sum_{i=1}^{K} q_i\right)\prod_{i=1}^{K} q_i^{\beta-1} \qquad \frac{d\xi(\beta)}{d\beta} \qquad \mathcal{P}(\xi(\beta))$$

Solution

Average over β — infinite Dirichlet mixtures.

$\beta \longrightarrow \xi$ Jacobian entropy prior

$$\mathcal{P}(\{q_i\};\beta) = \frac{1}{Z}\,\delta\left(1-\sum_{i=1}^{K} q_i\right)\prod_{i=1}^{K} q_i^{\beta-1} \qquad \frac{d\xi(\beta)}{d\beta} \qquad \mathcal{P}(\xi(\beta))$$

$$\widehat{S^m} = \frac{\int d\xi\,\rho(\xi,\{n_i\})\langle\, S^m[n_i]\,\rangle_{\beta(\xi)}}{\int d\xi\,\rho(\xi,[n_i])}$$

$$\rho(\xi,[n_i]) = \mathcal{P}\,(\xi)\,\frac{\Gamma(K\beta(\xi))}{\Gamma(N+K\beta(\xi))}\prod_{i=1}^{K}\frac{\Gamma(n_i+\beta(\xi))}{\Gamma(\beta(\xi))}\,.$$

Solution

Average over β — infinite Dirichlet mixtures.

$\beta \longrightarrow \xi$ Jacobian entropy prior

$$\mathcal{P}(\{q_i\};\beta) = \frac{1}{Z}\,\delta\left(1-\sum_{i=1}^{K} q_i\right)\prod_{i=1}^{K} q_i^{\beta-1} \qquad \frac{d\xi(\beta)}{d\beta} \qquad \mathcal{P}(\xi(\beta))$$

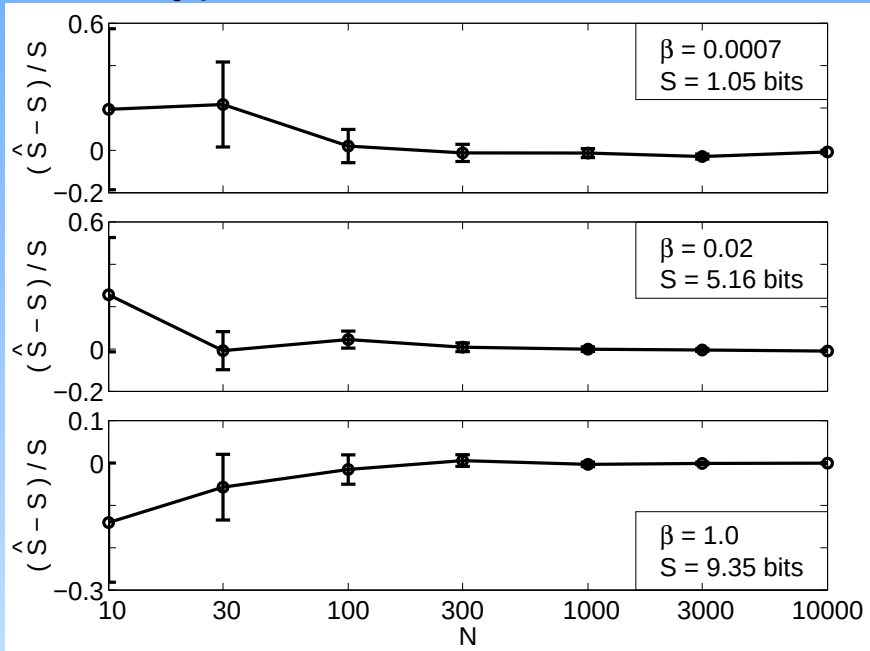
$$\widehat{S^m} = \frac{\int d\xi\,\rho(\xi,\{n_i\})\langle\, S^m[n_i]\,\rangle_{\beta(\xi)}}{\int d\xi\,\rho(\xi,[n_i])}$$

$$\rho(\xi,[n_i]) = \mathcal{P}\,(\xi)\,\frac{\Gamma(K\beta(\xi))}{\Gamma(N+K\beta(\xi))}\prod_{i=1}^{K}\frac{\Gamma(n_i+\beta(\xi))}{\Gamma(\beta(\xi))}\,.$$

- Smaller β — larger allowed volume in the space of $\{q_i\}$. Thus averaging over β is *Bayesian model selection*.
- $\langle\delta^2 S\rangle$ is dominated by $\langle\delta^2\beta\rangle$ (not $\langle\delta^2 S\rangle_\beta$) which is small if a particular β (model) dominates (is "selected")

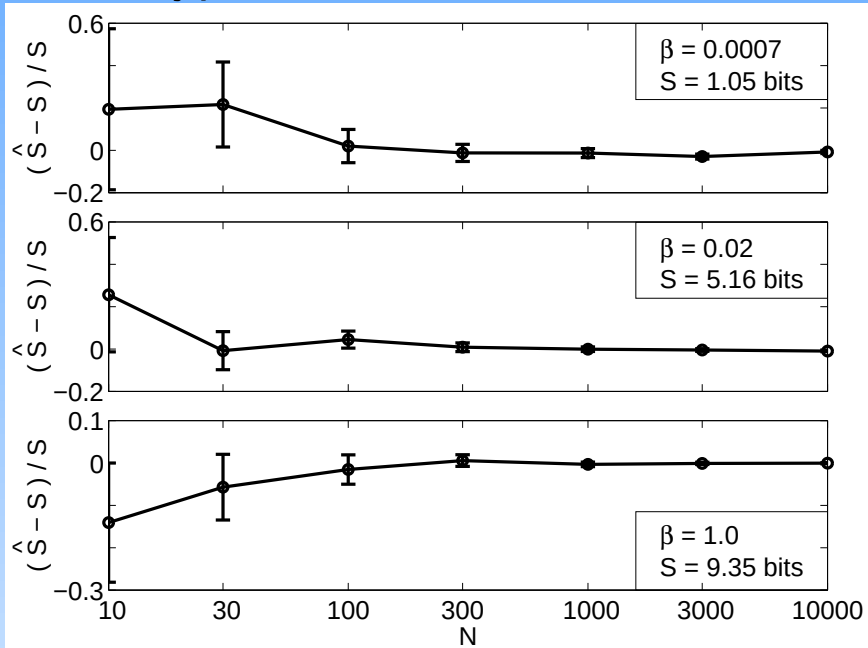
First attempts to estimate entropy

Typical distributions

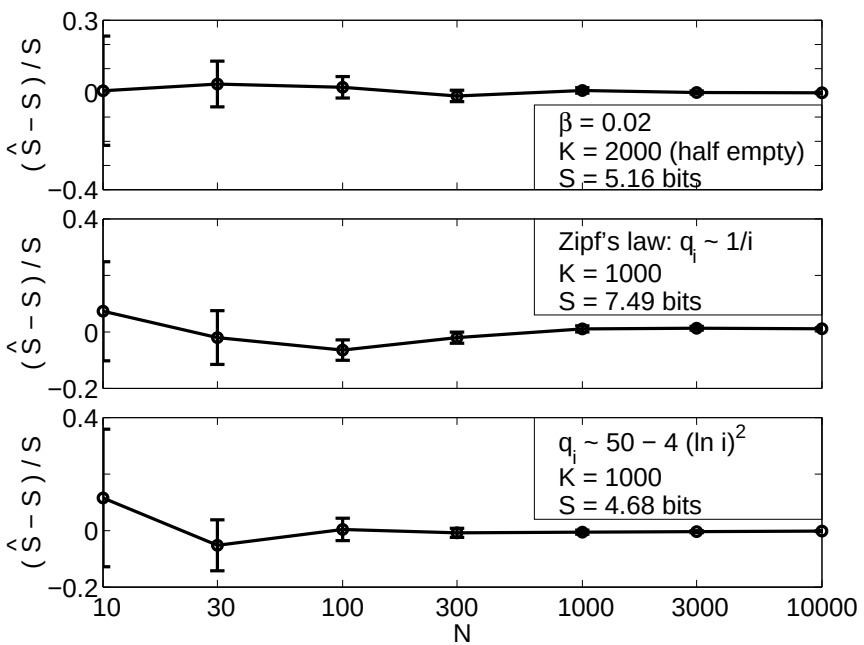


First attempts to estimate entropy

Typical distributions

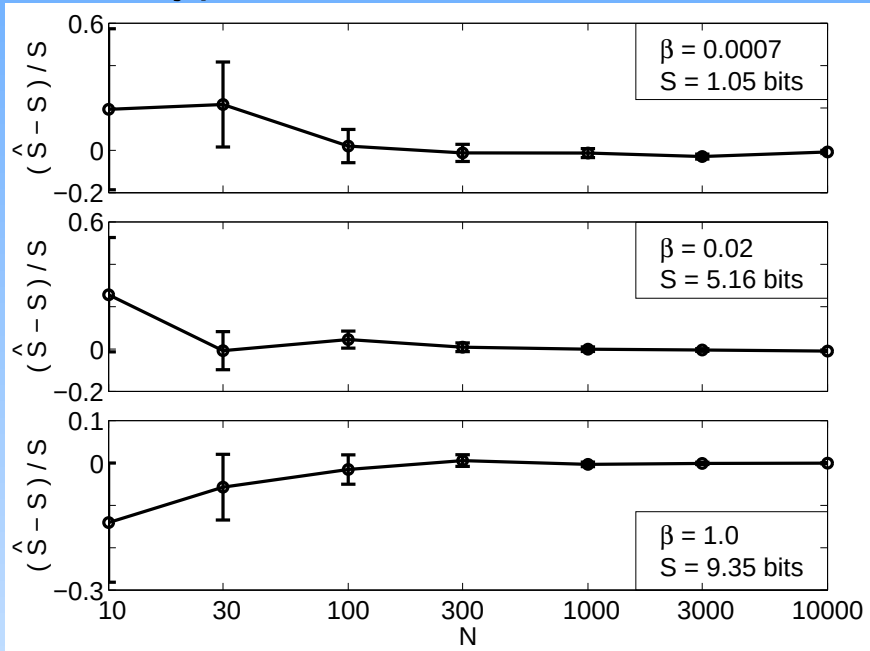


Atypical distributions

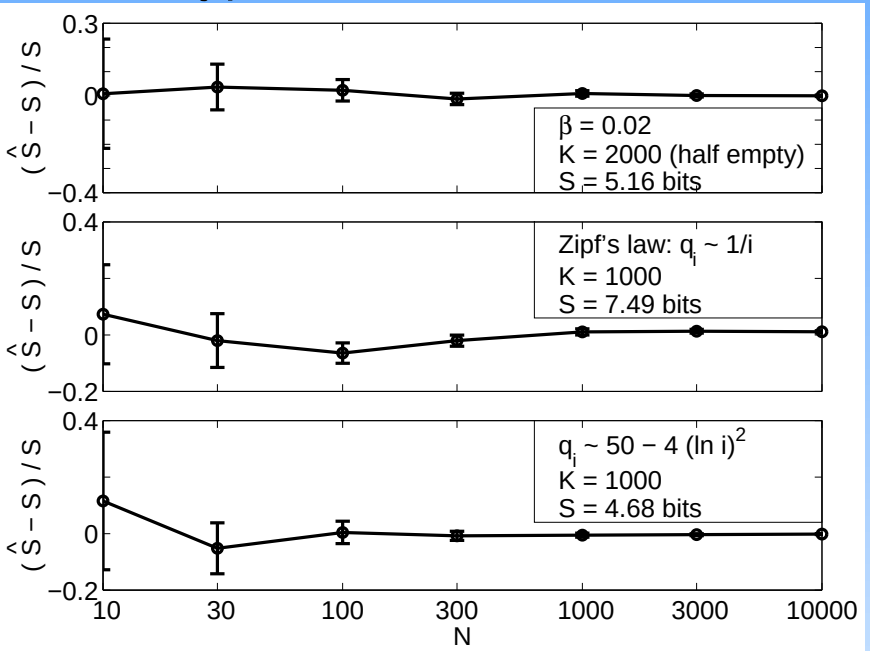


First attempts to estimate entropy

Typical distributions



Atypical distributions



Supports understanding that smoothness = speed of decay of Zipf plot.

Estimating entropy: first observations

- Relative error $\sim 10\%$ at N as low as 30 for $K = 1000$.
- Reliable estimation of posterior error.
- *Little bias*. Exception: too smooth distributions.
- Key point: *learn entropies directly without finding* $\{q_i\}$*!*
- The dominant β stabilizes for typical distributions; drifts down (to complex models) for rough ones and up (to simpler models) for too smooth cases.

Asymptotics

$K \gg 1$, $\Delta \equiv N - K_{\text{counts}>0} \gg 1$

- saddle point works
- $\left.\frac{\partial^2(-\log\rho)}{\partial\xi^2}\right|_{\xi(\beta^*)} = \Delta + NO([\Delta/N]^2)$

Asymptotics

$K \gg 1,\ \Delta \equiv N - K_{\text{counts}>0} \gg 1$

- saddle point works
- $\left.\frac{\partial^2(-\log\rho)}{\partial\xi^2}\right|_{\xi(\beta^*)} = \Delta + NO([\Delta/N]^2)$

$K, N \gg 1,\ \Delta \sim 1$

- $\widehat{S} \approx (C_\gamma - \ln 2) + 2\ln N - \psi_0(\Delta) + O(\frac{1}{N}, \frac{1}{K})$
- $\widehat{(\delta S)^2} \approx \psi_1(\Delta) + O(\frac{1}{N}, \frac{1}{K})$

Asymptotics

$K \gg 1,\ \Delta \equiv N - K_{\text{counts}>0} \gg 1$

- saddle point works
- $\left.\frac{\partial^2(-\log\rho)}{\partial\xi^2}\right|_{\xi(\beta^*)} = \Delta + NO([\Delta/N]^2)$

$K, N \gg 1,\ \Delta \sim 1$

- $\widehat{S} \approx (C_\gamma - \ln 2) + 2\ln N - \psi_0(\Delta) + O(\frac{1}{N}, \frac{1}{K})$
- $\widehat{(\delta S)^2} \approx \psi_1(\Delta) + O(\frac{1}{N}, \frac{1}{K})$

Remember Ma's estimate!

Estimator: Properties

- Uniform prior on S *and* Bayesian model selection

Estimator: Properties

- Uniform prior on S *and* Bayesian model selection
- K can be infinite
- Works for $\Delta \ll N$ if distribution is not atypically smooth.
- Δ matters, not K or N.
- The estimator is consistent.
- Thus correct if self-consistent for subsamples.
- When works, works for $N \sim \sqrt{2^S}$.

Estimator: Synthetic test

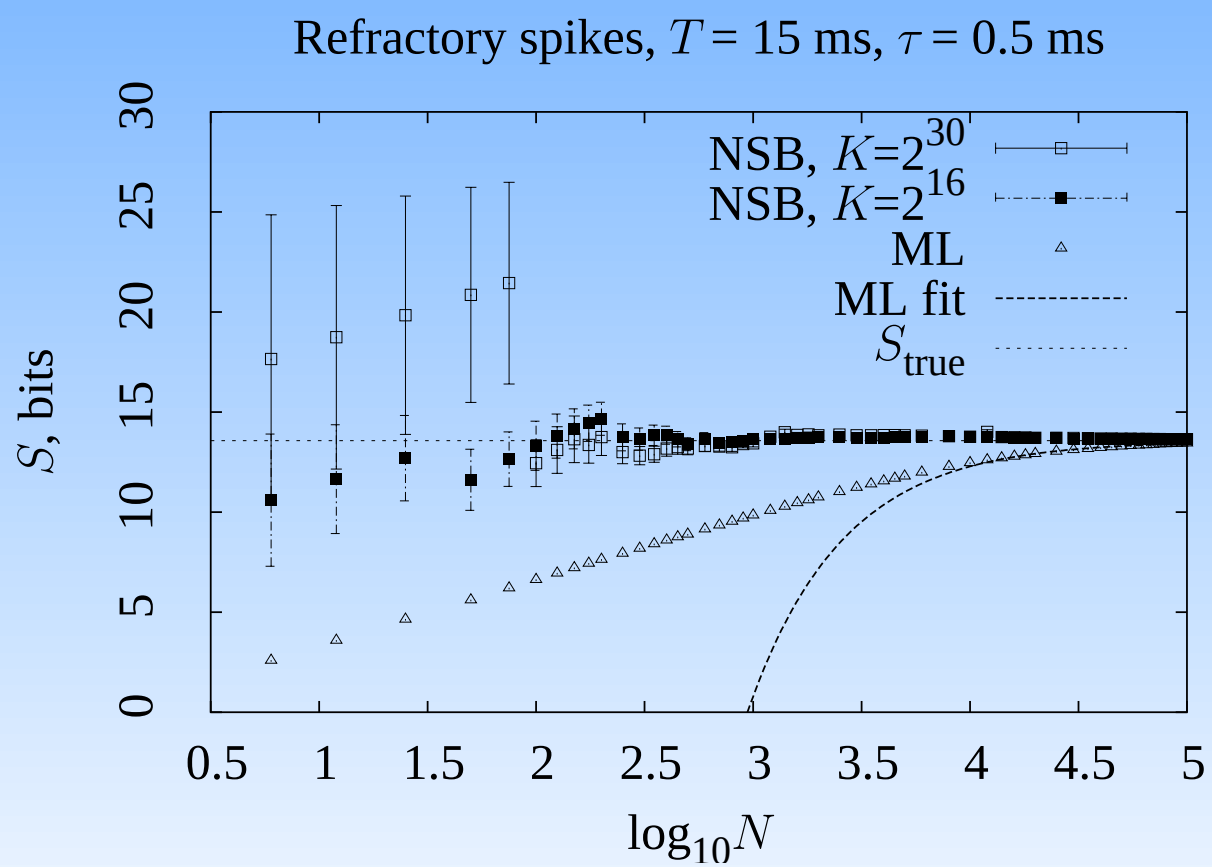
Refractory Poisson process: $r = 0.26\text{ms}^{-1}$, $R = 1.8\text{ms}$, $T = 15\text{ms}$, $\tau = 0.5\text{ms}$.

Estimator: Synthetic test

Refractory Poisson process: $r = 0.26\text{ms}^{-1}$, $R = 1.8\text{ms}$, $T = 15\text{ms}$, $\tau = 0.5\text{ms}$.
$K = 2^{30}$, $K_{\text{ref}} < 2^{16}$, $S = 13.57\text{bits}$.

Estimator: Synthetic test

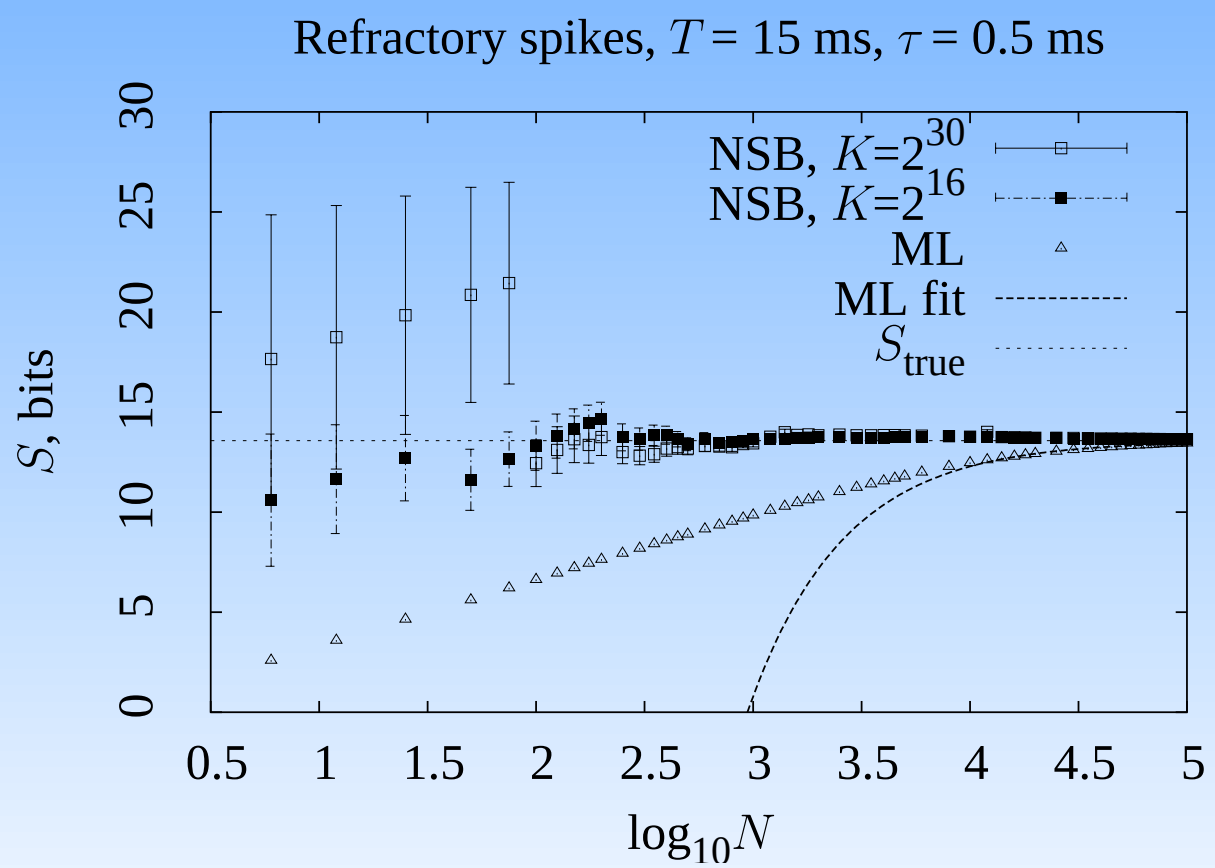
Refractory Poisson process: $r = 0.26\text{ms}^{-1}$, $R = 1.8\text{ms}$, $T = 15\text{ms}$, $\tau = 0.5\text{ms}$.
$K = 2^{30}$, $K_{\text{ref}} < 2^{16}$, $S = 13.57\text{bits}$.



True value reached within the error bars for $N^2 \sim 2^S$, when coincidences start to occur.

Estimator: Synthetic test

Refractory Poisson process: $r = 0.26\text{ms}^{-1}$, $R = 1.8\text{ms}$, $T = 15\text{ms}$, $\tau = 0.5\text{ms}$.
$K = 2^{30}$, $K_{\text{ref}} < 2^{16}$, $S = 13.57\text{bits}$.

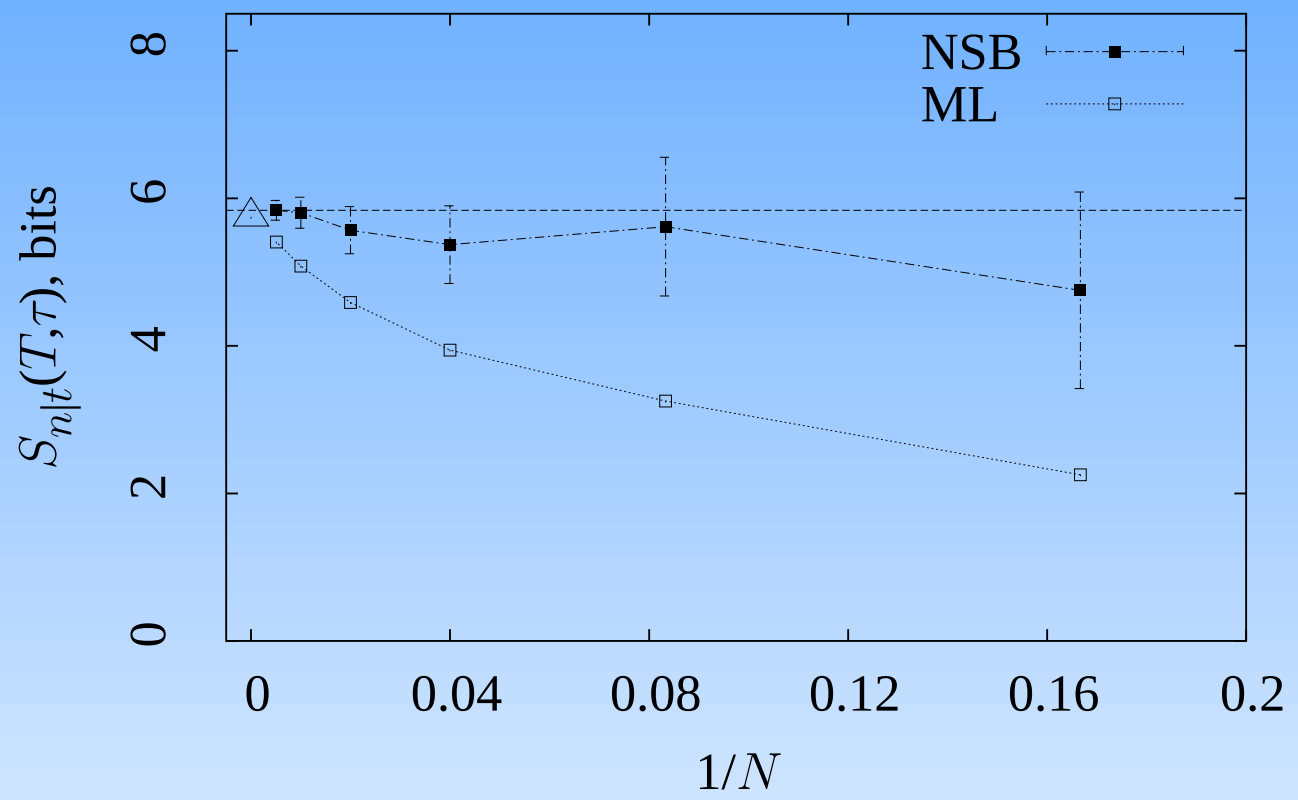


True value reached within the error bars for $N^2 \sim 2^S$, when coincidences start to occur.

Estimator is unbiased if it is consistent and agrees with itself for all N within error bars.

Natural data: Slice entropy vs. sample size

Slice at 1800 ms, τ = 2 ms, T = 16 ms

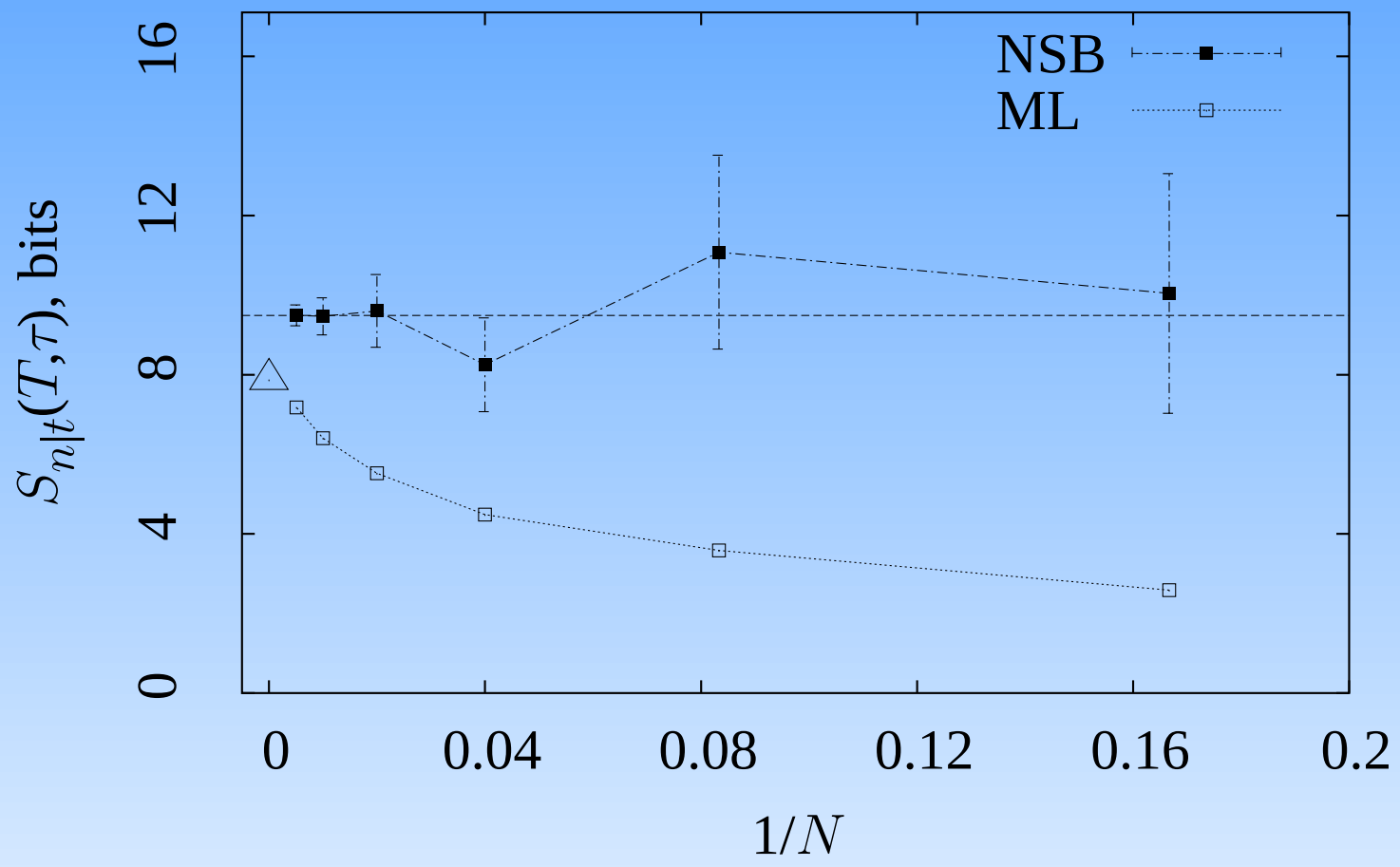


ML estimator converges with $\sim 1/N$ corrections.

NSB estimator is always within error bars.

$E(S^{\rm NSB} - S_{\rm ML})/\delta S^{\rm NSB} \approx 0$ if $S^{\rm ML}$ is reliably extrapolated ($N \gg 2^S$).

Slice at 1800 ms, τ = 2 ms, T = 30 ms



ML estimator cannot be extrapolated.

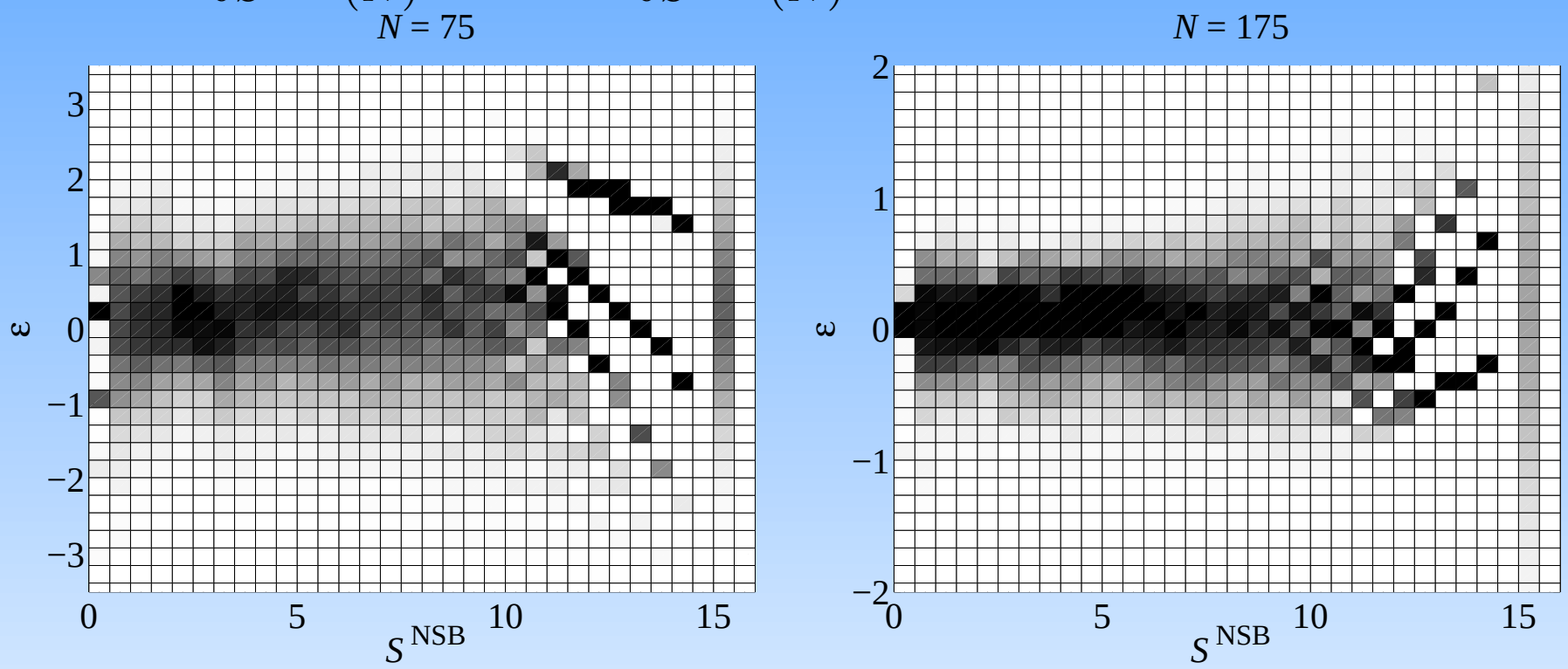
NSB estimator is always within error bars.

Natural data: Error vs. mean

$$\epsilon(N) \equiv \frac{S^{\mathrm{NSB}}(N) - S}{\delta S^{\mathrm{NSB}}(N)} \approx \frac{S^{\mathrm{NSB}}(N) - S^{\mathrm{NSB}}(196)}{\delta S^{\mathrm{NSB}}(N)}.$$ Remember: $\log_2 196 \approx 7.5\mathrm{bit}.$

Natural data: Error vs. mean

$$\epsilon(N) \equiv \frac{S^{\rm NSB}(N) - S}{\delta S^{\rm NSB}(N)} \approx \frac{S^{\rm NSB}(N) - S^{\rm NSB}(196)}{\delta S^{\rm NSB}(N)}.$$ Remember: $\log_2 196 \approx 7.5$bit.



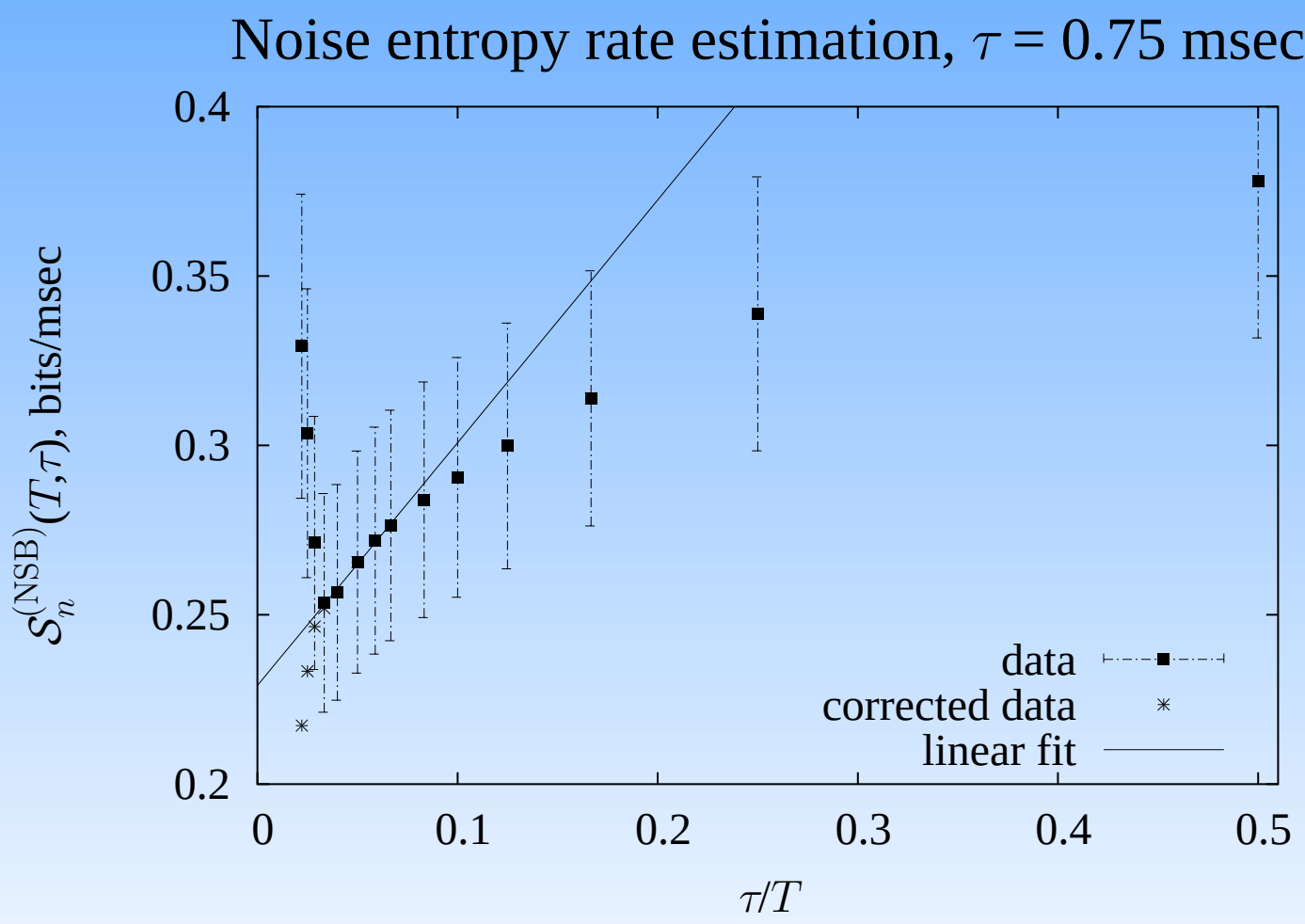
Almost no bias.

Empirical variance < 1 due to long tails in posterior, and $S \neq S^{\rm NSB}(196)$.

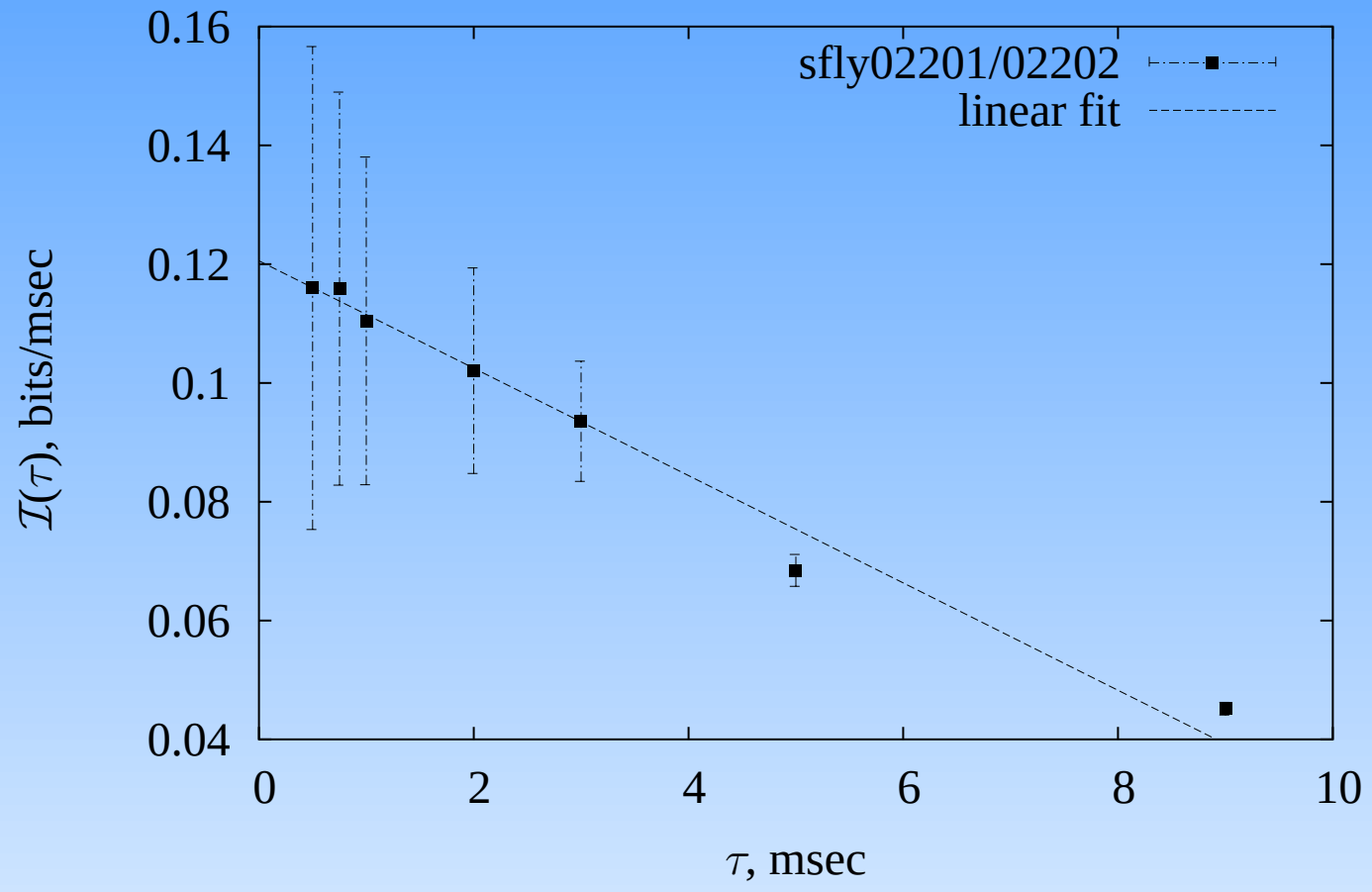
Bands are due to discrete nature of Δ.

Natural data: Rates

Further work is needed to properly estimate error bars due to signal correlations.



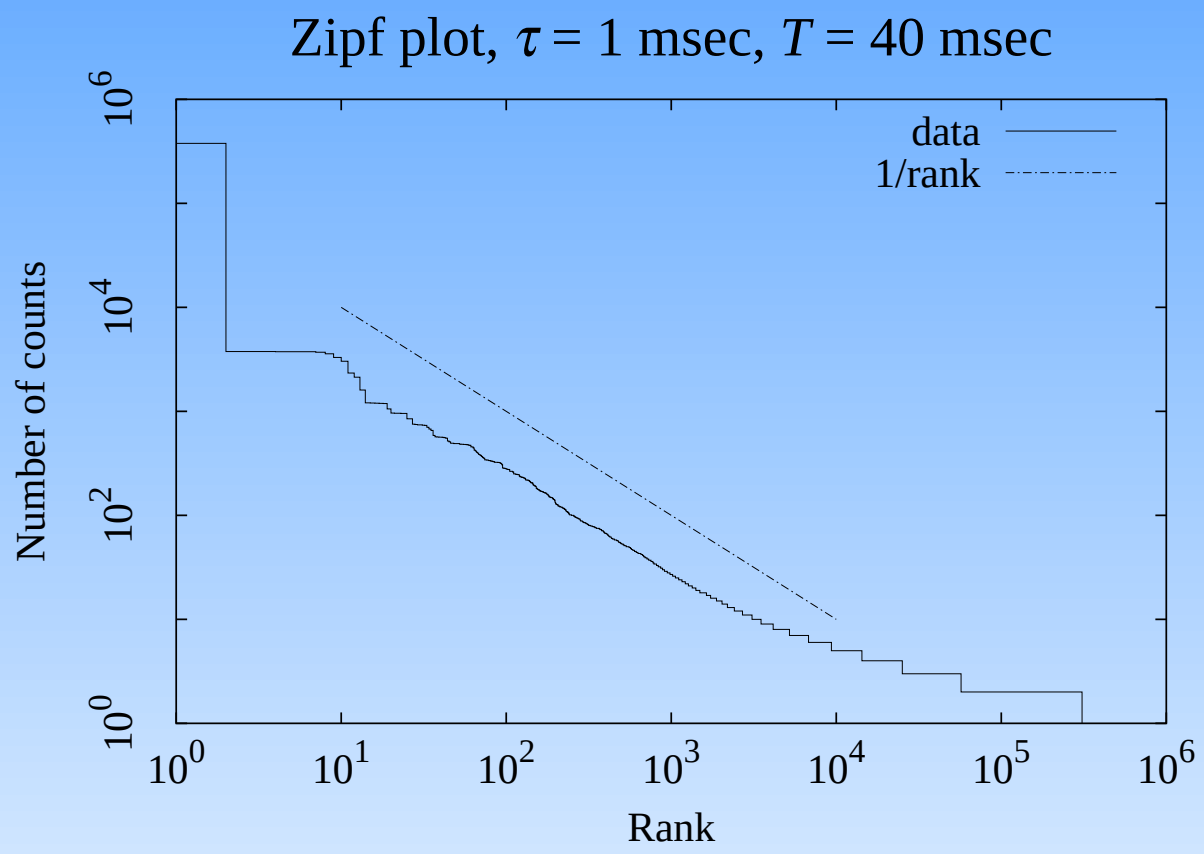
Information rate in the spike train



Conclusions

- Found new entropy estimator.
- Works in Ma regime.
- Produces error bars.
- Know if we should trust it.
- Neural data seems to be well matched to the estimator.
- Hope of similar progress on other biological data.

For amusement



Do not underestimate difficulty of working on real data!