

# Estimating information content of biological data

Ilya Nemenman

William Bialek, Fariel Shafee, Rob de Ruyter van Steveninck

(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

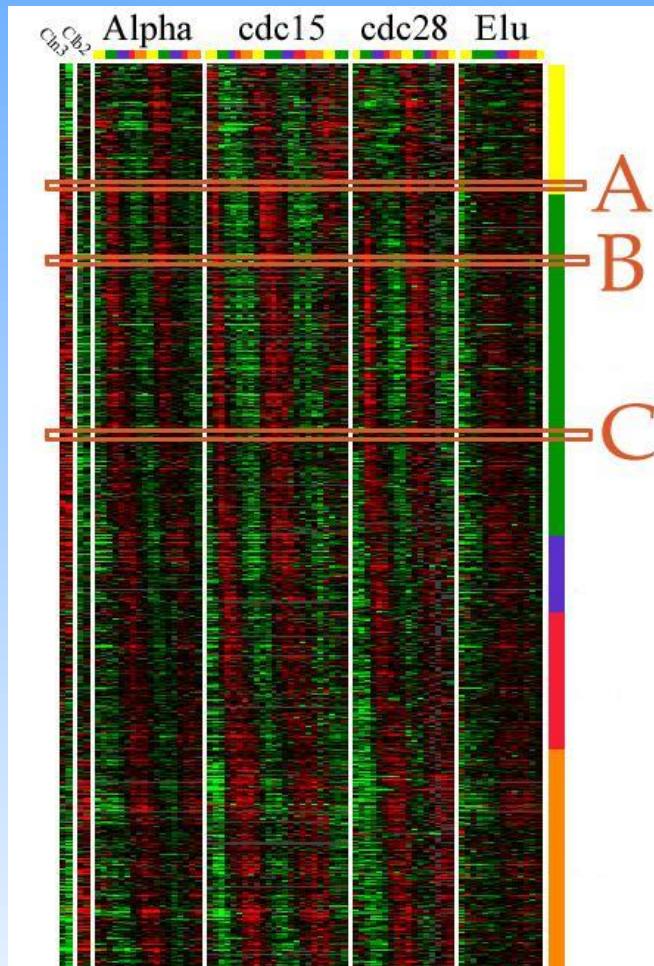
**Problem setup** Why bother?

**Developing intuition** Why hard?

**The method** An idea, analysis, asymptotics.

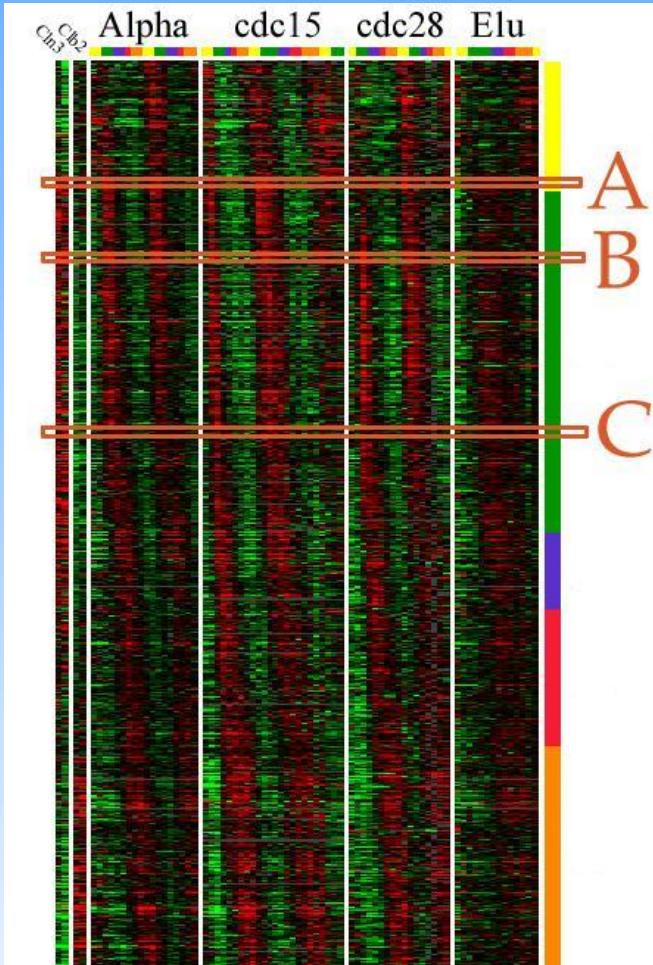
**Applications** Synthetic and natural neural data.

## A use: Inferring regulatory networks



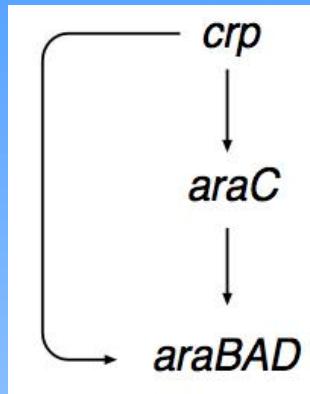
(Spellman et al., 1998)

## A use: Inferring regulatory networks

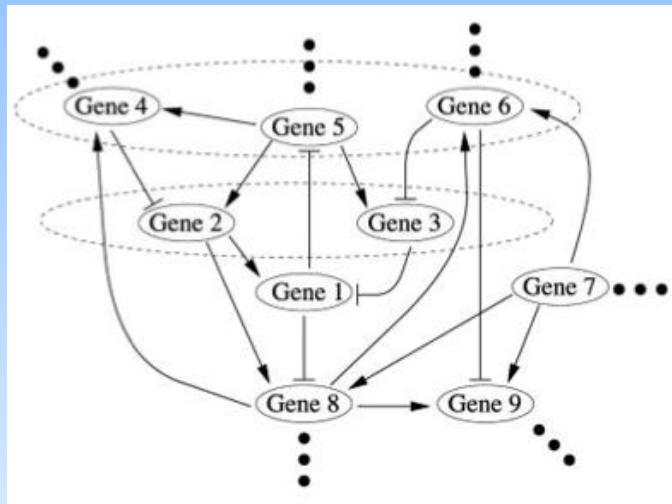


$$\begin{aligned} &\Rightarrow P(A, B, C, \dots) \\ &+ \delta P(A, B, C, \dots) \Rightarrow \end{aligned}$$

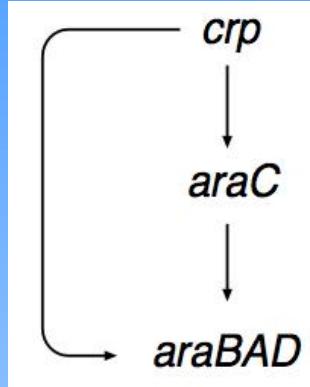
(Spellman et al., 1998)



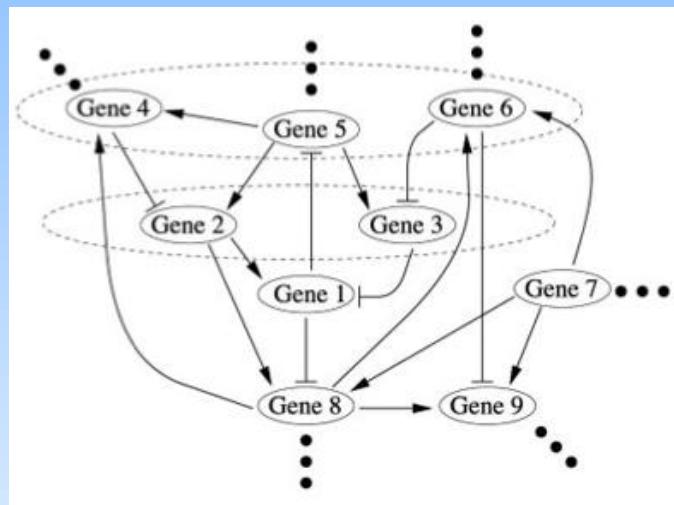
(Shen-Orr et al., 2002)



(Yeung et al., 2002)

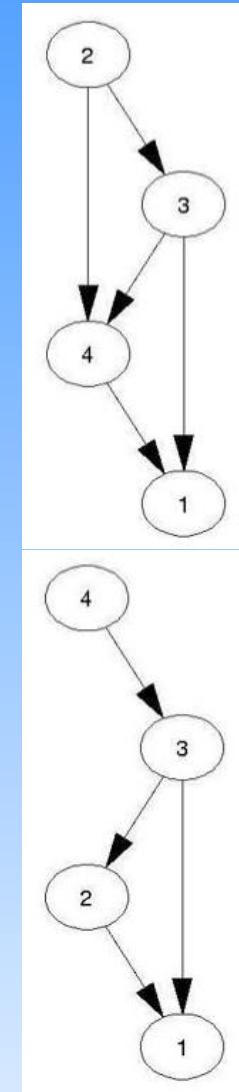


(Shen–Orr et al., 2002)

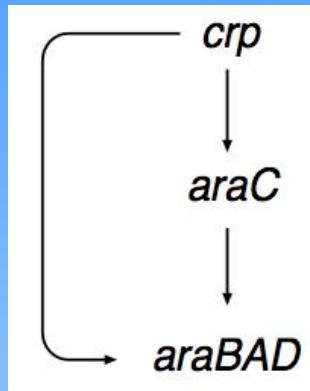


(Yeung et al., 2002)

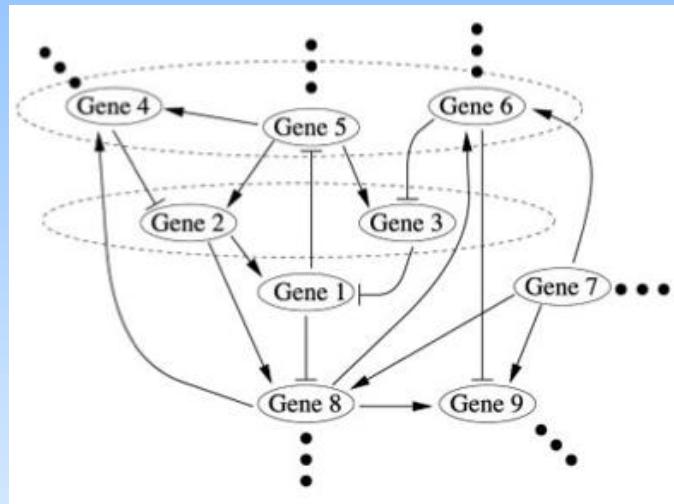
... but



(Ziv et al., 2003)

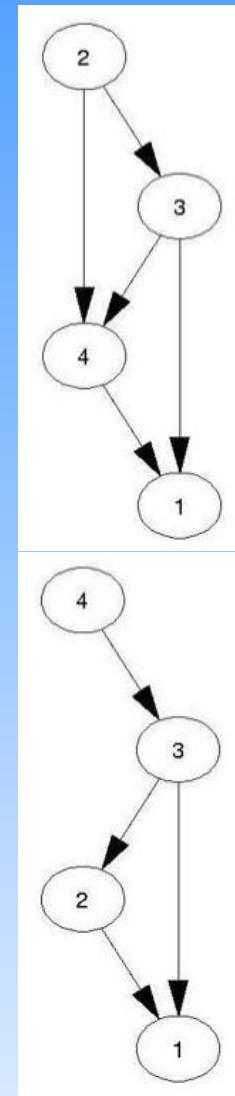


(Shen-Orr et al., 2002)

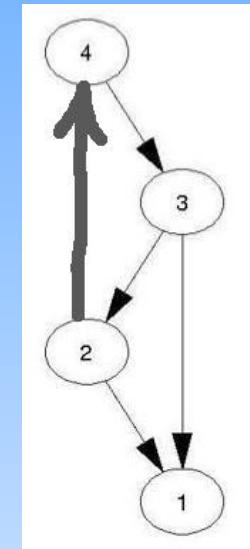


(Yeung et al., 2002)

... but



(Ziv et al., 2003)



Does the data support the dependence between 2 and 4?

# Entropy and information

Need assumptions-free measure of dependence.

# Entropy and information

Need assumptions-free measure of dependence.

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

# Entropy and information

Need assumptions-free measure of dependence.

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

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

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

# Entropy and information

Need assumptions-free measure of dependence.

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

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

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

Compare:  $I(x, y, z)$  to  $I(x, y) + I(x, z) + I(y, z)$ , to  
 $I(x, y) + I(x, z)$  etc. to distinguish topologies.

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

## Other uses

- phylogeny and haplotyping

	$\overbrace{\text{accGt}}^N$	ggtccatataaggaa
gccta	accAt	ggtccatatataggac
accta	accAt	ggtcgatataaggac

- length  $10^6 \dots 10^9$
  - $N$  up to 20
  - < 100 repeats
- Severe undersampling.

## Other uses

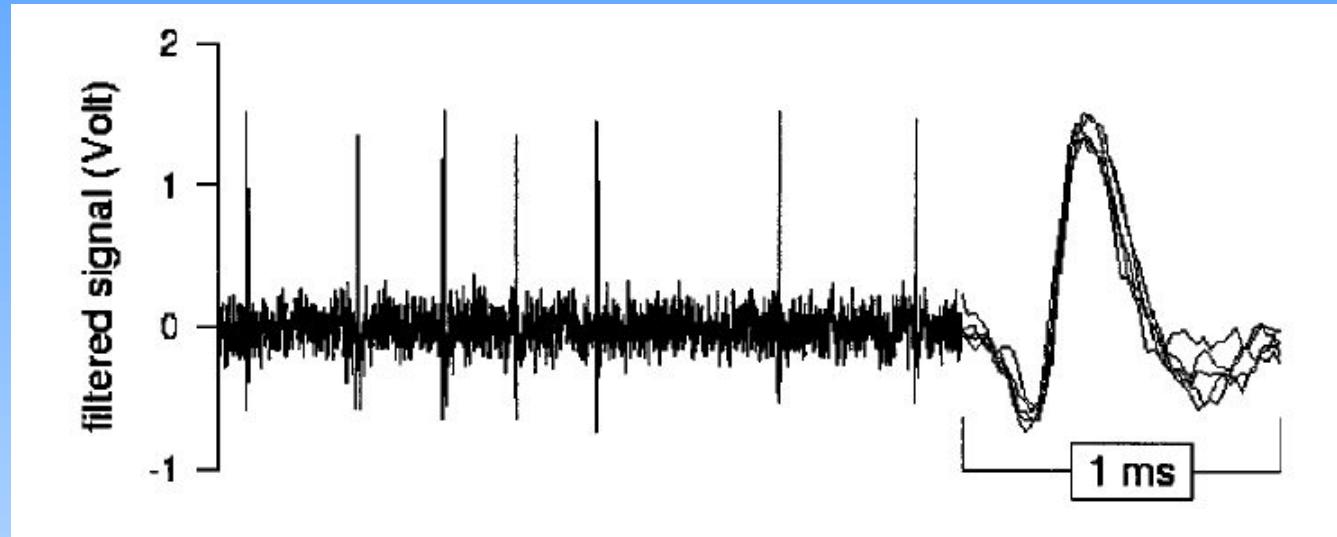
- phylogeny and haplotyping

	$\overbrace{\quad}^N$	
gccta	accGt	ggtccatatataaggaa
gccta	accAt	ggtccatataatggac
accta	accAt	ggtcgatataaggac

- length  $10^6 \dots 10^9$
- $N$  up to 20
- < 100 repeats
- Severe undersampling.

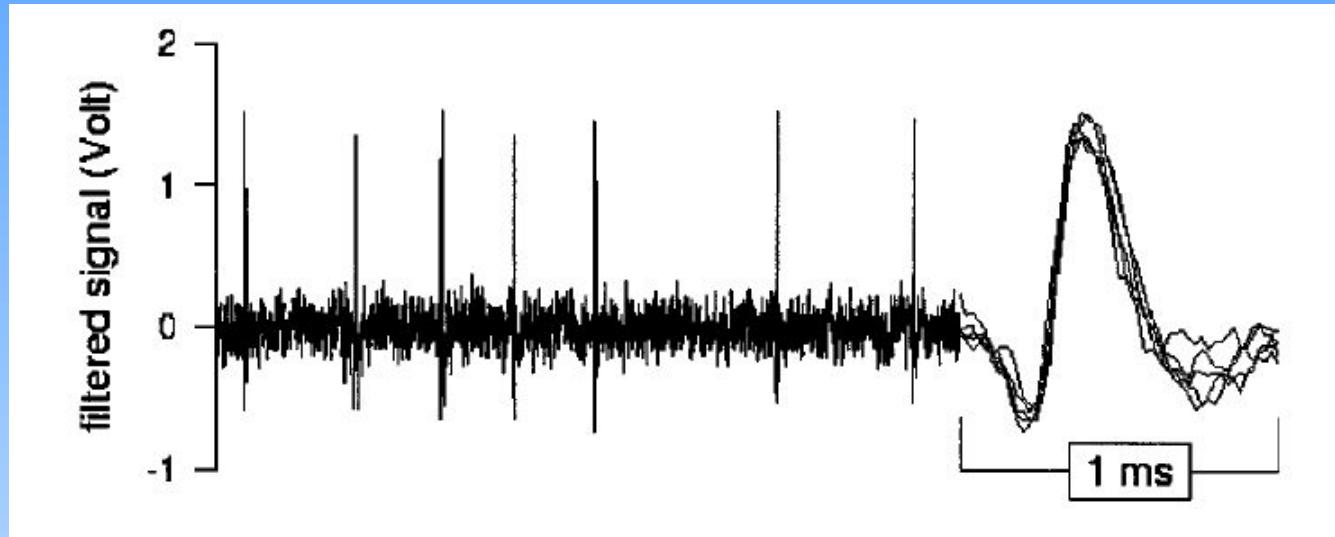
- motif search
- information content in molecular cell signals
- cross-compression: comparative texts analysis
- financial data and other prediction games
- dimensions of attractors in dynamical systems

# Neurophysiological applications



(Strong et al., 1998)

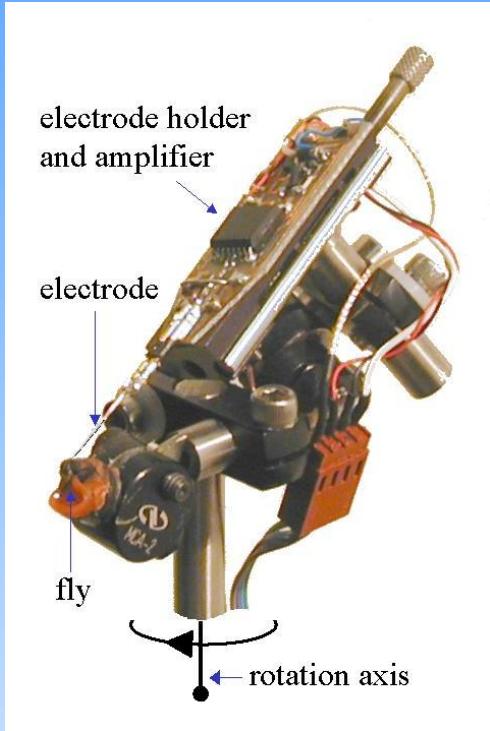
# Neurophysiological applications



(Strong et al., 1998)

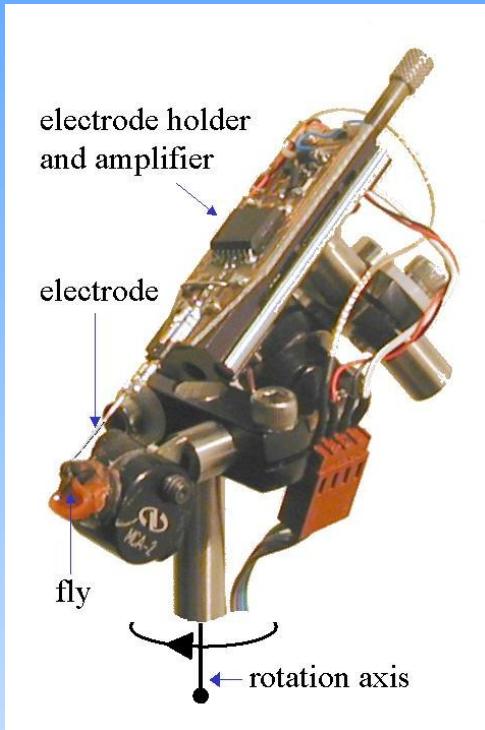
Neurons communicate by stereotypical pulses (spikes). Information is transmitted by spike rates and (possibly) **precise positions of the spikes.**

# Experimental setup

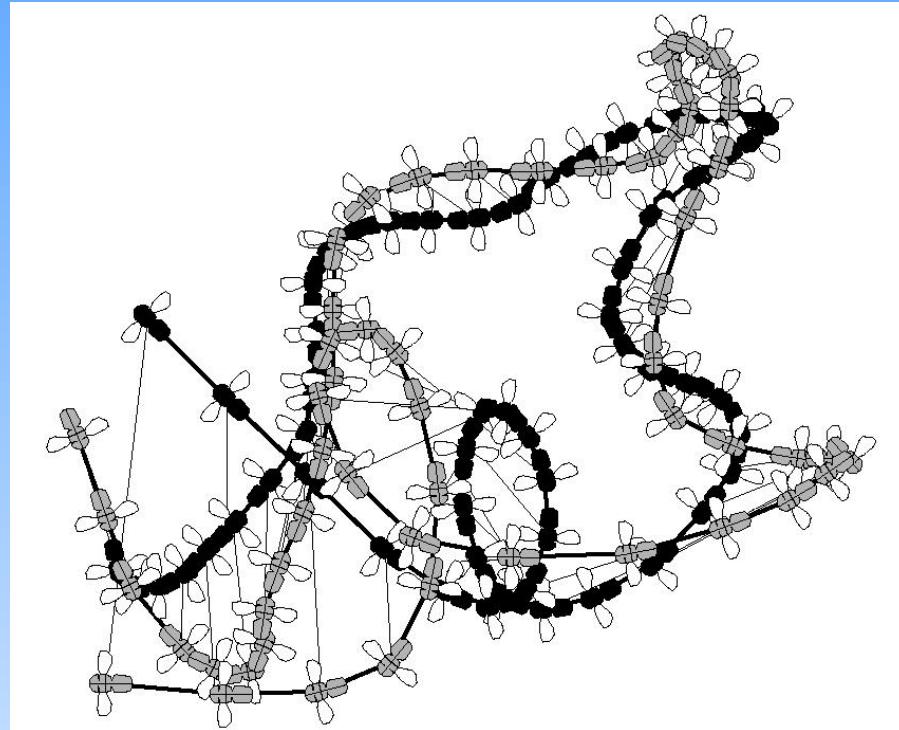


(Lewen, Bialek, and de  
Ruyter van Steveninck,  
2001)

# Experimental setup

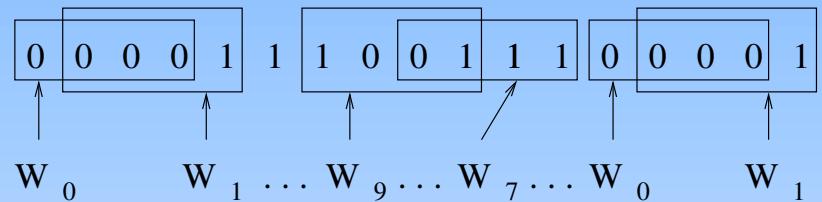


(Lewen, Bialek, and de Ruyter van Steveninck,  
2001)

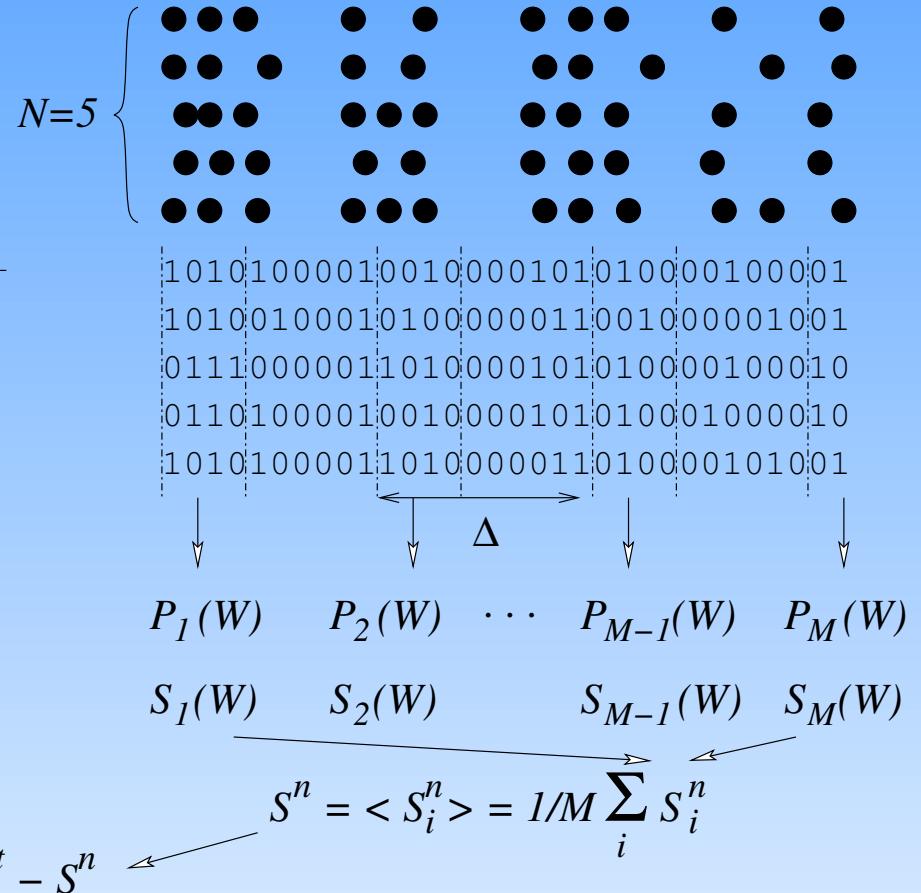


(Bialek and de Ruyter van Steveninck,  
2002; Land and Collett, 1974)

# Estimating information rate in spike trains

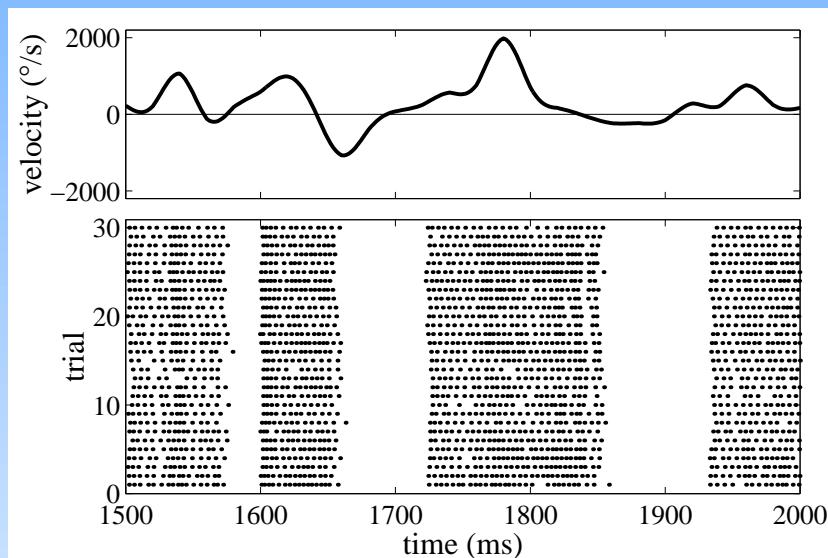
 $T=4$ 

$$P(W) \longrightarrow S(W) = S^t$$



# Recordings and problems

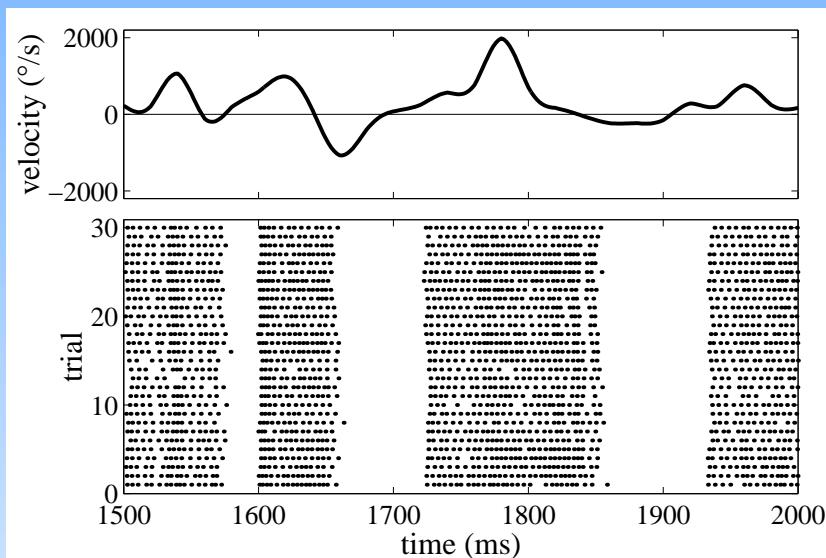
100–200 repeats of 5–10 s roller  
coasters rides



# Recordings and problems

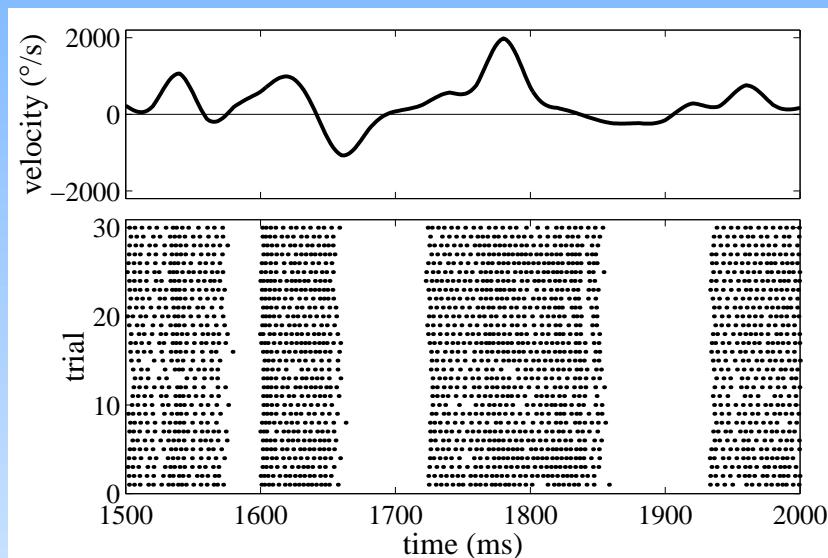
100–200 repeats of 5–10 s roller coasters rides

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



# Recordings and problems

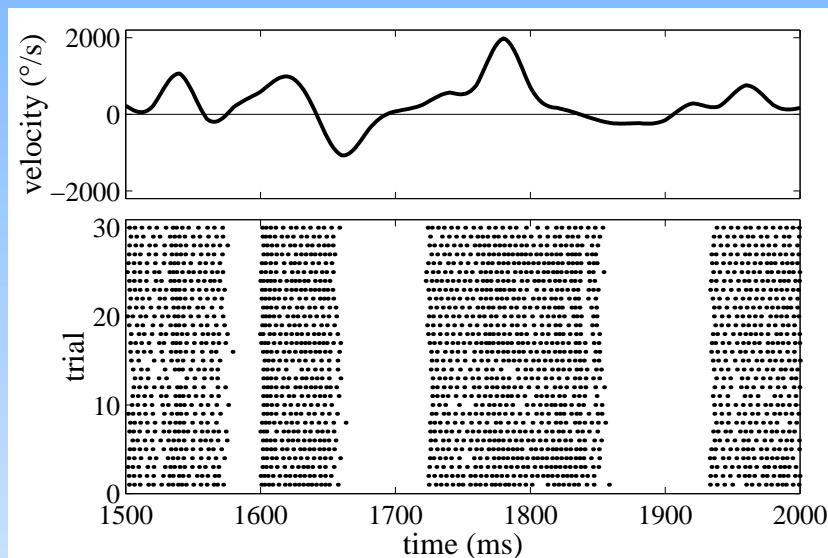
100–200 repeats of 5–10 s roller coasters rides



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

# Recordings and problems

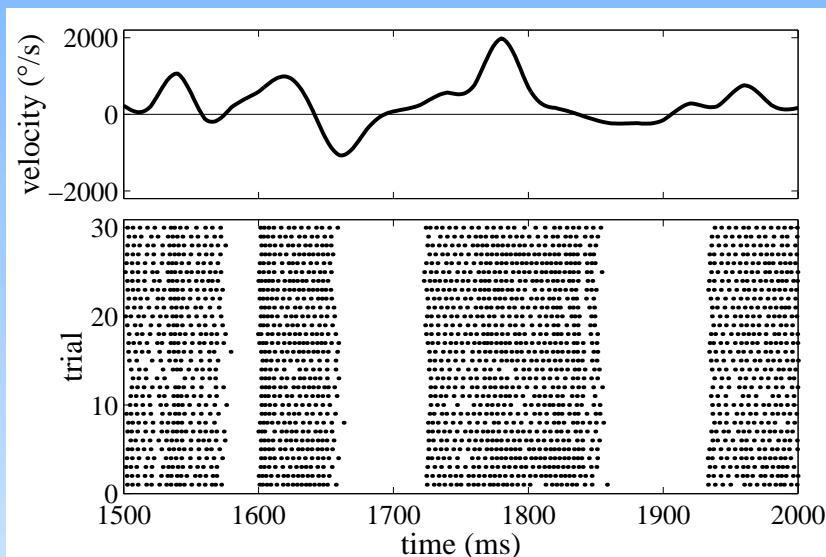
100–200 repeats of 5–10 s roller coasters rides



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

# Recordings and problems

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



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

Need to estimate entropies of words of length  $\sim 40$  from  $< 200$  samples.  
Undersampled!

## Why is this a difficult problem?

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

But for  $K \gg N$ ?

## Why is this a difficult problem?

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

But for  $K \gg N$ ?

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

improbable events but large entropy  
small errors in  $p$  but large errors in  $S$

## Why is this a difficult problem?

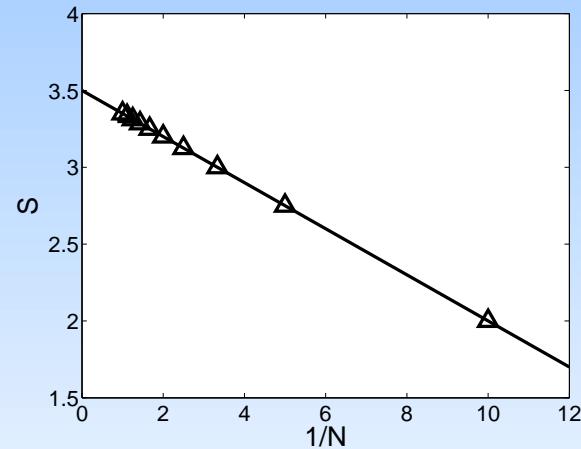
An asymptotically ( $K/N \rightarrow 0$ ) easy problem.

But for  $K \gg N$ ?

$$\begin{aligned} \lim_{p \rightarrow 0} \frac{p \log p}{p} &= \infty && \text{improbable events but large entropy} \\ &&& \text{small errors in } p \text{ but large errors in } S \\ S_{\text{ML}} &\equiv -\hat{p} \log \hat{p} - (1 - \hat{p}) \log(1 - \hat{p}) && \text{is convex} \\ \implies E S_{\text{ML}} < S(E \hat{p}) = S(p) &&& \text{unknown negative bias,} \\ &&& \text{variance is much smaller} \end{aligned}$$

- 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  $K/N \rightarrow \text{const}$ ,  $K \rightarrow \infty$  (Paninski, 2003)
- correcting for bias as a power series in  $2^S/N$

- 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  $K/N \rightarrow \text{const}$ ,  $K \rightarrow \infty$  (Paninski, 2003)
- 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  $2^S \ll N \ll K$



## The hope

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

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

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

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)

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

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

# Learning with nearly uniform priors

(ultra-local, Dirichlet priors)

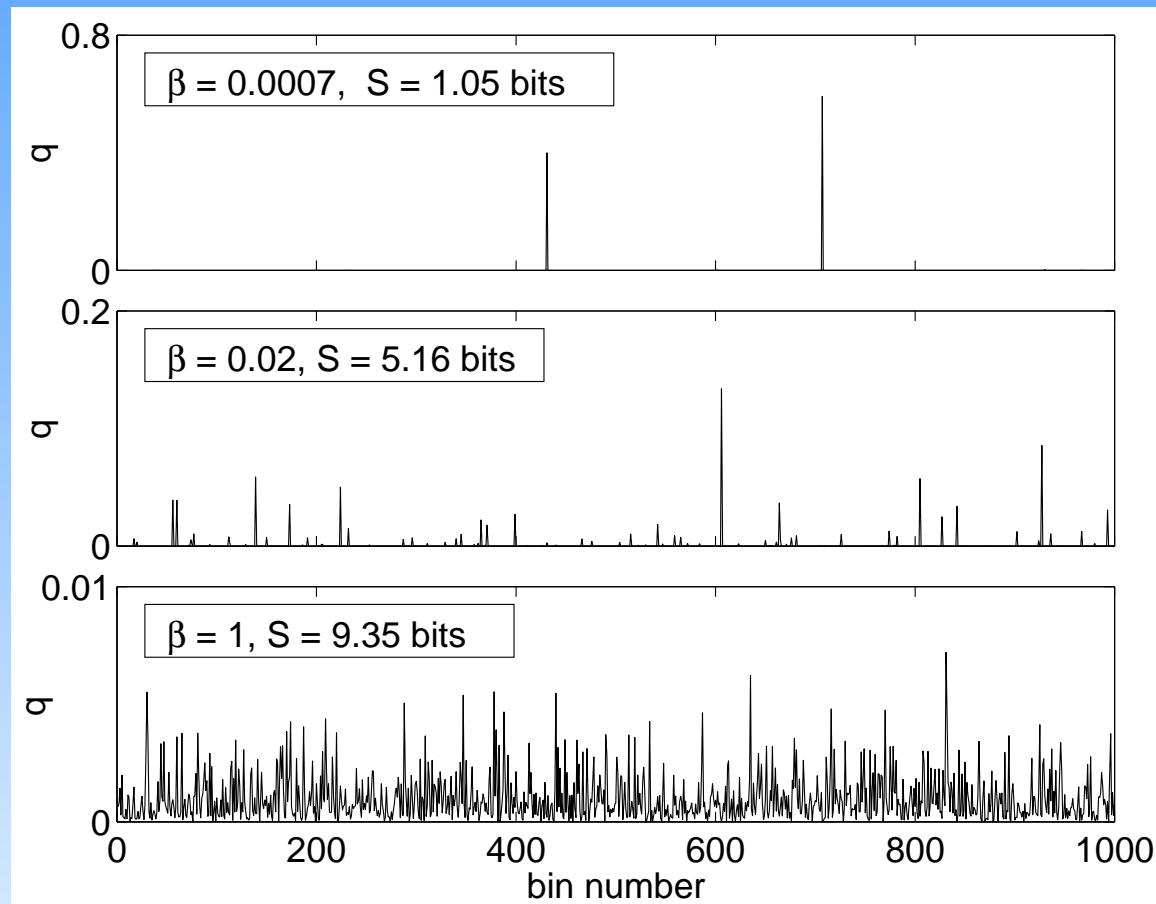
$\{q_i\}$ ,  $i = 1 \dots K$ :

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

Some common choices:

Maximum likelihood	$\beta \rightarrow 0$
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

$$P_\beta(\{q_i\}|\{n_i\}) = \frac{P(\{n_i\}|\{q_i\})\mathcal{P}_\beta(\{q_i\})}{P_\beta(\{n_i\})}$$

$$P(\{n_i\}|\{q_i\}) = \prod_{i=1}^K (q_i)^{n_i}$$

$$\langle q_i \rangle_\beta = \frac{n_i + \beta}{N + K\beta}$$

$$\langle S \rangle_\beta = \text{known } (\text{Wolpert and Wolf, 1995})$$

$$\langle \delta^2 S \rangle_\beta = \text{known}$$

## Bayesian inference with Dirichlet priors

$$P_\beta(\{q_i\}|\{n_i\}) = \frac{P(\{n_i\}|\{q_i\})\mathcal{P}_\beta(\{q_i\})}{P_\beta(\{n_i\})}$$

$$P(\{n_i\}|\{q_i\}) = \prod_{i=1}^K (q_i)^{n_i}$$

$$\langle q_i \rangle_\beta = \frac{n_i + \beta}{N + K\beta}$$

$\langle S \rangle_\beta$  = known (Wolpert and Wolf, 1995)

$\langle \delta^2 S \rangle_\beta$  = known

Equal pseudocounts added to each bin.

# Bayesian inference with Dirichlet priors

$$P_\beta(\{q_i\}|\{n_i\}) = \frac{P(\{n_i\}|\{q_i\})\mathcal{P}_\beta(\{q_i\})}{P_\beta(\{n_i\})}$$

$$P(\{n_i\}|\{q_i\}) = \prod_{i=1}^K (q_i)^{n_i}$$

$$\langle q_i \rangle_\beta = \frac{n_i + \beta}{N + K\beta}$$

$\langle S \rangle_\beta$  = known (Wolpert and Wolf, 1995)

$\langle \delta^2 S \rangle_\beta$  = known

Equal pseudocounts added to each bin.

Larger  $\beta$  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

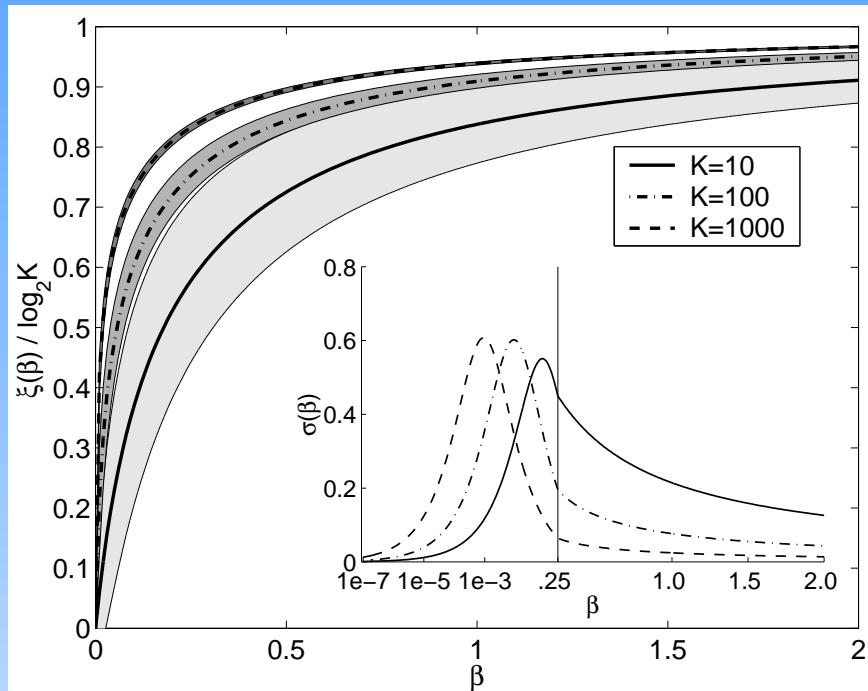
$$\mathcal{P}_\beta(S) = \int dq_1 dq_2 \cdots dq_K P_\beta(\{q_i\}) \delta \left[ S + \sum_{i=1}^K q_i \log_2 q_i \right]$$

$$\begin{aligned} \xi(\beta) &\equiv \langle S[n_i = 0] \rangle_\beta \\ &= \psi_0(K\beta + 1) - \psi_0(\beta + 1), \end{aligned}$$

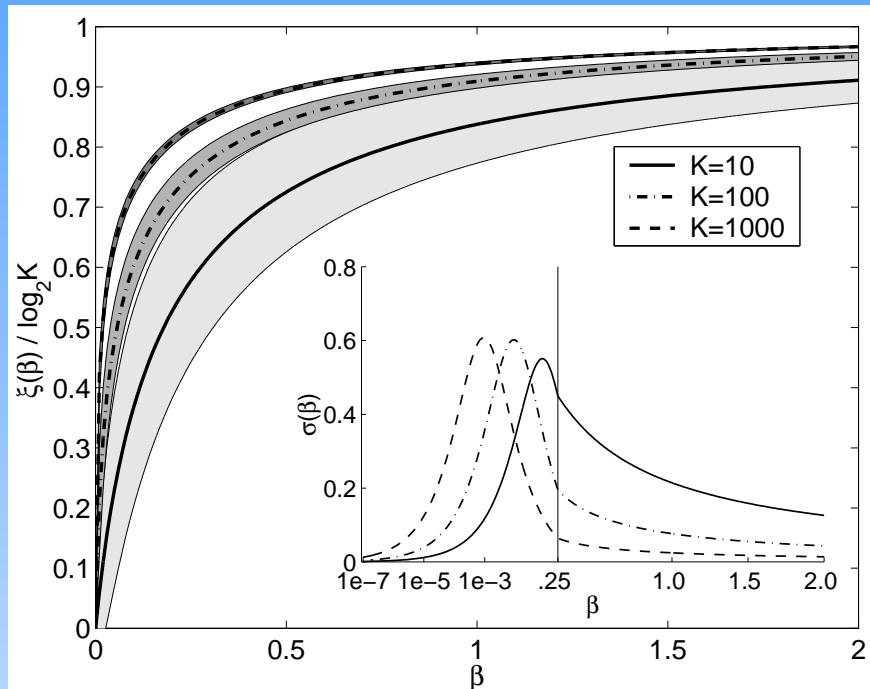
$$\begin{aligned} \sigma^2(\beta) &\equiv \langle (\delta S)^2[n_i = 0] \rangle_\beta \\ &= \frac{\beta + 1}{K\beta + 1} \psi_1(\beta + 1) - \psi_1(K\beta + 1) \end{aligned}$$

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

# The problem: Analysis

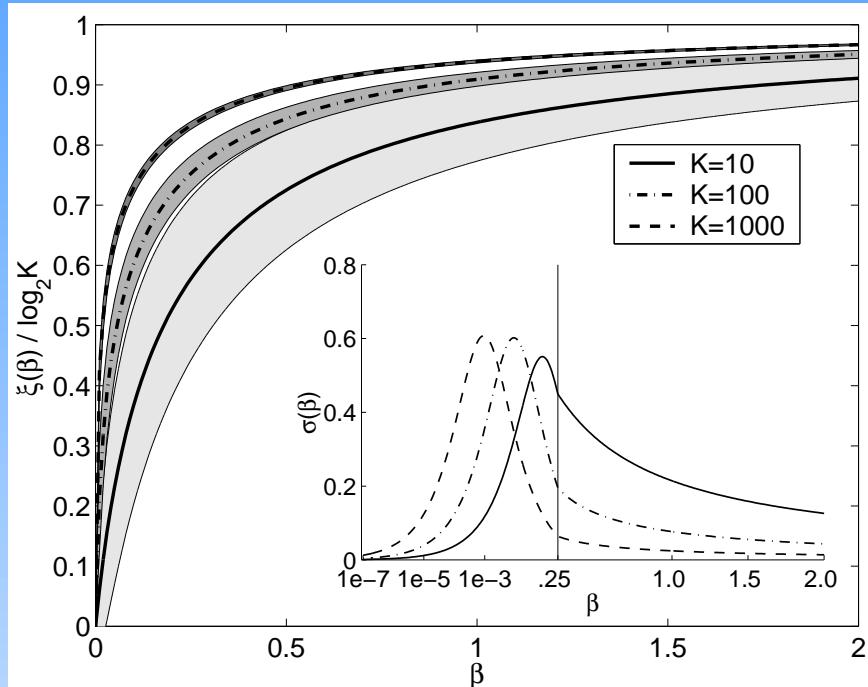


# The problem: Analysis



- Because of the Jacobian of  $\{q_i\} \rightarrow S$ , a priori distribution of entropy is strongly peaked.
- Narrow peak:  $\sigma(\beta) \propto 1/\sqrt{K\beta}$ ,  $\max \sigma(\beta) = 0.61$  bits.
- As  $\beta$  varies from 0 to  $\infty$ , 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  $\{q_i\} \rightarrow S$ , a priori distribution of entropy is strongly peaked.
- Narrow peak:  $\sigma(\beta) \propto 1/\sqrt{K\beta}$ ,  $\max \sigma(\beta) = 0.61$  bits.
- As  $\beta$  varies from 0 to  $\infty$ , 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  $\beta$ .
- Choosing  $\beta$  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, therefore, 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  $\delta$ -function!

# Solution

Average over  $\beta$  — infinite Dirichlet mixtures.

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

# Solution

Average over  $\beta$  — infinite Dirichlet mixtures.

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

$\beta \rightarrow \xi$  Jacobian entropy prior

# Solution

Average over  $\beta$  — infinite Dirichlet mixtures.

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

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

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

# Solution

Average over  $\beta$  — infinite Dirichlet mixtures.

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

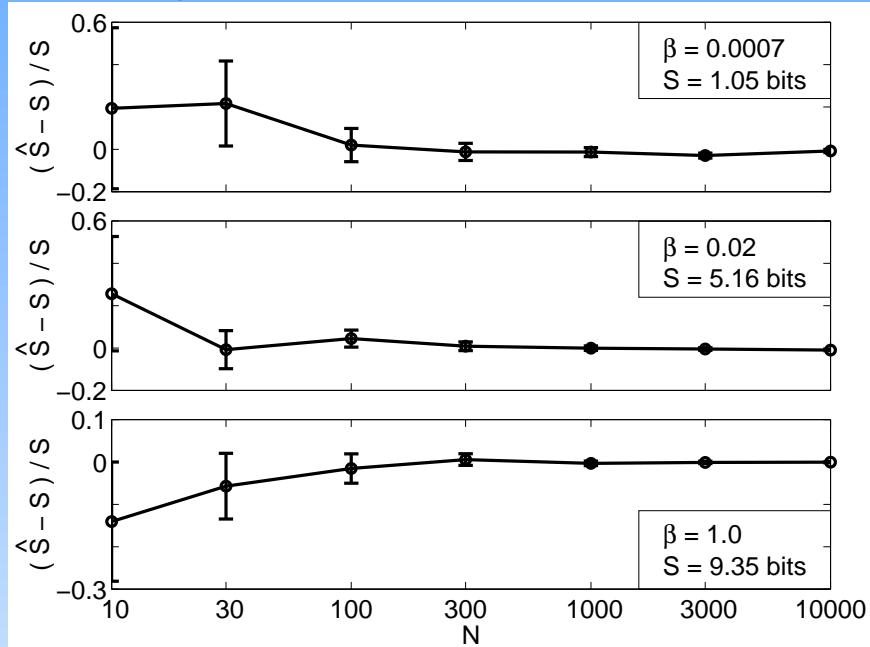
$$\widehat{S^m} = \frac{\int d\xi \rho(\xi, \{n_i\}) \langle S^m[n_i] \rangle_{\beta(\xi)}}{\int d\xi \rho(\xi, [n_i])}$$

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

- Smaller  $\beta$  means larger allowed volume in the space of  $\{q_i\}$ . Thus averaging over  $\beta$  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  $\beta$  (model) dominates (is “selected”)

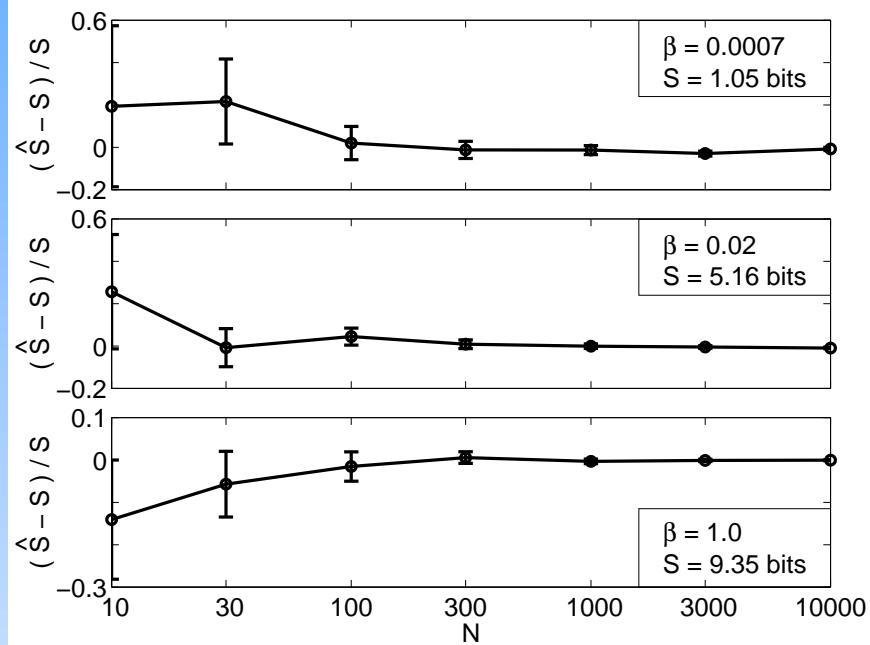
# First attempts to estimate entropy

Typical distributions

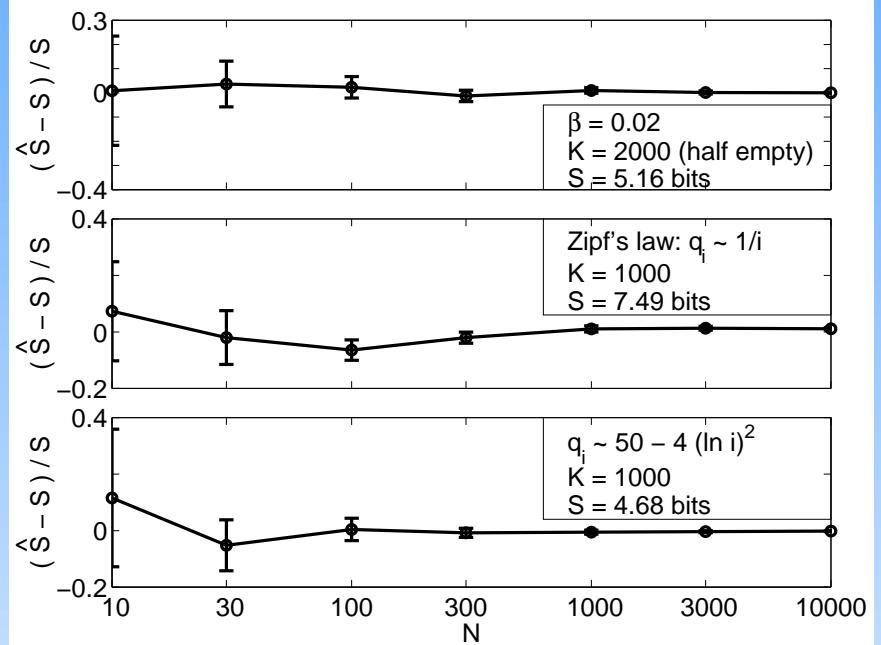


# First attempts to estimate entropy

Typical distributions

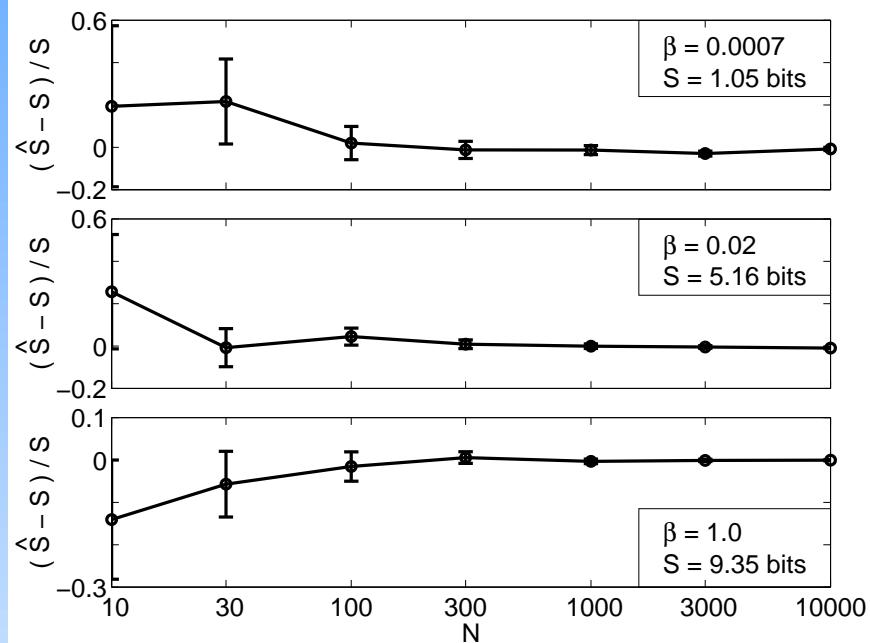


Atypical distributions

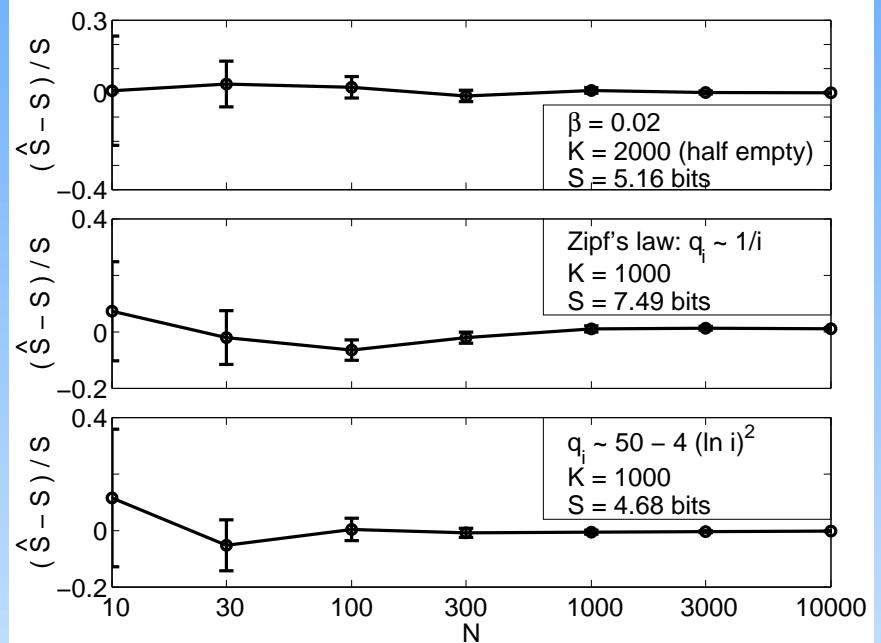


# First attempts to estimate entropy

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 error (posterior variance).
- *Little bias*, as it should be. Exception: too smooth distributions.
- Key point: *learn entropies directly without finding  $\{q_i\}$ !*
- The dominant  $\beta$  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
- $\frac{\partial^2(-\log \rho)}{\partial \xi^2}\Big|_{\xi(\beta^*)} = \Delta + NO([\Delta/N]^2)$

# Asymptotics

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

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

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

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

# Asymptotics

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

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

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

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

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.
- $\Delta$  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

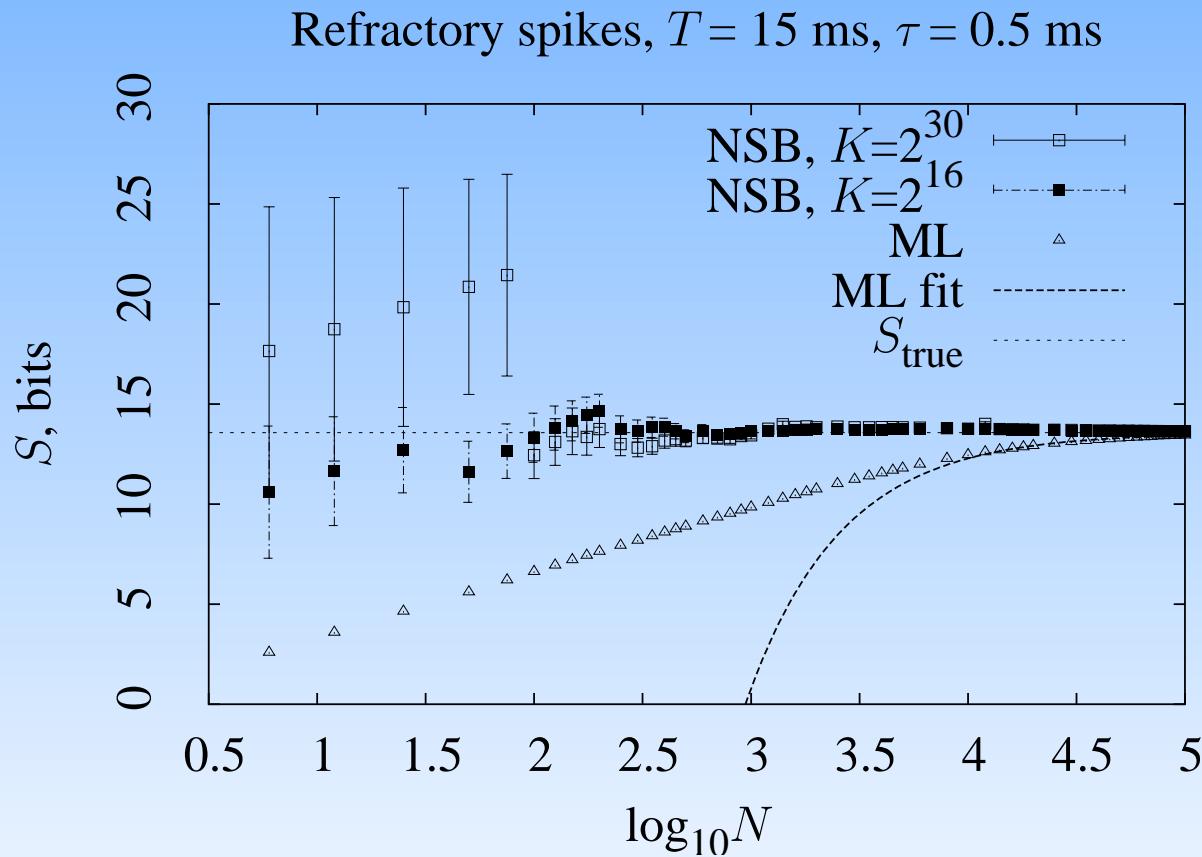
Refractory Poisson process:  $r = 0.26\text{ms}^{-1}$ ,  $R = 1.8\text{ms}$ ,  $T = 15\text{ms}$ ,  $\tau = 0.5\text{ms}$ .

## 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

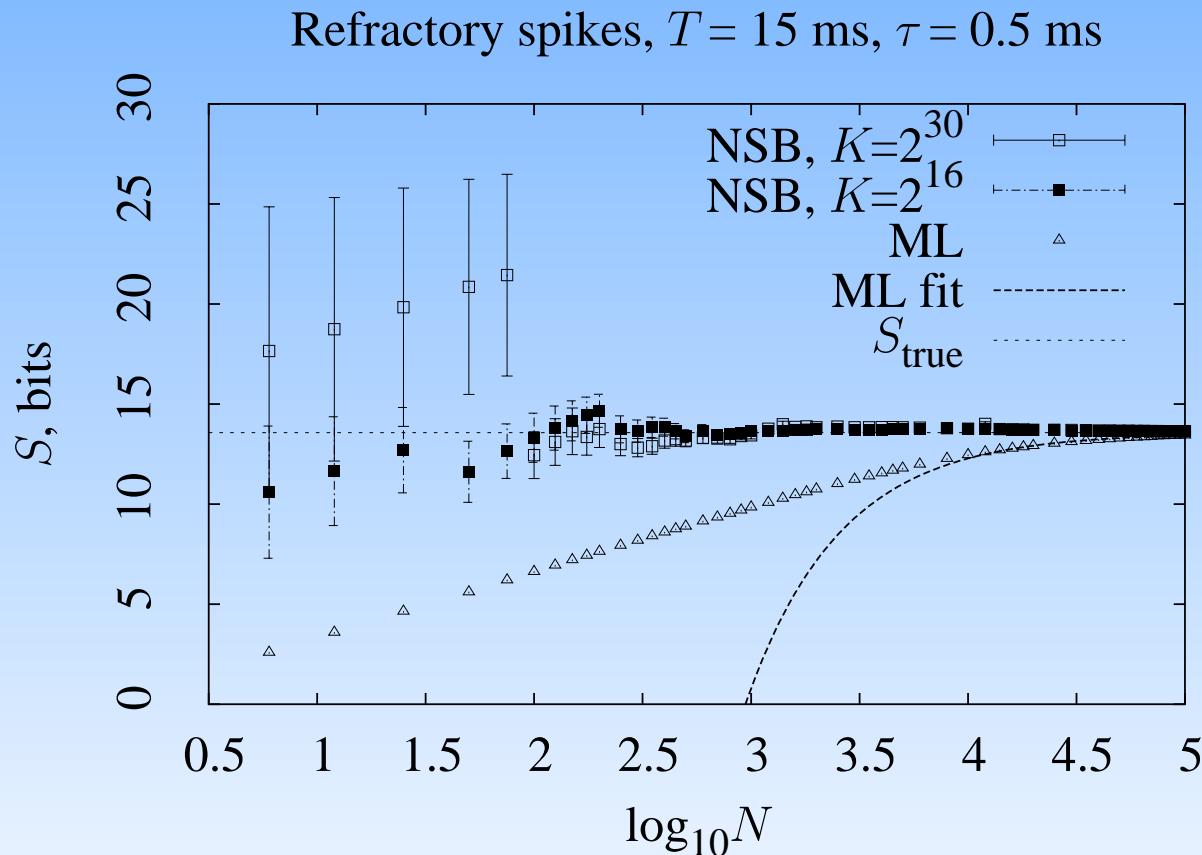
Refractory Poisson process:  $r = 0.26\text{ms}^{-1}$ ,  $R = 1.8\text{ms}$ ,  $T = 15\text{ms}$ ,  $\tau = 0.5\text{ms}$ .  
 $K = 2^{30}$ ,  $K_{\text{ref}} < 2^{16}$ ,  $S = 13.57\text{bits}$ .



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

## Estimator: Synthetic test

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

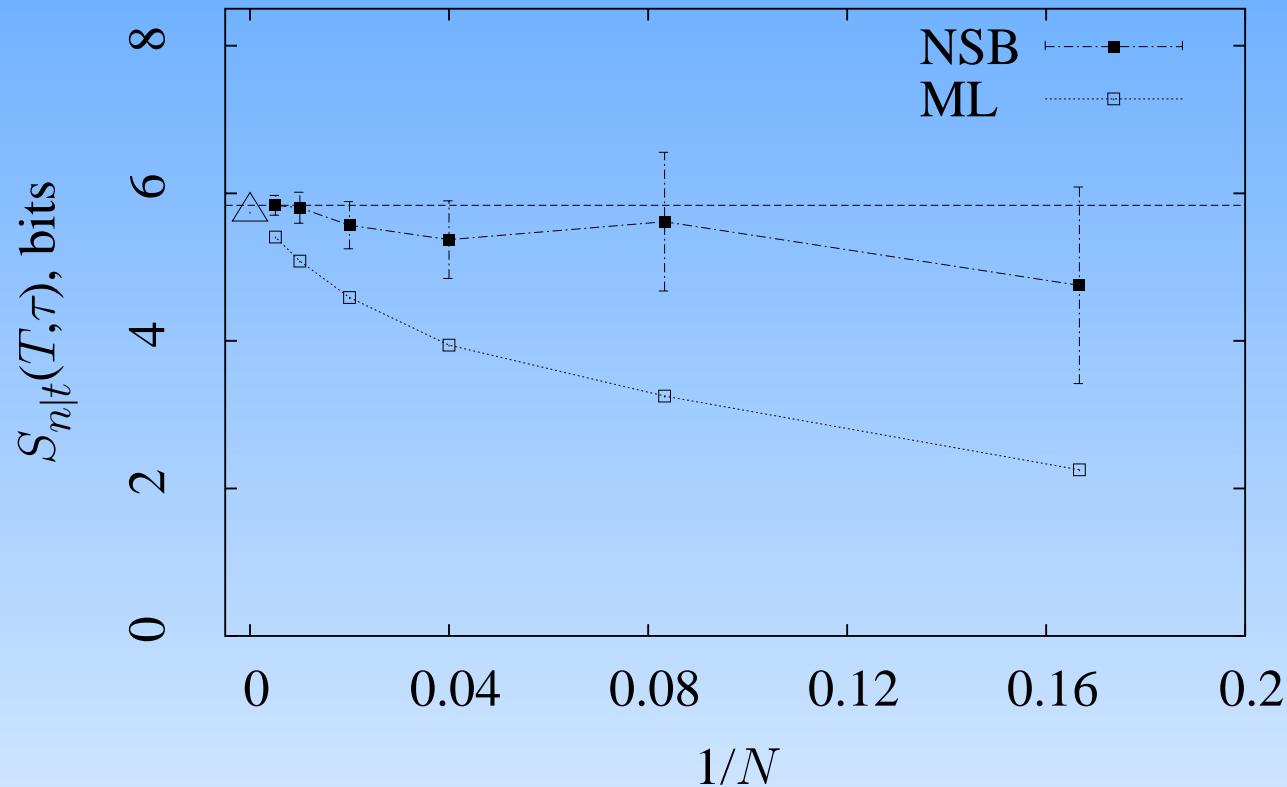


True value reached within the error bars for  $N^2 \sim 2^S$ , when coincidences start to occur.

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,  $\tau = 2$  ms,  $T = 16$  ms

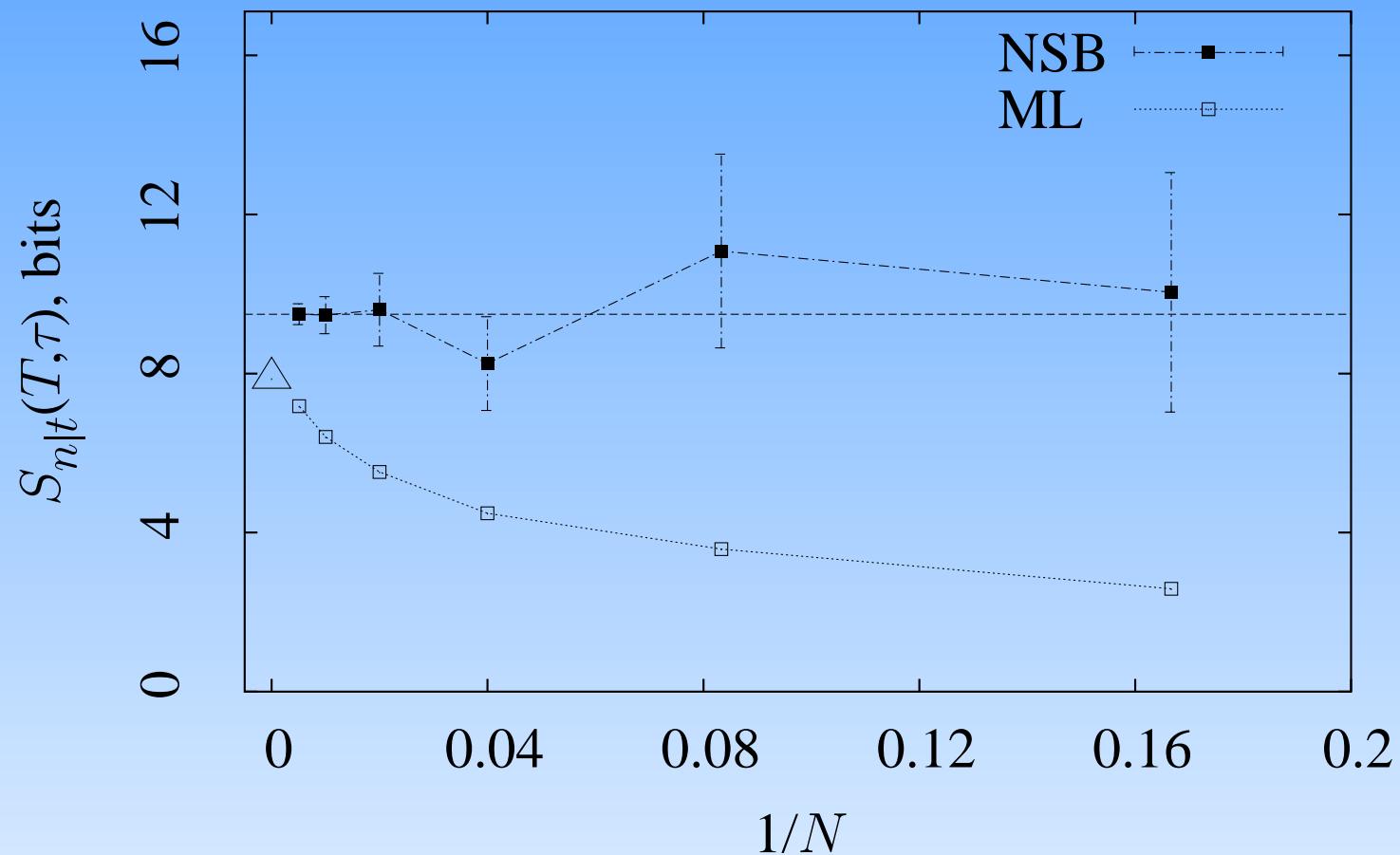


ML estimator converges with  $\sim 1/N$  corrections.

NSB estimator is always within error bars.

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

Slice at 1800 ms,  $\tau = 2$  ms,  $T = 30$  ms



ML estimator cannot be extrapolated.

NSB estimator is always within error bars.

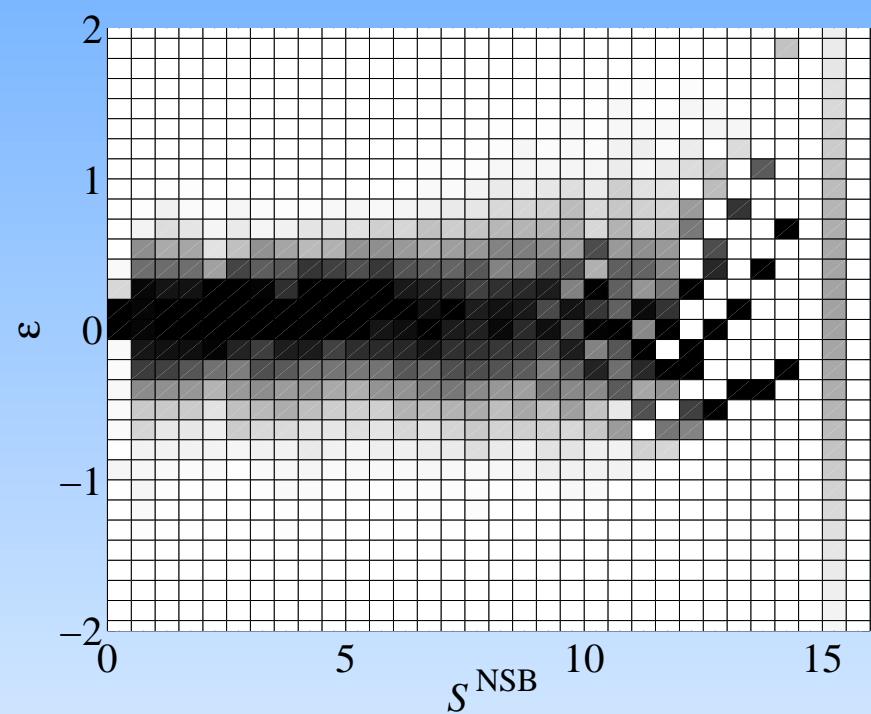
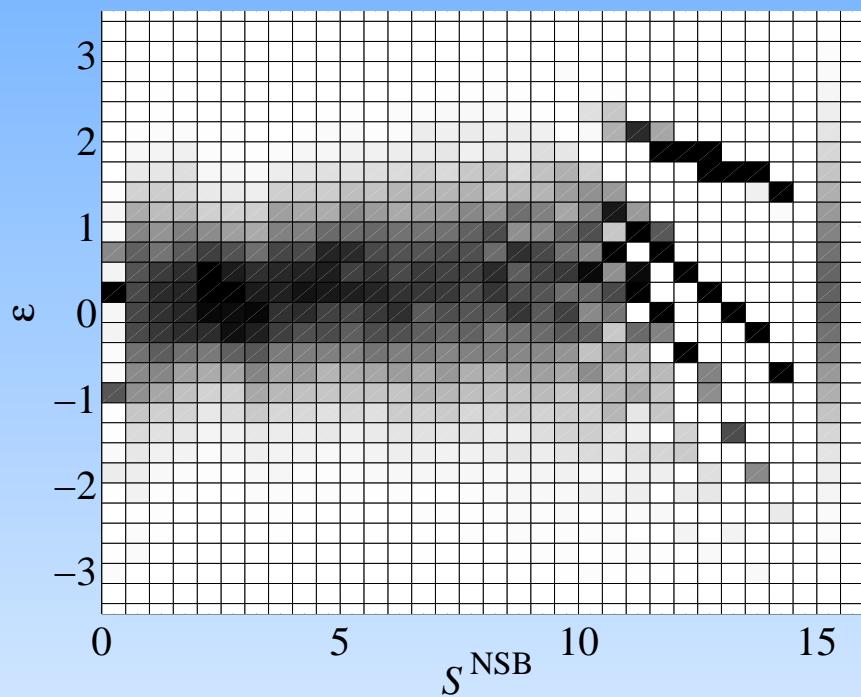
## Natural data: Error vs. mean

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

# Natural data: Error vs. mean

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

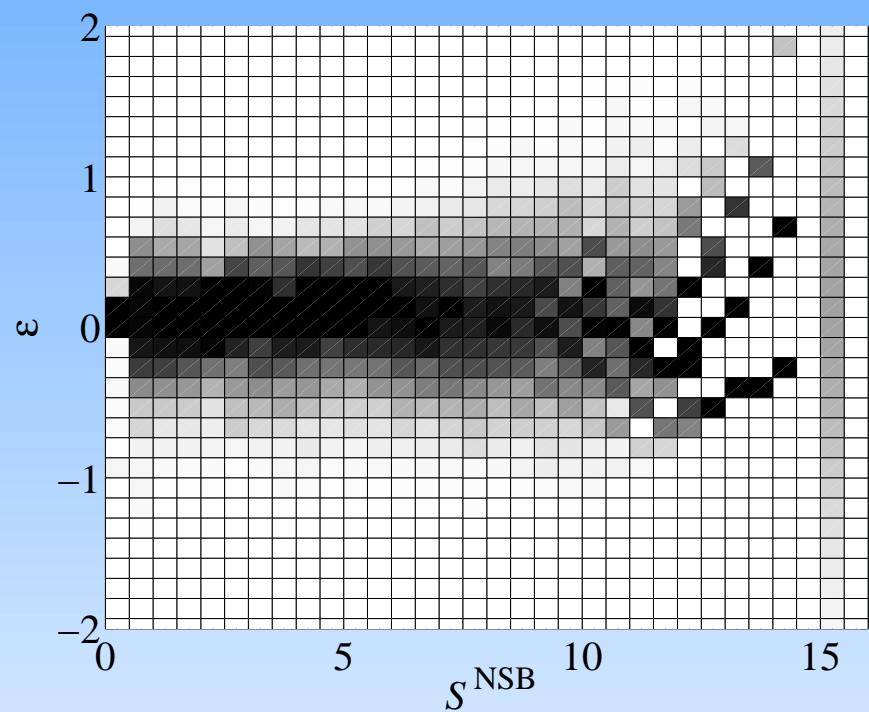
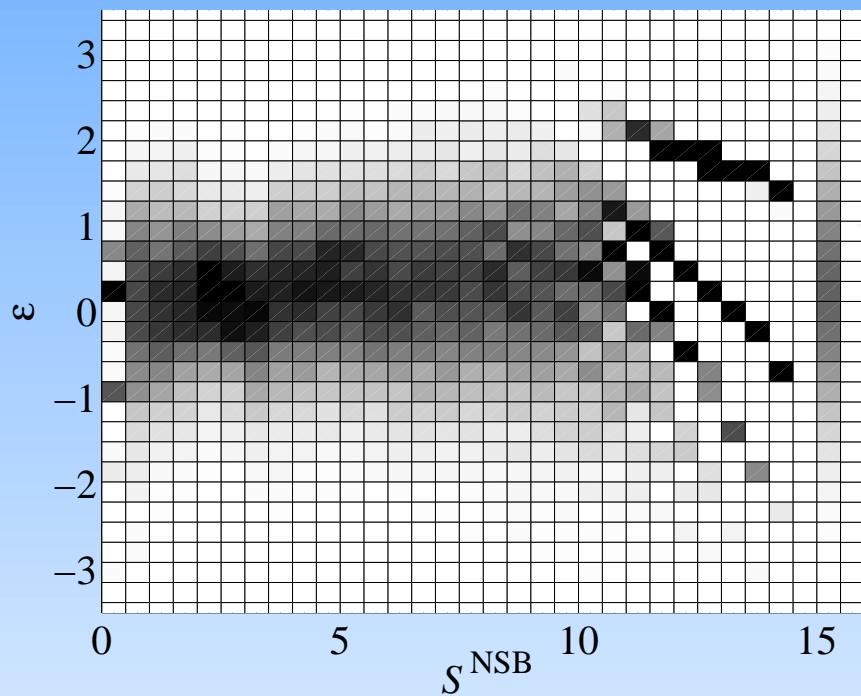
$N = 175$



# Natural data: Error vs. mean

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

$N = 175$



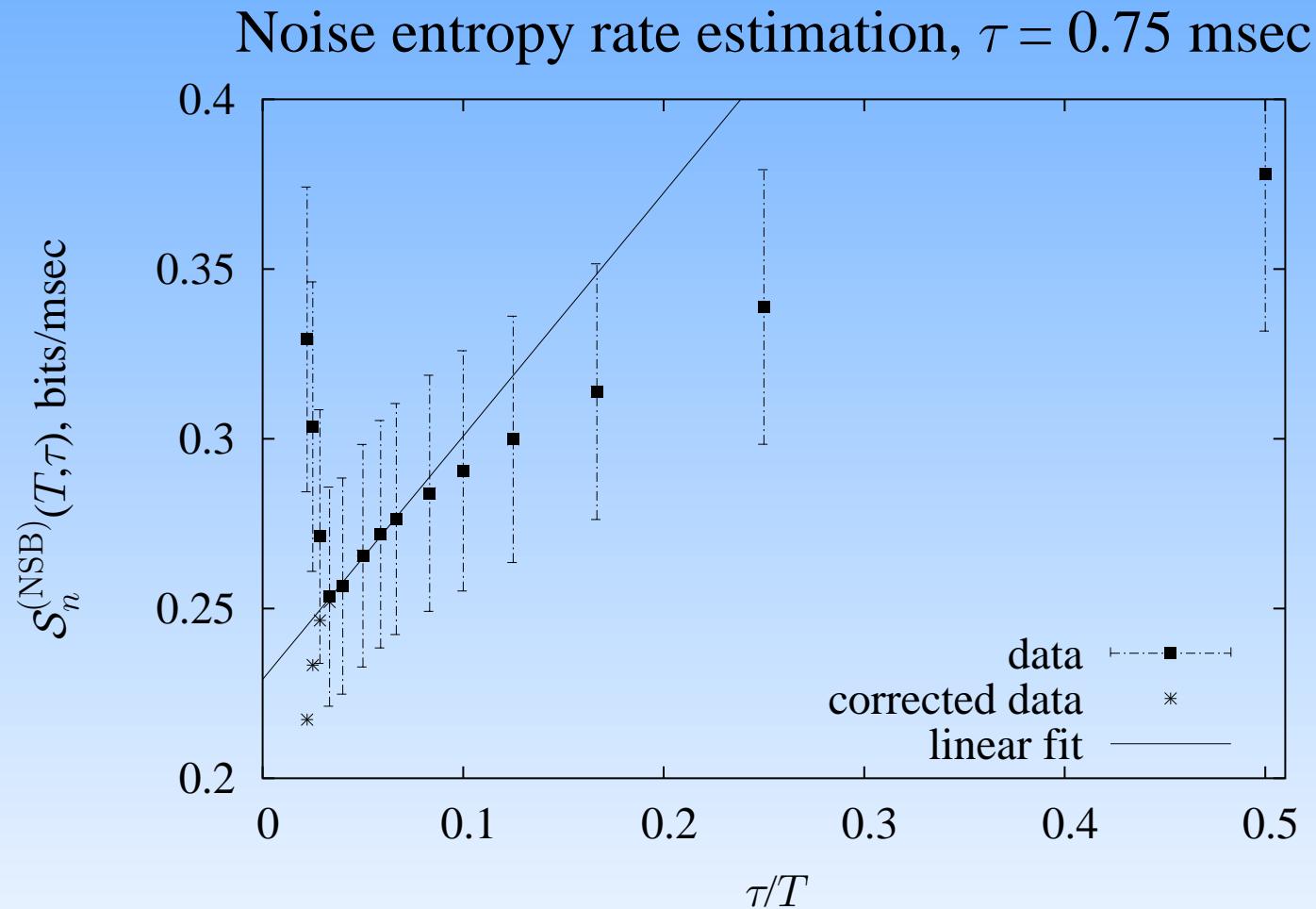
Almost no bias.

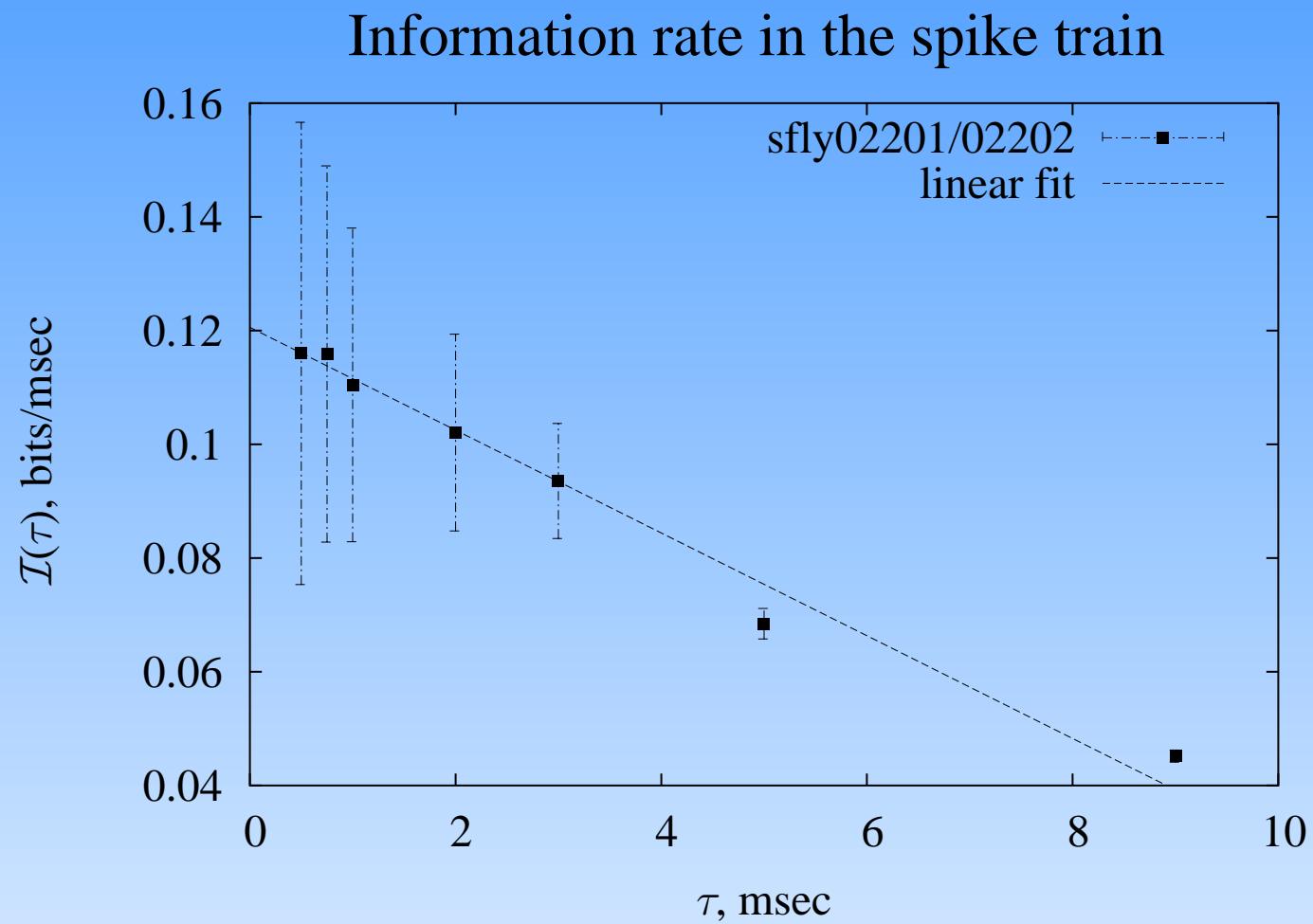
Empirical variance < 1 due to long tails in posterior, and  $S \neq S^{\text{NSB}}(196)$ .

Bands are due to discrete nature of  $\Delta$ .

## Natural data: Rates

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

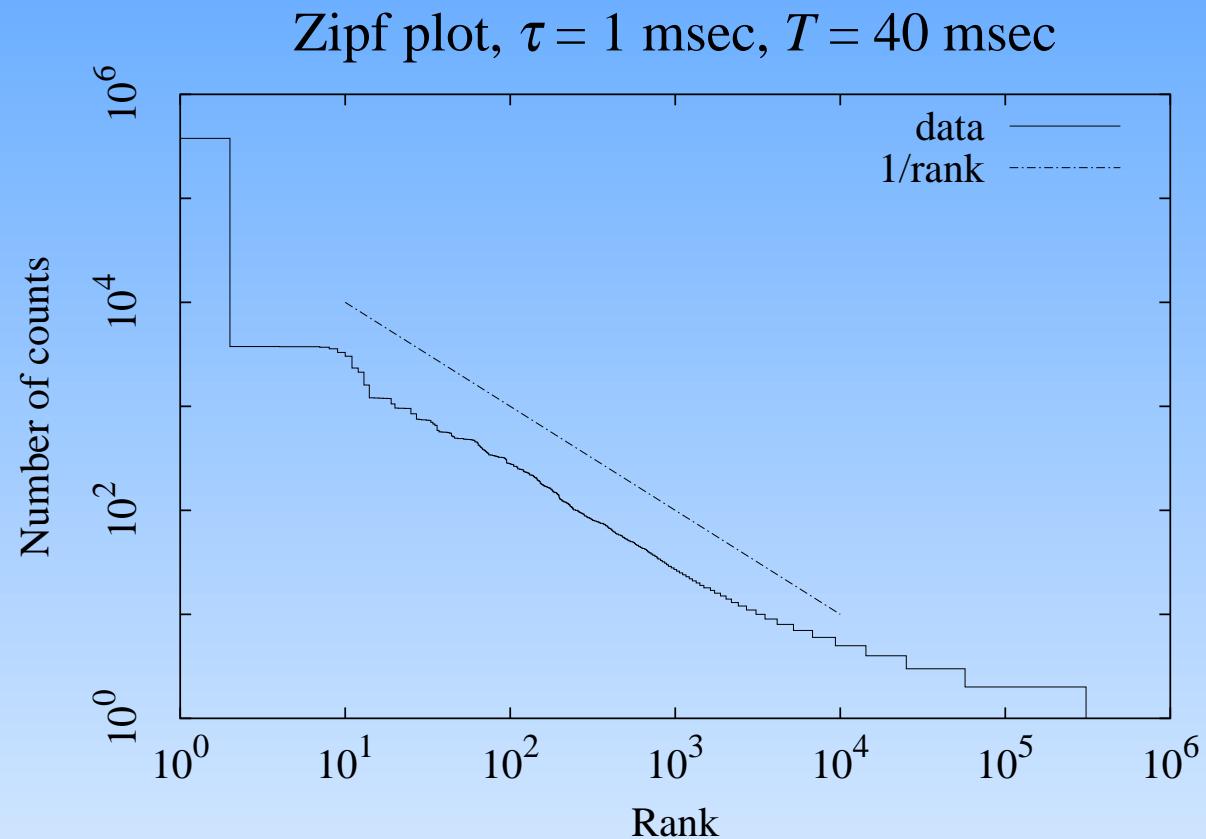




# Conclusions

- 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 genomic data.

## For amusement



Do not underestimate difficulty of working on real data!