Estimating Mutual Information by Bayesian Binning

Dominik Endres
University of St. Andrews, UK

The problem

- (Discrete) random variables: *X*, *Y*
- $x \in \{1;\ldots;K\}, y \in \{1;\ldots;C\}$
- In Neurophysiology:
  - *Y*: stimulus label
  - *X*: evoked response
- We would like to estimate the *mutual information*
  - $I(X;Y) = H(X) + H(Y) - H(X,Y)$
- from a sample of pairs $(x,y)_n$ of length N

The problem

- Where (Shannon 1948)

$$H(X) = -\sum_{k=1}^{K} P_k \ln(P_k) \qquad H(Y) = -\sum_{c=1}^{C} P_c \ln(P_c)$$

$$H(X,Y) = -\sum_{k=1}^{K}\sum_{c=1}^{C} P_{kc} \ln(P_{kc})$$

$$P_k = \sum_{c=1}^{C} P_{kc} \qquad P_c = \sum_{k=1}^{K} P_{kc}$$

When is this difficult/easy ?

- Easy:
  - If N >> CK, i.e. the joint distribution of X and Y is well sampled
  - Use *maximum-likelihood* estimate for P_{kc}, i.e.

$$\hat{P}_{kc} = \frac{n_{kc}}{N}, \quad \hat{H}(X,Y) = -\sum_{k=1}^{k} \sum_{c=1}^{C} \hat{P}_{kc} \ln(\hat{P}_{kc})$$

  - and add a bias correction term ∝ 1/N (Miller, 1955, Panzeri 1996) to the entropy estimate.
  - Similar result available for the variance of H (Paninski, 2003).

When is this difficult/easy ?

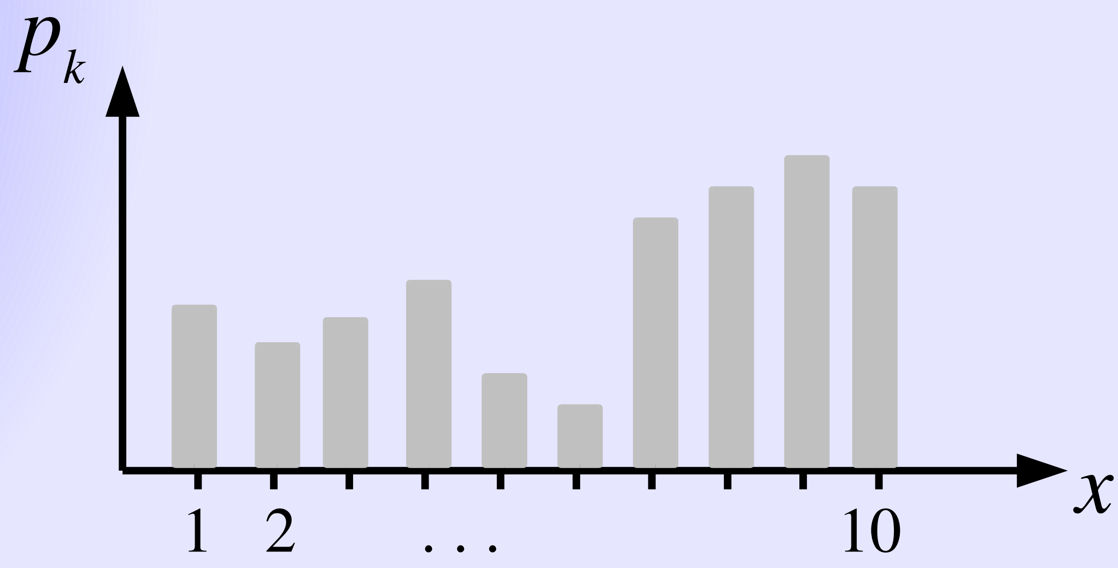
- Difficult:
  - N<CK
  - Uniformly consistent estimator for H exists, if N/(CK) is small, **but** N has to be large (Paninski, 2004).
  - For large CK, and small N/(CK) , Bayesian estimators of H can be heavily biased (Nemenman et al., 2003)

What is the interesting range ?

- In neurophysiological experiments, usually
  - N << CK
  - N small
- **Our approach: describe the P_k by a small number of bins.**

The model

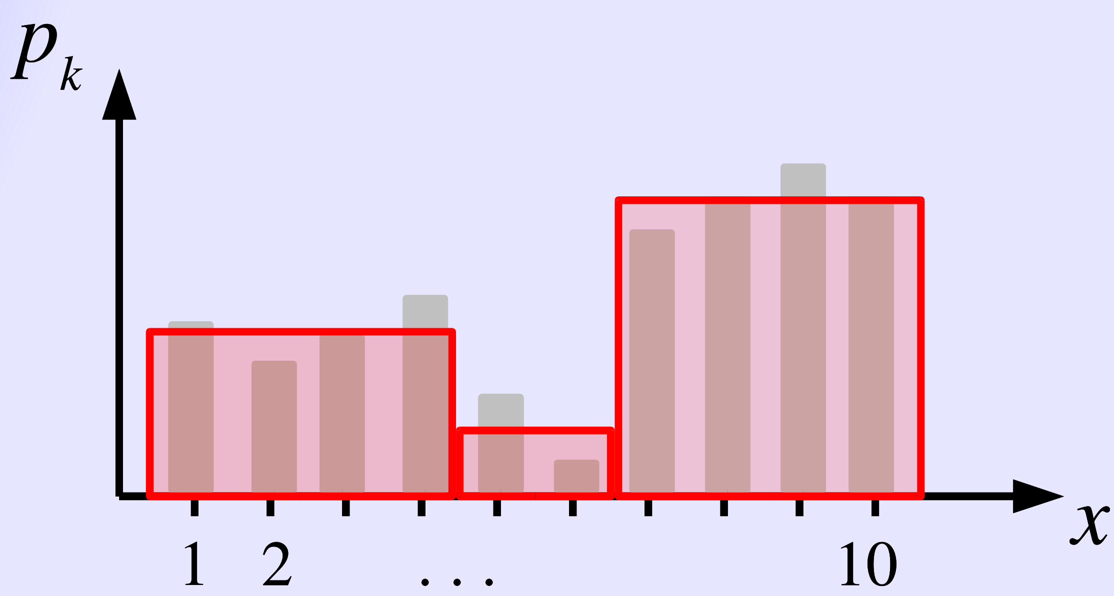
- For simplicity, assume C=1. Extension to C>1 later.
- Assume that instances of X can be **totally ordered**, and their similarity can be measured by x_i-x_j.



- Scale of X: **ordered metric or interval.**

The model

- If neighbouring P_k are 'similar enough', then they can be modeled by a **single** probability P_m, i.e. those P_k can be **binned** together.
- Here: only M=3 instead of K=10 probabilities need to be estimated. **M: number of bins.**



What is 'similar enough' ?

- Evaluate the **posterior probabilities** of all possible binnings **given** an i.i.d. sample of length N via **Bayesian inference.**
- Either pick the model with the highest posterior probability, or integrate out all unwanted parameters to get expectations.

Model parameters

- M: number of bins, K: number of support points of p_k.
- k_m: upper bound (inclusive) of bin m.
- P_m: probability in bin m.

$$\forall\, k_{m-1} < k \leq k_m : \quad \tilde{P}_k = \frac{P_m}{\Delta_m}$$

$$\Delta_m = k_m - k_{m-1}$$

Likelihood of a sample, length N

- D: the sample.
- n_m: number of sample points in bin m.

$$P\left(D \mid M, \{(k_m, P_m)\}\right) = \prod_{m=1}^{M} \left(\frac{P_m}{\Delta_m}\right)^{n_m}$$

Prior assumptions

- All M equally likely, given M≤K.
- All k_m equally likely, given that the bins are contiguous and do not overlap.
- P_m constant (e.g. equi-probable binning) or Dirichlet-prior over $\{P_m\}$.
- P_m and k_m independent *a priori*, and independent of M (except for their number).

Marginal expectations

- Posterior probability of M (e.g. for selecting the best M):

$$P(M|D) = \frac{P(D|M)\,P(M)}{P(D)}$$

$$P(D|M) \propto \sum_{k_1} \sum_{k_2} \ldots \sum_{k_{M-1}} P(D|M, \{(k_m, P_m)\})$$

- Problem: this takes $O(K^M)$ operations !!

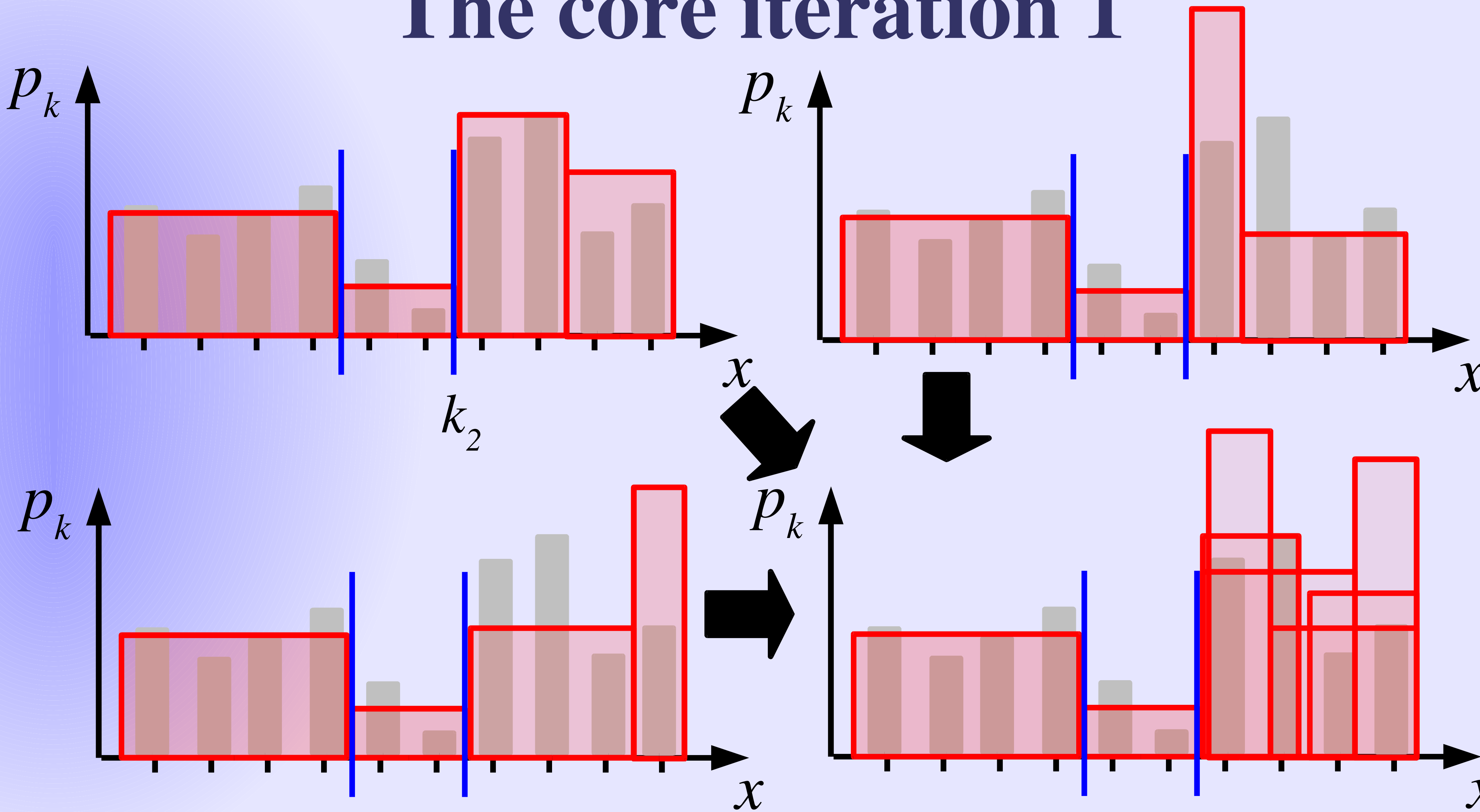
But there is hope...

- The likelihood factorizes into contributions for different bins.
- The k_m are ordered.

$$P(D|M) \propto \sum_{k_1}\left(\frac{P_1}{\Delta_1}\right)^{n_1}\sum_{k_2}\left(\frac{P_2}{\Delta_2}\right)^{n_2}\cdots\sum_{k_{M-1}}\left(\frac{P_{M-1}}{\Delta_{M-1}}\right)^{n_{M-1}}\left(\frac{P_M}{\Delta_M}\right)^{n_M}$$

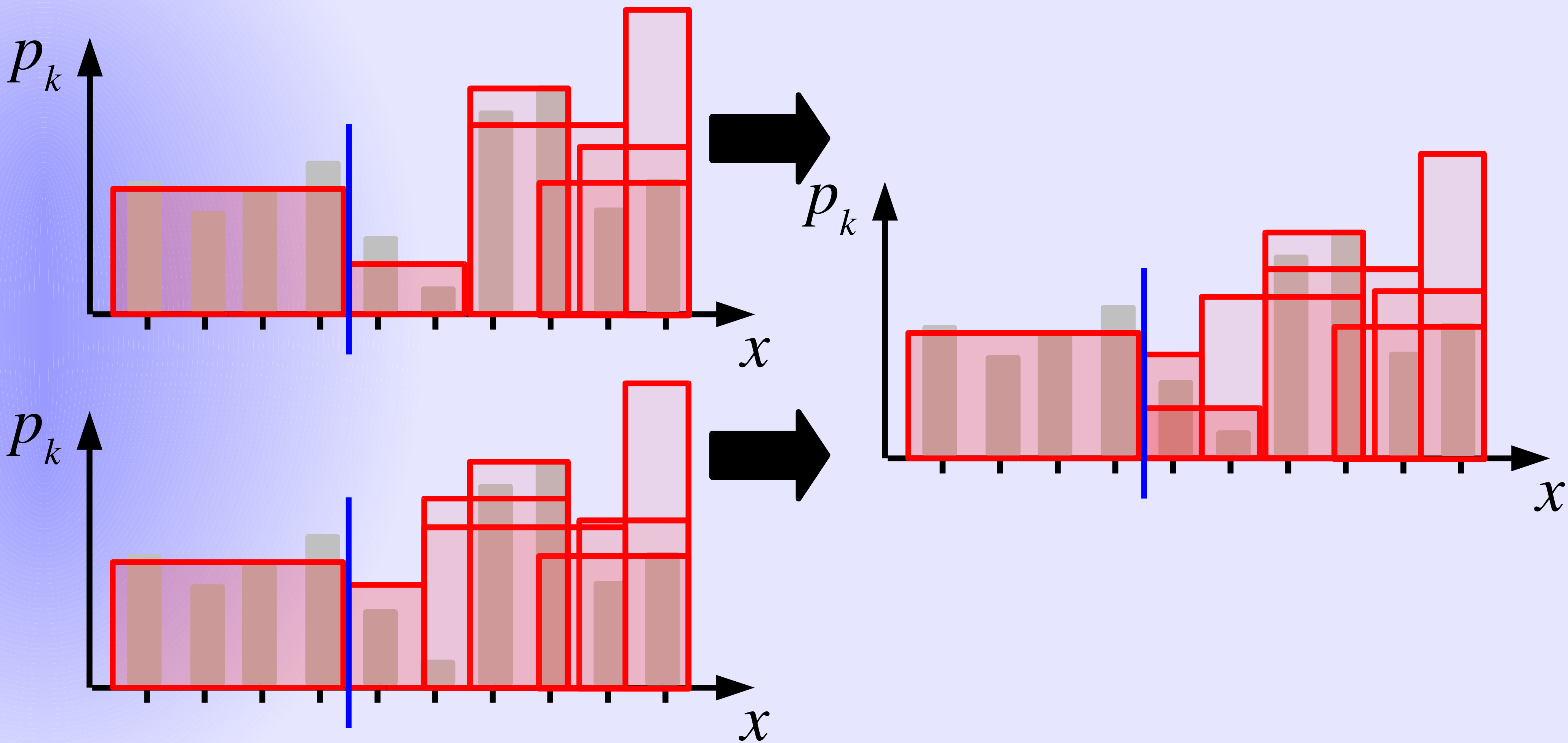
- Similar to sum-product algorithm (Kschischang et al., 2001), dynamic programming (Bellman, 1953).

The core iteration 1



1. For every fixed k_2, compute all contributions from bins 3 and 4, add and store the results: $O(K^2)$ operations, $O(K)$ memory.

The core iteration 2



2. Release k_2, and repeat the procedure for every fixed k_1: $O(K^2)$ operations. Reuse memory for new sub-results.

The algorithm

- Total computational cost: $O(MK^2)$ instead of the naïve $O(K^M)$.
- Instead of fixed $\{P_m\}$, a Dirichlet prior over $\{P_m\}$ can also be used:

$$p\left(\{P_m\} \mid M\right) \propto \prod_{m=1}^{M} P_m^{\theta-1} \delta\left(\sum_{m=1}^{M} P_m - 1\right)$$

- Advantage: P_m can be distributed freely across the bins.

The algorithm

- With a Dirichlet prior, we find

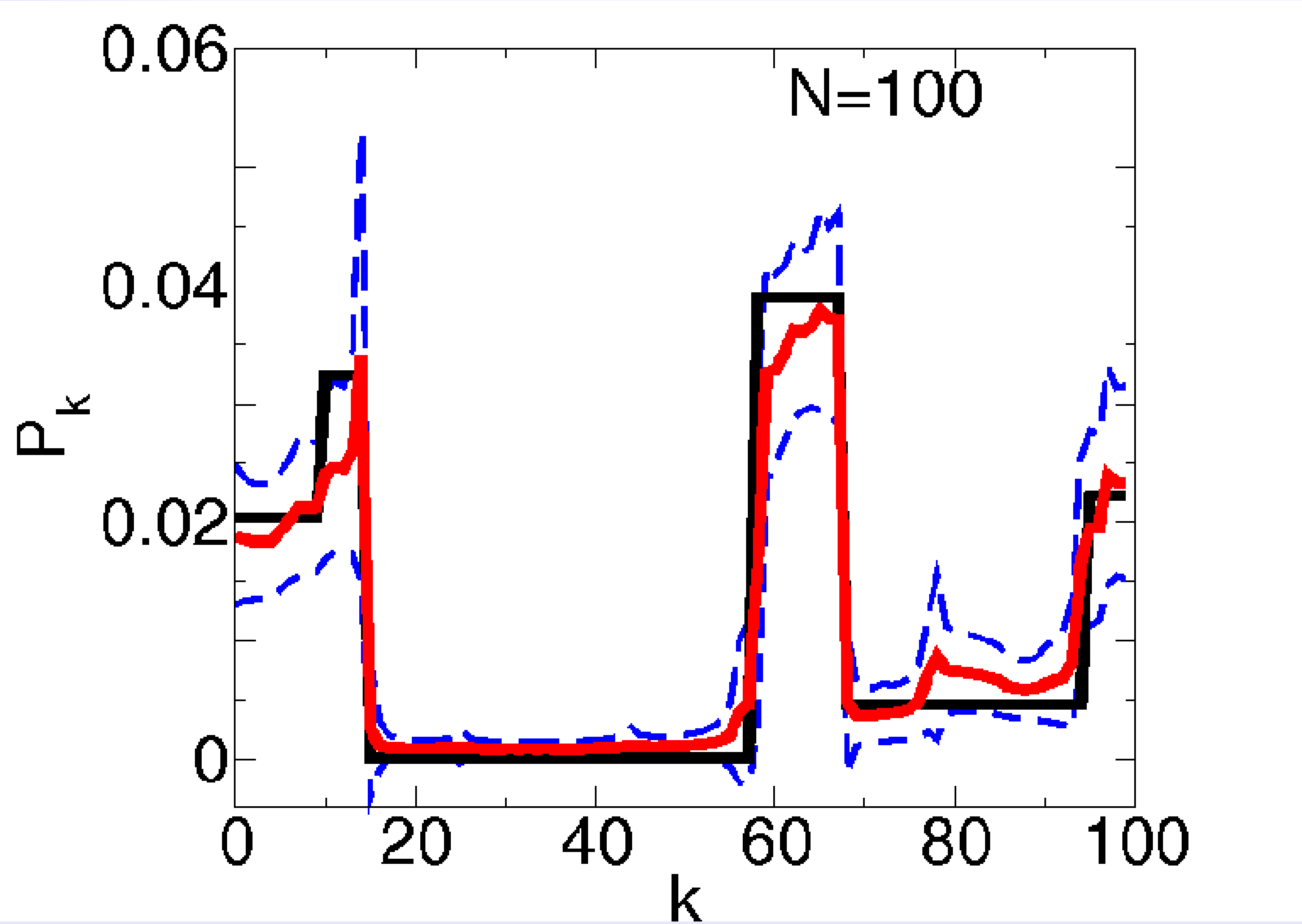
$$P(D|M,\{k_m\}) \propto \frac{\Gamma(M\theta)}{\Gamma(N+M\theta)} \prod_{m=1}^{M} \frac{\Gamma(n_m+\theta)}{\Delta_m^{n_m}\Gamma(\theta)}$$

- One factor per bin which depends **only** on the parameters of that bin.
- Thus, the same sum-product decomposition as before can be applied.

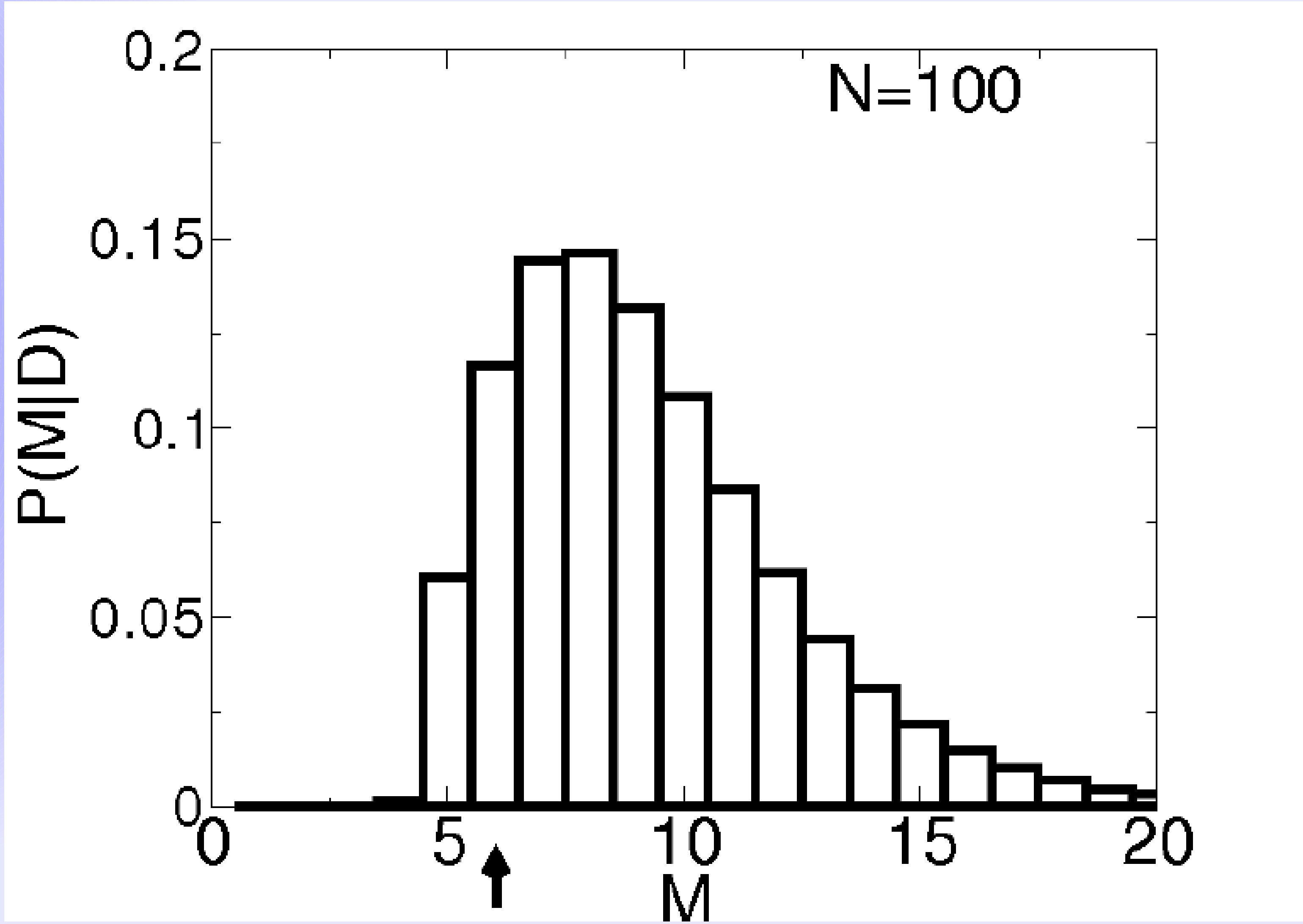
Computable expectations

- Instead of the marginal likelihood $P(D|M)$, we can also compute
  - The expectation of any function of X.
  - The expectations of various functions of the probabilities in the bins and the bin boundaries, such as
    - the predictive distribution and its variance,
    - the expected bin boundaries,
    - the entropy of X and its variance.

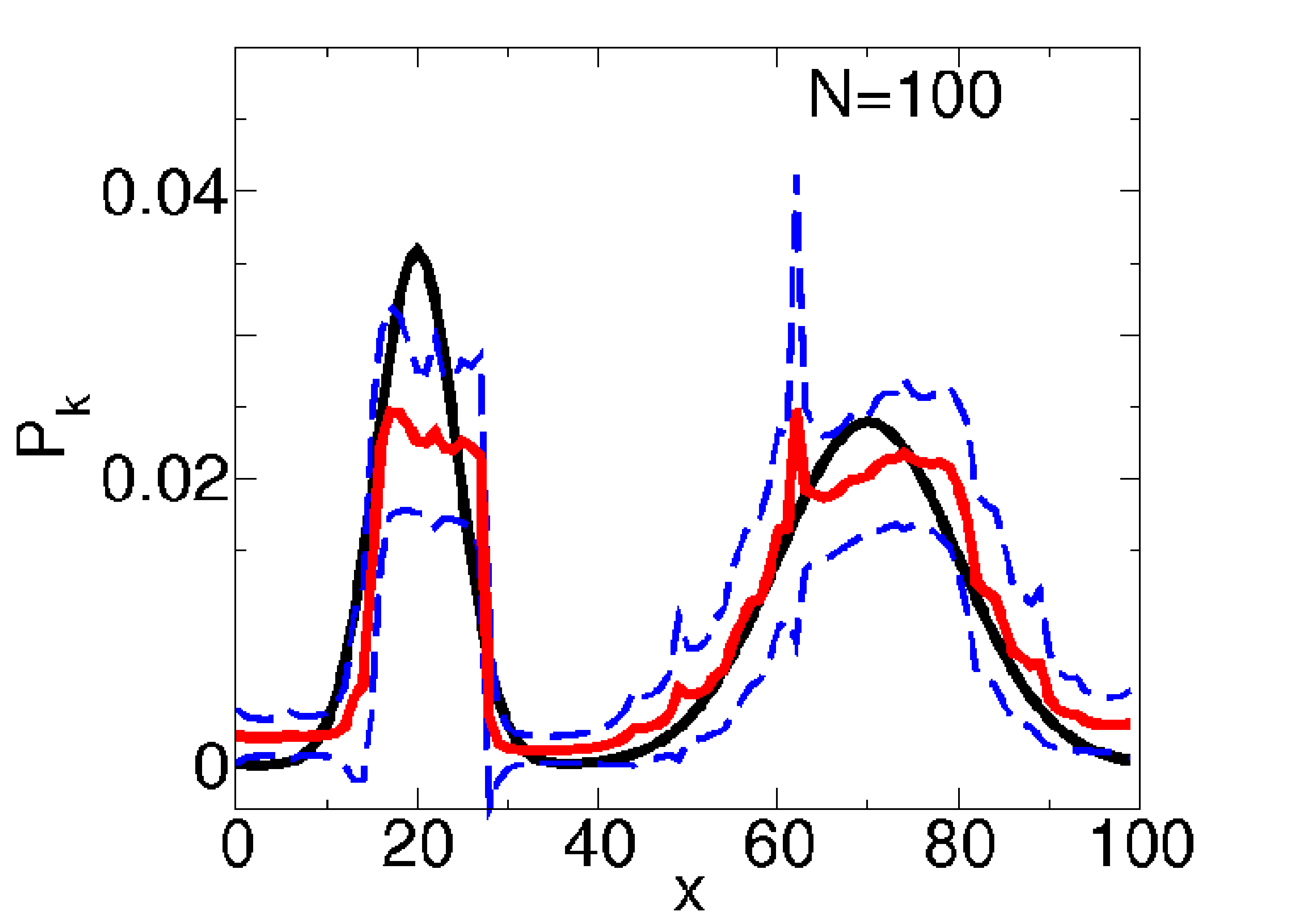
Predictive distribution



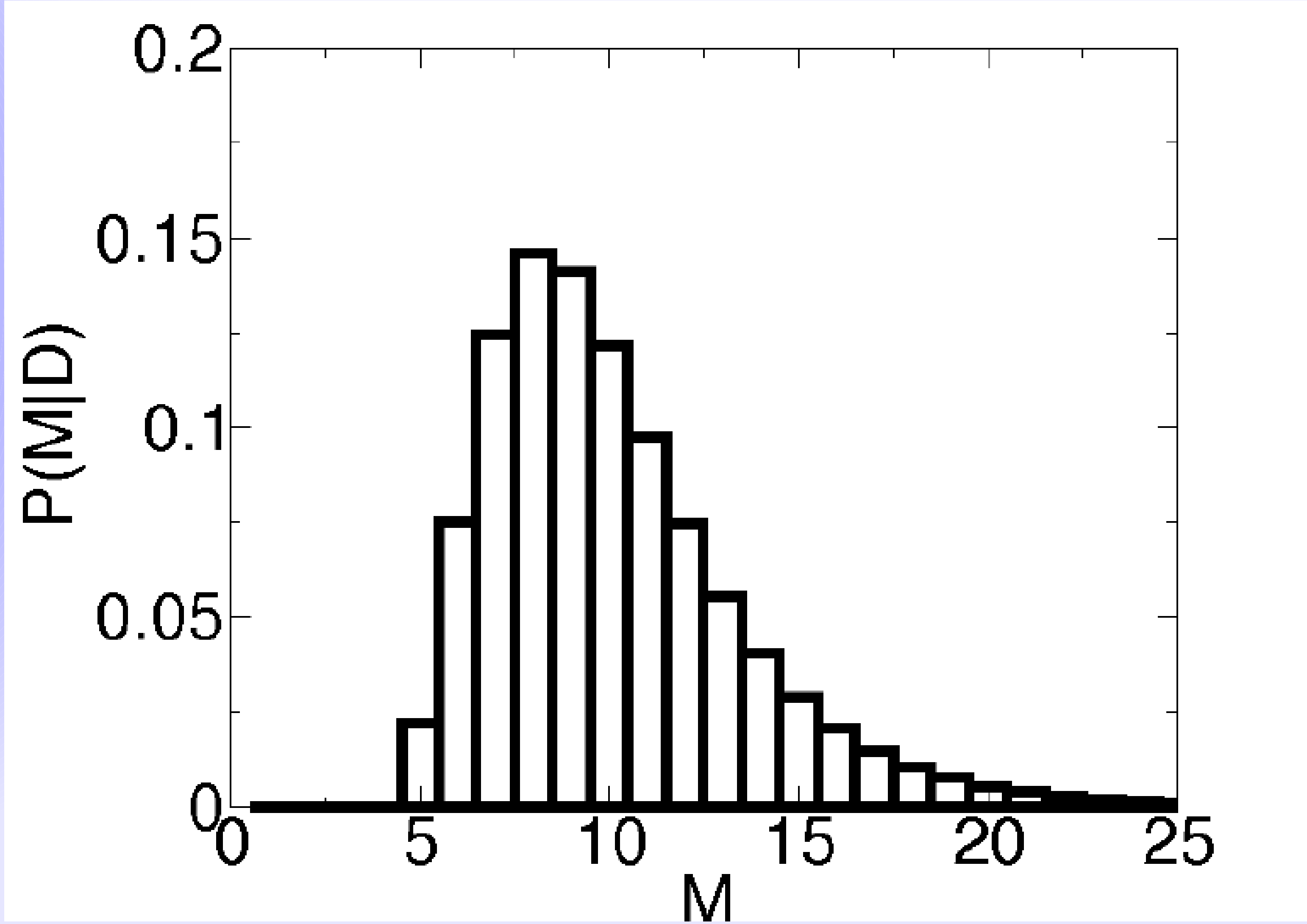
Posterior of the number of bins M



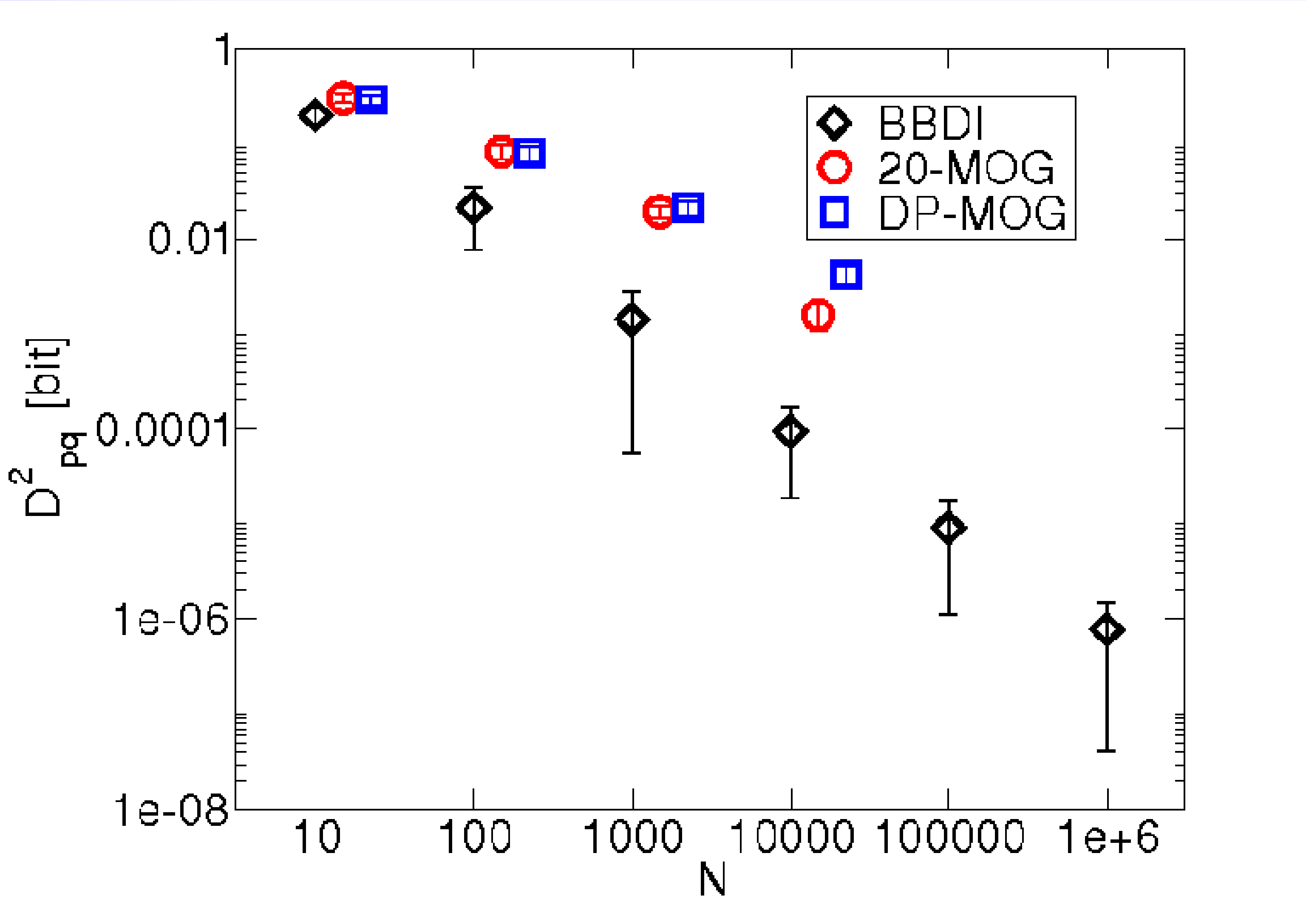
Predictive distribution



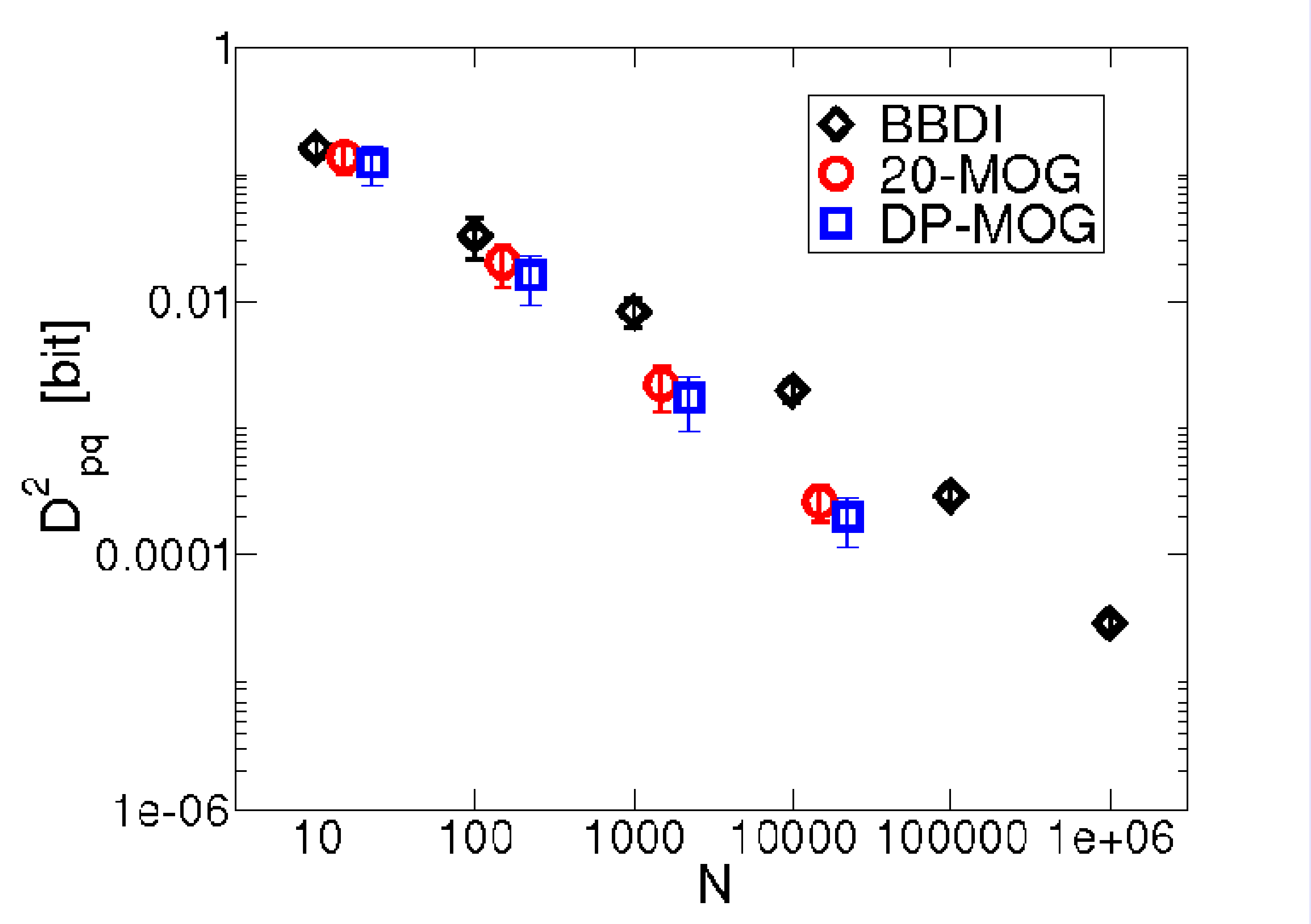
Posterior of the number of bins M



JS-distance between predictive and generating distributions – 6 bin distribution

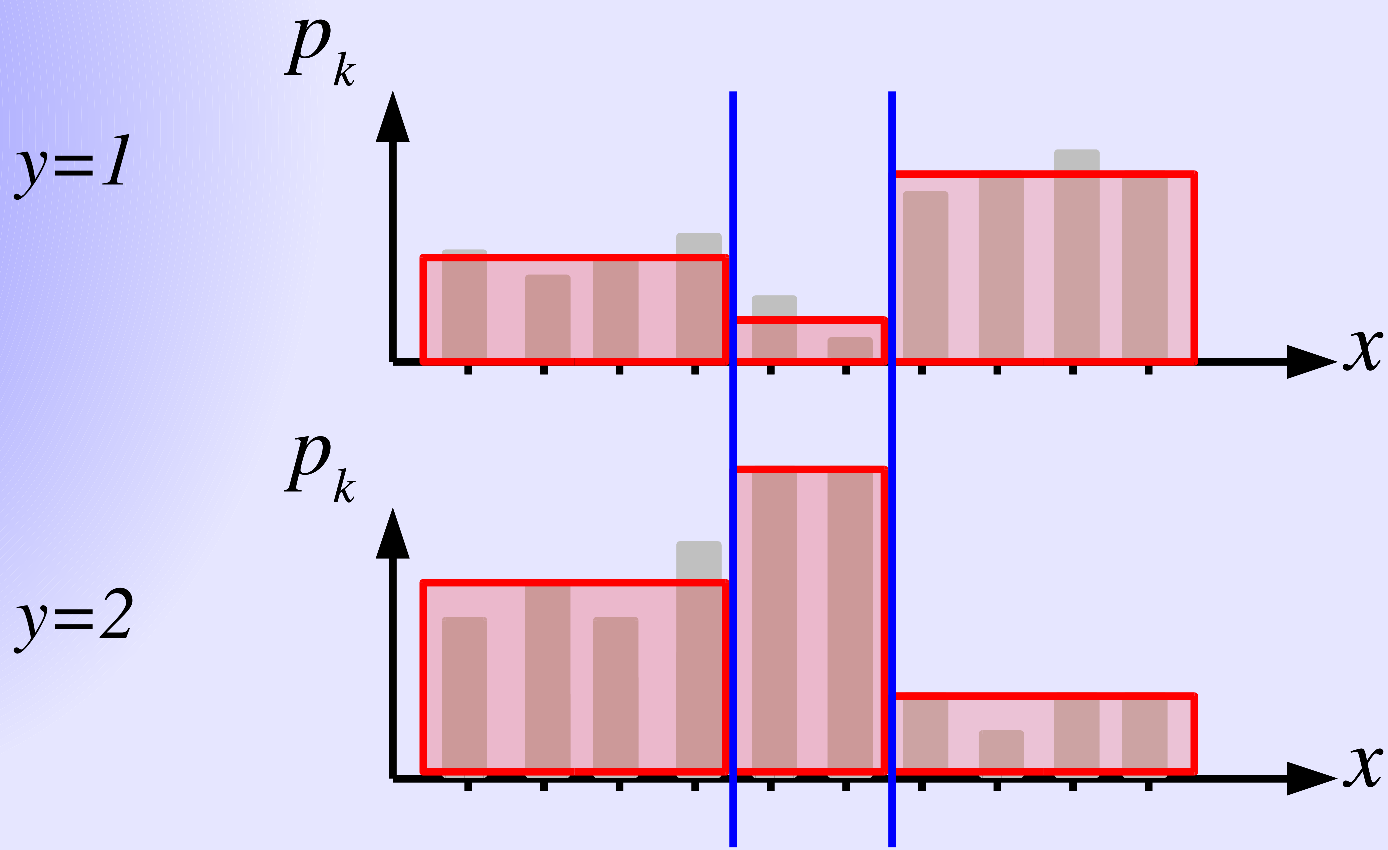


JS-distance between predictive and generating distributions – 2 MOG



Extension to C>1

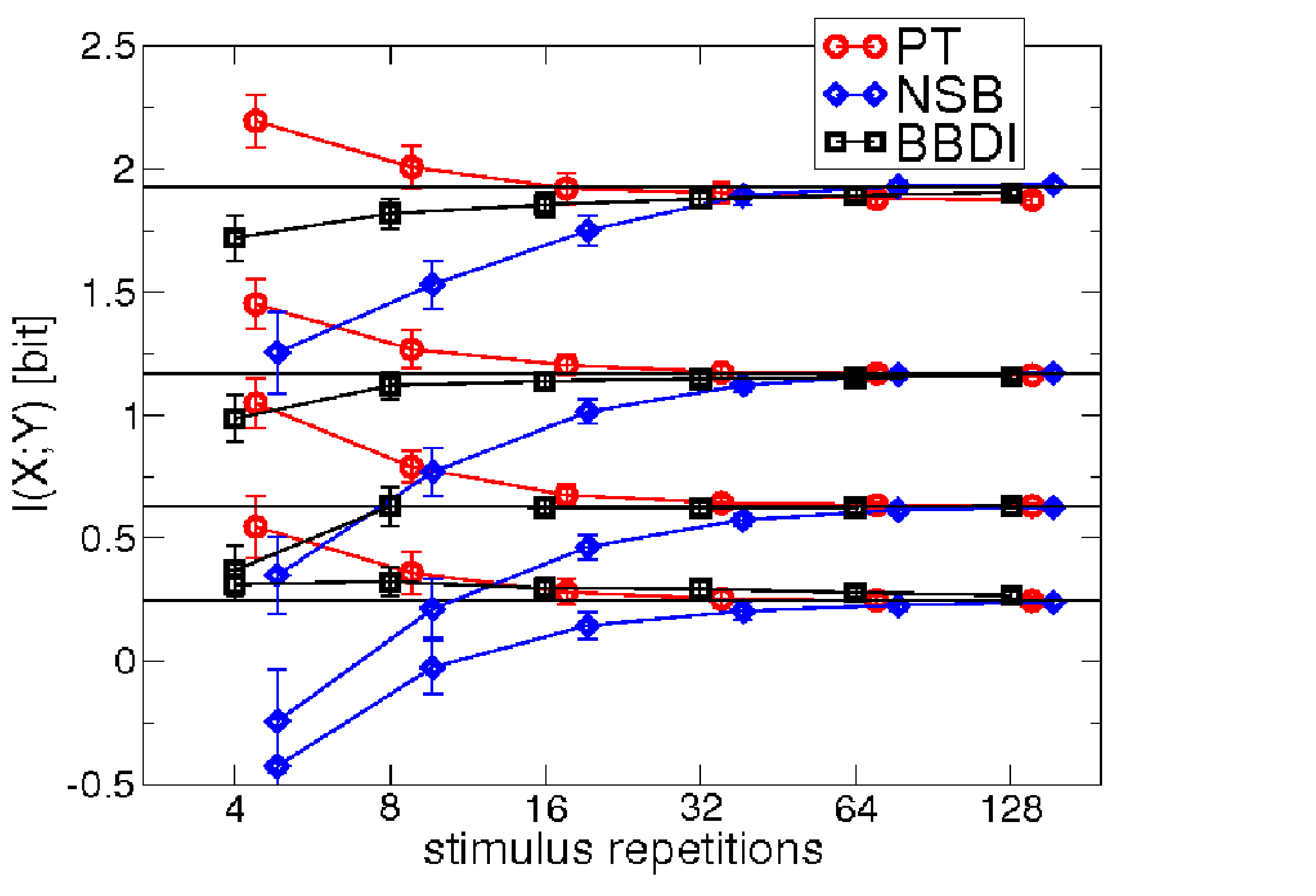
- The extension to C>1 is straightforward, given that the bin boundaries are the same for each *y*



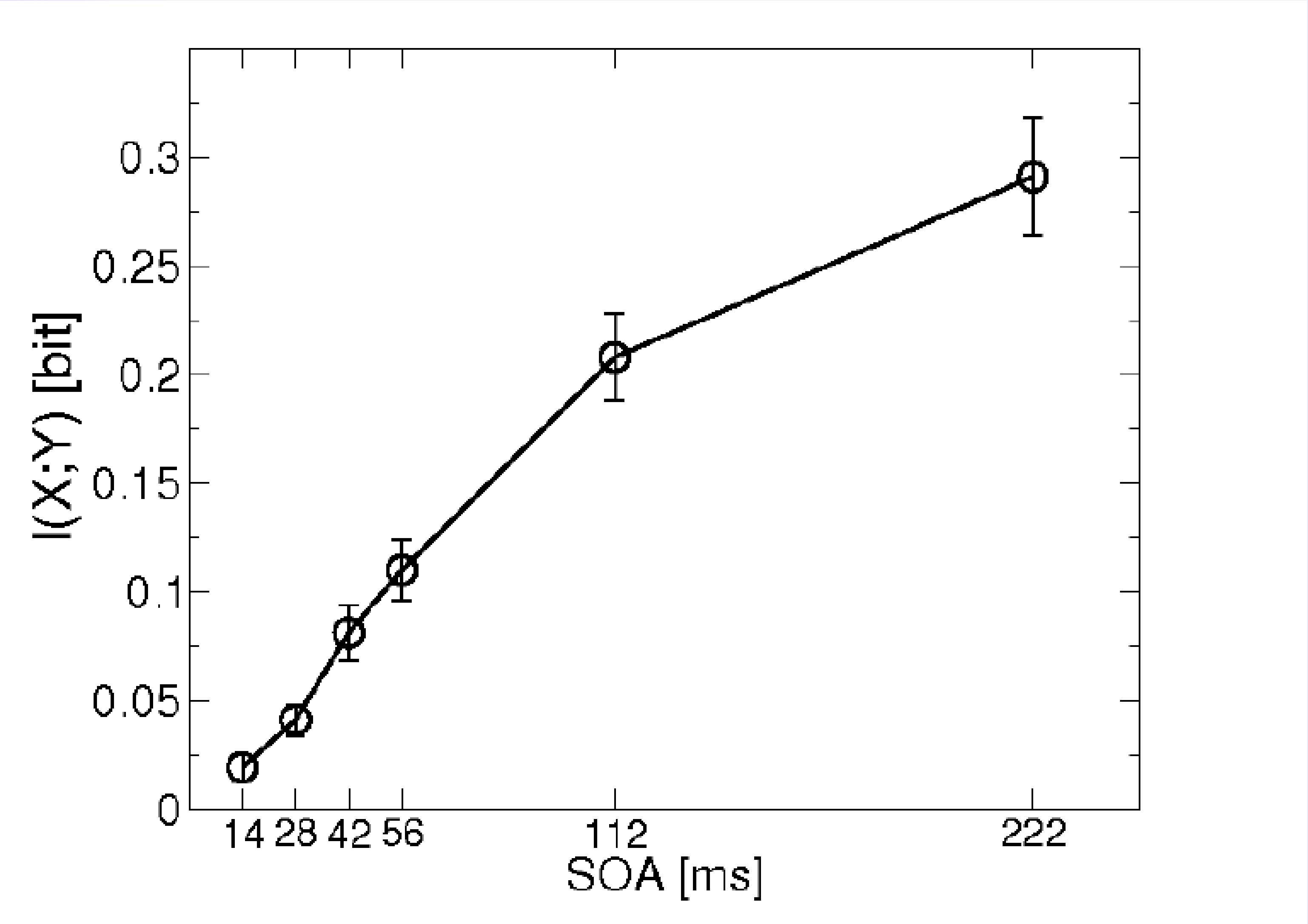
Computable expectations

- Joint entropy H(X,Y) and variance.
- Marginal entropies H(X), H(Y) and variances.
- Mutual information I(X;Y)=H(X)+H(Y)-H(X,Y).

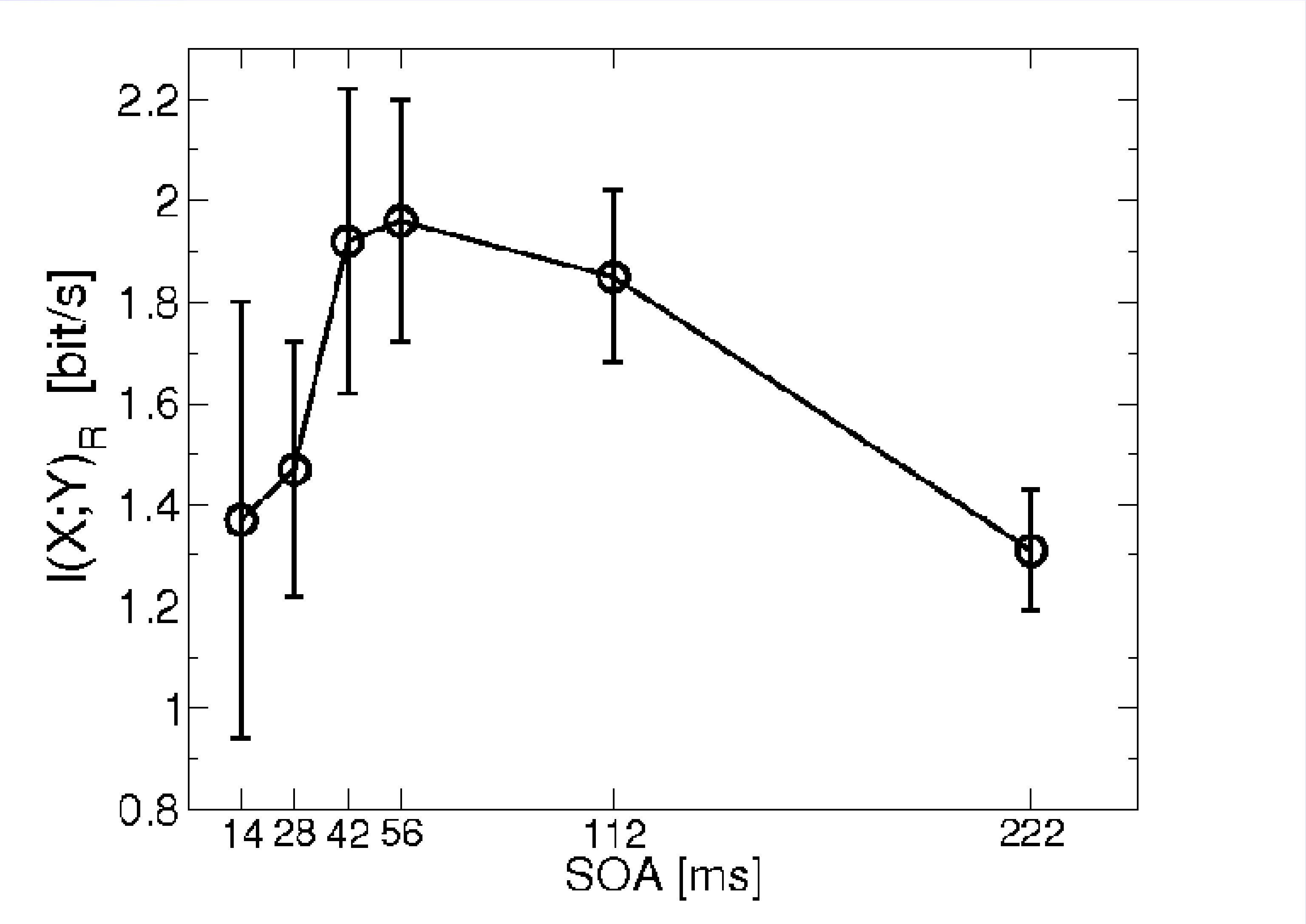
Comparison of I(X;Y) estimates to NSB and Panzeri-Treves



RSVP results



RSVP results



Upper bound on the variance of I(X;Y)

- Computing the variance of I(X;Y) is difficult.
- Simulations suggest:

$$Var\left(I(X;Y)\right) \leq Var\left(H(X,Y)\right)$$ for Dirichlet p(oste)riors

Bayesian Bin Distribution Inference

- Runtime $O(K^2)$ instead of $O(K^M)$ for *exact* Bayesian inference, if instances of X are totally ordered.
- Computable expectations: predictive distribution and variance, entropies and variances, expectation of mutual information.
- Available at:
  - D. Endres and P. Földiák, "Bayesian Bin Distribution Inference and Mutual Information", *IEEE Trans. Inf. Theo.*, 51(11), pp. 3766-3779, 2005.
  - http://www.st-andrews.ac.uk/~dme2