Toy models for validation of Bayesian inference of free energy and path selection

Alex H. Barnett

June 3, 2021

Abstract

We need to be able to test ideas for a Bayesian string method for path selection given cryo particle images. Here are some toy models and their tests.

Some standard Bayesian principles:

1. A model for what processes create the observations should be built, then unknowns in this model should be treated as random variables (if possible), and their posterior be sampled via inference on this same model.

2. Validation should be done from data simulated from the same model, to ensure that the inference procedure is consistent (the true values fall within the inferred posterior ranges). This is the probabilistic equivalent of testing a numerical algorithm on a known solution.

Note that the idea of extracting a 1D path from a higher-dim conformation space doesn't match principle 1, since the model (higher-dim conformation samples) differs from that which will be fit (conformations restricted to a 1D path). To do: fix this, maybe by allowing the

1 Warm-up toy problem: direct conformation observation

Let $x \in \mathbb{R}^n$ be the conformational space. If $n = 1$ then we have an equivalent problem to inference of a free energy function $G(x)$ of a single variable x (called s in our paper). If $n = 2$ (typical), then $x = (x_1, x_2)$, for instance (CMA,CMB). For general n, the model is: $G(x)$ is an unknown function over some region $x \in \Omega \subset \mathbb{R}^n$, assumed smooth. Conformations x_i are iid sampled from the Boltzmann (=Gibbs) density. Ie, $x_1, \ldots, x_N \sim \rho$, where

$$\rho(x) = Z^{-1} e^{-\beta G(x)} , \qquad x \in \Omega , \tag{1}$$

where $Z := \int_\Omega e^{-\beta G(x)} dx$. This distribution model basically defines the concept of free energy (including for a $n = 1$ path). The goal is to recover G, up to an additive constant, equivalent to recovering ρ.

Say we observe x_i directly. A probabilistic graphical model is a useful way to summarize this; see Fig. 1(a). Then we have a *density estimation* problem for ρ, very standard in statistics. There are nonparametric methods: histogram the samples into predetermined bins, smooth them with kernels, etc. This needs a lot of samples for $n > 1$ dimensions. There are parametric methods where ρ stays within a family, whose parameters are inferred (eg, a max-likelihood point estimate, or Bayesian posterior over parameters).

Parametric and nonparametric blur: Eg, histogramming may be viewed as a max-likelihood point estimate for a piecewise-constant parameterization of ρ. Eg, there are Bayesian kernel density methods (Sibisi–Skilling 96).

Todo: read Dirichlet processes, read Kathleen Rey 18 thesis whose GOLD method infers G above as a GP.

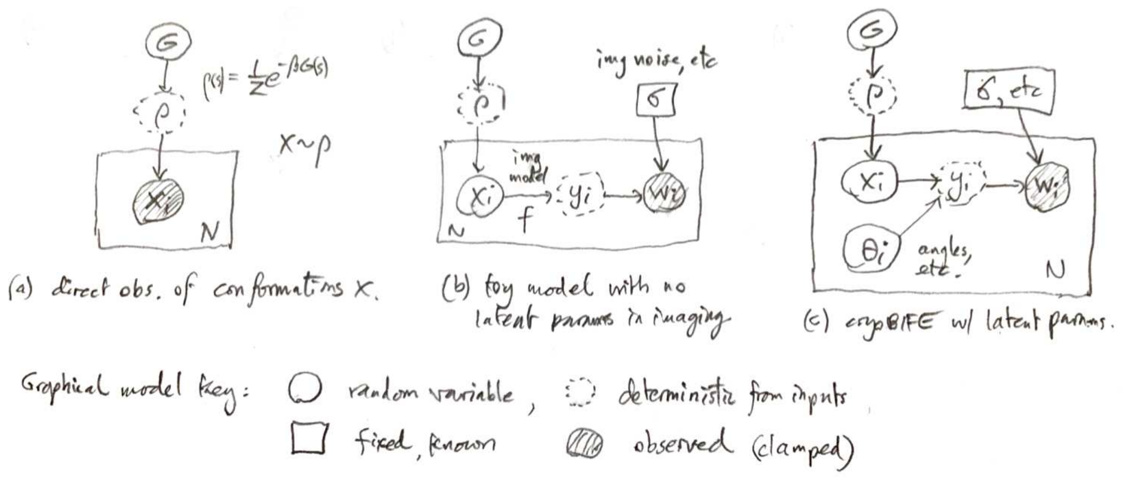

Figure 1: Probabilistic graphical models for toy and real problems. A box with N in the corner indicates iid repeated measurements $i = 1, \ldots, N$. See above for other key. Note that G and ρ are functions not just variables. (a) Direct observation of conformations. (b) Indirect observation of "images" given by a deterministic function $f(x)$ of the conformation plus noise. (c) Cryo-BIFE similar to (b) but with latent imaging parameters (θ is angles, translations, intensity offsets).