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

Alex H. Barnett

June 23, 2021

Abstract

We need to be able to test ideas for a Bayesian string method for path selection given cryo particle images. We also need to be able to understand why Pilar+Julian's path selection is giving strange results. Here are some toy models and their proposed tests.

Some standard Bayesian principles:

1. A model for what processes create the observations (data) should be built, then unknowns in this model should be treated as random variables (if possible), and their posterior is then prescribed by this same model given the data. The posterior on the unknowns should be sampled, or fit (via VI), or summarized (its MAP peak found, etc).

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 already 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: consider a fix by allowing a variable transverse covariance matrix to be fit along with the path—why is this not of interest?

1 Warm-up toy problem: direct observation of conformations

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), except its simpler because there's no projection down from higher dimensions. If $n = 2$ (typical for Pilar's group), then x could be (x, y), for instance (CMA,CMB For general n, the model is: $G(x)$, $x \in \Omega$ is an unknown smooth function over some known region $\Omega \subset \mathbb{R}^n$. Let $\beta > 0$ be the dimensionless inverse temperature. Conformations x_i are iid sampled from the Boltzmann (=Gibbs) density. Ie, $x_1, \dots, x_N \sim \rho$, 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. The goal is to recover G in Ω, up to an additive constant, equivalent to recovering ρ. We are not trying to restrict things to a path here, just work in the ambient dimension n that we believe captures the configurations.

The reason we can do this is that in this toy model we observe the data $D := \{x_i\}$ 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, a standard task in statistics.

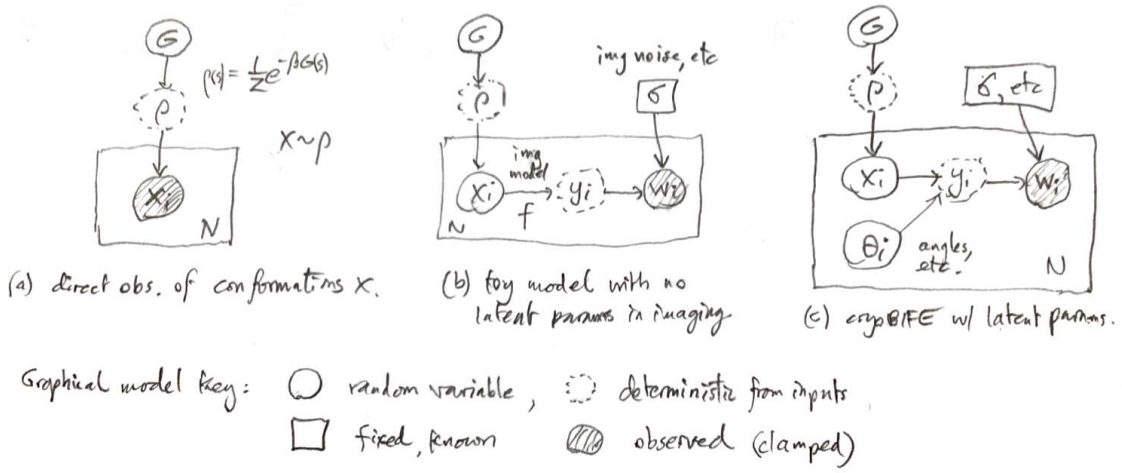

Figure 1: Probabilistic graphical models for toy and real problems. A box with N in the corner indicates iid repeated measurements $i = 1, \ldots, N$. See above for other key, which is standard in PGMs, apparently. Note that G and ρ are functions, not just variables. (a) Direct observation of conformations. (b) Indirect observation of "images" (vectors) given by some 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, etc).

Remark 1 (density estimation)**.** *There are nonparametric methods: histogram the samples into predetermined bins, smooth them with kernels, etc. This needs a lot of samples if we're in $n > 1$ 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). There is no hard line between parametric and nonparametric: eg, histogramming may be viewed as a max-likelihood point estimate for a piecewise-constant parameterization of ρ. Also there are Bayesian kernel density methods (eg Sibisi–Skilling '96). Todo: read Dirichlet processes.*

Remark 2 (no path optimization)**.** *To be clear, we don't get to test path optimization here, since if conformations come from $\mathbb{R}^n$, for $n > 1$, they are generally inconsistent with any 1D path without an intermediate noise model as in the next section.*

The likelihood is $p(D|G) = \prod_i \rho(x_i)$, so its negative log is

$$-\ln p(D|G) = \sum_i \ln \rho(x_i) = N \ln Z + \beta \sum_i G(x_i) \ . \tag{2}$$

Let's parameterize the free energy [1] with p fixed basis functions, in the $n = 1$ dim case on $[0, 1]$,

$$G(x) = \sum_{j=1}^{p} \alpha_j \phi_j(x) \ , \qquad 0 \le x \le 1 \ . \tag{3}$$

Then for the normalizing constant we use a Q-node quadrature approximation with nodes x_q and

[1] This is similar to GOLD method from Kathleen Rey '18 thesis, but there G is a GP.

weights v_q, $q = 1, \ldots, Q$,

$$Z = \int_0^1 e^{-\beta \sum_j \alpha_j \phi_j(x)} dx \tag{4}$$

$$\approx \sum_{q=1}^{Q} v_q e^{-\beta \sum_j \alpha_j \phi_j(x_q)} = \mathbf{v}^T \exp(-\beta B \boldsymbol{\alpha}) \tag{5}$$

where $\mathbf{v} := \{v_q\}_{q=1}^Q$ and $\boldsymbol{\alpha} := \{\alpha_j\}_{j=1}^p$ are column vectors, and $B \in \mathbb{R}^{Q \times p}$ a fixed basis-quadrature matrix with entries $B_{qj} := \phi_j(x_q)$. Here exp is taken to act elementwise.

In terms of the coefficient vector $\boldsymbol{\alpha}$, then, up to quadrature error,

$$-\ln p(D|G) \approx N \ln[\mathbf{v}^T \exp(-\beta B \boldsymbol{\alpha})] + \beta \mathbf{d}^T \boldsymbol{\alpha} =: L(\boldsymbol{\alpha}) \tag{6}$$

where $\mathbf{d}$ is a data-basis vector with entries $d_j := \sum_i \phi_j(x_i)$ that needs computing only once.

We want the posterior $\pi(G|D)$. Assuming a flat prior over $\boldsymbol{\alpha}$, the negative log posterior $U(\boldsymbol{\alpha}) := -\log \pi(G|D) = -\log p(D|G) + \text{const}$, which is (6) up to a constant. U becomes the potential function in MCMC sampling over $\boldsymbol{\alpha}$. The MAP estimate is $\boldsymbol{\alpha}_* := \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} U(\boldsymbol{\alpha})$. One could sample from the posterior $\pi(\boldsymbol{\alpha}) := e^{-U(\boldsymbol{\alpha})}$ by HMC, then summarize by iid samples or posterior mean, etc. There will be numerical speedups by exploiting locality of basis functions, or properties of exp.

To do: code up HMC here, using either ϕ_j continuous piecewise linear hat-functions, or something higher-order but still well-conditioned. Conditioning is crucial for success of MCMC. Check the posterior quantiles contain the true value the expected fraction of the time.

Remark 3 (Additive constant in G)**.** *Although adding c to G multiplies Z by $e^{-\beta c}$, in a predictable way, it is in general not true that G having zero mean implies $Z = 1$. Pilar is right that Z must be computed to perform MCMC correctly.*

2 Indirect observations of conformations via noisy "images" without latent parameters

Here we use "image" in the abstract to mean a vector $w \in \mathbb{R}^P$ derived from a conformation x with a known conditional pdf $p_{\text{im}}(w|x)$, which is the imaging plus noise model (likelihood). For an image, P is the number of pixels. The full model is now $x_i, \ldots, x_N \sim \rho$, iid, then $w_i \sim p_{\text{im}}(\cdot|x_i)$ for each i. Only w_i are observed, not x_i, ie the data is $D := \{w_i\}_{i=1}^N$.

For any additive-noise imaging model we have the likelihood

$$p_{\text{im}}(w|x) = p_{\text{noise}}(w - f(x))$$

where $y = f(x)$ is the deterministic *forward model* for the image. p_{noise} could be a Gaussian with zero mean and variance σ^2, another known parameter. See Fig. 1(b). Note that additive doesn't cover all cases of $p_{\text{im}}(w|x)$ such as a Poisson noise model.

The toy idea is to make P small, eg, $P = n$ (=1 or 2), and test consistency for a simple map f, which could be the identity map. This would also allow tests of path-selection methods, where the 1D path is selected in $\mathbb{R}^2$. *To do: write down a string method where the max likelihood for a given set of nodes has its negative log taken to give a "potential" for that string. Its grad wrt each node gives a "force" on that node. The string has a time-step taken with this force, then is re-interpolated to high order onto an equi-arc-length parameterization, as in the "improved" string method paper, or to Chebychev nodes.*

We first derive the posterior $\pi(G|D)$.

3 Observation of noisy images which also have latent (nuisance) parameters

The latent parameters summarized by θ include: rotation in SO(3), translation, and any other. Each particle image i gets a different draw θ_i. See Fig. 1(c). This is as in cryo-BIFE. It has the complication that θ must be marginalized over.