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

Alex H. Barnett

June 24, 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 with simulated data (eg, not finding lowest-energy paths). To think clearly, one needs to set up toy models, and test that known parameters can be inferred correctly by a proposed method. Here are some toy models and their proposed tests. One conclusion is that, since the generative model has configurations living in a higher-dim space, then any attempt to infer using configurations restricted to a 1D path will be inconsistent.

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). This is called inference on the model given the data.

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 and making sure you get the right answer.

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—this would at least be closer to the believed generative model for configurations.

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); however there is again no projection down to a 1D path. For general n, the model is: $G(x)$, $x \in \Omega$ is an unknown smooth function over some known region $\Omega \subseteq \mathbb{R}^n$. Let $\beta > 0$ be the dimensionless inverse temperature. Conformations x_i are iid sampled from the Boltzmann (= Gibbs) density. Ie, $x_1, \ldots, x_N \sim \rho$, where

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

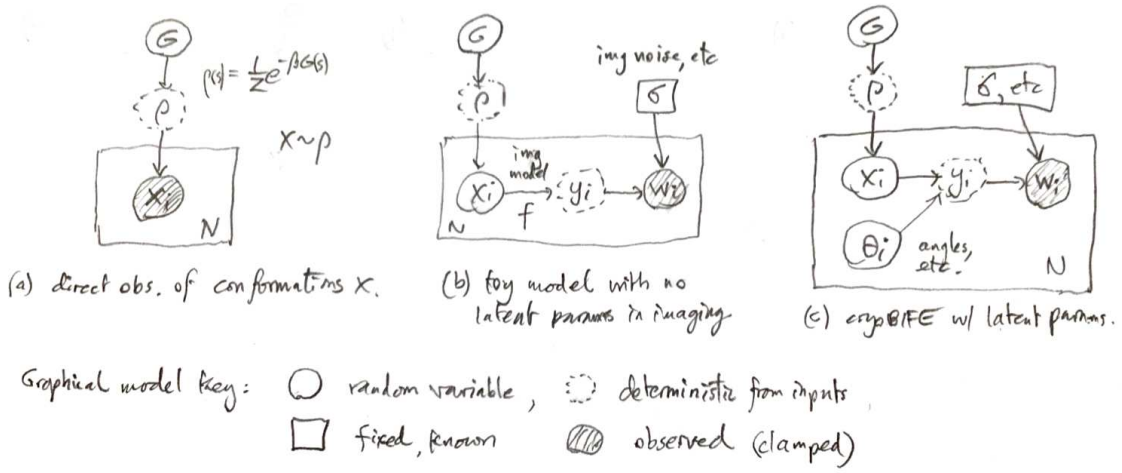

Figure 1: Probabilistic graphical models for toy and real problems. A box with N in the corner indicates iid repeated measurements $i = 1, \ldots, N$. See above for other key, which is standard in PGMs, apparently. Note that G and ρ are functions, not just variables. (a) Direct observation of conformations.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

where $Z := Z_G := \int_\Omega e^{-\beta G(x)} dx$. 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.

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 1D path)**.** *To be clear, we don't get to test fitting G on a 1D path, or doing 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. Pilar and Erik have said that learning G (ie ρ) over a $n = 2$ dim or higher is not of interest like learning it on a 1D path—I am not sure why not.*

Anyway, let's finish the description of inferring G in the full space. 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 weights v_q, $q = 1, \dots, Q$,

$$\begin{aligned} Z &= \int_0^1 e^{-\beta \sum_j \alpha_j \phi_j(x)} dx & (4) \\ &\approx \sum_{q=1}^{Q} v_q e^{-\beta \sum_j \alpha_j \phi_j(x_q)} = \mathbf{v}^T \exp(-\beta B \boldsymbol{\alpha}) & (5) \end{aligned}$$

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 $\pi(G|D)$, more concretely $\pi(\boldsymbol{\alpha}|D)$ because one must work in some basis for G. Assuming a flat prior over $\boldsymbol{\alpha}$, the negative log posterior $U(\boldsymbol{\alpha}) := -\ln \pi(G|D) = -\ln p(D|G) +$ 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 $x \in \mathbb{R}^n$ 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 generative model is now

$$x_i, \dots, x_N \sim \rho, \text{ iid, over } \mathbb{R}_n, \text{ then} \qquad w_i \sim p_{\text{im}}(\cdot|x_i), \quad i = 1, \dots, N \ . \tag{7}$$

Only the data $D := \{w_i\}_{i=1}^N$. are observed, not the set of conformations $X := \{x_i\}_{i=1}^N$.

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, given a conformation x. 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.

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

The toy idea is to make P small, eg, $P = n$, and test consistency for a simple map f, which could even be the identity map. This would also allow tests of path-selection methods, where the 1D path is selected in $\mathbb{R}^2$.

Say we use the *same model* for inference (as in principle 1 at the top), then we are not allowed to restrict to a 1D path. We derive the posterior. The model means the joint pdf factorizes as

$$p(D, X, G) = p(D|X)p(X|G)p(G) \tag{8}$$

Marginalizing over unobserved configurations X gives

$$p(D, G) = \int_{x_1 \in \mathbb{R}^n} \cdots \int_{x_N \in \mathbb{R}^N} p(D, X, G) dx_1 \dots dx_N \ . \tag{9}$$

So, also using the iid nature of (7), the posterior is generally

$$\pi(G|D) \propto p(D, G) = p(D|G)p(G) = p(G)Z^{-N} \prod_i \int_{\mathbb{R}^n} p_{\text{im}}(w_i|x_i) e^{-\beta G(x_i)} dx_i$$

One could insert a quadrature approximation over $\mathbb{R}^n$ as in the previous section. Concretely, in the basis (3), the negative log posterior would then be

$$-\ln p(G|D) \approx -\ln p(\boldsymbol{\alpha}) + N \ln[\mathbf{v}^T \exp(-\beta B \boldsymbol{\alpha})] + \sum_i \ln[\mathbf{l}_i^T \exp(-\beta B \boldsymbol{\alpha})] \tag{10}$$

where for each image $i = 1, \dots, N$ the data likelihood at quadrature nodes vector has entries

$$\mathbf{l}_i := \{v_q p_{\text{in}}(w_i|x_q)\}_{q=1}^Q, \tag{11}$$

which needs computing only once (independent of $\boldsymbol{\alpha}$). This big matrix $p_{\text{in}}(w_i|x_q)$ is like a simpler (no-latent variables) version of p_{BioEM} of Pilar.

That covers inference for G over $\mathbb{R}^n$ using the true generative model. If instead we fix a path $x(s)$, $s \in [0, 1]$, one could write a posterior for G (again in the $\boldsymbol{\alpha}$ basis) in an *incorrect* path model, which is the same as (10)–(11) but with x_q and v_q a quadrature scheme for the 1D path.

Toy test: choose $P = n = 2$, f the identity, and $p_{\text{noise}} \sim \mathcal{N}(0, \sigma^2 I)$, so

$$p_{\text{im}}(w|x) = (2\pi\sigma^2)^{-n/2} e^{-\|w-x\|_2^2/2\sigma^2}$$

Then "images" (data points) w_i are drawn from Gaussian blobs centered on configs x_i. The big matrix $p_{\text{in}}(w_i|x_q)$ is just the Gaussian pairwise interactions of data points with (path) quadrature nodes, cheap to fill. Configs that are further than σ from the path have little effect. I propose this should be tested via MCMC on $\boldsymbol{\alpha}$. If we believe that in BioEM the likelihoods drop off in some Gaussian way as you move a config away (in x space) from the one that produced a particular image, then this might be a good model for the full cryo-BIFE path optimization. It may address some of the weird behavior Julian was seeing in cryo-BIFE path optim.

As I keep reiterating, there is no "ground truth" test possible here where a G is input and checked to be recovered correctly, since the path model used for fitting is not the generative model. One can only check that the G recovered is some σ-blurred version of the true G along the path.

2.1 Path optimization for the path model

We wish to find a "force" vector in $\mathbb{R}^n$ on each quadrature node q on the path. I propose that the *negative log evidence* (NLE) $-\ln p(D)$ for a given path be treated as a potential energy functional for that path, and its grad wrt each node be taken to give the forces. The rest of the improved string

method of E et al '07 proceeds as written. Recall evidence for a model (here, a path and value of β) is the predictive density of the data D in the model (ie marginalizing over unknowns, here, $\boldsymbol{\alpha}$):

$$p(D) = \int_{\mathbb{R}^p} p(\boldsymbol{\alpha}, D) d\boldsymbol{\alpha} = \int_{\mathbb{R}^p} p(D|\boldsymbol{\alpha}) p(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \tag{12}$$

This is also the normalizing denominator for the posterior. As usual we write $\boldsymbol{\alpha}$ and G interchangeably.

However, the NLE isn't always easy to calculate numerically. In the Laplacian approximation it is: say the posterior $p(\boldsymbol{\alpha}|D)$ is nearly Gaussian, then its integral is well approximated by the desntiy at the MAP times the volume factor coming from the det of the Hessian of $-\ln p(\boldsymbol{\alpha}|D)$. That would be saying that G is well-specified by the data enough to be in a linear regime. A detail is that both these factors have to have their grad taken, and it's not obvious how to take the grad of a max (the MAP)—this might involve a Legendre xform or something.

As a cruder variant to warm up with, just use the peak of joint density

$$p(\boldsymbol{\alpha}^*, D) = p(D|\boldsymbol{\alpha}^*) p(\boldsymbol{\alpha}^*)$$

in place of the evidence $p(D)$. For flat prior, this is up to a constant just the max likelihood value $p(D|G^*)$. This is what I've suggested to Pilar's group in the full cryo-BIFE setting.

To finish.

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 of E et al '07, or to Chebychev nodes.

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

The latent parameters summarized by θ include: rotation in SO(3), translation, imaging system unknowns, etc. Each particle image i gets a different draw θ_i. The point really was really to draw Fig. 1(c) and compare its structure to the simpler toy models. This is as in cryo-BIFE. It has the complication that θ must be marginalized over when filling the big matrix $p_{\text{in}}(w_i|x_q)$, which is now p_{BioEM} of Pilar.

To test, a toy set of latent parameters could be used in a slight variant of the toy case from section 2.

To do.

3.1 Path optimization for the path model

To do.

In order to prevent the multi-hour computation of p_{BioEM} each time the path is changed, we have to precompute it to a set of x on a grid covering $\Omega \subset \mathbb{R}^n$, interpolate the log likelihoods from this grid.