Draft version June 14, 2022
Typeset using LaTeX **RNAAS** style in AASTeX63

Inference For Models That Produce Tracks

Benjamin D. Johnson

INTRODUCTION

Notes about inferring the parameters of models that produce tracks in the observable space using noisy observations of individuals in the space. This kind of model arises in the contexts of CMDs (isochrones), chemical evolution, and stellar streams. Definitions:

- D: Data. this could actually be inferred parameters like metallicity, age, effective temperature, or α enhancement. However, care must then be taken to divide out any priors on parameters that appear in θ when performing the inference; this data should represent a (marginalized) likelihood, not a posterior.
- $\mathcal{M}$: Model.
- θ: Model parameters.

We try to use uppercase script for matrices, uppercase for vectors, and lower case for scalars or index variables. Here are the assumptions:

- The track is infinitely thin. That is, in the absence of observational errors all observed data would fall perfectly on a line in J-dimensional space.
- The density *along* the track changes fairly slowly (this could be relaxed at the cost of more complicated integrals)
- The 'observational' uncertainties on the 'data' are described by a multivariate Gaussian.
- The sample selection is not a function of the observables (though this can be relaxed, see §4)

FULL LIKELIHOOD

The idea is to combine the likelihoods for N objects to produce a single likelihood for a given track. For each object i we will *marginalize* over its (unknown) true location along the track. Because each object is independent we can factorize the likelihood over all objects or observations as

$$\begin{aligned}\mathcal{L}(D\,|\,\theta) &= \mathcal{L}(D_1, D_2, \ldots, D_N\,|\,\theta) \\ &= \mathcal{L}(D_1\,|\,\theta)\,\mathcal{L}(D_2\,|\,\theta)\,\ldots\,\mathcal{L}(D_N\,|\,\theta) \\ &= \prod_{i=1}^{N} \mathcal{L}(D_i\,|\,\theta) \end{aligned} \tag{1}$$

INDIVIDUAL OBJECT LIKELIHOOD

Track models generally produce a set of K points $\mathcal{M}(\theta) = \{X_k\}_1^K$ in the J-dimensional space X of the observables. Thus the track can be represented as a $K \times J$ array of coordinates, along with an optional set of *weights* denoted w_k, all of which depend on the parameters θ. It is assumed that any observed datum comes from this track, and was then perturbed by observational noise to produce the observation. However, we don't necessarily know *where* on the track the the data came from, and the track is potentially complicated in shape. What we want is to integrate the likelihood of a particular observed data point along the entire track. This effectively marginalizes over some true, unknown position on the track.

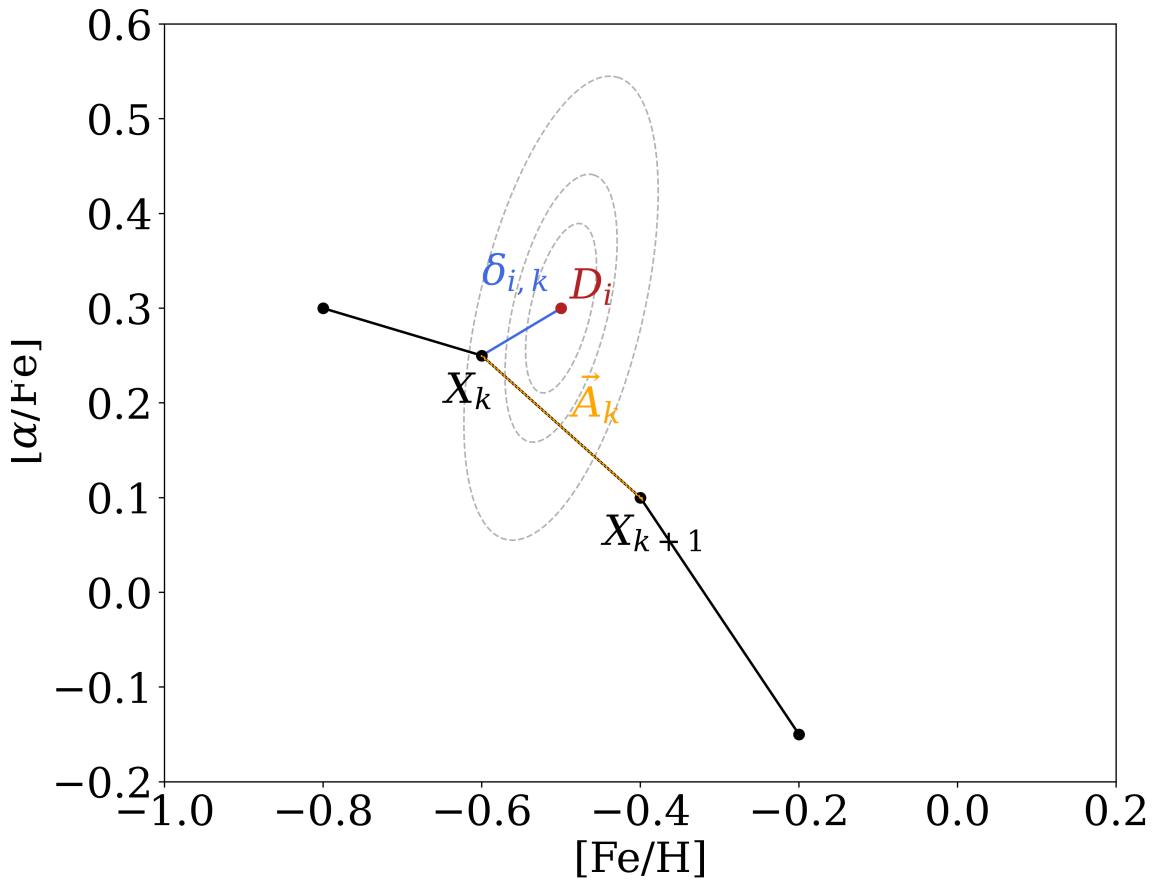

Figure 1: Illustration of the marginalization of the likelihood for a single observation i over the track segment k in a 2-dimensional space. The goal is to integrate the value of the Gaussian indicated in gray dashed ellipses over the line segment A_k. The integrals for every segment are then summed to produce a marginalized likelihood, integrated over the whole line for this single object.

For simplicity we treat the track as piece-wise linear between the points $\mathcal{M}_k$, which allows us to represent the track in a continuous form for each of the $K-1$ segments as

$$\begin{aligned} x_k(q) &= A_k\, q + B_k \\ B_k &= X_k \\ A_k &= X_{k+1} - X_k \\ 0 &< q < 1 \end{aligned}$$

where q is a continuous scalar, X_k is the the coordinate of the kth model point in the N-dimensional space, and x_k describes a line in N-dimensional space. We can then compute the likelihood for each segment and sum to generate the total likelihood of this datum data given the track.

Assuming Gaussian likelihoods, we have the usual thing that for a single 'true' location t along the track represented by k and q the likelihood of the data is

$$\begin{aligned}\mathcal{L}(D_i \,|\, k, q) &= \frac{w_k}{\sqrt{||2\pi\,\mathcal{C}_i||}}\, e^{-\frac{1}{2}\,\Delta_{i,k}{}^{\mathsf{T}}\,\mathcal{C}^{-1}\,\Delta_{i,k}} \\ \Delta_{i,k} &= D_i - x_k(q) \\ &= D_i - (A_k\,q + B_k) \\ &= \delta_{i,k} - A_k\,q\end{aligned}$$

where $\mathcal{C}_i$ is a $J \times J$ covariance matrix for the ith object, q is a scalar, i and k are index variables, w_k is the *weight* for this segment, and everything else is a $N \times 1$ vector. The weights are here because the likelihood that a datum is generated by a particular location on the track is proportional to the density at that location in the track.

Now, to marginalize over the true position on this segment, we should integrate this likelihood over the range $0 < q < 1$. But first we will do out the square and isolate the terms that depend on q:

$$\begin{aligned}\Delta_{i,k}{}^{\mathsf{T}}\,\mathcal{C}_i^{-1}\,\Delta_{i,k} &= \chi^2_{i,k} - 2\,q\,A_k{}^{\mathsf{T}}\,\mathcal{C}_i^{-1}\,\delta_{i_k} + q^2\,A_k{}^{\mathsf{T}}\,\mathcal{C}_i^{-1}\,A_k \\ \chi^2_{i,k} &= \delta_{i,k}{}^{\mathsf{T}}\,\mathcal{C}_i^{-1}\,\delta_{i,k}\end{aligned} \tag{2}$$

Clearly, when $A_k{}^{\mathsf{T}}\,\mathcal{C}_i^{-1}\,A_k$ is small, i.e. when the displacement between track points is small compared to the uncertainties, then the second and third terms become less important and this basically reduces to the χ^2 of each point along the track. However, if we consider the full line segment we need to take the integral over q, in the process making some variable substitutions

$$\begin{aligned}\mathcal{L}(D_i \,|\, k) &= \int_0^1 \mathrm{d}q\,\mathcal{L}(D_i \,|\, k, q) \\ &= \frac{w_k}{\sqrt{||2\pi\,\mathcal{C}_i||}}\, e^{-\frac{1}{2}\,\chi^2_{i,k}} \int_0^1 \mathrm{d}q\, e^{-\frac{1}{2}\,(a\,q^2 - 2\,b\,q)} \\ &= \frac{w_k}{\sqrt{||2\pi\,\mathcal{C}_i||}}\, e^{-\frac{1}{2}\,\chi^2_{i,k}} \sqrt{\frac{\pi}{2\,a}}\, e^{b^2/(2\,a)} \left(\mathrm{erf}\left(\frac{a-b}{\sqrt{2\,a}}\right) - \mathrm{erf}\left(\frac{b}{\sqrt{2\,a}}\right)\right) \\ a &= A_k{}^{\mathsf{T}}\,\mathcal{C}_i^{-1}\,A_k \\ b &= A_k{}^{\mathsf{T}}\,\mathcal{C}_i^{-1}\,\delta_{i_k}\end{aligned} \tag{3}$$

and for simplicity we can write this as

$$\mathcal{L}(D_i \,|\, k) = \frac{w_k\,\eta_{i,k}}{\sqrt{||2\pi\,\mathcal{C}_i||}}\, e^{-\frac{1}{2}\,\chi^2_{i,k}} \tag{4}$$

$$\eta_{i,k} = \sqrt{\frac{\pi}{2\,a}}\, e^{b^2/(2\,a)} \left(\mathrm{erf}\left(\frac{a-b}{\sqrt{2\,a}}\right) - \mathrm{erf}\left(\frac{b}{\sqrt{2\,a}}\right)\right) \tag{5}$$

We can then sum over all track segments k to obtain the total likelihood of an individual observation given the track

$$\mathcal{L}(D_i \,|\, \theta) = \sum_k \mathcal{L}(D_i \,|\, k, \theta) \tag{6}$$

NORMALIZATION

To really treat the track *density* correctly we need to consider the overall normalization. This will act to penalize tracks that produce significant predicted density where there are no objects. This does however mean that you need to have completeness or at least a well defined selection function in the space of the observables X.

So, the usual way to think about an ensemble of objects drawn from a density is as an inhomogeneous Poisson point process. This describes the probability that you'd draw a single object from an infinitesimal interval near a particular parameter value when the expected rate varies as a function of parameter. Because we have a one dimensional track,

we will express this as a one-dimensional density

$$p(t_1, t_2, ...t_N \,|\, \theta) = e^{-N_\lambda} \prod_i^N \lambda(t_k|\theta) \tag{7}$$

$$N_\lambda = \int_0^\infty \mathrm{d}t\, \lambda(t) \tag{8}$$

where t indicates position along the track. The density λ should be the predicted observed density and include any selection effects, e.g. as the product $\lambda(t_i \,|\, \theta) = \Lambda(t_i \,|\, \theta)\, S(t_i, \theta)$ where S is the selection function and Λ is the intrinsic density.

Now if we go back to the original full likelihood, and write out the individual likelihoods, we have something like

$$\begin{aligned}
\mathcal{L}(D \,|\, \theta) &= \mathcal{L}(D_1, D_2, \ldots, D_N \,|\, \theta) \\
&= \int_0^\infty \mathrm{d}T\, \mathcal{L}(D \,|\, T)\, p(T \,|\, \theta) && (9) \\
&= \int_0^\infty \mathrm{d}T\, e^{-N_\lambda(\theta)} \left(\prod_i \lambda(t_i \,|\, \theta) \right) \left(\prod_i \mathcal{L}(D_i \,|\, t_i) \right) && (10) \\
&= e^{-N_\lambda(\theta)} \prod_{i=1}^N \int_0^\infty \mathrm{d}t_i\, \mathcal{L}(D_i \,|\, t_i)\, \lambda(t_i \,|\, \theta) && (11)
\end{aligned}$$

where T is a vector of all true positions along the track $\{t_i\}_1^N$ (which we marginalize over), and in the last line we have factored the integral by exploiting the conditional independence of the individual object likelihoods, and that the integrated expected number of objects is not a function of t but only of θ. The value of $\lambda(t \,|\, \theta)$ should be used for the weight w_k in the individual likelihoods above.

One way to think about the e^{-N_λ} pre-factor is that it is a penalty for models that have extra probability where there are no observed instances; it encourages parsimony and models that explain the observed instances with as few total predicted instances as possible.

Scaling parameter

Let's consider the situation where we have a density with an unknown overall scaling that we will treat as a parameter. We can write

$$\begin{aligned}
\lambda(t \,|\, \theta, M) &= M\, \rho(t \,|\, \theta) && (12) \\
\int_0^\infty \mathrm{d}t\, \rho(t) &= 1
\end{aligned}$$

So here we are just replacing $N_\lambda(t)$ with the parameter M. Plugging this in for $\lambda(t)$ above we obtain

$$\begin{aligned}
\mathcal{L}(D \,|\, \theta, M) &= e^{-M} \prod_{i=1}^N \int_0^\infty \mathrm{d}t_i\, \mathcal{L}(D_i \,|\, t_i)\, M\, \rho(t_i \,|\, \theta) \\
&= e^{-M}\, M^N \prod_{i=1}^N \int_0^\infty \mathrm{d}t_i\, \mathcal{L}(D_i \,|\, t_i)\, \rho(t_i \,|\, \theta) && (13)
\end{aligned}$$

and in this case the weights w_k are derived from $\rho(t)$

FINAL LOG-LIKELIHOOD

So, bringing everything together, we have

$$\ln \mathcal{L}(D \,|\, \theta) = -N_\lambda(\theta) + \sum_i^N \ln \left(\sum_k^K w_k\, \eta_{i,k}\, e^{-\frac{1}{2}\, \chi^2_{i,k}} \right) - \sum_i^N \ln(\sqrt{||2\pi\, \mathcal{C}_i||}) \tag{14}$$

where $\eta_{i,k}$ is the optional correction factor for integrating the likelihood over the line segments, which can be set to 1 if the track is densely sampled compared to the uncertainties. The last term is a constant for a given dataset so can be dropped (or cached) before sampling.

If we have an overall normalization term M as a free parameter we can write this as

$$\ln \mathcal{L}(D \,|\, \theta, M) = N \ln M - M + \sum_i^N \ln \left(\sum_k^K w_k \, \eta_{i,k} \, e^{-\frac{1}{2} \chi^2_{i,k}} \right) - \sum_i^N \ln(\sqrt{||2\pi \, \mathcal{C}_i||}) \tag{15}$$

and if we set the gradient with respect to M to zero then clearly the maximum likelihood, regardless of θ, is obtained when $N = M$ so we can just use that. In this case, over-extended tracks are penalized by putting more of the *fractional* weight ρ further from the observed data points and at larger χ^2, thus lowering the overall total likelihood.