Automated recurrent integration

Computing the likelihood of a data set given a model is a keystone of probability theory as it is needed for maximum likelihood estimation and for Bayesian inference. By definition, *L*(model | data) = the probability of the data given the model P(data | model). In many cases, this computation requires the intervention of hidden variables within a conditional probability framework.

To illustrate this, we can consider a simple example of a particle in Brownian motion whose real positions R = (r1,r2,..) are estimated with a localization error (observed positions C=(c1, c2, …)). Here, we have Gaussian relationships between ri and ci (localization error) and between ri+1 and ri (displacements). In this example, the first thing that we can express is the joint probability of the track C and the real positions R. Using the principle of conditional probabilities, we get:

P(C,R) = P(r0 | c0) P(r1 | r0) P(r1 | c1) P(r2 | r1) P(r2 | c2) P(r3 | r2) … (eq 1)

= *f*(C,R) = *f*e(r0 - c0) *f*d(r1 - r0) *f*e(r1 - c1) *f*d(r2 - r1) *f*e(r2 - c2) *f*d(r3 - r2) …

with P(ri | ci) = *f*e(ri – ci) a Gaussian of ri, with mean ci and standard deviation *e* the localization error and P(ri+1 | ri) = *f*d(ri+1 – ri) a Gaussian of ri+1, with mean ri and standard deviation *d* the diffusion length per step.

To retrieve the probability of the data P(C), we need to integrate the joint probability P(C,R) over the possible real positions. This integration process can be done analytically using a fast recurrence formula described in our previous article <https://rupress.org/jcb/article/222/5/e202208059/213911/ExTrack-characterizes-transition-kinetics-and> .

The general case that we are considering for this article is the computation of the probability of observed data (variables) from the integration over all possible hidden variables of a joint probability which is a product of univariate Gaussians that depend on a linear combination of the observed and the hidden variables:

Example: integration over x1, x2 and x3 of the following joint probability f(x1, x2, x3, o1, o2)

= K \* *f*(*a*0,1+*a*0,1 *x*1+*a*0,2 *x*2+*a*0,3 *o*1) f(*a*1,0+*a*1,1 *x*1+*a*1,2 *x*3+*a*1,3 *o*1 + *a*1,4 *o*2) f(*a*2,0 + *a*2,1 *x*2+ *a*2,2 *x*3) f(*a*3,0+ *a*3,1 *x*3+*a*3,2 *o*2) (eq 2)

Here, *f* are standard Gaussians, *ai,j* are model coefficients and *K* is a constant. Our recurrence process is based on the fact that the product of two gaussians depending on *xi* can be ‘reshuffled’ into a product of two Gaussians, one being constant with regards to *xi* (but not necessarily with regards to the other *xj*). If we have *k* gaussians that depend on *xi*, we can repeat this to obtain *k*-1 gaussians independent of *xi* and 1 Gaussian dependent of *xi*. Then the integration with regards to *xi* becomes trivial as the integral of a single Gaussian of *xi* over all possible *xi* is 1 (independently of the other variables in this Gaussian). NB: during this process, it is possible that the coefficients cancel each other. In that case, the hidden variable *xi* is not a valid variable.

Our automated algorithm is already working for any function of the form shown in eq 2 (with any number of functions, observed variables and hidden variables) but we haven’t computed the scaling factor *K* yet. This scaling factor must verify that integrating over all (hidden and observed) variables results in a probability of 1.

The joint probability density function is actually equivalent to a multivariate Gaussian function of covariance matrix Σ (potentially sparce) and K is therefore essentially with k the number of variables (observed + hidden).

In case of temporal data, for each time point *i*, we can consider a set of observations and of hidden states whose relations are described by univariate gaussians that depend on a linear combination of temporally neighbor variables. For instance, eq 1 can be written as a product over the time steps *i*: .

More generally, we can compute the probability for any product of an arbitrary set of univariate gaussians that depend on linear combinations of variables .

In these cases, the joint probability can also be expressed as a multivariate Gaussian with a banded covariance matrix. Here, we need to find K (per time step).

One way that I see to find K for one time step is to actually perform the integration for the (hidden and observed) variables of one time step. I don’t really need help to do that but maybe you can find a smarter way to do it.

Overall, what we are missing and we think you could bring is the formalism which links the multivariate form of the expression to its form as a product of univariate gaussians (I’ll call this form the univariate form). First, for any multivariate form, we do have a univariate form that can be computed with conditional probabilities but do we have a multivariate form for any univariate form with arbitrary sets of coefficients *ai,j* ? From our iteration process, we already know that the coefficients of a given variable must not cancel each other but it can only happen at singular values in the space of reals so it’s not too concerning. Is their stronger constraints too ? If so, are these constraints easy to identify from the univariate form or do we actually need to start from the multivariate form and derive it into the univariate form to make sure it is a multivariate Gaussian? My intuition is that it is not necessary, and we would rather avoid starting from the multivariate form if possible as it is not required for our method and as in probabilistic models it is usually more natural to express the joint probability as a product of conditional probabilities.