Research!

June 14, 2016

In our model-selection problem we are given two (or more) phylogenetic population models, and aligned DNA data, and are required to select which model better describes the data. In this work we introduce the notion of a "Reference Model". This, in short, is a less restrictive, intermediate population model, which we utilize in the comparison of the target models –

IG: suggest 'phylogentic demographic models', instead of 'population'

$$K := \frac{P(X|M_r)}{P(X|M_h)}$$

A population model describes generation of DNA using a layer of hidden random variables. These are the model parameters (population sizes, divergence times and migration rates) and the loci genealogies.

IG: 'local genealogies' not 'loci genealogies'

The G-PhoCS framework uses an MCMC algorithm to simulate integration over the parameters of M_h. In our model-selection algorithm we attempt to use this integration on p_h (parameters of M_h) to also integrate over p_r. The following chapter lays down the justifications for integrating over a 'small' vector space (p_r) by employing integration on a larger vector space (p_h), this using an extension of the small space and a mapping between the two spaces.

IG: integration is done over all hidden RVs, not just parameters. I suggest that below this point you don't distinguish between parameters and other hidden RVs. Also, no need for p_h and p_r notation

Consider two models, M_h and M_r. Model M_h has i parameters, $y_{1..j}$ and model M_r has j parameters, $z_{1..j}$. Assume $j < i$. Given are j functions $\{f_j\}_{1..j}$, each mapping the parameters of M_h to a parameter of M_r — $f_j : \vec{y} \mapsto z_j$.

IG: I suggest a few notational conventions. See my alternative text given in blue below your text (next page)

Denote F the application of all $\{f_j\}$ functions —

$$F : \vec{y} \mapsto \langle f_1(\vec{y}), .., f_j(\vec{y}) \rangle$$

This F is a surjective ~~, differentiable, with continuous partial derivatives~~ mapping from M_h on to M_r. We intend to use F in a change of variables from $z_{1..j}$ to $\vec{y}$, so F must be turned into a bijection. We do this by creating $i - j$ new mock-parameters (parameters who have no effect on data likelihood) of M_r, $\{z'_{j+1..i}\}$, and $i - j$ new i-dimensional functions $f'_{j+1}, ..., f'_i$ onto the new parameters, such that F' is a bijection —

$$F' : \vec{y} \mapsto \langle f_1(\vec{y}), ..., f_j(\vec{y}), f'_{j+1}(\vec{y}), ..., f'_i(\vec{y}) \rangle$$

During the creation of the new random variables we are also given a conditional probability function $P(z'_{j+1..i}|z_{1..j})$, to be used when adding the mock-parameters to the integration.

We now apply these structures to our Model-Selection problem —

IG: my suggested text for the above

Consider a model M_h (the *hypothesis model*) with n hidden variables $\vec{Y} = (Y_1, \ldots, Y_n)$, and a *reference model*, M_r, with $m \leq n$ hidden variables $\vec{Z} = Z_1, \ldots, Z_m$. The mapping between the two models is defined by m functions $F = (f_1, \ldots, f_m)$, each mapping the n-dimensional vector $\vec{Y}$ to Z_j: $Z_j = f_j(\vec{Y})$. We assume that each f_j is surjective onto the domain of Z_j, implying that the combined mapping F is surjective onto the combined domain of $\vec{Z}$. However, when $m < n$ this mapping will typically not be a bijection, meaning that the mapping from the reference model back to the hypothesis model is ambiguous.

To enable inversion of this mapping, we assume that F can be appended by $n - m$ additional functions $\tilde{F} = (\tilde{f}_{m+1}, \ldots, \tilde{f}_n)$, and further assume that the combined mapping $\bar{F} = (F, \tilde{F})$ is invertible. Note that the additional mapping functions can be used to define $n - m$ random variables, $\tilde{Z}_j = \tilde{f}_j(\vec{Y})$, which we refer to as *mock variables* of the reference model. These can be thought of as hidden variables of the reference model whose values do not influence the observed data. Their sole purpose is to define an extension of the reference model that has the same dimension as the hypothesis model and can be mapped bijectively to it. To complete the setup, we define a probability distribution over the mock variables $\tilde{Z} = (\tilde{Z}_{m+1}, \ldots, \tilde{Z}_n)$ given the actual hidden variables: $P(\tilde{Z}|\vec{Z})$.

With these assumptions in place, the Bayes factor of M_h relatively to M_r can be expressed as follows:

IG: equations look fine. Just need to convert notations. I think that BF is better than K. No need for special notation PF'. Need to define notation for Jaacobian above.

$$
\begin{aligned}
\mathbf{K(M_h, M_r|X)} :=& \frac{\mathbf{P(X|M_r)}}{\mathbf{P(X|M_h)}} = \\
=& \int \frac{P(z_{1..j}, X|M_r)}{P(X|M_h)} dz_{1..j} = \int \frac{P(z_{1..j}, X|M_r)}{P(X|M_h)} \int P(z'_{j+1..i}|z_{1..j}) dz'_{j+1..i} dz_{1..j} = \\
=& \int \frac{P(F(\vec{y}), X|M_r)}{P(X|M_h)} P(f'_{j+1..i}(\vec{y})|f_{1..j}(\vec{y})) J_{F'} d\vec{y} =^* \\
=^*& \int \frac{P(F(\vec{y}), X|M_r)}{P(X|M_h)} PF'(\vec{y}) J_{F'} d\vec{y} = \\
=& \int \frac{P(F(\vec{y}), X|M_r)}{P(\vec{y}, X|M_h)} PF'(\vec{y}) J_{F'} P(\vec{y}|X, M_h) d\vec{y} = \\
=& E_{\vec{y}|X, M_h} \left[\frac{P(F(\vec{y}), X|M_r)}{P(\vec{y}, X|M_h)} PF'(\vec{y}) J_{F'} \right]
\end{aligned}
$$

$^* PF'(\vec{y}) := P(f'_{j+1..i}(\vec{y})|f_{1..j}(\vec{y}))$