

Bayesian Analysis Of Distance Estimation in the K2P Model.

Ilan Gronau

March 23, 2009

1 Posterior Distribution for Parameters of a Multinomial Random Variable.

Consider an experiment with s possible different outcomes x_1, \dots, x_s corresponding to (unknown) probabilities p_1, \dots, p_s (where $p_1 + \dots + p_s = 1$). The experiment is performed k times independently and the number of times each outcome occurred is counted – k_1, \dots, k_s (resp.). The probability of the observation (k_1, \dots, k_s) , given the unknown parameters p_1, \dots, p_s is:

$$\Pr[k_1, \dots, k_s | p_1, \dots, p_s] = \prod_i p_i^{k_i} .$$

The Bayesian approach focuses on the converse, i.e. the probability of the unknown parameters given the observed experiment results. Since this probability space is continuous, we are interested in the density function $f(p_1, \dots, p_s | k_1, \dots, k_s)$. Let Ω be an infinitesimal s -dimensional interval around the point (p_1, \dots, p_s) (where $p_1 + \dots + p_s = 1$). Then the probability, given the observations, that the parameters are within the interval Ω is:

$$\Pr[\Omega | k_1, \dots, k_s] = \frac{\Pr[k_1, \dots, k_s | \Omega] \cdot \Pr[\Omega]}{\Pr[k_1, \dots, k_s]} .$$

The expression $\Pr[k_1, \dots, k_s]$ is independent of Ω , and so it may be viewed as part of the normalizing constant (which will ensure that the integral of the density function f over its entire domain is 1). Now assume a uniform prior $\Pr[\Omega]$ (meaning that we do not give any a-priori preference to any specific parameter-set p_1, \dots, p_s). This means that $\Pr[\Omega]$ is proportional to the size of the interval Ω . Lastly, when the interval Ω is sufficiently small (around (p_1, \dots, p_s)) we get $\Pr[k_1, \dots, k_s | \Omega] = \prod_i p_i^{k_i}$. Summing this up, we get that

$$\Pr[\Omega | k_1, \dots, k_s] = \gamma \cdot |\Omega| \cdot \prod_i p_i^{k_i} .$$

Hence, the density function f is defined as follows:

$$f(p_1, \dots, p_s | k_1, \dots, k_s) = \gamma \prod_i p_i^{k_i}$$

(where γ normalizes f so that its integral over its entire domain is 1)

This implies that $(p_1, \dots, p_s | k_1, \dots, k_s)$ is distributed *Dirichlet* with parameters $(k_1 + 1, \dots, k_s + 1)$ [1]. Now, for $i = 1 \dots s$ denote $\bar{p}_i = \frac{k_i + 1}{k + s}$. Then the following facts are known about this distribution:

- $E[p_i] = \bar{p}_i$.
- $\text{VAR}[p_i] = \frac{\bar{p}_i(1-\bar{p}_i)}{k+s+1}$.
- $\text{COV}[p_i, p_j] = -\frac{\bar{p}_i \bar{p}_j}{k+s+1}$ (for $i \neq j$).

2 Bayesian Inference of Distance in the K2P Model.

Two k -long sequences which evolve according to the K2P substitution model can be described as k independent and identical experiments (corresponding to aligned site-pairs) with $s = 3$ possible outcomes:

1. No substitution: $A \leftrightarrow A$ or $G \leftrightarrow G$ or $C \leftrightarrow C$ or $T \leftrightarrow T$.
2. Transition-type substitution: $A \leftrightarrow G$ or $C \leftrightarrow T$.
3. Transversion-type substitution: $\{A, G\} \leftrightarrow \{C, T\}$.

Assuming the path connecting the two sequences has a K2P substitution matrix with parameters $p_{\text{nil}}, p_{\alpha}, p_{\beta}$, then the probability of the first outcome (no substitution) is p_{nil} , the probability of the second outcome (transition) is p_{α} , and the probability of the third outcome (transversion) is $2p_{\beta}$.

Assume the two observed sequences imply statistics $\widehat{p}_{\text{nil}}, \widehat{p}_{\alpha}, \widehat{p}_{\beta}$. These statistics induce a probability distribution over all substitution matrices in the K2P model. This distribution describes for each matrix its probability to have generated the observed sequence-pair. according to the analysis of the previous section, $(p_{\text{nil}}, p_{\alpha}, 2p_{\beta})$ are distributed Dirichlet with parameters $(k\widehat{p}_{\text{nil}} + 1, k\widehat{p}_{\alpha} + 1, 2k\widehat{p}_{\beta} + 1)$. Therefore,

$$\mathbb{E}[p_{\text{nil}}] = \overline{p}_{\text{nil}} \triangleq \frac{k\widehat{p}_{\text{nil}} + 1}{k + 3} \quad ; \quad \mathbb{E}[p_{\alpha}] = \overline{p}_{\alpha} \triangleq \frac{k\widehat{p}_{\alpha} + 1}{k + 3} \quad ; \quad \mathbb{E}[p_{\beta}] = \overline{p}_{\beta} \triangleq \frac{k\widehat{p}_{\beta} + \frac{1}{2}}{k + 3} .$$

Note that for large enough k , we get $\overline{p}_{\text{nil}} \cong \widehat{p}_{\text{nil}}$, $\overline{p}_{\alpha} \cong \widehat{p}_{\alpha}$, and $\overline{p}_{\beta} \cong \widehat{p}_{\beta}$ (see concluding discussion). Our estimated distance \overline{d} is then obtained by applying an SR function Δ on $\overline{\mathbf{P}}$, which is the K2P substitution matrix with parameters $\overline{p}_{\text{nil}}, \overline{p}_{\alpha}, \overline{p}_{\beta}$. In this view, the true distance is treated as a random variable $d = \Delta(\mathbf{P})$, whose distribution is determined by the Dirichlet distribution the observed alignment implies on $(p_{\text{nil}}, p_{\alpha}, 2p_{\beta})$. As in [2] (Section 4.2), we wish to estimate the expected error in distance estimation $\mathbb{E}[(\overline{d} - d)^2]$ implied by any SR function Δ . This is done by tracing the covariance matrix along the different steps of the distance estimation process (using Lemma 4.1 of [2]).

Step 1: The three parameters $\overline{p}_{\text{nil}}, \overline{p}_{\alpha}, \overline{p}_{\beta}$ are computed as detailed above. According to the properties of the Dirichlet distribution, the covariance matrix of the random variables $p_{\text{nil}}, p_{\alpha}, p_{\beta}$ is:

$$\Sigma \begin{pmatrix} p_{\text{nil}} \\ p_{\alpha} \\ p_{\beta} \end{pmatrix} = \frac{1}{k + 4} \begin{pmatrix} \overline{p}_{\text{nil}}(1 - \overline{p}_{\text{nil}}) & -\overline{p}_{\text{nil}} \overline{p}_{\alpha} & -\overline{p}_{\text{nil}} \overline{p}_{\beta} \\ -\overline{p}_{\text{nil}} \overline{p}_{\alpha} & \overline{p}_{\alpha}(1 - \overline{p}_{\alpha}) & -\overline{p}_{\alpha} \overline{p}_{\beta} \\ -\overline{p}_{\text{nil}} \overline{p}_{\beta} & -\overline{p}_{\alpha} \overline{p}_{\beta} & \frac{1}{2}\overline{p}_{\beta}(1 - 2\overline{p}_{\beta}) \end{pmatrix} \quad (1)$$

Step 2: The two distinct non-trivial eigenvalues $\overline{\lambda}_1, \overline{\lambda}_2$ of the K2P substitution matrix $\overline{\mathbf{P}}$ are computed:

$$\overline{\lambda}_1 = 1 - 4\overline{p}_{\beta} \quad , \quad \overline{\lambda}_2 = \overline{p}_{\text{nil}} - \overline{p}_{\alpha} .$$

Applying Lemma 4.1 gives us the covariance matrix of the random variables λ_1, λ_2 :

$$\Sigma \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \frac{1}{k + 4} \begin{pmatrix} 8\overline{p}_{\beta}(1 - 2\overline{p}_{\beta}) & 4\overline{p}_{\beta}(\overline{p}_{\text{nil}} - \overline{p}_{\alpha}) \\ 4\overline{p}_{\beta}(\overline{p}_{\text{nil}} - \overline{p}_{\alpha}) & \overline{p}_{\text{nil}}(1 - \overline{p}_{\text{nil}}) + \overline{p}_{\alpha}(1 - \overline{p}_{\alpha}) + 2\overline{p}_{\text{nil}} \overline{p}_{\alpha} \end{pmatrix} = \frac{1}{k + 4} \begin{pmatrix} 1 - \overline{\lambda}_1^2 & (1 - \overline{\lambda}_1)\overline{\lambda}_2 \\ (1 - \overline{\lambda}_1)\overline{\lambda}_2 & \frac{1}{2}(1 + \overline{\lambda}_1) - \overline{\lambda}_2^2 \end{pmatrix} \quad (2)$$

Step 3: The logarithms of the eigenvalues are taken to obtain the base components of the distance: $\overline{\delta}_1 = -\ln(\overline{\lambda}_1)$, $\overline{\delta}_2 = -\ln(\overline{\lambda}_2)$. The covariance matrix of the random variables δ_1, δ_2 is approximated using the following linear transformation of λ_1, λ_2 :

$$\delta_1 \cong 1 - \ln(\overline{\lambda}_1) - \frac{\lambda_1}{\lambda_1} - \ln(\overline{\lambda}_1) \quad , \quad \delta_2 \cong 1 - \ln(\overline{\lambda}_2) - \frac{\lambda_2}{\lambda_2} .$$

Applying Lemma 4.1 gives us the covariance matrix of these linear approximations:

$$\Sigma \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} \cong \frac{1}{k+4} \begin{pmatrix} \frac{1-\lambda_1^{-2}}{\lambda_1} & \frac{(1-\lambda_1)}{\lambda_1} \\ \frac{(1-\lambda_1)}{\lambda_1} & \frac{(1+\lambda_1)}{2\lambda_1^2} - 1 \end{pmatrix} = \frac{1}{k+4} \begin{pmatrix} e^{8\bar{\beta}} - 1 & e^{4\bar{\beta}} - 1 \\ e^{4\bar{\beta}} - 1 & \frac{1}{2}e^{4\bar{\alpha}}(e^{4\bar{\beta}} + 1) - 1 \end{pmatrix} \quad (3)$$

($\bar{\alpha}, \bar{\beta}$ are the transition and transversion rate of the K2P rate matrix $\bar{\mathbf{R}}$ corresponding to $\bar{\mathbf{P}}$.)

Step 4: The linear combination of $\bar{\delta}_1, \bar{\delta}_2$ is taken according to the SR coefficients c_1, c_2 of Δ :

$$\bar{d} = c_1 \bar{\delta}_1 + c_2 \bar{\delta}_2.$$

Applying Lemma 4.1 gives us the following approximation of the variance of the true distance d :

$$\text{VAR}[d] \cong \frac{1}{k+4} \left(c_1^2 (e^{8\bar{\beta}} - 1) + 2c_1 c_2 (e^{4\bar{\beta}} - 1) + \frac{1}{2} c_2^2 (e^{4\bar{\alpha}} (e^{4\bar{\beta}} + 1) - 2) \right). \quad (4)$$

The linear approximation we applied for δ_1, δ_2 implies that $\mathbf{E}[d] \cong \bar{d}$ (see [2], Section 4.1 for more details). Hence, $\text{VAR}[d] \cong \mathbf{E}[(d - \bar{d})^2]$. Now, the Bayesian optimal SR function for the evolutionary path under inspection is the one minimizing expected normalized difference between the (random) true distance d and the (Bayesian) observed distance \bar{d} which is approximated as follows:

$$\mathbf{E} \left[\left(\frac{d - \bar{d}}{\bar{d}} \right)^2 \right] \cong \frac{\text{VAR}[d]}{\bar{d}^2} \cong \frac{c_1^2 (e^{8\bar{\beta}} - 1) + 2c_1 c_2 (e^{4\bar{\beta}} - 1) + \frac{1}{2} c_2^2 (e^{4\bar{\alpha}} (e^{4\bar{\beta}} + 1) - 2)}{(k+4)(4c_1 \bar{\beta} + 2c_2 (\bar{\alpha} + \bar{\beta}))^2} \quad (5)$$

Similar analysis to the one done in Lemma 5.1 of [2] gives us that coefficients c_1, c_2 with the following ratio minimize the right-hand expression above:

$$\frac{c_1}{c_2} = \frac{\frac{e^{4\bar{\beta}} + 1}{e^{4\bar{\beta}} - 1} (e^{4\bar{\alpha}} - 1) \bar{\beta} - \bar{\alpha}}{(e^{4\bar{\beta}} - 1) \bar{\beta} + (e^{4\bar{\beta}} + 1) \bar{\alpha}}. \quad (6)$$

3 Discussion

First, note that in this Bayesian approach the distances are estimated a bit differently than in the regular approach. Originally, the distance estimate \hat{d} was based on the maximum-likelihood estimations $\widehat{p}_{\text{nil}}, \widehat{p}_\alpha, \widehat{p}_\beta$. On the other hand, in the Bayesian approach, it makes more sense to use the means $\bar{p}_{\text{nil}}, \bar{p}_\alpha, \bar{p}_\beta$, and obtain a distance estimate \bar{d} . However, note that $\bar{p}_{\text{nil}} = \frac{k\widehat{p}_{\text{nil}} + 1}{k+3} = \widehat{p}_{\text{nil}} - \frac{3\widehat{p}_{\text{nil}} - 1}{k+3}$, implying that \bar{p}_{nil} and \widehat{p}_{nil} are very close (as long as the sequence length k is not too small), and the same goes to $\bar{p}_\alpha, \widehat{p}_\alpha$ and $\bar{p}_\beta, \widehat{p}_\beta$. Therefore, as long as k is not too small, \bar{d} is actually very close to \hat{d} . Similarly, due to the difference between $\bar{\alpha}, \bar{\beta}$ and $\hat{\alpha}, \hat{\beta}$, the Bayesian optimal SR function (obtained by §6) might be a bit different than the adaptive SR function described in [2] (Section 5.3). But, again, we expect these two SR functions to be very similar (for large enough k).

In conclusion, the Bayesian approach is very appealing because it is data driven. Rather than treating the observed data as a random variable, we treat it as a given constant that implies a probability distribution over the possible generating models (substitution matrices). Other than this conceptual advantage, it has the added advantage of allowing the introduction of meaningful priors to the inference process. This might be relevant when we believe some substitution matrices to be a-priori more likely than others. In the framework presented here, this can be easily done using a Dirichlet prior on $(p_{\text{nil}}, p_\alpha, 2p_\beta)$.

References

- [1] M. Evans, N. Hastings, and B. Peacock. *Statistical distributions*. Wiley interscience, 3rd edition, 2000.
- [2] I. Gronau, S. Moran, and I. Yavneh. Towards optimal distance functions for stochastic substitutions models. http://www.cs.technion.ac.il/~ilangr/papers/optimal_distances09.pdf (submitted March, 2009).