INFERENCE OF DIVERGENCE TIMES AS A STATISTICAL INVERSE PROBLEM

STEVEN N. EVANS, DON RINGE, AND TANDY WARNOW

1. Introduction

A familiar complaint about statisticians and applied mathematicians is that they are the possessors of a relatively small number of rather elegant hammers with which they roam the world seeking convenient nails to pound, or at least screws they can pretend are nails. One all too often hears tales of scholars who have begun to describe the details of their particular research problem to a statistician, only to have the statistician latch on to a few phrases early in the conversation and then glibly announce that the problem is an exemplar of a standard one in statistics that has a convenient, pre-packaged solution – preferably one that uses some voguish, recently developed technique (bootstrap, wavelets, Markov chain Monte Carlo, hidden Markov models,...)

To some degree, this paper continues that fine tradition. We will observe that various facets of the inference of linguistic divergence times are indeed familiar nails to statisticians. However, we will depart from the tradition by being less than sanguine about whether statistics possesses, or can ever possess, the appropriate hammers to hit them. In particular, we find the assertion of (Forster & Toth, 2003) that

> Phylogenetic time estimates ... are statistically feasible once the language tree has been correctly reconstructed, by uncovering any recurrent changes of the items.

and similar optimistic uses of statistical methodology for dating purposes elsewhere in the historical linguistics literature need much more justification before they can be accepted with any degree of confidence.

We begin with a discussion about inverse problems in Section 2, and continue in Section 3 with a description of stochastic models of evolution that have been proposed for biomolecular or linguistics. The critical issues involved in parameter estimation under these models are presented in Section 4, while specific issues involving data selection are discussed in Section 5.

SNE supported in part by NSF grants DMS-0071468 and DMS-0405778.
DR supported in part by NSF grant BCS-0312911.
TW supported by NSF grant BCS-0312830.

We then turn to the specific issues involved with estimating dates at internal nodes in Section 6. We make some concluding remarks in Section 7.

2. Inverse problems

It has been recognized for a long time that phylogenetic reconstruction in linguistics (and, of course, biology) can be viewed as a statistical inference problem: we have a collection of possible stochastic models for the past evolution of languages and we are trying to determine which of those models best fits some observed data. However, phylogeny is clearly rather different from the statistical problems that typically confront experimental scientists, such as determining from a long sequence of *independent and identical trials* what the mortality will be for mice that are injected with a particular toxin. Phylogeny raises issues of model complexity and adequacy of the data (both in terms of amount and structure) that are usually not so germane to standard experimental situations.

Phylogenetic inference is an instance of a *statistical inverse problem* (see (Evans & Stark, 2002) for a survey of this general area). This term doesn't have a clear, agreed-upon definition. Rather, it is a little like U.S. Chief Justice Potter Stewart's definition of obscenity, 'I know it when I see it.' It is therefore best to give a standard example that shares many features with phylogenetic inference, but for which those features are more immediately comprehensible. However, before we give an example of a statistical inverse problem, we give an example of a corresponding *forward* problem.

Consider monitoring earthquakes using seismographic apparatus. If we knew the detailed composition and structure of the earth, then we could, in principle, solve the relevant partial differential equations to compute what the measurements for amplitude, time-of-travel and so forth would be at a particular seismic monitoring station when an earthquake of a given magnitude occurs at some other location. This is an example of a *forward problem*: we know how a system is built and we want to calculate what the output of the system will be from a given input. In practice, an exact deterministic model of the earth's interior is often replaced by a stochastic model that attempts to mimic the finer details of the interior's heterogeneity using a random structure that is more tractable to compute with.

Of course, the forward problem is not the situation that actually confronts us. Based on seismographic observations we want to *reverse engineer* the earth and determine its internal structure from a knowledge of various inputs and outputs. In essence we have some universe of potential interiors for the earth; for each of these we can solve the forward problem, and we want to determine which of the potential interiors best fits our observed data.

We hope that the correspondence between this problem in geophysics and the problem of linguistic phylogeny is starting to form in the reader's mind. Instead of standing on the surface of the earth and attempting to infer the nature of its interior, we are standing at the present and attempting to infer the nature of the linguistic past. We will have an ensemble of models for the possible course of linguistic evolution. These models have a stochastic component because we believe that the *typical* outcome of an appropriate random process will somehow act as a suitable proxy for a detailed description of historical events and because it is usually rather straightforward to compute the predictions of these models for our data. We then wish to determine which model from our ensemble does the best job of explaining the data.

This reverse engineering process is fraught with difficulties in both seismology and phylogeny. We will go into more detail later, but we can give a brief overview of some of the difficulties as follows.

To begin with, the mathematics of wave propagation is such that the reconstruction of the earth's interior from seismographic observations is an *ill-posed* problem: there can be different structures for the interior that would lead to the same connection between inputs (that is, earthquake magnitudes and locations) and outputs (that is, seismographic measurements around the earth). This is as we might expect: it should be difficult to determine the conditions inside a country from just watching people entering and leaving its borders. Even if there wasn't this degeneracy, reasonable models for the interior can contain infinite dimensional or very high dimensional features (for example, the boundary between the core and the mantle is some surface of potentially arbitrary *complexity*), whereas we only have a finite number of low dimensional observations on which to base our inferences. In essence, we are trying to constrain a high dimensional object with low dimensional observations and, in essence, we run up against a basic mathematical problem of *too few equations in too many unknown variables.*

The counterpart of this issue in phylogeny is what statisticians usually call *lack of identifiability*. The most extreme instance of this difficulty is when two different models make the same predictions about the state of the present. However, even if different models do make different predictions about the present, our actual data might be too meager to notice the difference because of the too few equations in too many unknowns phenomenon. Thus two genuinely different historical processes simply may not be distinguishable on the basis of our data.

It should be stressed that this problem of inadequacy of data is rather different to what we usually encounter in experimental settings. There we repeat the 'same' experiment and the purpose of collecting more data is just to reduce variability in our inferences about nature. Whether we inject 10 mice or 1,000, our estimate of mortality will be correct 'on average' in both cases, but injecting more mice means that the probability distribution of the estimated mortality will be more tightly clustered around its expected value.

In an inverse problem, we don't usually make repeated observations under the same conditions. New observations may enable us to probe the object we are interested in from different directions and hence have the dimensional coverage of our data set approach that of the object we are investigating. Simply repeating the 'same' observations may increase our certainty about what is present in the directions we probe, but this won't help in constraining the object in the 'perpendicular' directions. Furthermore, if our data are observational rather than experimental, then we don't have control of the directions in which such probing occurs, and so in certain directions we may never see data that will constrain the model. Earthquakes occur in a limited set of positions and their effects can only be measured at a sparse set of locations that, for example, only include places on dry land, and this prevents certain features of the earth's interior from being inferred unless we are willing to assume *a priori* a certain amount of smoothness or homogeneity. This does not mean that the inferential task is completely hopeless: there may still be facets of the model for which the data are adequate even if they are inadequate for others. For example, in historical linguistics we may be able to estimate the tree topology of the model with some confidence even if the data are not available which would enable us to resolve certain divergence times with any degree of accuracy.

In any use of statistics, great care needs to be taken to ensure valid sampling strategies and, in particular, that missing data are treated appropriately. The usual probabilistic models for linguistic evolution are meant to describe *generic* characters from some population that is being sampled, and inferences are only justified to the extent that this is a reasonable assumption. Essentially, we have the following *Gedankenexperiment*: there is a population of possible states for a character that are akin to tickets in a box, some states appear on more tickets than others in proportion to how *likely* the state is to be exhibited by the character, and we imagine that nature has somehow shaken up the box and chosen a ticket at random to give us the observed state of the character. If someone was allowed to rummage through the box and discard tickets before the drawing took place or we are able to look at a ticket after the drawing and can accept or reject it, then the proportions of tickets originally in the box no longer describe the experiment and we need to consider another, perhaps substantially more complex, box that somehow incorporates this *a priori* or *a posteriori* winnowing. In essence, if we have a model that describes the probability of a character exhibiting a particular value, then we need to be able to tell a reasonable story about why the state of the character can be seen as a draw from a box of possible states with a given composition and we have to be careful about any pre- or post- processing we do that may invalidate this story. This is a fundamental concern in statistics, and is usually covered early on in elementary courses under the rubrics of *random versus convenience samples*, *ascertainment bias*, or *selection bias*.

We can do no better than quote (Berk & Freedman, 2003) on this point

> If the random-sampling assumptions do not apply, or the parameters are not clearly defined, or the inferences are to a population that is only vaguely defined, the calibration of uncertainty offered by contemporary statistical technique is in turn rather questionable. Thus, investigators who use conventional statistical technique turn out to be making, explicitly or implicitly, quite restrictive behavioral assumptions about their data collection process. By using apparently familiar arithmetic, they have made substantial empirical commitments; the research enterprise may be distorted by statistical technique, not helped.
>
> ... researchers may find themselves assuming that their sample is a random sample from an imaginary population. Such a population has no empirical existence, but is defined in an essentially circular way – as that population from which the sample may be assumed to be randomly drawn. At the risk of the obvious, inferences to imaginary populations are also imaginary.

One way in which a sampling strategy can violate the implicit assumptions of the model is via its treatment of missing data. For example, social scientists have long been aware that non-respondents to questionnaire items can't simply be scored as though they would have replied in the same proportions as respondents, the so-called *missing at random* assumption. A standard monitory example is a poll conducted just prior to the 1936 U.S. presidential election by *Literary Digest*. Based on 10,000,000 post card surveys, the magazine predicted that Republican Alfred Landon would win 57 percent of the popular vote to convincingly defeat Franklin Delano Roosevelt, whereas Roosevelt actually polled 62.5 percent of the vote. About 80 percent of the people who received the mail-in questionnaires did not respond and the mailings went to people who, in the depths of the Great Depression, had current automobile registrations or were listed in telephone books – hardly a sampling scheme designed to fairly capture low-income supporters of Roosevelt's New Deal. The moral is that there is not much point having lots of data if they aren't the right data for the question we want to answer.

Linguistic data often undergo a great deal of processing. A character may be removed from a word list after it is discovered that certain languages don't have a representative in that lexical slot. Similarly, in methods based on estimated evolutionary distances, if a language is missing an item, then that item may be simply ignored in computations. The latter procedure seems to be recommended by (Embleton, 2000)

> Some languages are missing a translation for a particular item on the Swadesh-list (e.g. 'ice' or 'swim' in some languages). Normally this is best dealt with just as a statistician would with any missing data, namely as a blank and reduce N accordingly in relevant comparisons.

As we have remarked, the validity of this seemingly innocuous practice rests on definite assumptions and is not what a statistician would (or at least should) do without careful thought and justification.

Alternatively, a language is sometimes scored as exhibiting a fictitious unique state of the character if the character is not present for that language. This procedure is suspect, as it treats the 'missing' states in the same manner as observed states and, in particular, treats them as being the outcome of the same substitution mechanism. At the very least, this practice will typically inflate estimates of the variability of the substitution process (and hence effect estimates of divergence times) even if it doesn't have significant effect on estimates of the tree topology.

Such matters of sampling adequacy and treatment of missing data do not seem to have been considered in the historical linguistics literature at the length they deserve to justify the subsequent use of statistical methodology. In short, statistical methods are often applied without any clear sense of the population that the data are meant to represent and whether they actually do so.

Lastly, we come to the question of producing probability models for our whole corpus of data from models for individual data points. The standard practice in both linguistic and biological phylogenetic inference has been to assume that successive characters behave statistically independently. For example, biologists will sometimes erroneously use each position in a stretch of DNA sequence and treat neighbouring positions as independent, even though this is generally not supportable in light of our understanding of the mechanisms by which DNA evolves, but if characters are chosen more carefully (for instance, from separated areas of the genome, or perhaps from non-coding regions), then the independence assumption is usually more tenable. The contrast with linguistic data is drawn by (McMahon & McMahon, 2000)

> Second, individual mutations in biology are generally independent, but this is not necessarily the case in linguistics. Sometimes

> the effects of contact can set up predisposing conditions for further influence; in other cases, one change in a system will cause knock-on effects in a classically structuralist fashion. These interdependencies cause extreme difficulty for our prospects for quantification.

If one wishes to treat characters as independent for the sake of inference, then one needs to argue that any common mechanisms can be satisfactorily captured by simply positing similar evolution dynamics (that is, by the operation of stochastic models with shared parameters but independent noise).

3. Models in phylogeny

We present a mathematical exposition of the basic ingredients of stochastic models of evolution. (Readers interested in obtaining a deeper or more extensive discussion about models should consult any standard textbook in the field; however, readable accounts with extensive bibliographies of probability models in biological phylogenetics and related inferential issues may be found in (Felsenstein, 2003; Kim & Warnow, 1999; Holmes, 1999).) Most stochastic models in phylogeny consist of two ingredients – a putative phylogenetic tree (for the sake of simplicity we will not consider *reticulate* situations such as word borrowing) and various numerical parameters that describe the evolution of a given character through time.

The tree is a rooted directed tree, with edges directed away from the root. The leaves of the tree correspond to taxa for which we can observe character states, the interior nodes of the tree correspond to divergence events in lineages, and the root corresponds to the most recent common ancestor of the set of taxa. Given an edge in the tree connecting two nodes, we will refer to the node closer to the root as the *tail* of the edge and to the node further from the root as the *head* of the edge.

Each node has a time. If the node is a leaf, then the time is the date when the leaf was observed (so that if all of our taxa are contemporary, then all leaves will have the same time). For an interior node, the time is the date at which the corresponding divergence of lineages occurred. There is a significant amount of debate over this so-called *Stammbaum* structure, which views the divergence of lineages as a clear-cut event that can be localised in time, but we will not address these issues here. The assignment of times to nodes induces an assignment of durations to edges: the duration of an edge is the difference between the time of its head and the time of its tail. The duration of the edge e is denoted by t_e.

Each character c will have a state space S_c that is the collection of possible values of the character. In biological models, the choice of S_c is usually fairly straightforward. For example, it will be something like the set of 4

DNA nucleotides or 20 amino acids. For certain linguistic characters (for example, lexical characters), it is not clear how one delineates *a priori* the entire universe of possible states of the character. Using just the set of states observed in the data is insidiously self-referential and complicates the inference process because our model is implicitly conditional on certain features of the data and it is unclear how to interpret the results of statistical procedures within the usual frequentist or Bayesian paradigms. More importantly, such a choice of state-space rules out the possibility of ancestral forms that aren't present in the data, thereby placing quite severe constraints on the model.

Given a state-space, the next step is to build a model for the states observed at the taxa for a given character. This step is usually divided into two parts.

Firstly, one builds a model for both the observed states of the taxa and the unobserved states of the ancestral forms present at each interior node in the tree. This model is usually of the *Markov random field* type, which simply means that the random states exhibited at two nodes (either leaf nodes or interior nodes) are conditionally independent given the random state exhibited at any node on the path through the tree joining the two nodes. Informally, this is the same as saying that the course of evolution on two lineages is independent after the lineages diverge.

Secondly, one obtains a model for the observed states of the taxa by taking the appropriate marginal distribution in the above notional model for both observed and unobserved forms. That is, we 'sum over' the possibilities for the unobserved ancestral states.

The specification of the Markov random field model is equivalent to specifying a probability distribution for the state exhibited at the root and specifying for each edge in the tree a conditional distribution for the state exhibited at the head of the edge given the state exhibited at the tail. If the state space S_c of character c is finite with k states, this means that there is a probability vector π_c of length k giving the distribution of the state exhibited at the root and a $k \times k$ stochastic matrix (that is, a matrix with rows that are probability vectors) $P_{c,e}$ for each edge e giving the family of conditional distributions for e. That is, $\pi_c(i)$ is the probability that the state i is exhibited at the root, and $P_{c,e}(i, j)$ is the conditional probability that the state exhibited at the head of edge e is j given that the state exhibited at the tail is i.

Furthermore, it is usual to think of the matrix $P_{c,e}$ as arising from time-homogeneous Markovian dynamics. That is, there is a rate matrix $Q_{c,e}$ such that $P_{c,e}$ is the matrix exponential $\exp(t_e Q_{c,e})$. The matrix $Q_{c,e}$ has non-negative off-diagonal entries and rows that sum to 0. The interpretation is that $-Q_{c,e}(i, i)$ is the infinitesimal rate at which substitutions away from state i occur and $-Q_{c,e}(i,j)/Q_{c,e}(i,i)$ is the probability that such a substitution will be to state j for $i \neq j$.

This special form for $P_{c,e}$ is somewhat hard to justify, as it posits that the the dynamics of substitution are constant throughout the edge of a tree, but may change discontinuously at a divergence time. However, some structure of this sort appears necessary if the edge durations t_e are to appear explicitly in the model and hence for the model to be of use for dating purposes, but it is not necessary if one is primarily interested in obtaining the tree topology.

Note that, in general, the rate $-Q_{c,e}(i,i)$ may depend on i, and this should indeed be the case if there no reason to believe that all states of the character c are equally mutable. When $-Q_{c,e}(i,i)$ takes the same value for all states i, say $r_{c,e}$, then we can say that 'character c evolves at rate $r_{c,e}$ on edge e'. Otherwise, such a rate does not have any clear-cut meaning other than simply the expected number of substitutions of character c on edge e divided by the time duration t_e, but this quantity will then typically be a rather complicated function of the root distribution π_c along with the edge durations $t_{e'}$ and rate matrices $Q_{c,e'}$ for all the edges e' on the path through the tree from the root to e.

In biological settings, the choice of the rate matrices $Q_{c,e}$ is a mixture of mathematical convenience and biological insight about the *geometry* of the state space. For example, the four DNA nucleotides $\{A, G, C, T\}$ come in two families, the purines $\{A, G\}$ and the pyrimidines $\{C, T\}$. For biochemical reasons, substitutions that stay within a family, $A \leftrightarrow G$ or $C \leftrightarrow T$, (so-called transitions) are easier than substitutions that move between families (transversions). Many commonly-used models of DNA nucleotide evolution incorporate this fact by allowing different rate parameters for transitions and transversions.

The geometry of linguistic state spaces doesn't seem to be understood as well and there don't appear to be models that incorporate structure inherent in the landscape of possible states for a character. Rather, existing models seem to be 'black boxes' that either adopt a 'one-size-fits-all' strategy by allowing the entries of the matrices $Q_{c,e}$ to be arbitrary, or adopt Procrustean solutions such as treating all character substitutions as being equally likely.

The issue of appropriate models of evolution has been addressed in (McMahon & McMahon, 2000)

> The difficulty for this point ... which requires that we understand the mechanisms of change and transmission, lies primarily in the relevant forces of selection. While, as with biological change, language change rests on the bedrock of mutation and consequent variation, the subsequent selection for language change is often under (at least subconscious) human control, since variants

> adopted frequently depend on the situation in which the speaker finds herself and the impression she wishes to make. In other words, although the initial processes of mutation may be random (and note that, at present, we understand the mechanisms creating variants much better in sound change than in any other area, and in semantic change, for instance, these are still particularly opaque), it is hard to conceive of randomness or neutrality when the shift from variation to incipient change often reflects the acquisition of social meaning to a variant, which then becomes manipulable in context.

Moreover, most of the models used in biology are *time-reversible* in the sense that the evolution of the model in equilibrium is the same whether time is run forwards or backwards (that is, we would be unable to distinguish between a videotape of the process on 'fast forward' or 'rewind'). Many processes in linguistic evolution don't have this feature. For example, phonemic mergers are irreversible, and though the exact reversal of a change in inflectional morphology is theoretically possible, actual examples are rare and confined to specific types of development. Thus, reversible models will not be appropriate for morphological and phonological characters.

In connection with our comments above about the choice of the state space S_c, it is worth pointing out that there are problems inherent in dodging the issue of just what the 'real' state space should be by simply scoring character states in terms of the presence or absence of one or more features (that is, by collapsing the state space down to a set of binary vectors, where each binary vector can represent several states of the original state space – a binary vector is just a sequence of zeros and ones representing the answers to a corresponding list of yes-or-no questions). The chief difficulty with this approach is that if a model of the form above with its concomitant, interpretable Markovian substitution dynamics holds for the original state space, then such a model will, in general, no longer hold for the collapsed model: when one clumps the states of a Markov random field together the resulting object is typically no longer a Markov random field. Thus taking a multi-state character and shoe-horning it into an off-the-shelf model from, say, biology by clumping states together will typically lead to a model that doesn't obey any sort of Markovian dynamics with attendant, understandable rate parameters.

Moreover, if one does code a multi-state character as a binary vector (or a collection of binary characters) and model the evolution of this vector with a Markov process, then it is critical that the Markovian dynamics reflect the fact that one is looking at the evolution of the answers to a number of related yes-or-no questions about the **same** object. For example, forcing the coordinates to evolve independently, as seems to be done for such a coding in (Gray & Atkinson, 2003), is patently inappropriate: at the very least, such a model assigns positive probability to binary vectors that contain several ones, even though the coding procedure used to turn data into binary vectors will only produce binary vectors that contain a single one.

This point was made by Bill Poser (Poser, 2004) in a *Language Log* posting where one can find some illustrative graphics, but is easily seen in the following example. Suppose one takes a lexical character c with three states, indicated by A, B, and C. The binary encoding of this multi-state character produces three binary characters, c_A, c_B, c_C. Now suppose the character c changes its state on an edge in the tree from A to B. Then c_A changes from 1 (indicating presence) to 0 (indicating absence), c_B changes its state from 0 (indicating absence) to 1 (indicating presence), and c_C doesn't change at all. Similarly, if one traces just the evolution of c_A on the tree and observe that c_A changes state from 0 to 1 on an edge e, then both c_B and c_C *must* both be 0 at the end of that edge. That is, it is *impossible* for two of the three binary characters to both be 1 at any node in the tree, since every node can have only one state of the character c. In other words, these characters are highly dependent.

While independence may not be absolutely a valid assumption for genuinely different linguistic characters (rather than the ersatz binary "characters" produced by this encoding of a single multi-state character), extreme violations of the independence assumption that arise from this type of binary encoding need to be avoided unless thorough theoretical and/or simulation studies have been done to support a claim that there is enough data with a strong enough signal to overcome even such gross imperfections in the model.

Lastly, we come to the question of how such single character models may be combined into a model for a whole data set consisting of many characters. As we discussed in Section 2, the only really feasible way of doing this is to justifiably assume that the evolution of different characters is stochastically independent. Any attempt to work with dependent characters would introduce another layer of complexity to the model and involve some hard thinking about how one could sensibly model any dependencies present. Moreover, any model that allows dependence would entail yet more parameters that must then be estimated from the relatively meager amount of data available in most linguistic settings.

4. Parameter estimation

Once one has a model of the form described in Section 3, the next question that arises is how one goes about estimating the various parameters in the model, that is, the tree, the times of the nodes, the root distributions π_c, and the matrices $Q_{c,e}$.

In order to understand the issues involved, it is helpful to consider a generic statistical problem, that of non-parametric regression, where analogous issues arise.

Suppose that our data consist of pairs of real numbers

$$(x_1, y_1), \ldots, (x_n, y_n),$$

where the x_i are thought of inputs that are under an experimenter's control whilst the y_i are the corresponding outputs and have a stochastic component to them (for example, measurement error may be present).

A reasonable model in such circumstances is that $y_1, \ldots, y_n$ are realisations of independent random variables $Y_1, \ldots, Y_n$, for Y_i of the form

$$Y_i = f(x_i) + \epsilon_i,$$

where f is an unknown function (the *regression function*) that needs to be determined and the ϵ_i are independent random variables with some known distribution (for example, the normal distribution with mean zero and variance σ^2 for a known value of σ). The function f is the parameter in this model that must be estimated.

A standard procedure for constructing estimates of parameters in statistical models is that of *maximum likelihood* whereby one writes down the probability of the data under the model (or, more correctly in this continuous case, the probability density – the term likelihood subsumes both cases) and attempts to find the choice of parameters that maximises the likelihood. Maximum likelihood is not the only method for constructing sensible estimates, but most of the difficulties we discuss are also present with other approaches, so we will confine our discussion to this approach.

The likelihood in this case is given by

$$L(y_1, \ldots, y_n; x_1, \ldots, x_n, f) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f(x_i))^2}{2\sigma^2}\right).$$

Note first of all that there is a degeneracy problem with the likelihood function. Any two regression functions f' and f' that satisfy $f'(x_i) = f''(x_i)$ for $1 \le i \le n$ will have

$$L(y_1, \ldots, y_n; x_1, \ldots, x_n, f') = L(y_1, \ldots, y_n; x_1, \ldots, x_n, f'')$$

for all values of $y_1, \ldots, y_n$. Thus the probability model for the data is the same for f' and f'' and we have no hope of deciding which of these two regression functions is behind the data. This is an instance of the problem of identifiability that we discussed in the Introduction.

In particular, we see that the likelihood will be maximised for any regression function f such that $f(x_i) = y_i$ for $1 \leq i \leq n$. Rather than being a peak with a unique summit, the likelihood surface is a flat-topped mesa.

The problem here is one of dimensionality. The space of possible regression functions is infinite-dimensional, while we only have finite-dimensional data.

However, this is not simply a problem of infinite versus finite dimensions. Suppose that instead of allowing arbitrary regression functions we insisted that the regression function be a polynomial of degree at most m. That is, f is of the form

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m$$

for some choice of real parameters a_j, $1 \leq j \leq m$. As soon as m is larger than n there are two distinct such polynomials f' and f'' for which $f'(x_i) = f''(x_i) = z_i$ for any given choice of z_i, $1 \leq i \leq n$, and so the model is no longer identifiable and again there is no unique regression function that maximises the likelihood. That is, we run into difficulties once the dimension m of the parameter set is greater than the dimension n of the data.

Even if $m \leq n$ and the model is identifiable, the maximum likelihood estimate may not be a good estimate of f if m is large compared to n because there is not enough *independent replication* present for a law of large numbers effect to take over and the noise in the data to become *averaged out*. That is, when the n is not much greater than m, relatively small changes in the observations can have significant effects on the estimates of the parameters, and so the noise in the data leads to substantial variability in the estimates. In other words, the estimates are correct on average, but for any particular data set there is a substantial probability that the estimates are quite far from the true values. It may be comforting to the statistician to know that in a life-time of applying such procedures he/she will be correct on the average, but this is of little comfort to the practitioner who wants make substantially correct inferences for a particular data set.

These same issues arise in phylogeny: it would be nice to have maximal flexibility by allowing the matrices $Q_{c,e}$ to vary from character to character and edge to edge and for each $Q_{c,e}$ to be completely unconstrained within the class of rate matrices on the state space of the character c. However, even if such rich models are identifiable they may well have so little *replication per parameter* that parameter estimates are unacceptably variable. As (Embleton, 2000) remarks with reference to the use of complex models in linguistic dating

> All those variables/parameters in those elegant models then have to be estimated, and that is where the problems begin. It is virtually impossible to do even with languages for which we are blessed with both extensive written records over a long time-span and an overwhelming amount of scholarly attention, for example, Germanic or Romance ... There will simply be too many parameters to estimate, making it either impossible to proceed or else only possible to proceed in a somewhat circular manner.

It should be stressed that the problem of insufficient replication may not affect all parameters equally. For example, even in cases where the rate matrices $Q_{c,e}$ or edge durations t_e can't be estimated that well, it may still be possible to estimate the tree topology quite accurately.

For example, in molecular phylogenetics, under even fairly complex models of evolution (such as the *General Markov model* (Steel, 1994)), the tree topology can be reconstructed with high probability from not that many characters; this is studied theoretically in (Erdős *et al.*, 1999), and confirmed in simulation studies (Nakhleh *et al.*, 2001). Therefore, for some models and some reconstruction methods, the tree topology can be estimated quite well, even though the other parameters (edge durations and substitution matrices) may be hard to estimate.

There are various options for dealing with this problem. The simplest is to restrict the class of regression functions to some low-dimensional subspace. For example, one could insist that f is linear (that is, $m = 1$ in the above polynomial regression). The analogue in phylogeny problems would be to insist that the the matrices $Q_{c,e}$ come from some low-dimensional class. Molecular or lexical clock hypotheses are of this form, in that they essentially posit that $Q_{c,e}$ is independent of the edge e. The problem with such approaches is that the low-dimensional class may simply not be rich enough to effectively capture the features of the data: in the regression example, a scatter plot of the pairs (x_i, y_i) might have a pronounced U-shape and hence it is apparent that a linear choice of f is simply not appropriate. The molecular and lexical clock hypotheses have been discredited for a wide range of biological and linguistic data (see, for example, (Bergsland & Vogt, 1962; Embleton, 2000; Arbogast *et al.*, 2002; Felsenstein, 2003)).

The next simplest option is, in statistical parlance, to replace *fixed effects* by *random effects*. In the polynomial regression example, one could, for instance, take the coefficients $a_0, \ldots, a_m$ to themselves be realisations of independent normal random variables with mean μ and variance τ^2. The effect of this move is to turn a problem with $m+1$ parameters into one with two parameters while not restricting f to lie in a low-dimensional space. Analogous approaches have been tried in phylogeny where rate parameters (either across characters or edges) have been taken to be independent, identically distributed realisations of random variables with common distribution from some parametric family such as the gamma distributions or as the output of an unseen Markov random field on the tree (so we have a *hidden Markov model*). A discussion of these approaches with references to the literature may be found in Chapter 16 of (Felsenstein, 2003). Proponents of these models typically do not propose that this structure should be taken too literally. Rather, they assert informally that the range of 'proxy' parameter values that are likely to be produced by such a mechanism can match the heterogeneity seen in the 'actual' parameters. Hence this approach is essentially a somewhat *ad hoc* device for mathematical convenience in situations where the various numerical parameters are *nuisance parameters* that are not the chief focus of scientific interest and the practitioner is most interested in establishing the correct tree topology. Note that this sort of random effects model is still effectively imposing something like a molecular or lexical clock: the independent, identically distributed rates model forces the now random rates to have the same distribution (and hence, for example, the same mean) on each branch of the tree, and hidden Markov models also force a great deal of homogeneity amongst branches.

Random effects models implicitly replace the original likelihood function by one in which there is a penalty for the parameters being too heterogeneous: the parameters are forced to look like a typical realisation of the random generating mechanism. An approach which tries to do the same thing more explicitly is that of *penalised likelihood*. In the regression example, one could look for functions that don't maximise the likelihood but rather a new function of the data and the regression function f that incorporates both the likelihood and a *penalty* for functions that are too 'rough'. For example, one could index the data so that $x_1 < x_2 < \cdots < x_n$ and seek to maximise

$$\log L(y_1, \ldots, y_n; x_1, \ldots, x_n, f) - C \sum_{i=1}^{n-1} [(f(x_{i+1}) - f(x_i))/(x_{i+1} - x_i)]^2,$$

where C is a positive *tuning constant*. The larger C is the more preference is given to 'smooth' functions f, and different choices of C will typically lead to different estimates. There is no canonical choice of C and the choice of this constant is essentially a matter of taste for the practitioner.

Analogous approaches have been proposed in phylogeny in (Sanderson, 1997; Sanderson, 2002). There the logarithm of the likelihood is modified by the addition of a term that penalises daughter branches for having rates that are far from the rate of their parent, so that that some sort of local smoothness of rates is rewarded. Such approaches are *ad hoc* in the extreme.

To begin with, the object being maximised is no longer a likelihood for any probability model. Instead it is an apples-and-oranges hybrid that attempts to combine two incommensurable quantities, and the appealing rationalisation of the maximum likelihood method is lost. Moreover, the choice of roughness penalty and the particular manner in which it is combined with the likelihood are arbitrary and there is no particularly convincing reasoning why one choice is better than another. Sanderson offers a somewhat heuristic explanation of his method by claiming that it is, in some vague sense, a proxy for maximum likelihood in a random effects model in which there is local statistical dependence between the random rate parameters for adjoining branches, but this is not justified by actually proposing a formal random effects model that incorporates such a dependence structure. As with a random effects model, the penalised likelihood approach is a mathematical device rather than a procedure driven by a clear and convincing modelling rationale. Unlike random effects, however, penalised likelihood is not a model, but simply an inferential procedure applied to a model that one could fit to data in other ways. This point seems to be a cause of some confusion. For instance, (Gray & Atkinson, 2003) speak of the 'penalized-likelihood model'.

5. Data issues

The structures of the datasets that are typically used in linguistic cladistics are not obviously appropriate as a basis for statistical inferences. Most are based on the famous 'Swadesh list' of basic vocabulary. A brief examination of the Swadesh list will suggest where pitfalls can be expected to lie.

Morris Swadesh originally constructed his comparative list with several objects in mind (see e.g. (Swadesh, 1955)). He attempted to include only lexical items which are psychologically 'basic' and culturally universal; he also hoped to include lexemes which are maximally resistant to replacement. The criterion of 'basicness' raises no clear issues for the statistician, but the others are clearly problematic for at least the following reasons.

It proves to be impossible to construct a list of even a hundred lexical meanings that are genuinely universal (Hoijer, 1956). Languages not uncommonly use the same word for 'come' and 'go', or for 'yellow' and 'green'; 'who' and 'what' are often morphologically derived from the same root (e.g. in Indo-European languages); 'this' and 'that' are likewise often morphologically derived from the same root, or are not distinguished at all, or the language's system of demonstratives is more complex– and so on. These and similar problems virtually guarantee that there will be some duplication among the items of even the most basic wordlist, which will of course affect any statistical inferences based on the list. The alternative is to tailor the lists for individual languages or families; but in that case we no longer have a standard list to be used for all experiments.

This last outcome would not be a problem if the wordlists used were in any sense representative random samples from an explicit, well-defined population; but a Swadesh list is anything but a random selection of lexical items from some *a priori* prescribed wider universe with an "empirical existence" in the words of the quote from (Berk & Freedman, 2003) given in Section 2. Most importantly, the empirical distribution of rates of replacement of the items over time is heavily skewed toward the retentive end of the scale. Words that are fairly seldom or even very seldom replaced (such as pronouns, numerals, and body-part terms) are well represented on the list, while those that are replaced very often are scarcely represented at all. The consequences for statistical inferences of this skewness remains unclear.

Finally, there is the problem of sample size. Swadesh began with a list of 200 words, then reduced it to 100 in the hope of achieving maximal universality and a minimal average rate of replacement. But a glottochronological experiment by Johann Tischler has shown that the shorter list gives much more widely varying dates of divergence (Tischler, 1973). This is prima facie evidence that sample size is a problem. Even the longer Swadesh list may not be substantial enough to give statistically reliable results; as of now we just don't know.

It would seem that experiments of a different kind, using lexical items selected randomly from a significantly larger 'universe', are needed.

6. Limitations with dating internal nodes

Much of what we have said has focused on two issues: one is formulating appropriate stochastic models of character evolution (by formally stating the properties of the stochastic processes operating on linguistic characters), and the other is inferring evolutionary history from character data under stochastic models.

As noted before, under some conditions it may be possible to infer highly accurate estimations of the tree topology for a given set of languages. In these cases, the problem of dating internal nodes can then be formulated as: *given the true tree topology, estimate the divergence times at each node in the tree*. This approach is implicit in the recent analyses in (Gray & Atkinson, 2003; Forster & Toth, 2003), although they used different techniques to obtain estimates of the true tree for their datasets.

The problems with estimating dates on a fixed tree are still substantial.

Firstly, dates do not make sense on unrooted trees, and so the tree must first be rooted, and this itself is an issue that presents quite significant difficulties.

Secondly, if the tree is wrong, the estimate of even the date of the root may have significant error.

Thirdly, and most importantly perhaps, except in highly constrained cases it simply may not be possible to estimate dates at nodes with any level of accuracy. Recall that it is usual to model the data for each character with a tree that is common to all characters plus stochastic evolution mechanisms on the edges of the tree that may be character and edge specific. The evolution mechanism for character c on edge e is described by a matrix $P_{c,e}$ of conditional probabilities for possible substitutions. Different characters are usually assumed to be independent. In order for such a model to be useful for dating, it is necessary that there is some connection between $P_{c,e}$ and the duration t_e of the edge e: this is usually accomplished by positing the existence of a time-homogeneous Markov chain evolution with rate matrix $Q_{c,e}$ so that $P_{c,e} = \exp(t_e Q_{c,e})$. Note that there is some ambiguity in such a model because one can multiply t_e by some number r and divide each entry of $Q_{c,e}$ by r to arrive at the same value of $P_{c,e}$. Thus the best we can hope for is to identify the *relative* values of the durations t_e; that is, we can only get at ratios $t_{e'}/t_{e''}$ for different edges e' and e''. We need some external calibration of at least one *absolute* edge duration t_e to estimate the remaining absolute edge durations. Moreover, if the rates $Q_{c,e}$ can vary arbitrarily, then the model may be too parameter-rich for the edge durations to be estimated successfully. Even supposing we could construct the true tree, in order to estimate the t_e we would need to overcome this over-parameterisation by placing constraints on how the rates $Q_{c,e}$ can vary across characters and across edges. All attempts we know of to estimate edge durations and so estimate times at nodes are based upon either some kind of explicit assumptions about how rates can vary (such as assuming that the $Q_{c,e}$ are independent of c and e or are the output of some further homogeneous stochastic mechanism), or else they try to minimise the variation of the estimated values of the $Q_{c,e}$ in some *ad hoc* way that is supposed to be justified by implicit assumptions about the degree to which rates vary. Currently, we just do not know that any of these assumptions are sufficiently valid to suggest that such attempts are reasonable.

Unlike biological phylogenetics, in historical linguistics the amount of data we have is rather meagre and we aren't going to get much more of it. It is thus unsatisfactory to cross one's fingers and hope for the best when using data, models, or inference procedures that have obvious imperfections in the hope that there is enough signal to overcome these flaws or that further data can be collected which will confirm or discredit conclusions made using questionable methods.

Therefore we propose that rather than attempting at this time to estimate times at internal nodes, it might be better for the historical linguistics community to seek to characterise evolutionary processes that operate on linguistic characters. Once we are able to work with good stochastic models that reflect this understanding of the evolutionary dynamics, we will be in a much better position to address the question of whether it is reasonable to try to estimate times at nodes. More generally, if we can formulate these models, then we will begin to understand what can be estimated with some level of accuracy and what seems beyond our reach. We will then have at least a rough idea of what we still don't know.

7. Conclusion

In this paper we have shown that phylogenetic estimation, and in particular the dating of divergence times in historical linguistics, are instances of a statistical inverse problem, and many of the issues that complicate the proper treatment of other inverse problems are also present there. We have also argued that development of better models in linguistic evolution are needed before the dating of internal nodes can be done with any degree of accuracy and/or reliability.

8. Epilogue

As we know,
There are known knowns.
There are things we know we know.
We also know
There are known unknowns.
That is to say
We know there are some things
We do not know.
But there are also unknown unknowns,
The ones we don't know
We don't know.

– Donald Rumsfeld, U.S. Secretary of Defense

9. Acknowledgments

TW would like to acknowledge the Radcliffe Institute for Advanced Study and the Program in Evolutionary Dynamics at Harvard University, which provided generous support for this research. We also thank the McDonald Institute for inviting us to the conference, and to submit this paper.

REFERENCES

Arbogast, Brian S., Edwards, Scott V., Wakeley, John, Beerli, Peter, & Slowinski, Joseph B. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annu. Rev. Ecol. Syst.*, **33**, 707–740.

Bergsland, Knut, & Vogt, Hans. 1962. On the validity of glottochronology. *Current Anthropology*, **3**, 115–153.

Berk, Richard A., & Freedman, David A. 2003. Statistical assumptions as empirical commitments. *Pages 235–234 of:* Blomberg, T.G., & Cohen, S. (eds), *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger*, 2nd edn. Aldine de Gruyter.

Embleton, Sheila. 2000. Lexicostatistics/glottochronology: from Swadesh to Sankoff to Starostin to future horizons. *Pages 143–165 of:* Renfrew, Colin, McMahon, April, & Trask, Larry (eds), *Time Depth in Historical Linguistics*, vol. 1. The McDonald Institute for Archaeological Research.

Erdős, Peter L., Steel, Michael A., Székély, Laszlo, & Warnow, Tandy. 1999. A few logs suffice to build almost all trees - I. *Random Structures and Algorithms*, **14**, 153–184.

Evans, Steven N., & Stark, Philip B. 2002. Inverse problems as statistics. *Inverse Problems*, **18**, R55–R97.

Felsenstein, Joseph. 2003. *Inferring Phylogenies*. Sinauer Associates.

Forster, Peter, & Toth, Alfred. 2003. Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proc. Natl. Acad. Sci. USA*, **100**, 9079–9084.

Gray, Russell D., & Atkinson, Quentin D. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, **426**, 435–439.

Hoijer, Harry. 1956. Lexicostatistics: a critique. *Language*, **32**, 49–60.

Holmes, Susan P. 1999. Phylogenies: An Overview. *Pages 81–119 of:* Halloran, M. Elizabeth, & Geisser, Seymour (eds), *Statistics and Genetics*. The IMA Volumes in Mathematics and its Applications, vol. 112. Springer Verlag.

Kim, Junhyong, & Warnow, Tandy. 1999. *Tutorial on Phylogenetic Tree Estimation*. Presented at ISMB (Intelligent Systems for Molecular Biology) 1999, Heidelberg, Germany. Available electronically at `http://kim.bio.upenn.edu/~jkim/media/ISMBtutorial.pdf`

McMahon, April, & McMahon, Robert. 2000. Problems of dating and time depth in linguistics and biology. *Pages 59–73 of:* Renfrew, Colin, McMahon, April, & Trask, Larry (eds), *Time Depth in Historical Linguistics*, vol. 1. The McDonald Institute for Archaeological Research.

Nakhleh, Luay, St. John, Katherine, Roshan, Usman, Sun, Jerry, & Warnow, Tandy. 2001. Designing fast converging phylogenetic methods. *Bioinformatics*, **17**, S190–S198.

Poser, William. 2004. Gray and Atkinson – Use of Binary Characters. `http://itre.cis.upenn.edu/~myl/languagelog/archives/000832.html`

Sanderson, Michael J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.*, **14**, 1218–1231.

Sanderson, Michael J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.*, **19**, 101–109.

Steel, Michael. 1994. Recovering a tree from the leaf colourations it generates under a Markov model. *Appl. Math. Lett.*, **7**, 19–24.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *IJAL*, **21**, 121–137.

Tischler, Johann. 1973. *Glottochronologie und Lexikostatistik*. Innsbruck: IBS.

STEVEN N. EVANS, DEPARTMENT OF STATISTICS #3860, UNIVERSITY OF CALIFORNIA AT BERKELEY, 367 EVANS HALL, BERKELEY, CA 94720-3860, U.S.A

E-mail address: `evans@stat.Berkeley.EDU`

DON RINGE, DEPARTMENT OF LINGUISTICS, 619 WILLIAMS HALL, UNIVERSITY OF PENNSYLVANIA, PHILADELPHIA, PA 19104-6305, U.S.A.

E-mail address: `dringe@unagi.cis.upenn.edu`

TANDY WARNOW, DEPARTMENT OF COMPUTER SCIENCES, UNIVERSITY OF TEXAS AT AUSTIN, AUSTIN, TX 78712, U.S.A.

E-mail address: `tandy@cs.utexas.edu`