Report on JSTAT_002P_0321

Date: April 6, 2021

Author(s): Edwin Rodriguez Horta, Alejandro Lage, Martin Weigt, Pierre Barrat-Charlaix

Title: Global multivariate model learning from hierarchically correlated data

Received: 2021-03-02 14:31:30.0

Report of referee 1

This paper looks at inference problems defined on a tree. The concrete model is a standard Ornstein-Uhlenbeck process, but the application in the back of the authors' minds is the inference of protein structures on the basis of coevolutionary data. In evolutionary protein structure determination, the fact the that data comes from an evolutionary tree is often neglected, which in principle makes a general look at inference on trees a timely enterprise. The paper is well written and should be published.

A clear-eyed look at what has been done on the inference of stochastic processes on evolutionary trees would be helpful. Only a small part of the literature is on protein structure - and there the effects may be small, at least when papers are titled "Phylogenetic correlations have limited effect on coevolution-based contact prediction in proteins". The mathematical literature contains contains many examples under the name tree stochastic processes. The machine learning community has also looked extensively at diffusion processes on random trees, a pointer to the literature may be Knowles, Ghahramani "Pitman Yor diffusion trees". Another strand of the literature is in the evolution of gene expression levels (continuous variables). Work by Khaitovich may be interesting, a detailed inference method using an OU-type model appears in "Pervasive adaptation of gene expression in Drosophila" (Nourmohammad et al., 2015).

A very minor comment is on Fig 2, which does not say what the x-axis is. Captions should be self-contained. Also, the font for the axis labels in the actual figure makes it harder than necessary to figure out what the axis really is.

Report of referee 2

This manuscript by E. R. Horta and co-authors addressed the inference problem on hierarchically corrected data samples which follow an equilibrium multivariate Gaussian distribution. Assuming the phylogenetic (binary) tree of branching process from a common ancestor is given, and assuming all the observed samples correspond to the leaves of this binary tree, the authors present a theoretically solid method to infer the correlation matrix of the Gaussian distribution based on this phylogenetic tree and the data.

The method is checked to be successful for low-dimensional systems (e.g., dimension L of order 10). Extending the methodology to higher-dimensional systems is listed as an computation challenge for future investigations.

The multivariate Gaussian distribution, assumed in this work, corresponds to a particularly simple energy landscape with only a single minimum. This may be a good approximation for homologous protein sequences.

The manuscript is clearly written. It is an interesting new step in the field of inverse statistical mechanics. I think the work is a very helpful contribution and will be beneficial to future theoretical and algorithmic studies along this same direction. I recommend publication of this work in JSTAT in its present form.

Report of referee 3

General comments:
Horta et al. propose a new methodological approach to inferring the parameters of a Ornstein-Uhlebeck (OU) process defined on a tree structure. They first develop the method in a rather detailed way from the mathematical perspective and next they show the gain in accuracy in a simple synthetic system. I believe this is important and novel work that deserves publication in your journal. It is motivated by a concrete, open issue when modelling biological systems, i.e., the presence of phylogenetic correlations, it addresses this issue in a mathematically sound way (while currently mainly empirical corrections are applied), paving the way to additional theoretical and computational work in this direction. I have minor formal/technical remarks that I hope will help the authors improve the clarity and completeness of the explanation of their findings.

Main text:
- I think the authors should add some comparison to other approaches of inference for OU processes (for example in the introduction), to allow the reader to understand how the work sits in the context of dynamical inference

methods.
- I think that 'statement of the problem' is an important part that probably deserves to be a section in itself instead of a subsection of 'Methods'.
- Line 191: I would state explicitly why the inference becomes impossible.
- The extreme cases discussed at lines 193-195 seem the same discussed at lines 144-147. If this is the case, the authors could re-think whether it is worth to repeat this point or to establish explicitly the connection to what has been discussed before.
- Line 240: should 'we construct the direct product of vectors s_{a} and u_{ka}, defining vectors S_{ka} of dimensions L × N' be 'we construct the direct product of vectors u_{ka} and s_{a}, defining vectors S_{ka} of dimensions N × L'?I would also state explicitly that the superindex refers to the components of u.
- Is it necessary to introduce a new symbol, z, for the expression (13)? Note that z is used also in the Appendix for something with a different meaning (e.g. line 637)
- Line 278: double brackets around 17.
- Is there a minus in front of r.h.s of (19)?
- I think it is not immediately clear what the 'Results' section will address. I think one could immediately state (like at line 316) that the goal is to assess the accuracy of Cmax compared to the real C and that such estimate will be compared to the alternativa estimate provided by the empirical C. The definition of empirical C could be also explicitly written down. Later on, at line 374, the authors mention the 'empirical coupling matrix' but it is not clear how it is derived, is this simply then the inverse of the empirical C? Clarifying this would help better understand why
the fact that C is close to singular (line 349) has an impact especially on the error on J (Fig. 3 right).
- I thought also that it is not clear how the predictions for an 'i.i.d sample' (lines 355, 376) are obtained: do they serve as a random control giving the random expectation in for the errors in e.g. Fig. 3? I think this should be stated more clearly, also in the caption, maybe along with the numerical value itself of the i.i.d. error (that is difficult to read from the plots's scale).
- Line 332: I would add a comment on the choice of the range for gamma (why not even smaller or bigger) - this is probably due to the fact that all the variation concentrates in the range considered but I would state it.
- Line 336: It is written that the sampling is repeated 100 times, it is not clear if also the inference is repeated 100 times or if this is done to increase N, the sample size. In relation to this, it would be important to write somewhere what N is considered in the computational tests and hence what is the computational complexity given this N (and L=4,10).

- Line 346: As far as I see the l2-error is not defined anywhere - I think it would be important to have it defined somewhere (could be also in a caption or in the appendix)
- Line 370: Please add some references for the problem of contacts in proteins.
- Line 421: Please add some references for the evidence that protein landscapes are well described by pairwise models.
- Regarding the results, do the authors believe that inferring \gamma could be interesting in terms of biological applications? Could it inform about typical evolutionary timescales?

Figures
Fig. 3: I do not understand what the legend symbol for C_iid refers to (left panel). Is this simply the dotted horizontal line at the bottom of the plot? If yes I do not see why the legend has an additional central dot. In the caption, I do not understand 'The inset in both panels', isn't the inset only in the right panel? Same comment applied to caption of Fig. S9.
Fig 4: I think the caption should state clearly what the 'Perfect' case refers to. Same comment applies to caption of Fig. S10.

Bibliography
I think the author of Ref. [15] was repeated twice by mistake.

Appendix
- Section 1: it is probably worth repeating how the values for \Delta t are chosen (it is mentioned at line 318 but I think it's good to have all the technical details on generating artificial data in the corresponding section of the appendix).
- I do not understand why the comment at lines 549-552 should be useful, it tells that Eq. (2) is different from (A1) but does not really tell precisely what is the main difference, and I am confused because clearly they should be related otherwise they are not describing the same OU process. In particular, isn't (A1) simply a discretized version of the solution of (2) with an elementary time step \Delta t ?
- I think the title of section 2 ('Initializing parameters') is a bit misleading, given that this section describes how to calculate the empirical C and the empirical \gamma, which I understand are considered as alternative estimates of the true C and \gamma. I would therefore clarify this point and then mention that these values are also used to initialize the max likelihood search. The expression for the empirical C could also be placed in the main text, see comment above.

- Line 566: I think the ensemble of data was denoted by a capital X with an arrow, not as a bold X, please check for consistency of notation.
- Line 583: I did not understand what prevented the authors using the indices \alpha, \beta for the angles \theta as done in the main text.
- Expression (A5): I would suggest to use dots to indicate missing elements in the matrix
- Eqs. (A6)-(A10): It wasn't completely clear to me the order of the steps, maybe they could be presented in reverse order (that should be the order by which they are implemented)?
- Eq. (A13): What is the dot written as subindex of T? It should be clarified. Or is it a mistake? (probably one should write a 'k' in its place?)
- Line 624: 'this allows to' ->'this allows one to'
-Line 638: 'parametrization in term' ->'parametrization in terms'
Here I would write a sentence to remind what is this parametrization used for in connection to the inference problem of the main text.
- Several properties of the tree structure considered are not specified anywhere while I think it would be useful to the reader. For instance, what is a 'fully balanced' tree? How is this property related to the fact of being 'binary, symmetric and completely homogeneous'? (lines 653-654).
Clarifying what is meant by binary and symmetric would be useful, probably this could be done by a small visual sketch where also the number of levels K is indicated. Please note that it would be useful to clarify these denominations already in the main text, e.g. line 317 or when the tree structure (Fig. 1) is presented.
- Line 663: Does one need an arrow symbol in the argument of \Gamma?
- Line 666: I don't understand why it is relevant to comment on the fact that 'The eigenvectors ... can be normalized and arranged horizontally into a matrix U' given that I do not see the matrix U explicitly used anywhere else.
- When introducing the Hessian matrix (line 673) I would give its definition. The acronym LBFGS (line 678) should be spelt out (and possibly a reference provided).

Supplementary Figures
It seems to me that Figures S6-S8 are not mentioned in the main text. In Fig. S7 it is important to say what is L (otherwise it is not clear in what respect the fig. is different from Fig. 2).