LINKÖPING UNIVERSITY

Linköping, March 20, 2021

PhD student in Statistics (Ref IDA–2021–00030)

Krzysztof Bartoszek
Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University, Linköping, Sweden
Contact: krzysztof.bartoszek@liu.se

The Division of Statistics and Machine Learning (STIMA) at Linköping University is expanding. In this call we are looking for a new PhD student to work on the development and mathematical analysis of statistical inference methods for stochastic processes in phylogenetics, as part of an ELLIIT Call C collaboration grant with the Department of Biology, Lund University. This document briefly presents the research at the division and the background of the PI (Section 1). This is followed by a few concrete suggestions for research topics for the PhD position (Section 2). Some administrative details are:

- **Application deadline:** May 3, 2021
- **Application procedure:** All applications are to be submitted via Linköping University's online application system, available via this link:
  `https://liu.se/en/work-at-liu/vacancies?rmpage=job&rmjob=15865&rmlang=UK` (English)
  `https://liu.se/jobba-pa-liu/lediga-jobb?rmpage=job&rmjob=15647&rmlang=SE` (Swedish)
- **Questions:** If you have any questions do not hesitate to contact Krzysztof via e–mail.

1 The Division of Statistics and Machine Learning

STIMA—The Division of Statistics and Machine Learning belongs to the Department of Computer Science at Linköping University. This fact makes us unique in Sweden, and we like to view ourselves as Sweden's most modern division of Statistics with a clear focus on state–of–the–art data analysis, prediction and decision making in complex systems. We are engaged in basic methodological research motivated by a wide range of problems in areas that span from journalism and psychology to computational biology and climatology.

Krzysztof Bartoszek, who will be the main supervisor for the PhD students recruited in this call, is a Docent and Senior Lecturer (Universitetslektor) in Statistics at STIMA. His research background is in stochastic processes with a special focus on phylogenetic comparative methods and other applications of probabilistic modelling in life science problems. In particular, his research has focused on probabilistic models and the development of computational algorithms for learning and reasoning about these models. Using probability theory, probabilistic models are able to systematically represent and cope with the uncertainty that is inherent to most data. This is of central importance in many applications of machine learning.

The group at STIMA has a wide network of strong international collaborators all around the world. The intended project work will give the student direct possibilities of working with scientists from the Łódź University, Medical University of Gdańsk, Stockholm University, University of Alabama, University of Gdańsk, University of Gothenburg, University of Oslo, University of Toruń, and Uppsala University. Other members of the Division have joint projects with researchers for example at the University of British Columbia (Vancouver), University of Cambridge, and the University of Oxford. We strive for all PhD students to get a solid international experience during their PhD studies.

2 Potential Research Topics

The research projects for the advertised position will be in the area of "Developing core–technologies for tree based models", Krzysztof Bartoszek's and Prof. Niklas Wahlberg's (Department of Biology, Lund University) ELLIIT Call C Grant that funds the PhD position (and a parallel one at Lund University). The two PhD candidates will be working closely together. A few examples of potential research topics are briefly outlined below. In brief the high–level description is that one has a tree structure and on top of this tree, i.e. along the tree's branches some stochastic processes evolves. The values of the process are observed only at the tips of the tree and based on such a sample one is to make inference about the data generating stochastic process and possibly the tree itself. The approach has as its goal to model the evolution of traits, e.g. body size, on a between–species level. In the scope of the project the focus will be on methods for modelling and inference on evolutionary biology applications.

As an applicant you are not required to specify a specific research topic in your application, but you are of course welcome to do so if you want. Below we list topics that are within the scope of the ELLIIT Call C and we welcome your own initiatives around them. Depending on the interests of the students the topics can be approached from either a mathematical perspective, a software development/computational one or a mixture of the two. Apart from the main methods development work the PhD student will have the opportunity to collaborate directly with the "parallel" PhD student at the Department of Biology, Lund University, and furthermore to interact with biological researchers from various backgrounds and institutions, testing the methods and contributing to the analyses of real data.

Exploring the likelihood surface: Contemporary developments have resulted in fast likelihood calculation methods (e.g. **PCMBase** `R` package with its `C++` backend, **PCMBaseCpp**) for a very general class of (multivariate) Gaussian processes, evolving on a tree. This finally allows for a detailed exploration of the likelihood surface under multivariate Gaussian models, starting with the Ornstein–Uhlenbeck process. We actually do not even know if it is uni– or multimodal. This work could be done in collaboration with my current PhD student.

Punctuated equilibrium models of evolution: While most modelling approaches assume gradual evolution, i.e. the trajectory of the traits is continuous, biological theory and the fossil record point to the possibility of jumps (i.e. punctuated equilibrium). Due to computational difficulties including jumps is an area of promising research. It is clear from current studies that inference methods for and development of Lévy process models is desired.

In this project we will try study inference for traits that can jump at speciation points. A possible starting point for accomplishing this goal is to use currently available models and software, modified with a jump possibility. An open question is whether, based on only a contemporary sample, the jumps effects can be singled out.

PPL based modelling: Current phylogenetic comparative methodologies are not sufficient for more complex modelling setups. Hence, an approach based on probabilistic programming

languages (PPLs) has seen recent development in phylogenetics, e.g. Prof. Fredrik Ronquist's (Stockholm Natural History Museum) `RevBayes` or `TreePPL`. PPLs are a very flexible modelling approach, where probabilistic models are constructed by overloading standard programming operations to have probabilistic meanings. Hence, a domain expert, e.g. a biologist, can define a probabilistic model by with standard code, defining certain variables to be unknown and random. PPLs are very generic, and hence the proposed phylogenetic inference motivated research, could have far-reaching consequences for a wide range of applications. As part of this work one could develop Sequential Monte Carlo-based back-end inference engines for phylogenetic applications of PPLs. This work could be done in collaboration with Dr. Fredrik Lindsten (STIMA, Linköping University) and Prof. Ronquist's group.

Sparse interactions: Each species (i.e. tip in a phylogeny) can be described by a huge number of traits (e.g. gene expression levels for more than tens of thousands of genes). However, one knows that the dependency structure between the variables is sparse—genes only interact with a small number of genes. Hence, versions of penalized estimation methods for sparse covariance structures for dependent data are much desired.