Master thesis proposal
Kronecker–structure covariance estimation in the small–sample–high–dimension case with phylogenetic applications (2 topics)

Krzysztof Bartoszek

September 27, 2025

Background

Closed form maximum–likelihood formulæ for estimation under a Kronecker–structured covariance model are known. However, when the number of dimensions of the data is higher than the number of observations they might not be valid and alternative approaches might be necessary. Some methods, e.g., based on the Frobenius norm or entropy loss function have been developed by Dr. Monika Mokrzycka (Institute of Plant Genetics, Polish Academy of Sciences), see for example [2]. Such a situation is natural in non–standard phylogenetic comparative methods, e.g., when analyzing gene expression levels or microbiome composition. The fundamental model for phylogenetic comparative methods is the branching Brownian motion, which conditioned on the phylogeny has a Kronecker–structured covariance matrix.

Thesis project

The aim of the thesis is to investigate behaviour of various methods of estimation for Kronecker–structured covariance models in the phylogenetic setting. Due to the available data two topics are formulated

Topic I

Human lung cancer gene expression—measurements of gene expression levels are available for a small number of mutated clones of human lung cancer cells [1]. The aim would be to estimate the covariance structure and see if the methods provide the same grouping of genes as in [1] or does it provide new insights. Alternatively, some a priori structure can be superimposed on the covariance matrix to be estimated based on known interactions between genes. An additional element here would the creation of the phylogeny from the sequences of the clones. Due to the nature of the data the phylogeny would not have an

evolutionary interpretation—rather would be a measurement of the similarity of the clones and hypothesized similarity of gene expression behaviour.

Topic II

An exciting direction of current research is the analysis of the microbiome composition of various tardigrade species [4]. This phylum has well known for being able to survive in extreme conditions and it is hypothesized that their microbiome plays a role in this ability. Hence, question would be whether including a phylogenetic analysis allows for identifying groups of microbiome OTUs that would be interesting for further downstream analyses. Alternatively, some a priori structure can be superimposed on the covariance matrix to be estimated based on known dependencies between OTUs. An additional element here will be that each tardigrade species is described by a vector of OTUs, with many zeroes. Hence, some transformations/further modelling might be useful to align with the assumptions behind using a branching Brownian motion model. One possible direction is are excess/structural zero models [5, 7, 6].

Goals

1. Investigate whether methods based on Kronecker–structured covariance models are applicable in the high dimensional phylogenetic setting.

2. Compare if Kronecker–structured covariance based methods are able to outperform already proposed methods, e.g. [3, 8].

3. Investigate if methods based on Kronecker–structured covariance model are able to confirm found groups of genes, OTUs.

Data

Data will be provided for Topic I by Prof. Marcin Okrój (Department of Cell Biology and Immunology, Intercollegiate Faculty of Biotechnology, University of Gdańsk and Medical University of Gdańsk) and for Topic II by Dr. hab. Monika Mioduchowska (Department of Genetics and Biosystematics, Faculty of Biology, University of Gdańsk). Work will be done in collaboration with the respective data provider and also with Dr. Jolanta Pielaszkiewicz from our Institution.

References

[1] A. Felberg, M. Bieńkowski, T. Stokowy, K. Myszczyński, Z. Polakiewicz, K. Kitowska, R. Sądej, F. Mohlin, A. Kuźniewska, D. Kowalska, G. Stasiłojć, I. Jongerius, R. Spaapen, M. Mesa-Guzman, L. M. Montuenga, A. M. Blom, R. Pio, and M. Okrój. Elevated expression of complement factor i in lung cancer cells associates with shorter survival—potentially via non-canonical mechanism. *Transl. Res.*, 261:1–13, 2024.

[2] K. Filipiak, D. Klein, A. Markiewicz, and M. Mokrzycka. Approximation with a Kronecker product structure with one component as compound symmetry or autoregression via entropy loss function. *Linear Algebra and its Applications*, 610:625–646, 2021.

[3] E. W. Goolsby. Likelihood-based parameter estimation for high-dimensional phylogenetic comparative models: Overcoming the limitations of "distance-based" methods. *Syst. Biol.*, 65(5):852–870, 2016.

[4] Ł. Kaczmarek, M. Roszkowska, I. Poprawa, K. Janelt, H. Kmita, M. Gawlak, E. Fiałkowska, and M. Mioduchowska. Integrative description of bisexual *Paramacrobiotus experimentalis* sp. nov. (Macrobiotidae) from republic of Madagascar (Africa) with microbiome analysis. *Mol. Phyl. Evol.*, 145:106730, 2020.

[5] A. Kaul, O. Davidov, and S. D. Peddada. Structural zeros in high-dimensional data with applications to microbiome studies. *Biostatistics*, 18(3):422–433, 2017.

[6] A. Kaul, S. Mandal, O. Davidov, and S. D. Peddada. Analysis of microbiome data in the presence of excess zeros. *Frontiers in Microbiology*, 8:2114, 2017.

[7] J. Li. Classification of microbiome data with structural zeroes and small samples, 1 2021. Master thesis in Statistics, Division for Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, Sweden.

[8] S. Wang, T. T. Cai, and H. Li. Hypothesis testing for phylogenetic composition: a minimum-cost flow perspective. *Biometrika*, 108(1):17–36, 2020.