Computing the Fixed-Point Densities of Phylogenetic Tree Balance Indices

Hao Chi Kiang[1]

[1] *Division of Statistics and Machine Learning, Linköping University, Sweden.*
hao.chi.kiang@liu.se

One way of detecting evolutionary pressure is to test whether a rooted phylogenetic tree G were drawn from the *Yule model*.[1] A convenient way to test such hypothesis is to use the distribution of some summary statistics $T(G)$ of the observed tree G, and reject the hypothesis if $T(G)$ is too extreme if G were Yule. For instance, when $T(G)$ is the sum of the number of edges between each leaf and the root, it is the widely-used Sackin's index [2]; and Colless' index [3] and the cophenetic index [4] are similarly defined statistics which involves summing up the edges in different ways. These indices, all of which designed to capture the 'balance' of the tree, shares a very interesting property: letting n be the number of tips in the random Yule tree G_n, when $n \to \infty$, the distribution F_n of $n^{-k}[T(G_n) - \mathbb{E}T(G_n)]$, for some k dependent on which index T is, converges weakly to the fixed point F^* of a contraction mapping S in the Wasserstein metric space of probability distributions. Furthermore, denoting by $L(Y)$ the law of any r.v. Y, the contraction S can be written as

$$S(F) = L\left[\sum_{i=1}^{M} g_i(\tau) X_i + C(\tau)\right],$$

where $(X_i)_{i=1}^M$ are i.i.d. with distribution F, the r.v. τ is Unif(0,1)-distributed independent of any of X_i, and $(g_i)_{i=1}^M$, C are some 1-dimensional functions.

My work concerns the problem of accurately computing the tail of the density of F^*, which can be used for large-sample hypothesis testing. In particular, I presented a new iterative algorithm that approximates the fixed point density directly via successive numerical integrations, reviewed some other existing methods including approximate simulation, as well as attempted to speed up a very slow rejection algorithm. Numerical results show that the new algorithm converges well in Wasserstein distance.

References

[1] Yule GU (1925) II.—A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. *Philosophical transactions of the Royal Society of London Series B, containing papers of a biological character, 213(402-410)*, 21-87

[2] Sackin M (1972) "Good" and "bad" phenograms. *Systematic Biology, 21(2)*, 225–226

[3] Colless, D. (1982). *Systematic Zoology, 31(1)*, 100-104. doi:10.2307/2413420

[4] Mir A, Rosselló F, et al. (2013) A new balance index for phylogenetic trees. *Mathematical biosciences 241(1)*, 125–136

Cumulative hasard estimators

Haris Fawad[1] and Kjetil Røysland[2]

[1] *Department of Biostatistics, University of Oslo, Norway, haris.fawad@medisin.uio.no*
[2] *Department of Biostatistics, University of Oslo, Norway, kjetil.roysland@medisin.uio.no*

The goal of causal inference in survival analysis is to estimate the effect of an intervention, such treatment, on patient survival. Using the framework of counting processes, interventions can be modelled as changes to the intensity processes. In practical terms, this leads to a weighted population of the observational data at hand, which represents data that would have been recorded had the intervention been implemented. To estimate relevant outcome measures such as survival, we can transform estimates of the cumulative hasard [1]. A consistent estimator for the cumulative hasard in a weighted population is the weighted Nelson-Aalen estimator, which is given as

$$\hat{A}_t^{(n)} := \sum_i^n \int_0^t \frac{R_{s-}^{i,(n)}}{Z_s^{(n)}} dN_s^i,$$

where $R_t^{i,(n)}$ is a consistent estimate of individual i's weight, N_t^i is a counting process that records the event of interest (death), and $Z_t^{(n)} := \sum_i^n R_{t-}^{i,(n)} Y_t^i$, where Y_t^i denotes the at-risk indicator function.

We have considered two alternative estimators for the cumulative hasard, which are given by

$$\tilde{A}_t^{(n)} := \hat{A}_t^{(n)} - \sum_i^n \int_0^t \frac{\tilde{A}_{s-}^{(n)}}{Z_s^{(n)}} dR_s^{i,(n)},$$

and

$$\bar{A}_t^{(n)} := \sum_{i=1}^n \int_0^t \frac{Y_s^i R_{s-}^{i,(n)} X_{s-}^{i,\top}}{Z_s^{(n)}} dH_s^{(n)},$$

where $H_t^{(n)} = \int_0^t \beta_s^{(n)} ds$ are estimated coefficients of an additive regression model such as $\alpha_t = Y_t X_{t-}^\top \beta_t$, where α_t denotes the hasard rate of the event of interest (death).

Based on a simulation study, the recursive estimator ($\tilde{A}$) and the regression estimator ($\bar{A}$) seem to behave similarly to the Nelson-Aalen estimator ($\hat{A}$). Intuitively, both the recursive estimator and the regression estimator utilise more information from the data compared to the Nelson-Aalen estimator. Yet, all three estimators display the same level of efficiency; they all have similar point-wise confidence intervals (bootstrapped).

References

[1] Ryalen, Pål C and Stensrud, Mats J and Røysland, Kjetil (2018). Transforming cumulative hazard estimates. *Biometrika*, **105-4**, 905–916.

Change point detection in environmental data using quantile regression

Svetlana Aniskevich[1]

[1] *University of Latvia, Latvia, aniskevich.s@gmail.com*

Nowadays with constantly increasing number of datasets the change point detection in series is becoming an actual task. A change point – a time period when there is a change in data distribution, could be of different nature: a change point that comes from a dataset itself, e.g., pauses in a speech recording, or could arise from an external source. In environment monitoring observations are a key source of information about processes, yet the subject to external influence. Usually, such datasets are gathered from a network of stations, that might be working for centuries and inevitably are influenced by various changes in the landscape, measurement methods and instruments.

In order to detect a change in dataset there are a lot of methods: tests and routines [1], and also there are some procedures that are specifically developed to identify such features in climatic data [2]. In addition, in environmental applications, it is necessary not only to identify a break point, but also to perform a data correction, so the data would be representative and usable in any further analysis. Therefore, it was assumed that additional information on change behaviour could be of use. For example, an observation station that is overgrown with time would potentially show less high wind values, still the minimum values could be similar as before. In order to characterize such heteroscedasticity, a quantile regression behaviour was observed.

The quantile regression models with change points were used for simulated and applied data [3],[4] before and in this study the historical wind speed data in Latvia was observed. As environmental data usually are non-stationary due to various natural trends, it is also necessary to perform a detrending, for example, by calculating observation differences from the closest stations. Further it is possible to detect the changes by analysing regression coefficients or test statistics.

In addition, environmental data usually are supplemented with metadata – historical records of changes in observation stations. It was noted that some documented historical changes are also identified by the change point detection method, pointing out that there is a significant shift in the data, still some metadata records weren't confirmed by the applied method, and the other way around – it identified the change points, that weren't mentioned in metadata. Usually, bigger changes were identified in 0.25 and 0.75 quantiles, showing that further corrections should have more focus on the tails. Further it is planned to examine also other parameters, e.g. temperature.

References

[1] Aminikhanghahi, S., Cook, D.J. (2017). A Survey of Methods for Time Series Change Point Detection, *Knowledge and Information Systems*, **51(2)**, 339 – 367.

[2] Ribeiro, S., Caineta, J., and Costa, A.C. (2016). Review and discussion of homogenisation methods for climate data, *Physics and Chemistry of the Earth, Parts A/B/C*, **94**, 167 – 179.

[3] Zhou, M., Wang, H. J., and Tang, Y. (2015). Sequential change point detection in linear quantile regression models, *Statistics & Probability Letters*, **100(C)**, 98 – 103.

[4] Li, C., Dowling, N. M., and Chappell, R. (2015). Quantile regression with a change-point model for longitudinal data: An application to the study of cognitive changes in preclinical alzheimer's disease, *Biometrics*, **71(3)**, 625 – 635.

A shared parameter model for accounting for drop-out not at random in a predictive model for hypertension using the HUNT study

Aurora Christine Hofman [1], Lars Fredik Espeland [2], Ingelin Steinsland [3], and Emma Ingeström [4]

[1] *The Norwegian University of Science and Technology, Norway, aurorach@stud.ntnu.no*
[2] *The Norwegian University of Science and Technology, Norway, larsespeland1@gmail.com*
[3] *The Norwegian University of Science and Technology, Norway, ingelin.steinsland@ntnu.no*
[4] *The Norwegian University of Science and Technology, Norway, emma.ingestrom@ntnu.no*

This work proposes and evaluates a shared parameter model (SPM) to account for data being missing not at random(MNAR). The method is evaluated on a large cohort using data from the Nord-Trøndelag Health Study (HUNT) for a predictive model of systolic blood pressure ten years ahead based on current observations.

The proposed SPM consists of a linear model for the systolic blood pressure and a logistic model for the drop-out process connected through a shared random effect. Both models use the current systolic blood pressure, age, sex, and body mass index as explanatory variables. In the logistic model age is included as a smooth effect, while the other effects are assumed to be linear. This is a Bayesian latent Gaussian model and hence it is suitable for the integrated nested Laplace approximation (INLA) methodology for approximate Bayesian inference. This is done using R-INLA.

To evaluate the SPM we compare the parameter estimates and predictions of the SPM with a naive linear Bayesian model using the same explanatory variables while ignoring the drop-out process. This corresponds to assuming data to be missing at random (MAR). Both models are trained on data from HUNT1 (1984-86) and HUNT2 (1995-97) where the explanatory variables are from HUNT1 and the response is from HUNT2.

In addition, two simulation studies are performed in which the naive model and the SPM are tested on data with known parameters when missingness is assumed to be both MNAR and MAR.

Fitting the SPM and the naive model to HUNT data results in different parameter estimates. The SPM indicates that participants with high systolic blood pressure at HUNT2 have a higher probability of dropping out, suggesting that the data are MNAR. This effect is supported by the simulation studies.

Lower-dimensional Bayesian Mallows Model (lowBMM) for variable selection applied to gene expression data

Emilie Eliseussen Ødegaard[1] **and Valeria Vitelli**[2]

[1] *Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Norway, emilie.odegaard@medisin.uio.no*

[2] *Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Norway, valeria.vitelli@medisin.uio.no*

Variable selection is crucial in statistical genomics since in high-dimensional omics-based analyses it is biologically reasonable to assume that only a small piece of information is relevant for prediction or subtyping. However, a priori ad hoc selection of genes according to in-sample statistics or outer literature information is still a quite common practice, which heavily affects the solidity and reproducibility of the results. Rank-based methods, making use of the ordering of data points instead of the actual continuous measurements, increase the robustness of conclusions when compared to classical statistical methods. We develop a novel extension of the Bayesian Mallows model for variable selection that allows for a full probabilistic analysis, leading to coherent quantifications of uncertainties. We test our approach on simulated data using several data generating procedures, demonstrating the method's versatility and robustness under different scenarios. Specifically, we run our method on gene expression data from ovarian cancer samples and verify our findings with a gene set enrichment analysis showing its usefulness in the context of signature discovery for cancer genomics, with uncertainty quantification playing a key role for subsequent biological investigation.

A joint Bayesian framework for measurement error and missing data

Emma Sofie Skarstein[1] **and Stefanie Muff**[2]

[1] *NTNU, Norway, emma.s.skarstein@ntnu.no*
[2] *NTNU, Norway, stefanie.muff@ntnu.no*

Despite measurement error and missing data being a common problem for most applied scientists, and despite a number of resources (see for instance, [1], [2] and [3]), many researchers are still not routinely accounting for varying types of measurement error in their data. We aim to develop and make easily available new techniques of dealing with measurement error in covariates, that have previously been less explored. We aim to develop a method for viewing missing data as a limiting case of measurement error, allowing scientists to account for all their measurement errors in the same framework, as well as developing methods for dealing with measurement error in categorical covariates. A Bayesian hierarchical structure provides a flexible and convenient framework that also allows us to incorporate prior knowledge we may have about the nature of the measurement error. The methods can then be efficiently implemented for potentially complex models via integrated nested Laplace approximations (INLA). In addition to developing these methods, we will examine in depth in which way and to what extent different types of measurement error affect inferences, by studying examples in published research, and suggest ways of dealing with this.

References

[1] Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006) *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*, CRC Press

[2] Buonaccorsi, J.P. (2010) *Measurement Error: Models, Methods, and Applications*, CRC Press

[3] Yi, G.Y. (2017) *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*, Springer New York

Issues that complicate the model selection process when determining neuronal tuning using event count models

Fredrik Nevjen[1] **and Benjamin Dunn**[2]

[1] *Norwegian University of Science and Technology, Norway, fredrik.nevjen@ntnu.no*
[2] *Norwegian University of Science and Technology, Norway*

All models are wrong, but maybe we should be more specific. Event count models such as the Poisson generalized linear model are often used to study the role of singular neurons in the brain, which in certain cases exhibit increased activity as a function of one or few external covariates. However, the selection of time scale and model components, imbalance in tracked covariates, and improper division of data into training and validation sets are all issues that can lead to misleading results that the scientific community can mistake for truth. Identifying, disclosing, and resolving the potential statistical issues in neural data is crucial for the advancement of neuroscience. Our overarching goal is to develop robust methods for model selection in this setting, thus requiring these issues to be thoroughly sorted out.

Marker-based estimation of additive genetic variance using dimension-reduction techniques – adaptions from animal breeding adapted to wild study systems using a Bayesian framework.

Janne Cathrin Hetle Aspheim[1,2], Stefanie Muff[1,2], Henrik Jensen[1,3], Jane M. Reid[1,3], Geir Bolstad[4], Robert B. O'Hara[1,2], Thomas Kvalnes[1,3]

[1] *Centre for Biodiversity dynamics, Norwegian University of Science and Technology, Norway*
[2] *Department for Mathematical Sciences, Norwegian University of Science and Technology, Norway*
[3] *Department of biology, Norwegian University of Science and Technology, Norway*
[4] *Norwegian Institute for Nature Research*

Our world is rapidly changing and it is of great interest to investigate the potential for and speed of adaptive evolutionary change in a population. Adaptive evolutionary change relies on additive genetic variance (VA), which we seek to estimate in a precise and effective manner. In order to obtain a good estimate of VA, information on the relatedness structure in the population is crucial. The relatedness can be either pedigree-based or from genomic data, and it is encoded in a relatedness matrix. There are known advantages and disadvantages with both methods [1]. We therefore want to utilize advances in animal breeding within a flexible Bayesian framework with adaptions to the additional challenges of a wild study system.

Here we will investigate matrix reduction techniques such as singular value decomposition [2]. To this end, marker-based regression will be performed with integrated nested Laplace approximations (INLA), which we believe will aid on the challenges with relatedness matrixes, as well as being more efficient with regard to computational efficiency. We also seek to answer questions of practical relevance, such as how many genetic markers and principal components are needed to achieve satisfactory estimates of VA, as well as suggestions for suitable priors.

The methods are developed and tested using data from a free-living house sparrow population in northern Norway [3], as well as with simulated data. We hope to provide useful guidelines for estimating VA not only with the data available today, but also when the number of genotyped individuals continue to increase. To make the results practically accessible, R packages with ready-made functions will be made available.

References

[1] Steinsland, I. and Jensen, H. (2010). Utilizing Gaussian Markov Random Field Properties of Bayesian Animal Models. *Biometrics*, **66**, 763 – 771.

[2] Ødegård, J., Indahl, U., Strandén, I. and Meuwissen, T.H.E. (2018). Large-scale genomic prediction using singular value decomposition of the genoytpe matrix. *Genetics, Selection, Evolution*, **50**, 1 – 12.

[3] Jensen, H., Sæther, B-E., Ringsby, T. H., Tufto, J., Griffith, SC., and Ellegren, H. (2003). Sexual variation in heritability and genetic correlations of morphological traits in house sparrow (Passer domesticus). *Journal of evolutionary biology*, **16**, 1296–1307.

A spline-based latent variable model, with applications to neural spike train data

Martin Bjerke

Norwegian University of Science and Technology, Norway, martin.bjerke@ntnu.no

Recent advances in neural data recording have given researchers the ability to monitor a multitude of neurons simultaneously, resulting in an increasing demand for models that can handle large neural populations, while also being able to extract important, unobservable features. We propose a new model in the family of Latent Variable Models, utilizing a parametric B-Spline to model the profile of the single-neuron response function. This allows for the incorporation of known structures related to neural tuning. As an example, we demonstrate how sharing features across neurons results in improved inference of single-neuron response functions and the latent trajectory, while also enabling more informed initializations and efficient latent variable discovery.

Detecting statistical interactions in immune receptor datasets

Thomas Minotto[1] **and Ingrid Hobæk Haff**[2]

[1] *University of Oslo, Norway, thomamin@math.uio.no*
[2] *University of Oslo, Norway, ingrihaf@math.uio.no*

Recent progresses in the understanding of immune receptors have suggested that complex interactions between the amino acids are at work in the process of binding to antigens. However, current methods focus mostly on constructing good predictors for the data with complex models, and less on the understanding of the underlying processes. In this work, we study how it is possible to retrieve high-order statistical interactions in immune receptor data, with different methods and for different types of interactions. We test methods based on constructing models with interactions in mind [1], [2], [3], [4], and methods that retrieve interactions from a complex model [5], [6]. We focus on two-way to four-way interactions, implemented through various functions of the covariates in simulated immune-like data. We compare methods' performances at retrieving them, and investigate which type of interaction is easier to detect.

References

[1] Ruczinski, I., Kooperberg, C. & LeBlanc, M. (2003), Logic Regression, *Journal of Computational and Graphical Statistics*, **12**(3), 475 – 511.

[2] Hubin, A., Storvik, G. & Frommlet, F. (2020), A Novel Algorithmic Approach to Bayesian Logic Regression (with Discussion), *Bayesian Analysis*, **15**(1), 263 – 333.

[3] Rendle, S. (2010), Factorization Machines, *2010 IEEE International Conference on Data Mining*, 995 – 1000.

[4] Shah, Rajen D. (2016). Modelling Interactions in High-dimensional Data with Backtracking. *Journal of Machine Learning Research*, **17**, 1 – 31.

[5] Tsang, M., Cheng, D. & Liu, Y. (2018). Detecting statistical interactions from neural network weights. *ICLR 2018*.

[6] Louppe, G., Wehenkel, L., Sutera, A. & Geurts, P. (2013). Understanding variable importances in forests of randomized trees.

Autonomous Oceanographic Sampling Using Excursion Sets and Expected Integrated Bernoulli Variance

Yaolin Ge[1] and Jo Eidsvik[2]

[1] *Norwegian University of Science and Technology, Norway, yaolin.ge@ntnu.no*
[2] *Norwegian University of Science and Technology, Norway, jo.eidsvik@ntnu.no*

To understand complex spatio-temporal phenomena in our ocean, there have recently been increased efforts in using numerical process modeling, methods for data assimilation, novel computing and sensor technology. In particular, autonomous robots with onboard computing resources provide rich opportunities for oceanographic sampling. Ideas from statistical design are highly useful in this field, because they can help guide the robot to informative locations. In our case, we are focusing on an application to river plumes, where the goal is to use the autonomous underwater vehicle (AUV) to explore the boundary between fresh water and the saline ocean water. We construct statistical methods and algorithms for sampling design strategies that are embedded onboard the AUV. With the AUVs limiting computing resources, a Gaussian random field (GRF) model serves as a statistical proxy models for the spatial salinity field $X(\mathbf{s})$ at locations $\mathbf{s} \in \mathcal{R}^3$ (north, east, depth). This GRF model is fast to update onboard the AUV and in our case it also helps in efficient planning of sampling designs.

The main contribution of this work is a 3D full-scale adaptive (myopic) sampling strategy. The main computational tasks onboard the AUV are associated with the evaluation of design criteria which over survey time instructs the AUV engines to move in promising directions and azimuths, based on the currently available data which is integrated in the onboard GRF model for salinity. We use a criterion based on the excursion set (ES) of high salinity which separates the fjord water from the river. This ES is defined by $\{\mathbf{s} : X(\mathbf{s}) > t\}$ for a given salinity threshold t. We show closed form expressions for the Expected Integrated Bernoulli Variance (EIBV) associated with the ES for several sampling designs at each adaptation point [1]. The EIBV is then used as a criterion for improved mapping of the ES in the sense that it pulls the probabilities in the ES towards 0 or 1, and in this way reduces the uncertainty.

Via simulation studies and real data from the Nidelva river plume in Trondheim, we study the properties of the EIBV sampling plans in the 3D domain. There is added value of having intelligence and autonomy in the AUV path planning compared with pre-scripted lawn-mower designs and yoyo patterns. We also discover challenges associated with the myopic strategy that was applied in our work.

References

[1] Olav Fossum, T., Travelletti, C., Eidsvik, J., Ginsbourger, D., Rajan, K. 2020. Learning excursion sets of vector-valued Gaussian random fields for autonomous ocean sampling. arXiv e-prints.