Enclose it or Lose it! Computer-aided Proofs in Statistics

Principal Investigator:

Dr. Raazesh Sainudiin, Department of Mathematics and Statistics, University of Canterbury, NZ

Associate Investigators:

Professor Michael Nussbaum, Department of Mathematics, Cornell University, US

Professor Warwick Tucker, Department of Mathematics, Uppsala University, SW

Professor Ziheng Yang, F.R.S., Department of Biology, University College London, UK

January 18, 2010

Abstract

Enclosure methods are a class of computer-aided proofs used in analysis. They are used increasingly to solve open problems in mathematics. The proposed project will use enclosure methods to address two open statistical decision problems:

1. rigorous parameter estimation in a chaotic statistical experiment, and
2. rigorous point estimation and exact posterior sampling in phylogenetics.

To address these problems, we will adapt and extend recent developments in contractor programming, interval constraint propagation, algebraic statistical constraints and employ a novel mapped sub-paving arithmetic. A `C++` class library that can harness UC's super computing power for such computer-aided proofs will be made publicly available along with a database of solutions.

5A. ABSTRACT OF RESEARCH PROPOSAL

Enclosure methods that rely on machine interval arithmetic — validated computer arithmetic that encloses or bounds all numerical errors — have become an important tool in computer-aided proofs in analysis. Some examples where these methods have been applied include proofs of the Feigenbaum conjectures[1], the double bubble conjecture[2], the existence of the Lorenz attractor[3] and the Kepler conjecture[4]. However, computer-aided proofs have rarely been applied to validate heuristic solutions to challenging decision problems in statistics. The aim of the proposed project is to adapt and extend enclosure methods to address two challenging statistical decision problems:

(1) rigorous parameter estimation in a chaotic statistical experiment, and

(2) rigorous point estimation and exact posterior sampling in phylogenetics.

Though these open problems arise in distinct disciplines, the algorithmic innovations we propose to address them are similar. Further, the theory, algorithms and `C++` implementation from this study can be easily generalised to other problems of these classes. This will be achieved by an international team of experts in machine interval experiments (PI Sainudiin), computer-aided proofs in analysis (AI Tucker), asymptotic statistics (AI Nussbaum) and computational phylogenetics (AI Yang).

(1) Rigorous Parameter Estimation in a Chaotic Statistical Experiment.

Chaotic or complex non-linear systems pose some of the most challenging decision problems in engineering science. One of the simplest systems of this type is the **double pendulum (DP)**. The DP exhibits rich dynamics and chaos at certain energies[5], thus making it challenging to model and measure for parameter estimation[6;7]. Rigorous parameter estimation in such systems must account for the physical limits of the sensors' empirical resolution and the computer's numerical resolution. Past experiments with DP systems[6] [8] were not rigorous; they neither enclosed the uncertainty in DP's angular positions nor employed validated numerical methods.

The PI and AI Tucker have recently designed a measurable DP[9] (see Fig.1) that successfully accounts for the limit of the sensors' empirical resolution. The proposed project will use data enclosures of the state trajectories of each arm of this custom-built DP[9] (Fig.2) and recent advances in interval constraint propagation[10] to obtain the first fully rigorous enclosures of point estimates for this system. We will also study set-valued extensions of classical decision-theoretic properties of our estimator[11], such as *identifiability*, *consistency* and *efficiency*, over machine-representable filtrations using the formalism of machine interval experiments[12] developed by the PI. The estimators will be cast as contractors[13] of the initial parameter space[14;15]. They exclude parameters that are not compatible with enclosures of each arm's observed angular positions through time (Fig.3).

Fig.1: DP Schematic

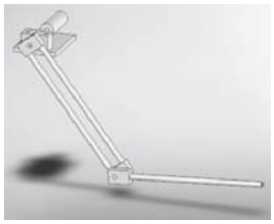


Fig.2: Streaming DP data

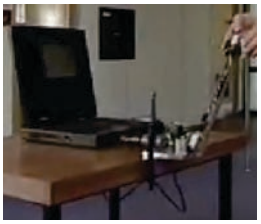
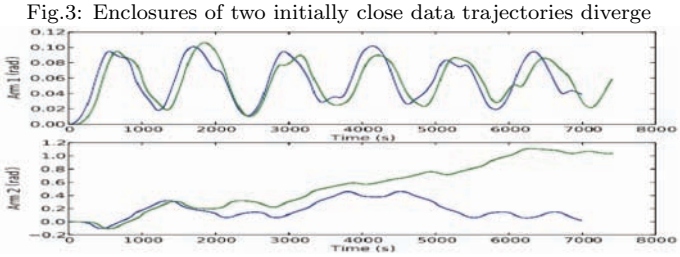


Fig.3: Enclosures of two initially close data trajectories diverge



In Year 1, the PI and AI Tucker plan to adapt a practical spline-based method[14] to contract the parameter space to a set of boxes that enclose candidate point estimates. This method will reduce our complex parameter estimation problem to a simpler problem of solving algebraic equations. In Years 2 and 3, we will adapt and extend the existing intrinsic methods developed by AI Tucker[15] for our DP problem. Unlike the spline-based method[14], intrinsic methods produce direct enclosures of derivatives from model and data alone. The PI and AI Nussbaum intend to use *deficiency distances*[11] and random dynamical systems[16] to address the asymptotic

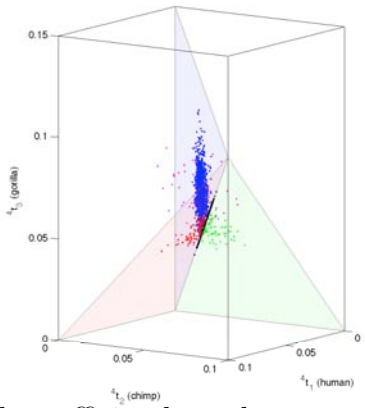
behaviour of our estimators indexed by partially ordered filtrations. These filtrations arise from (i) the number of independent trials with possibly different initial conditions, (ii) the sensor resolutions and (iii) the set of measurement times. Thus, we will conduct the first rigorous parameter estimation using classical decision-theoretic notions in a quintessential chaotic statistical experiment. The theory and algorithms developed for this model system have general implications for rigorous parameter estimation in chaotic systems.

(2) Rigorous Point Estimation and Exact Posterior Sampling in Phylogenetics. Evolutionary biologists rely on practical local search and dependent sampling algorithms to estimate evolutionary relationships between organisms heuristically[17–20] from their DNA sequences. These estimated gene trees are the basic ingredients in phylogenetic estimation of evolutionary trees between species[21] or taxa and in population genetic estimation of model parameters[22;23]. The exactness of gene tree estimation is fundamental to decision problems in disease mapping, agricultural breeding and conservation genetics. We cannot distinguish between large rooted, unrooted, 2-mixture and k-mixture gene tree(s) without first distinguishing its sub-gene trees with three[24], four[25], six[26] or more[27] taxa, respectively. Current methods do not provide rigorous estimates of gene trees with more than four taxa. Thus, obtaining exact estimates of small gene trees from finite DNA sequences is a necessary first step toward robust estimation, provided the probability model for the data is also statistically identifiable[20;28;29]. Developing rigorous point estimates and exact samplers for small gene trees from finite DNA sequence data through novel enclosure methods is the second aim of our proposed study.

Rigorous Point Estimation. AI Yang analytically obtained exact maximum likelihood point estimates (MLEs) of three-taxa trees for any data pattern[30]. Exact MLEs were later produced by computational algebraic methods that used statistical invariants in constrained optimization[31–35]. Enclosure methods developed by the PI[12;36] have solved the MLE problems with machine interval arithmetic. However, none of these approaches can solve the MLE problem exactly for five or more taxa, even for the simplest mutation models on 2 or 4 character states.

Exact Posterior Sampling. Dependent samplers, such as Markov chain Monte Carlo (MCMC) algorithms, cannot be systematically guaranteed to sample from the desired stationary distribution for any given data, even for small gene trees[17–19]. The PI developed the first exact sampler for gene trees capable of producing independent and identical samples from the posterior distribution over phylogenetic tree spaces for 3 and 4 species[37]. Currently, the PI is extending this exact sampler to a more challenging setting where possibly multi-furcating gene trees with different numbers of real-valued edge weights are part of the parameter space (Fig.4). The PI and AI Yang have observed MCMC convergence problems in rooted four-taxa trees and plan to extend the PI s exact sampler[37] in a more efficient manner for this case.

Fig.4



Our objective is to solve the MLE and the exact sampling problem efficiently and systematically for 3, 4, 5 and possibly 6 taxa multi-furcating trees. This will be achieved by obtaining tighter enclosures of gene tree likelihoods by combining machine interval arithmetic[36;37] with algebraic statistical constraints called phylogenetic invariants[32] that can be obtained from symbolic algebra software. In Years 1 and 2, the PI and AI Tucker will adapt interval constraint propagation[10] to achieve this novel fusion of symbolic and rigorous numeric methods. These tighter likelihood enclosures will be represented efficiently by a multi-dimensional metric data-structure called a regular sub-paving[38]. The PI and AI Yang will adapt these efficient likelihood enclosures for MLE and exact sampling in biologically interesting applications. In Year 3, a message-passing module for the proposed exact small tree estimation will be created and added to the PI s `C++` class library[39]. This module will harness UC s unique super-computing power in NZ to build a public database of exact estimates of small gene trees from data patterns.

References

[1] O.E. Lanford. A computer-assisted proof of the Feigenbaum conjectures. *Bull. Amer. Math. Soc. (N.S.)*, 6(3):427–434, 1982.

[2] J. Hass and R. Schlafly. Double bubbles minimize. *Ann. of Math.*, 151(2):459–515, 2000.

[3] W. Tucker. A rigorous ODE solver and Smale's 14th problem. *Foundations of Computational Mathematics*, 2(1):53–117, 2002.

[4] T.C. Hales. A proof of the Kepler conjecture. *Ann. of Math.*, 162:1065–1185, 2005.

[5] T. Shinbrot, C. Grebogi, J. Wisdom, and J.A. Yorke. Chaos in a double pendulum. *Am. J. Phys.*, 60:491–496, 1992.

[6] R.B. Levien and S.M. Tan. Double pendulum: An experiment in chaos. *Am. J. Phys.*, 61(11):1038–1044, 1993.

[7] Y. Liang and B. Feeny. Parametric identification of a chaotic base-excited double pendulum experiment. *Nonlinear Dynamics*, 52(1):181–197, 2008.

[8] M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

[9] P. Lawrence, M. Stuart, R. Brown, W. Tucker, and R. Sainudiin. A mechatronically measurable double pendulum. Distributed at `http://www.math.canterbury.ac.nz/~r.sainudiin/lmse/double-pendulum/`, 2010.

[10] H. Schichl and A. Neumaier. Interval analysis on directed acyclic graphs for global optimization. *J. of Global Optimization*, 33(4):541–562, 2005.

[11] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, 1986.

[12] R. Sainudiin. *Machine Interval Experiments: Accounting for the Physical Limits on Empirical and Numerical Resolutions*. LAP Academic Publishers, 2009.

[13] G. Chabert and L. Jaulin. Contractor programming. *Artif. Intell.*, 173(11):1079–1100, 2009.

[14] C. Michalik, B. Chachuat, and W. Marquardt. Incremental global parameter estimation in dynamical systems. *Ind. Eng. Chem. Res.*, 48:5489–5497, 2009.

[15] T. Johnson and W. Tucker. Rigorous parameter reconstruction for differential equations with noisy data. *Automatica*, 44(9):2422–2426, 2008.

[16] L. Arnold. *Random Dynamical Systems*. Springer, 1998.

[17] E. Mossel and E. Vigoda. Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science*, 309:2207–2209, 2005.

[18] E. Mossel and E. Vigoda. Limitations of Markov chain Monte Carlo algorithms for Bayesian inference of phylogeny. *Ann. Appl. Probab.*, 16(4):2215–2234, 2006.

[19] D. Stefankovic and E. Vigoda. Pitfalls of heterogeneous processes for phylogenetic reconstruction. *Syst. Biol.*, 56(1):113–124, 2007.

[20] D. Stefankovic and E. Vigoda. Phylogeny of mixture models: Robustness of maximum likelihood and non-identifiable distributions. *Journal of Computational Biology*, 14(2):156–189, 2007.

[21] J.H. Degnan and N.A. Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecol. Evol.*, 24(6):332—340, 2009.

[22] R. Sainudiin, A. Clark, and R. Durrett. Simple models of genomic variation in human SNP density. *BMC Genomics*, 8(1):146, 2007.

[23] A. Siepel. Phylogenomics of primates and their ancestral populations. *Genome Research*, 19(11):1929–1941, November 2009.

[24] A. V. Aho, Y. Sagiv, T. G. Szymanski, and J. D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal on Computing*, 10(3):405–421, 1981.

[25] M. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9:91–116, 1992.

[26] F.A. Matsen, E. Mossel, and M. Steel. Mixed-up trees: the structure of phylogenetic mixtures. *Bulletin of Mathematical Biology*, 70(4):1115–1139, 2008.

[27] P.J. Humphries. *Combinatorial Aspects of Leaf-Labelled Trees*. PhD dissertation, University of Canterbury. Mathematics and Statistics, Christchurch, New Zealand, 2008.

[28] F.A. Matsen and M. Steel. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Systematic Biology*, 56(5):767–775, 2007.

[29] E.S. Allman, S. Petrović, J.A. Rhodes, and S. Sullivant. Identifiability of 2-tree mixtures for group-based models. *arXiv:0909.1854v2*, 2009.

[30] Z. Yang. Complexity of the simplest phylogenetic estimation problem. *Proceedings Royal Society London B Biol. Sci.*, 267:109–119, 2000.

[31] S. Hosten, A. Khetan, and B. Sturmfels. Solving the likelihood equations. *Found. Comput. Math.*, 5(4):389–407, 2005.

[32] Pachter L. and Sturmfels B., editors. *Algebraic Statistics for Computational Biology*. Cambridge University Press, 2005.

[33] A. Khetan, B. Chor, and S. Snir. Maximum likelihood on molecular clock comb: Analytic solutions. *Journal of Computational Biology*, 13(3):819 – 837, 2006.

[34] B. Chor, M.D. Hendy, and S. Snir. Maximum likelihood Jukes-Cantor triples: Analytical solutions. *Mol. Biol. Evol.*, 23(3):626–632, 2006.

[35] B. Chor and S. Snir. Molecular clock forks: Symbolic mathematical analysis. *Mathematical Biosciences*, 208(2):347 – 358, 2007.

[36] R. Sainudiin and R. Yoshida. Applications of interval methods to phylogenetics. In Pachter and Sturmfels[32], pages 359–374.

[37] R. Sainudiin and T. York. Auto-validating von Neumann rejection sampling from small phylogenetic tree spaces. *Algorithms for Molecular Biology*, 4(1), 2009.

[38] L. Jaulin, M. Kieffer, O. Didrit, and É. Walter. *Applied Interval Analysis: with Examples in Parameter and State Estimation, Robust Control and Robotics*. Springer-Verlag, 2004.

[39] J. Harlow, R. Sainudiin, W. Tucker, and T. York. `MRS`: A `C++` class library for statistical set processing. Distributed at `http://www.math.canterbury.ac.nz/~r.sainudiin/codes/mrs`, 2009.