Research Statement

Hien Duy Nguyen

October 24, 2019

My core research program can be broken up into three key areas: mathematical statistics, computational statistics, and applied statistics. I will discuss in detail my achievements and progress in each of the three topics.

Mathematical Statistics

My primary work in mathematical statistics revolve around the asymptotic theory of estimators and the approximation capabilities of important functional classes in statistics, such as the class of convex combinations of certain families of density functions, or finite mixture models, for the uniform approximation of more general classes probability densities. My publications in this direction of research include the articles [31, 40, 32, 44]. In [31], my coauthors (which will now be taken as an understood acknowledgement when referring to joint work) and I demonstrate how the class of mixtures of triangular distributions can be used to arbitrarily approximate any twice-differentiable density function on the unit interval. In this paper, I prove both an approximation rate under which uniform approximation occurs, as well as a convergence rate regarding the maximum likelihood estimator of such a mixture models.

In [32], we provide a literature review of the current state of the affairs regarding the estimation and approximation rate of the maximum likelihood estimator for general location-scale finite mixture models. The results in [40] provides a convergence result regarding the use of finite mixture models with randomized parameters for density estimation on a convex domain. Finally, and most significantly, my PhD student and I, along with esteemed coauthors, were able to obtain L^p approximation results for $p \in [1, \infty]$ regarding the use of finite mixtures of continuous densities that are vanishing at infinity. Our results are more general than those previously reported in the literature, and have powerful features, such as allowing for approximation of densities on non-compact domains.

There are still a number of significant open problems that stem from this research path. Firstly, we are still investigating whether there are weaker conditions than those that we have imposed that lead to uniform and L^p approximation results. Secondly, we are investigating whether there are simpler proof methods than the convolution and Riemann integration techniques that we have used in our previous results. And thirdly, we seek weak conditions under which we may establish rates of convergence of our constructed classes for use when estimating density functions from data.

In addition to the works above, regarding the use approximation capabilities of finite mixture models, I am also actively investigating the use of mixtures of experts for functional approximation. To this end, I have used the famous Stone-Weierstrass theorem to obtain results regarding the approximation capabilities of the mean functional of mixture of experts models for both univariate and multivariate approximation, in [23] and [17], respectively. In [17], we also established conditions under which approximation in Kullback-Leibler divergence regarding conditional densities could be established. Further work in this direction will revolve around the establishment of approximation and estimation rates of certain classes of mixture of experts models.

Regarding more general estimation theory, I have made some contributions towards the study of divide-and-conquer type estimators, when data are not independent [30]. I have also been keenly interested in the study of maximum pseudolikelihood estimators (MPLEs). For example, in [14], I provided a general result regarding the consistency of the MPLE for discrete data generating processes. A specific study of the asymptotic normality of the so-called fully visible Boltzmann machine (FVBM) was conducted in [42].

A characterization of an autoregressive mixture of Laplace distributions is provided along with MPLE theorems in [36], and an asymptotic normality result regarding feasible least squares of an autocorrelated linear model is established in [15].

Miscellaneously, I established some new results regarding an interesting maximum likelihood problem regarding the triangle distribution in [29]. Here, I investigated an interesting phenomenon regarding the expected time complexity for the maximum likelihood estimator of the triangle distribution. I have also made some exploration of the concept of strict sub-Gaussianity for bounded random variables in [1].

Computational Statistics

The majority of my work in computational statistics stems from my PhD studies, which was primarily on the estimation of finite mixture models and regression variants of such models [12]. In particular, my work focuses on the development of EM (expectation–maximization) and MM (majorization/minorization–minimization/maximization) type algorithms for such models, and beyond.

Regarding finite mixture models, in [24], I developed a new EM algorithm for the popular multivariate normal mixture model. Using MM algorithm techniques for optimization of non-differentiable objectives, I proposed a method for maximum likelihood estimation (MLE) of mixtures of triangle distributions in [27].

In [26], I proposed a method for using mixture models in order to draw non-parametric inference regarding linear mixed effects models, where one cannot make assumptions regarding the random effect distribution, other than symmetry. Mixtures of linear mixed effects models were further considered for the purpose of classification and clustering of bivariate functional data in [39]. A novel MM algorithm for estimation of LASSO penalized mixtures of linear regression models was proposed in [9]. Further work in this direction includes the exploration of generalized linear model constructions as well as the investigation of asymptotic and finite sample results regarding the nature of penalized mixture regression models.

For time series clustering and classification, I further proposed two the use of mixtures of autoregressive models, estimated using MM algorithms. The papers [37] and [35] are variations on this theme, which utilized different pseudolikelihood constructions for the characterization of the estimation objective. Both sets of results have been successfully applied to the clustering and classification of time series data that arise in neuroscience. In the future, this work can be expanded to take into account inter class variation via random effects as well as time series constructions other than autoregressive models.

Similar to finite mixture of regression models are their generalization, the class of mixtures of experts. On this front, I proposed an MM algorithm for the MLE of the mixture of Laplace experts model, as well as investigated its robustness properties in [25]. My further work in mixture of experts modeling are reported in [16] and [3]. Mixture of experts models are an important probabilistic neural network and understand and estimating them will only become more important in the current era of deep network architectures.

Apart from mixture models applications, I have also made numerous research contributions to the study of EM and MM algorithms, more generally. For example, I have contributed the following tutorial paper [13] regarding the use of MM algorithms. An application of the MM algorithm framework towards the problem of fitting FVBM appears in [41]. MM algorithms for some support vector machine (SVM) variants were proposed in [28], and an MM algorithm for the estimation of heteroskedastic regression models was considered in [22].

The need for stochastic approximation optimization algorithms in order to handle estimation problems pertaining to large data sets is an important consideration of modern statistical analysis. Towards this direction, I have contributed a stochastic approximation algorithm for the MLE of a subclass of normal mixture models in [20]. Stochastic MM algorithms for the estimation of some SVM variants were considered in [21]. Recently, I studied the use of a stochastic EM algorithm for the mini-batch estimation of normal mixture models in [18]. This area of research is an important one and I hope to make further contributions towards the derivation and understanding of stochastic approximation type algorithms, when applied to non-convex and non-smooth estimation problems in the future.

Importantly, I believe that statistical methods should be made easily available, as to best benefit the scientific community. To this end, I have been diligent in making novel methodology available via software implementations, specifically in the R statistical language. For example, I have coauthored the packages BoltzMM, logKDE, lowmemtkmeans, SAGMM, and SSOSVM, that are all available on the comprehensive R archive network (CRAN). Furthermore, I believe that it is important that software adhere to community standards, such as those set out by the Journal of Open Source Software. Thus, I have also strived to prepare my software to meet these more stringent goals. Example of such preparation are my R packages that are detailed in [5, 7, 8].

Applied Statistics

Besides my computational and mathematical work, I have also endeavored to engage with the scientific community, via applied work that are both primary and consultative in nature. The first example of my applied work are towards the better understanding of proteomics experiments via distribution-robust permutation testing methodology that are constructed specifically for the data type. My work in this direction includes the descriptive articles [19] and [34], as well as the development of a web application for the application of such methodology in [4].

Further works regarding hypothesis testing are those that are applied to the various false discovery rate (FDR) control problems in neuroscience. Here, I provided a framework for model-based FDR control, under spatial correlation for structural magnetic resonance imaging (MRI) data in [33]. I further suggest a method for FDR control of quantized data, which are common in MRI studies, in [43]. Another work relating to the analysis of brain imaging data is [45], which considers the problem of generalizability of machine learning studies between different imaging cohorts. In [38], I also considered the application of mixture models to the problem of clustering time series data that arise from calcium imaging studies of zebrafish. Further research in this direction has been rapid, as I am currently collaborating with multiple neuroscience groups regarding both human imaging problems as well as problems that arise from the study of animal models.

I have also actively collaborated on research in fishery science. For example, in [11], I helped in characterizing a random effects model that lead to the improved estimation of size transition matrices of aquatic species that are sampled via catch–recapture studies. In [10], I contributed to the study of multigenerational data arising form crab fisheries studies.

In other works, I have for instance to the analysis of Australian political science data using FVBMs in [2]. I have also consulted on various applied problems such as towards the study chemical pathways via proteomics [6], the analysis of behavioral economics experiments [46], and the study of pedestrian safety for accident prevention [47]. I believe that a part of doing good statistics is to appreciate the plethora of available problems that can be studied through the statistical lens. As such, I am always open and welcoming of new and interesting applied projects.

References

[1] J Arbel, O Marchal, and H D Nguyen. On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables. *ESAIM: Probability and Statistics*, to appear, 2019.

[2] J Bagnall, A Jones, N Karavarsamis, and H Nguyen. The fully-visible Boltzmann machine and the Senate of the 45th Australian Parliament in 2016. *Journal of Computational Social Science*, to appear, 2019.

[3] F Chamroukhi and H D Nguyen. Model-based clustering and classification of functional data. *WIREs Data Mining and Knowledge Discovery*, e1298, 2019.

[4] D Chen, A Shah, H Nguyen, D Loo, K Inder, and M Hill. Online quantitative proteomics p-value calculator for permutation-based statistical testing of peptide ratios. *Journal of Proteomics Research*, 13:4184–4191, 2014.

[5] D Fryer, H Nguyen, and P Orban. studentlife: tidy handling and navigation of a valuable mobile-health dataset. *Journal of Open Source Software*, 4(40), 2019.

[6] K L Inder, Y Z Zheng, M J Davis, H Moon, D Loo, H Nguyen, J A Clements, R G Parton, L J Foster, and M M Hill. Expression of PRTF in PC-3 cells modulated cholesterol dynamics and actin cytoskeleton impacting secretion pathways. *Molecular and Cellular Proteomics*, 11(M111.012245), 2012.

[7] A T Jones, J J Bagnall, and H D Nguyen. BoltzMM: an R package for maximum pseudolikelihood estimation of fully-visible Boltzmann machines. *Journal of Open Source Software*, 4:1193, 2019. https://doi.org/10.21105/joss.01193.

[8] A T Jones, H D Nguyen, and G J McLachlan. logKDE: log-transformed kernel density estimation. *Journal of Open Source Software*, 3:870, 2018.

[9] L R Lloyd-Jones, H D Nguyen, and G J McLachlan. A globally convergent algorithm for lasso-penalized mixture of linear regression models. *Computational Statistics and Data Analysis*, 119:19–38, 2018.

[10] L R Lloyd-Jones, H D Nguyen, G J McLachlan, W Sumpton, and Y-G Wang. Mixture of time dependent growth models with an application to blue swimmer crab length-frequency data. *Biometrics*, 72:1255–1265, 2016.

[11] L R Lloyd-Jones, H D Nguyen, Y-G Wang, and M F O'Neill. Improved estimation of size-transition matrices using tag-recapture data. *Canadian Journal of Fisheries and Aquatic Sciences*, 71:1385–1394, 2014.

[12] H D Nguyen. *Finite mixture models for regression problems*. PhD thesis, University of Queensland, 2015. https://doi.org/10.14264/uql.2015.584.

[13] H D Nguyen. An introduction to MM algorithms for machine learning and statistical estimation. *WIREs Data Mining and Knowledge Discovery*, 7(e1198), 2017.

[14] H D Nguyen. Near universal consistency of the maximum pseudolikelihood estimator for discrete models. *Journal of the Korean Statistical Society*, 47:90–98, 2018.

[15] H D Nguyen. Asymptotic normality of the time-domain generalized least squares estimator for linear regression models. *Stat*, to appear, 2019.

[16] H D Nguyen and F Chamroukhi. An introduction to the practical and theoretical aspects of mixture-of-experts modeling. *WIREs Data Mining and Knowledge Discovery*, e1246, 2018.

[17] H D Nguyen, F Chamroukhi, and F Forbes. Approximation results regarding the multiple-output mixture of linear experts model. *Neurocomputing*, to appear, 2019.

[18] H D Nguyen, F Forbes, and G J McLachlan. Mini-batch learning of exponential family finite mixture models. 2019.

[19] H D Nguyen, M M Hill, and I A Wood. A robust permulation test for quantitative SILAC proteomics experiments. *Journal of Integrated OMICS*, 2(80-93), 2012.

[20] H D Nguyen and A T Jones. Big data-appropriate clustering via stochastic approximation and Gaussian mixture models. In *Data Analytics: Concepts, Techniques, and Applications*. CRC Press, 2018.

[21] H D Nguyen, A T Jones, and G J McLachlan. Stream-suitable optimization algorithms for some soft-margin support vector machine variants. *Japanese Journal of Statistics and Data Science*, 1:81–108, 2018.

[22] H D Nguyen, L R Lloyd-Jones, and G J McLachlan. A block minorization-maximization algorithm for heteroscedastic regression. *IEEE Signal Processing Letters*, 23:1031–1135, 2016.

[23] H D Nguyen, L R Lloyd-Jones, and G J McLachlan. A universal approximation theorem for mixture-of-experts models. *Neural Computation*, 28:2585–2593, 2016.

[24] H D Nguyen and G J McLachlan. Maximum likelihood estimation of Gaussian mixture models without matrix operations. *Advances in Data Analysis and Classification*, 9:371–394, 2015.

[25] H D Nguyen and G J McLachlan. Laplace mixture of linear experts. *Computational Statistics and Data Analysis*, 93:177–191, 2016.

[26] H D Nguyen and G J McLachlan. Linear mixed models with marginally symmetric nonparametric random-effects. *Computational Statistics and Data Analysis*, 106:151–169, 2016.

[27] H D Nguyen and G J McLachlan. Maximum likelihood estimation of triangular and polygonal distributions. *Computational Statistics and Data Analysis*, 106:23–36, 2016.

[28] H D Nguyen and G J McLachlan. Iteratively-reweighted least-squares fitting of support vector machines: a majorization-minimization algorithm approach. In *Proceedings of the 2017 Future Technologies Conference (FTC)*, 2017.

[29] H D Nguyen and G J McLachlan. Progress on a conjecture regarding the triangular distribution. *Communications in Statistics - Theory and Methods*, 46:11261–11271, 2017.

[30] H D Nguyen and G J McLachlan. Chunked-and-averaged estimators for vector parameters. *Statistics and Probability Letters*, 137:336–342, 2018.

[31] H D Nguyen and G J McLachlan. Some theoretical results regarding the polygonal distribution. *Communications in Statistics - Theory and Methods*, 47:5083–5095, 2018.

[32] H D Nguyen and G J McLachlan. On approximation via convolution-defined mixture models. *Communications in Statistics - Theory and Methods*, 48:3945–3955, 2019.

[33] H D Nguyen, G J McLachlan, N Cherbuin, and A L Janke. False discovery rate control in magnetic resonance imaging studies via Markov random fields. *IEEE Transactions on Medical Imaging*, 33:1735–1748, 2014.

[34] H D Nguyen, G J McLachlan, and M M Hill. Permutation tests with false discovery corrections for comparative-profiling proteomics experiments. In *Methods in Molecular Biology: Proteomics Bioinformatics*. Springer, 2017.

[35] H D Nguyen, G J McLachlan, P Orban, P Bellec, and A L Janke. Maximum pseudolikelihood estimation for a model-based clustering of time series data. *Neural Computation*, 29:990–1020, 2017.

[36] H D Nguyen, G J McLachlan, J F P Ullmann, and A L Janke. Laplace mixture autoregressive models. *Statistics and Probability Letters*, 110:18–24, 2016.

[37] H D Nguyen, G J McLachlan, J F P Ullmann, and A L Janke. Spatial clustering of time-series via mixture of autoregressions models and Markov Random Fields. *Statistica Neerlandica*, 70:414–439, 2016.

[38] H D Nguyen, G J McLachlan, J F P Ullmann, V Voleti, W Li, E M C Hillman, D C Reutens, and A L Janke. Whole-volume clustering of time series data from zebrafish brain calcium images via mxiture modeling. *Statistical Analysis and Data Mining*, 11:5–16, 2018.

[39] H D Nguyen, G J McLachlan, and I A Wood. Mixtures of spatial spline regressions for clustering and classification. *Computational Statistics and Data Analysis*, 93:76–85, 2016.

[40] H D Nguyen, D H Wang, and G J McLachlan. Randomized mixture models for probability density approximation and estimation. *Information Sciences*, 467:135–148, 2018.

[41] H D Nguyen and I A Wood. A block successive lower-bound maximization algorithm for the maximum pseudolikelihood estimation of fully visible Boltzmann machines. *Neural Computation*, 28:485–492, 2016.

[42] H D Nguyen and I A Wood. Asymptotic normality of the maximum pseudolikelihood estimator for fully visible Boltzmann machines. *IEEE Transactions on Neural Networks and Learning Systems*, 27:897–902, 2016.

[43] H D Nguyen, Y Yee, G J McLachlan, and J P Lerch. False discovery rate control for grouped or discretely supported p-values with application to a neuroimaging study. *SORT*, to appear, 2019.

[44] T T Nguyen, H D Nguyen, F Chamroukhi, and G J McLachlan. Approximation by finite mixtures of continuous density functions that vanish at infinity. https://arxiv.org/abs/1903.00147, 2019.

[45] P Orban, C Dansereau, L Desbois, V Mongeau-Perusse, C-E Giguere, H Nguyen, A Mendrek, E Stip, and P Bellec. Multisite generalizability of schizophrenia diagnosis classification based on functional brain connectivity. *Schizophrenia Research*, 192:167–171, 2018.

[46] C Oyarzun, A Sanjurjo, and H Nguyen. Response functions. *European Economic Review*, 98:1–31, 2017.

[47] L Truong, H Nguyen, H Nguyen, and H Vu. Pedestrian overpass use and its relationship with digital and social distractions, and overpass characteristics. *Accident Analysis and Prevention*, 131:234–238, 2019.