A Festschrift for Geoff McLachlan

Hien Duy Nguyen, Sharon Lee, Florence Forbes

Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **64**(2), 2022, 111–116 doi: 10.1111/anzs.12372

A Festschrift for Geoff McLachlan

Hien Nguyen[1,*], Sharon Lee[1] and Florence Forbes[2]

University of Queensland and Inria Grenoble Rhône–Alpes

Summary

This article introduces a special issue of the Australian and New Zealand Journal of Statistics, dedicated as a Festschrift for Geoff McLachlan on the occasion of his 75th birthday.

1. Background of a lengthy career in statistics

Geoff was born in Rockhampton on 3 October 1946, and was educated at the Rockhampton Grammar School, matriculating as dux of the School in 1964. He subsequently undertook studies in a BSc (Hons) degree at the Department of Mathematics, University of Queensland (UQ), where he graduated with First Class Honours. With a Commonwealth Postgraduate Scholarship, he then proceeded to undertake PhD studies in Statistics within the same department, in the field of discriminant analysis, under the supervision of Professor Stephen Lipton, who had come to Australia from the famed Rothamsted Experimental Station, in England.

Geoff wrote his PhD thesis on the topic of estimating error rates of discriminant functions, which was a topical area of research at the time. He recalls that Professor Tony Lachenbruch had written a PhD thesis in 1965 at the University of North Carolina on a similar topic (see Lachenbach & Mickey 1968), while Hills (1966), based on the author's PhD thesis under the guidance of Professor Peter Armitage from the London School of Hygiene and Tropical Medicine, was read as a discussion paper at the Royal Statistical Society. Geoff also recalls that during his PhD candidature, he occupied a room next to Mildred Prentice (née Barnard), who had returned to teaching after a long absence from academia while raising her family. She had written her PhD thesis some 40 years earlier, at the Rothamsted Experimental Station, under the guidance of none other than Sir Ronald Fisher. Geoff recalls warm memories of Mildred's guidance.

In 1972, Geoff completed his thesis, and subsequently he graduated in 1973. In the same year, he was married to Beryl Seymour, with whom he has two sons and four granddaughters. Geoff held a tutorship at UQ until the beginning of 1974, when he moved to the University of New England to take up a lectureship. He subsequently returned to UQ at the end of the same year to take up an unexpected opening at the Department of Mathematics. Geoff was promoted to a Senior Lectureship in 1978, a Readership in 1984, and a Personal Chair in 1998. Prior to the latter appointment, he was also awarded a Doctor of Science by UQ in 1994.

*Author to whom correspondence should be addressed.
[1]School of Mathematics and Physics, University of Queensland, St. Lucia, Australia
[2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Inria Grenoble Rhone-Alpes, 655 av. de l'Europe, Grenoble, 38335 Montbonnot, France. e-mail: h.nguyen7@uq.edu.au

Geoff was awarded an Australian Research Council (ARC) Professorial Fellowship for 2007–2011, and over the period of 2003–2011, he held a part-time appointment at the Institute of Molecular Biosciences at UQ, returning to a full-time position at the School of Mathematics and Physics as a UQ Vice-Chancellor Senior Research Fellow in 2013–2015. From 2016 onwards, Geoff has returned into his present Teaching and Research position at the department.

His professional service has included chairing of the Mathematics group from 2008 to 2010, after its amalgamation into the School of Mathematics and Physics, as well as numerous committees both within and outside of UQ, with the notable inclusion of his service as an ARC College of Experts member for 2008–2010. Most pertinent to the Statistical Society of Australia (SSA) is Geoff's service as the inaugural Secretary and second President of the Queensland Branch of the SSA. Over the period 1992–1997 Geoff was also an Associate Editor of this publication (then, the Australian Journal of Statistics), when it was still under the sole proprietorship of the SSA.

Over the years, Geoff has mentored many students of Statistics at UQ, being associated with the projects of numerous Honours and PhD students, with perhaps the most famous being Professor Peter Donnelly (Oxford University), who was knighted in 2019, for his services to Statistics and Genomics. Geoff has provided supervision for 15 PhD theses, with the most recent two being two of the authors of this paper, Drs. Sharon Lee and Hien Nguyen, who are lucky to have both recently returned to UQ, where they now continue their work with Geoff.

Facilitating his research, Geoff has travelled extensively overseas, to Asia, Europe and the USA, where he has spent time visiting the facilities of his many collaborators and where has has delivered many of his invited conference talks, seminars, and workshops. He recalls fondly his first sabbatical in 1980, visiting the Department of Statistics at Stanford for six months, during which time he did some summer teaching. Among his other memorable trips were his 6 months visit to the Issac Newton Institute at the University of Cambridge in 2008, as part of the Future Directions in High-Dimensional Data Analysis program, and more recently, sabbaticals to the Department of Statistics at the University of Bologna.

2. An overview of Geoff McLachlan's research

As noted by Geoff, at the time of his candidature, he had not imagined nor anticipated the sustained interest in classification and, more broadly, in machine learning over the subsequent years, whereupon classification techniques in applied disciplines have become so commonplace as to be prosaic. Geoff's now over five decades of research has contributed not only to the area of classification, but also to the related domains of cluster analysis, machine learning and various fields of statistical inference. In particular, Geoff has dedicated much of his research efforts towards the theory and application of finite mixture models and towards their maximum likelihood estimation via the expectation–maximisation (EM) algorithm. His methodological developments have also lead to applications in many areas, including bioinformatics, biostatistics, flow cytometry, image analysis and survival analysis.

To date, Geoff has authored or coauthored over 300 peer-reviewed articles, including 270 papers in international journals or conference proceedings, and 34 book chapters. Additionally, he has written six monographs, including five as volumes of the prestigious Wiley Series in Statistics and Probability. These monographs have been tremendously influential and extensively cited both within and outside of the statistics literature:

- McLachlan (1992) is widely recognised as a standard reference on discriminant analysis and supervised clustering, and was reprinted in 2004 as part of the Wiley Classics Library.
- His joint manuscripts: McLachlan & Basford (1988) and McLachlan & Peel (2000a) have also become standard references on the topic of finite mixture models.
- Written together with Professor Krishnan at the Indian Institute of Statistics, McLachlan & Krishnan (2008) (First Edition, published in 1997) has been an ever-popular resource for practitioners, not only in statistics but across the numerical sciences.
- Motivated by his interest in the many applications of classification in bioinformatics, Geoff wrote McLachlan, Do & Ambroise (2004) together with former UQ Honours student Kim-Anh Do, who is now Professor at the MD Anderson Cancer Centre, University of Texas, and Professor Christophe Ambroise, now at the University of Evry.

The impact of Geoff's research is most easily illustrated by his more than 50,000 citations on Google Scholar, as well as the many awards and plaudits that he has received. Most notably, Geoff was awarded the Pitman Medal by the SSA in recognition of his 'outstanding achievement in, and contribution to, the discipline of Statistics' in 2010, his IEEE ICDM Research Contribution Award for 'influential contributions to the field of data mining' in 2015, his election to a Fellowship of the Australian Academy of Science in 2015, and his International Federation of Classifications Societies Research Medal for Outstanding Research Achievements in 2017. Geoff notes that other awardees of the medal include Professors Brad Efron and David Hand.

Specifically, Geoff's research can be summarised into four themes, as follows. His pioneering works on the application and estimation of mixture models has been especially influential. Of particular note is his work on **(i)** mixture modelling for inference and clustering via normal and Student t-mixtures (McLachlan & Peel 2000a; Peel & McLachlan 2000; Quinn, McLachlan & Hjort 1987), extensions of such models to account for latent factors (Baek, Mclachlan & Flack 2010; Lee, Lin & McLachlan 2018; Viroli & McLachlan 2019), and non-normal variants, including those accounting for non-elliptical and long-tailed clusters (Lee & McLachlan 2013, 2022); **(ii)** applications and extensions of the EM algorithm, especially for the estimation of complex multivariate distributions related to clustering problems (Lee & McLachlan 2018; Lee, Leemaqz & McLachlan 2018); **(iii)** accurate error rate estimation for classifiers and diagnostic tests (McLachlan 1976, 1986); and **(iv)** the analysis of high-dimensional gene-expression and medical data sets (McLachlan, Bean & Peel 2002).

Geoff notes particular fondness and satisfaction for the works: McLachlan (1974a), McLachlan (1974b) and McLachlan (1974c), where he provides a theoretical accounts of the relative performances of error rate estimators for classifiers; McLachlan (1975), where he introduces an iterative reclassification scheme that can be seen as a hard-threshold version of the subsequent EM algorithm applied to classification; Ambroise & McLachlan (2002), where he and Professor Ambroise communicated the dangers of overfitting in high-dimensional classification, under selection bias, in genetic experiments; McLachlan, Bean & Ben-Tovim Jones (2006), Nguyen *et al.* (2014) and Ng *et al.* (2015), where he and collaborators explored the application of finite mixture for false discovery rate control and multiple testing corrections in large scale testing problems; Pyne *et al.* (2009), where he, author Lee, and collaborator Saumyadipta Pyne (then at the Broad Institute, MIT and Harvard University) made substantial

progress towards the clustering of blood cell markers in flow cytometry data; and most recently Ahfock & McLachlan (2020), where he and current Post-doctoral Research Fellow Daniel Ahfock explore the efficiency gains from modelling missingness mechanisms in semi-supervised learning problems.

3. Contributions to the Festschrift

The papers appearing in this Festschrift present a microcosm of the topics that Geoff has spent his career researching. We order the works according to the four themes noted in the previous section.

The leading paper of Tomarchio, Ingrassia & Melnykov (2022) follows naturally from Geoff's interest in the study of mixture of normal distributions for vectorial data (**Theme (i)**), and extends such an approach to the problem domain of matrix data clustering. The works of Hui & Nghiem (2022) and Scrucca (2022) can also be viewed as contributions to this theme, with particular focus on the use of latent space representations via dimensionality reduction in order to facility better mixture-based clustering outcomes. Other contributions to **Theme (i)** include the works of Durand *et al.* (2022), Greve *et al.* (2022), and Hennig & Coretto (2022), who each provide differing perspectives and solutions to the problem of clustering and mixture model estimation when the underlying number of clusters is unknown. Here, Durand *et al.* (2022) and Greve *et al.* (2022) provide Bayesian solutions for spatial regression data and vectorial data, respectively, whereas Hennig & Coretto (2022) consider an approach based on optimally tuned robust improper maximum likelihood estimation.

Nguyen & Forbes (2022) then provide a contribution towards **Theme (ii)**, where the authors consider the verification of assumptions that are necessary for the application of online versions of the EM algorithm for the maximum likelihood estimation of broad classes of statistical models and their mixtures. **Theme (iii)** is represented by the work of Stewart (2022), who considers the error rate-related problem of hypothesis test detection boundary characterisation, when data may arise from finite mixture families under the alternative hypothesis.

Lastly, the Festschrift finishes with the pair of works of Arief *et al.* (2022) and Zhang, Swallow & Gupta (2022), who both contribute towards **Theme (iv)** via their statistical analyses of gene-expression data. In the former work, the authors study genomic prediction of plant genotype data, and in the latter work, the authors use a mixture modelling approach to study subpopulations in human genome-wide association data.

We thank all of the contributors for their articles, as well as the reviewers and editorial staff whose hard work enabled the production of this Festschrift. Finally, we wish Geoff McLachlan a very happy 75th Birthday, and hope that he enjoys this Issue and is able to use it as the seed for years of fruitful work to come.

References

Ahfock, D. & McLachlan, G.J. (2020). An apparent paradox: A classifier trained from a partially classified sample may have smaller expected error rate than that if the sample were completely classified. *Statistics and Computing* **30**, 1779–1790.

Ambroise, C. & McLachlan, G.J. (2002). Selection bias in gene extraction on the basis of microarray gene expression data. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6562–6566.

ARIEF, V.N., DELACY, I.H., PAYNE, T. & BASFORD, K.E. (2022). Visualising the pattern of long-term genotype performance by leveraging a genomic prediction model. *Australian and New Zealand Journal of Statistics* **64**, 297–312.

BAEK, J., MCLACHLAN, G.J. & FLACK, L. (2010). Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualisation of high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**, 1298–1309.

DURAND, J.B., FORBES, F., PHAN, C.D., TRUONG, L., NGUYEN, H.D. & DAMA, F. (2022). Bayesian nonparametric spatial prior for traffic crash risk mapping: A case study of Victoria, Australia. *Australian and New Zealand Journal of Statistics* **64**, 171–204.

GREVE, J., GRÜN, B., MALSINER-WALLI, G. & FRÜWIRTH-SCHNATTER, S. (2022). Spying on the prior of the number of data clusters and the partition distribution in Bayesian cluster analysis. *Australian and New Zealand Journal of Statistics* **64**, 205–229.

HENNIG, C. & CORETTO, P. (2022). An adequacy approach for deciding the number of clusters for OTRIMLE robust Gaussian mixture-based clustering. *Australian and New Zealand Journal of Statistics* **64**, 230–254.

HILLS, M. (1966). Allocation rules and their error rates. *Journal of the Royal Statistical Society B* **28**, 1–20.

HUI, F.K.C. & NGHIEM, L.H. (2022). Sufficient dimension reduction for clustered data via finite mixture modelling. *Australian and New Zealand Journal of Statistics* **64**, 133–157.

LACHENBACH, P.A. & MICKEY, M.A. (1968). Estimation of error rates in several-population discriminant analysis. *Technometrics* **10**, 1–10.

LEE, S.X. & MCLACHLAN, G.J. (2013). On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification* **7**, 241–266.

LEE, S.X. & MCLACHLAN, G.J. (2018). Emmixcskew: An r package for the fitting of a mixture of canonical fundamental skew t-distributions. *Journal of Statistical Software* **83**. Article 3.

LEE, S.X. & MCLACHLAN, G.J. (2022). An overview of skew distributions in model-based clustering. *Journal of Multivariate* **88**. Article 104853.

LEE, S.X., LIN, T.I. & MCLACHLAN, G.J. (2018). Mixtures of factor analyzers with scale mixtures of fundamental skew symmetric distributions. *Advances in Data Analysis and Classification* **15**, 481–512.

LEE, S.X., LEEMAQZ, K.L. & MCLACHLAN, G.J. (2018). A block EM algorithm for multivariate skew normal and skew t-mixture models. *IEEE Transactions on Neural Networks and Learning Systems* **29**, 5581–5591.

MCLACHLAN, G.J. (1974a). The asymptotic distributions of the conditional error rate and risk in discriminant analysis. *Biometrika* **61**, 131–135.

MCLACHLAN, G.J. (1974b). The relationship in terms of asymptotic mean square error between the separate problems of estimating each of the three types of error rate of the linear discriminant function. *Technometrics* **16**, 569–575.

MCLACHLAN, G.J. (1974c). An asymptotic unbiased technique for estimating the error rates in discriminant analysis. *Biometrics* **30**, 239–249.

MCLACHLAN, G.J. (1975). Iterative reclassification procedure for constructing and asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association* **70**, 365–369.

MCLACHLAN, G.J. (1976). The bias of the apparent error rate in discriminant analysis. *Biometrika* **63**, 239–244.

MCLACHLAN, G. (1986). Assessing the performance of an allocation rule. *Computers & Mathematics with Applications* **12A**, 261–272.

MCLACHLAN, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.

MCLACHLAN, G.J. & BASFORD, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

MCLACHLAN, G.J. & KRISHNAN, T. (2008). *The EM Algorithm and Extensions*, 2nd edn. Hoboken, NJ: Wiley.

MCLACHLAN, G.J. & PEEL, D. (2000a). *Finite Mixture Models*. New York: Wiley.

MCLACHLAN, G.J., BEAN, R. W. & PEEL, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18**, 413–422.

MCLACHLAN, G.J., DO, K.A. & AMBROISE, C. (2004). *Analyzing Microarray Gene Expression Data*. Hoboken, NJ: Wiley.

MCLACHLAN, G.J., BEAN, R.W. & BEN-TOVIM JONES, L. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* **22**, 1608–1615.

NG, S.K., MCLACHLAN, G.J., WANG, K., NAGYMANYOKI, Z., LIU, S. & NG, S.W. (2015). Inference on differential expression using cluster-specific contrasts of mixed effects. *Biostatistics* **16**, 9–112.

NGUYEN, H.D. & FORBES, F. (2022). Global implicit function theorems and the online expectation—maximisation algorithm. *Australian and New Zealand Journal of Statistics* **64**, 255–281.

NGUYEN, H.D., MCLACHLAN, G.J., CHERBUIN, N. & JANKE, A.L. (2014). False discovery rate control in magnetic resonance imaging studies via Markov random fields. *IEEE Transactions on Medical Imaging* **33**, 173–1748.

PEEL, D. & MCLACHLAN, G.J. (2000). Robust mixture modelling using the *t* distribution. *Statistics and Computing* **10**, 339–348.

PYNE, S., HU, X., WANG, K., ROSSIN, E., LIN, T.I., MAIER, L.M., BAECHER-ALLAN, C., MCLACHLAN, G.J., TAMAYO, P., HAFLER, D.A., DE JAGER, P.L. & MESIROV, J.P. (2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 8519–8524.

QUINN, B.G., MCLACHLAN, G.J. & HJORT, N.L. (1987). A note on the Aitkin-Rubin approach to hypothesis testing in mixture models. *Journal of the Royal Statistical Society B* **49**, 311–314.

SCRUCCA, L. (2022). Model clustering on PPGMMGA projection subspace. *Australian and New Zealand Journal of Statistics* **64**, 158–170.

STEWART, M.I. (2022). Detection boundary for a sparse gamma scale mixture model. *Australian and New Zealand Journal of Statistics* **64**, 282–296.

TOMARCHIO, S.D., INGRASSIA, S. & MELNYKOV, V. (2022). Modelling students' career indicators via mixtures of parsimonious matrix-normal distributions. *Australian and New Zealand Journal of Statistics* **64**, 117–132.

VIROLI, C. & MCLACHLAN, G.J. (2019). Deep Gaussian mixture models. *Statistics and Computing* **29**, 43–51.

ZHANG, H., SWALLOW, B. & GUPTA, M. (2022). Bayesian hierarchical mixture models for detecting non normal clusters applied to noisy genomic and environmental datasets. *Australian and New Zealand Journal of Statistics* **64**, 313–337.