For NSERC office use only

Form 101 - Application for a Grant

Send to NSERC with your attachments, if applicable

Reference Number: 125543286

Applicant: Nancy Reid
Toronto

NSERC PIN: 12406

Program: Discovery Grants - Individual

Application Title: Likelihood inference for complex data

Nancy Reid

Form 101 - Application for a Grant

Electronic Attachments:

- Budget Justification - Budget Justification
- Research Support - Relationship to other research support
- Proposal - Proposal
- References - References
- Research Contributions 1 - Mean log-likelihood, to appear in Biometrika
- Research Contributions 2 - Default priors, revision submitted to JRSS B
- Research Contributions 3 - Chapter 8 of Applied Asymptotics
- Research Contributions 4 - Paper with Yun Yi to appear in Statistica Sinica

Nancy Reid

F100/Personal Data Form

Electronic Attachments:

- Contributions - Contributions

Natural Sciences and Engineering Research Council of Canada | Conseil de recherches en sciences naturelles et en génie du Canada

1508

Institutional Identifier

System-ID (for NSERC use only): 125543286

FORM 101
Application for a Grant
PART I

Date: 2009/10/26

Family name of applicant	Given name	Initial(s) of all given names	Personal identification no. (PIN)
Reid	Nancy	NM	**Valid** 12406

Institution that will administer the grant: Toronto

Language of application: [X] English [ ] French

Time (in hours per month) to be devoted to the proposed research / activity: 80

Type of grant applied for: Discovery Grants - Individual

For Strategic Projects, indicate the Target Area and the Research Topic; for Strategic Networks and Strategic Workshops indicate the Target Area.

Title of proposal: Likelihood inference for complex data

Provide a maximum of 10 key words that describe this proposal. Use commas to separate them.

asymptotic theory, composite likelihood, conditional inference, correlated data, estimating equations, longitudinal data, multi-level models, multivariate analysis, observational studies, sample surveys

Research subject code(s)		Area of application code(s)	
Primary	Secondary	Primary	Secondary
3009	3001	1211	1212

CERTIFICATION/REQUIREMENTS

If this proposal involves any of the following, check the box(es) and submit the protocol to the university or college's certification committee.

Research involving : Humans [ ] Human pluripotent stem cells [ ] Animals [ ] Biohazards [ ]

Does any phase of the research described in this proposal a) take place outside an office or laboratory, or b) involve an undertaking as described in Part 1 of Appendix B?

[X] NO [ ] If YES to either question a) or b) – Appendices A and B must be completed

TOTAL AMOUNT REQUESTED FROM NSERC

Year 1	Year 2	Year 3	Year 4	Year 5
84,600	84,600	84,600	84,600	84,600

SIGNATURES (Refer to instructions "What do signatures mean?")

It is agreed that the general conditions governing grants as outlined in the NSERC *Program Guide for Professors* apply to any grant made pursuant to this application and are hereby accepted by the applicant and the applicant's employing institution.

Applicant

Applicant's department, institution, tel. and fax nos., and e-mail

Statistics
Toronto
Tel.: (416) 978 5046
FAX: (416) 978 5133
reid@utstat.utoronto.ca

Head of department

Dean of faculty

President of institution (or representative)

The information collected on this form and appendices will be stored in the Personal Information Bank for the appropriate program.

PROTECTED WHEN COMPLETED

Version française disponible

Canada

Personal identification no. (PIN)	Family name of applicant
Valid 12406	Reid

SUMMARY OF PROPOSAL FOR PUBLIC RELEASE (Use plain language.)

This plain language summary will be available to the public if your proposal is funded. Although it is not mandatory, you may choose to include your business telephone number and/or your e-mail address to facilitate contact with the public and the media about your research.

Business telephone no. (optional): (416) 9785046

E-mail address (optional): reid@utstat.utoronto.ca

Modern technology has enabled the collection of large and complex sets of data, and these are being used to answer important research questions in many fields of science and social science. Statistical methods are used to understand both the structure and the noise in this data, and new methods are being rapidly developed by statisticians working in collaboration with biologists, physicists, epidemiologists, social scientists and many others. These new methods typically involve quite complex modelling, and are computationally intensive in their implementation. In many cases these new methods are defined algorithmically in each context. The theory of statistical inference provides us with a set of guidelines for tackling new problems, provides a framework for assessing approaches developed in particular contexts, and searches for the common structure in what may seem to be very diverse problems. Statistical theory has been very successfu in finding basic structure and suggesting new solutions, and as the field develops more and more sophisticated approaches to data, there is an accompanying need for understanding the basis for the analysis.

The research proposed here is to develop the theory of inference to very complex settings by investigating statistical methods in detail, both mathematically and in practical application. The mathematical analysis provides insight into the process of inference from data, suggesting what information is available and how this information may be extracted. The statistical interpretation of the mathematics can suggest new ways for scientists to analyse their data. This research program emphasizes the development and dissemination of both new theory and new statistical methods.

Other Language Version of Summary (optional).

Personal identification no. (PIN)	Family name of applicant
Valid 12406	Reid

Before completing this section, **read the instructions** and consult the *Use of Grant Funds* section of the NSERC Program Guide for Professors concerning the eligibility of expenditures for the direct costs of research and the regulations governing the use of grant funds.

TOTAL PROPOSED EXPENDITURES (Include cash expenditures only)

	Year 1	Year 2	Year 3	Year 4	Year 5
1) Salaries and benefits					
a) Students	35,100	35,100	35,100	35,100	35,100
b) Postdoctoral fellows	44,000	44,000	44,000	44,000	44,000
c) Technical/professional assistants	0	0	0	0	0
d)	0	0	0	0	0
2) Equipment or facility					
a) Purchase or rental	2,000	2,000	2,000	2,000	2,000
b) Operation and maintenance costs	1,000	1,000	1,000	1,000	1,000
c) User fees	4,000	4,000	4,000	4,000	4,000
3) Materials and supplies	2,500	2,500	2,500	2,500	2,500
4) Travel					
a) Conferences	9,000	9,000	9,000	9,000	9,000
b) Field work	0	0	0	0	0
c) Collaboration/consultation	6,000	6,000	6,000	6,000	6,000
5) Dissemination costs					
a) Publication costs	1,000	1,000	1,000	1,000	1,000
b)	0	0	0	0	0
6) Other (specify)					
a)	0	0	0	0	0
b)	0	0	0	0	0
TOTAL PROPOSED EXPENDITURES	104,600	104,600	104,600	104,600	104,600
Total cash contribution from industry (if applicable)					
Total cash contribution from university (if applicable)					
Total cash contribution from other sources (if applicable)	20,000	20,000	20,000	20,000	20,000
TOTAL AMOUNT REQUESTED FROM NSERC (transfer to page 1)	84,600	84,600	84,600	84,600	84,600

Explanation of expenditures

1. Salaries and benefits

   (a) Students

   The requested amount is for partial support of 2 PhD students under my primary supervision, two PhD students under co-supervision, one MSc student and two undergraduate research students. I have regularly supervised between 2 and 4 PhD students during the past five years and expect this rate to continue. The funding for MSc students is used for research projects for students proceeding to the PhD program, during their summer preceding entry into the program. The summer undergraduate research students are funded through the USRA award program, which is increased with funds from my Discovery grant. The amount budgeted for PhD students is $21,000 per year plus 10% benefits, In 2009-2010 these students are Wei Lin and Ximing Xu. Lin is in her first year of the PhD program, and is completing required course work and exams. Xu started his research with me in June, 2009. I have allocated $7,000 plus 10% benefits for co-supervision of two PhD students per year. Support for an MSc student is budgeted at $5,000 per year plus 10% benefits, and USRA top-up money at $1,500 $\times$ 2. The total amount budgeted for student support in support of this proposal is $70,100, but I am requesting $35,100 from NSERC. The remaining funds are comprised of external scholarship support to some of the students, research funds from my Canada Research Chair, and opportunities that arise through the Accelerate Canada internship program. CRC funds are shown under "Other sources of support", but the student scholarship and internship moneys are not.

   (b) Postdoctoral Fellow: $44,000 ($40,000 salary and $4,000 benefits). From 2007–2009 my research group consisted of two PDFs, two to four PhD students, occasional faculty visitors, one or more MSc students, and two or three advanced undergraduates. We met regularly to discuss research in likelihood inference, often jointly with colleagues in the Department of Statistics. This model worked extremely well because of the range of expertise available, but also very importantly because the PDFs played a leadership role in organizing the group and presenting material. In recent years PDFs have become increasingly important in statistical science, and I receive several requests for supervision from potential PDFs each year. All the students who graduated under my supervision recently took up PDF positions upon graduation, and this is becoming the norm in our discipline. The PDF gains a valuable opportunity to concentrate on research for two years, and the graduate students learn a great deal from interacting with the PDFs, not only in research, but in how to work in teams and communicate results effectively. As statistical science becomes increasingly interdisciplinary, this type of training is very important. Going forward I plan to have one or two PDFs each year under my supervision, supported through a combination of funds from my Discovery Grant and other sources of funding, including co-supervision with colleagues, and funds from the University Professor research fund at the University of Toronto.

2. Equipment or facility

   (a) $2,000 on average year for computer hardware: laptop, printer, and printer cartridges.

(b) $1,000 for software and software upgrades

(c) Research of students and postdoctoral fellows, and problems requiring substantial computing resources are developed on the departmental computing system, a network of Suns. User fees for this system contribute to the infrastructure needs, including the salary support of the system manager. The fees have averaged $4,000 per year for the past three years.

3. Materials and supplies

   phone $1,000; photocopy $1,000; miscellaneous materials $500

4. Travel

   (a) Conferences

   - One overseas conference for myself and one for a PhD student or PDF each second year. In 2010 this will be the Annual Meeting of the IMS in Gothenburg, Sweden. $3,000 per year.
   - Two North American conferences annually: $2,000 for myself and $2,000 for a PDF. In 2010 these will be the SSC Annual Meeting in Quebec City and the Joint Statistical Meetings in Vancouver, BC.
   - four graduate students to the SSC Annual Meeting $2,000.

   (b) Collaboration

   - Cristiano Varin to visit Toronto for joint work on composite likelihood in 2010: $2,000
   - Rahul Mukerjeee to visit Toronto for joint work on likelihood asymptotics in 2010: $4,000

5. Dissemination costs

   (a) Publication costs

   - page charges for self and students: $1,000

Other sources

There is a $10,000 pa research grant attached to my University Professorship and a $10,000 pa research grant attached to my Canada Research Chair, which appear in the budget as cash contributions from other sources.

Relationship to other support

I am a co-applicant on the MRS proposal for the Banff International Research Station. No funding from this proposal flows to the co-applicants: all funding is directed to programs taking place at BIRS, and these programs are funded through an international peer-reviewed competition.

I am a co-investigator on our Department's equipment grant, which is used for computer hardware in support of graduate student research.

I am a member of the project "Statistical methods for complex survey data", funded by the National Center of Excellence on "Mathematics and Information Technology for Complex Systems" (MITACS). This project was co-funded by the National Program on Complex Data Structures from 2003 to 2008. The funds from this project were used for partial funding for PhD student Zheng Zheng, who completed a four-month internship at Statistics Canada, and later for research assistant Lequn Zeng, who completed his M.Sc. in August, 2009. For September to December 2009 Zeng is working for me as a research assistant on a project related to the current proposal, and is funded by my Discovery Grant.

There is a small ($10,000) grant associated with the University Professor position, and another ($10,000) with the Canada Research Chair. These funds are used in support of the research described in this proposal, and are included in the budget pages under "Cash contribution from other sources".

Proposal

Recent progress and current research *References to papers on Form 100 are given as [n].*

My main accomplishments during the current grant period have been the publication of a book on higher order approximations with Davison and Brazzale, new work with Fraser, Staicu, and Sun on aspects of the theory of higher order asymptotics, with particular emphasis on the study of the overlap between Bayesian and nonBayesian approaches, and new work in the area of composite likelihood, starting with the publication of a paper on composite likelihood with Cox. This work has been published in *Biometrika*, *Statistica Sinica*, *JRSS* B, and other more specialized journals. Three students have completed their PhD theses (Staicu, Iglesias-Gonzalez, Jin), two students are currently engaged in PhD research under my direction, and I supervised J.-F. Plante, an NSERC postdoctoral fellow (PDF) and co-supervised PDF Y. Sun, jointly with D. Fraser, for two years (2007-2009). I published two review papers on higher order asymptotics; the Wald lectures [13], and a paper in a *Festschrift* for D.R. Cox [18].

The book *Applied Asymptotics* [27] presents a concise but detailed account of higher order asymptotic theory for likelihood inference in Chs. 2 and 8, with an emphasis in the other chapters on the application of these methods to models and problems that arise in practical settings. The goal was to provide illustrations on common classes of problems, along with computer code, to make higher order methods accessible to applied statisticians. A second goal was to illustrate both the practicality of higher order methods and their extreme accuracy on a wide variety of problems. The book has been favorably reviewed in *Short Book Reviews*, *J. Appl. Statist.* and *JRSS A*.

The asymptotic theory that my colleagues and I use is based on saddlepoint and Edgeworth expansions, interpreted in a likelihood setting. While the detailed derivations are somewhat technical, the essential point is that an approximation to the p-value function for a scalar parameter of interest, ψ, is completely determined by a pair of functions $\{\ell(\theta;y^0), \varphi(\theta;y^0)\}$, and their derivatives with respect to the parameter θ. Here $\ell(\theta;y)$ is the log-likelihood function for θ based on a response $y = (y_1, \ldots, y_n)$, y^0 is the observed value of the response, and $\varphi(\theta;y^0)$ is a re-parameterization of the model, at the observed data point, that gives an exponential model approximation to the original model, in a neighbourhood of y^0. The re-parameterization $\varphi(\theta;y^0)$ is in turn developed from a location-model approximation to the original model, and incorporates conditioning on an approximately ancillary statistic. The approximation is referred to briefly as the r^* approximation, as an inference quantity r^* can be computed from the pair $\{\ell(\theta;y^0), \varphi(\theta;y^0)\}$. The distribution of r^* is approximately standard normal, with relative error $O(n^{-3/2})$, and the formula for r^* is $r^* = r + (1/r)\log(q/r)$, where $r = [2\{\ell(\hat\theta;y^0) - \ell(\hat\theta_\psi;y^0)\}]^{1/2}$, q is a function of $\varphi(\theta)$ and $\ell(\theta)$ and their derivatives with respect to θ, $\hat\theta$ and $\hat\theta_\psi = (\psi, \hat\lambda_\psi)$ are the unconstrained and constrained maximum likelihood estimates of $\theta = (\psi, \lambda)$, ψ is a scalar parameter of interest and λ is a vector nuisance parameter.

We are continuing to develop and extend this higher order approximation method to more general settings and to study the connections between Bayesian and frequentist inference. In [7] we showed how this approach can be applied to discrete models, where it gives $O(n^{-1})$ accuracy. We are adapting the directional test of Fraser & Massam (1985) to construct approximations for inference for vector parameters of interest: first steps were developed in [8] and further work refining this is described below. PhD student Staicu undertook a detailed study of approximations for binomial data, obtaining analytical expressions for expansions to $O(n^{-3/2})$ that illuminate the connections between various approximation methods, and explain the accuracy of the method based on r^* ([22]). In [6], based on Ch. 4 of [28], we showed that the class of probability matching priors developed

by Tibshirani (1989) is essentially unique, when used in higher order approximations to Bayesian marginal posterior distributions. In Fraser & Reid (2002) we developed a theory of strong matching by defining a data-dependent prior for which the posterior probability is equal to the likelihood based p-value, to a high order of approximation. We have carried this theory much further in [15], where we study the structure of strong matching priors in detail, and propose default priors for scalar and for vector parameters. In [5], which was invited for a *Festschrift* volume for Professor Akahira, we show how higher order approximations can be used to easily assess the sensitivity of posteriors to priors, and also show that using flat priors for α and β in a logistic regression with $\text{logit}(p_i) = \alpha + \beta x_i$ regression leads to poorly calibrated inference for the ED_{50} parameter α/β. This is to be expected in light of theoretical results in Fraser & Reid (2002), but does not seem to be widely appreciated in applications of Bayesian inference.

In [3] Fraser and I showed that an $O(n^{-1})$ approximation suggested by Skovgaard (1996, 2001) can be obtained from a pair $\{\ell(\theta; y^0), \tilde{\varphi}(\theta)\}$ in exactly the same way as the r^* approximation is obtained from $\{\ell(\theta; y^0), \varphi(\theta; y^0)\}$, where the calculation of $\tilde{\varphi}$ involves simply the derivative of the information function $I(\theta_0; \theta) = E_{\theta_0}\{\ell(\theta; y)\}$: $\tilde{\varphi}(\theta) = \partial I(\hat{\theta}; \theta)/\partial\hat{\theta}$. Since I averages over the distribution of y, $\tilde{\varphi}$ does not depend on an approximately ancillary statistic. This makes it simpler to calculate in many problems: for example in normal theory linear models with fixed and random effects (Lyons & Peters, 2000). Iglesias-Gonzalez ([29]) developed expressions for the more accurate version of r^* for this model, and showed that the third order version was just slightly more accurate, although more complex to implement.

With [11] I began work in a new area of research, that of composite likelihood, and this research is continuing in joint work with PhD students Jin and Xu, with PDF Plante, and with colleagues Yun-Yi, Varin and Firth. A composite likelihood is a product of marginal or conditional likelihoods, sometimes suitably weighted: in its most general form $CL(\theta; y) = \prod_{s\in\mathcal{S}} f_s(y_s; \theta)^{w_s}$, where each component f_s is either a marginal or conditional density for the components of y that fall in the subset s. In [11] we studied the pairwise likelihood $\prod_{r<r'} f(y_{r'}, y_r; \theta)$, based on the joint marginal density for each pair of components of y. Composite likelihood is a generalization of Besag's (1974) pseudo-likelihood, and was defined and studied in Lindsay (1988). More recently it has found application in a wide range of applied problems, including both discrete and continuous longitudinal data, frailty models in survival data, generalized linear mixed models, spatial models, state-space models and statistical genetics. The maximum composite likelihood estimator is consistent, although not efficient (Lindsay, 1988), as the sample size $n \to \infty$ with the dimension p fixed. In a great many settings it has been verified, usually through simulations, that the efficiency loss is relatively small, and the computational benefits are substantial. An overview of these results is given in Varin (2008). However, as we showed in [11], if $p \to \infty$ with n fixed, as is applicable typically to time series and genetics applications, composite likelihood will not in general give consistent estimates. A workshop held in April 2008 on composite likelihood summarized the current state of understanding and raised a number of open problems, and my research objectives include several problems related to composite likelihood, discussed in the next section.

Short and long term objectives

In the short term my first research objective is to understand the properties of composite likelihood inference. While efficiency of composite likelihood estimators has been well established in a number of applications, it is not at all clear why these estimators are so efficient. They are related to estimators obtained by generalized estimating equations (GEE), and in simple models can be shown to be either more or less efficient than GEE, depending on the context. Likelihood ratio type tests can be formed from composite likelihood, but their distribution is difficult to obtain in practical

settings. However, likelihood ratio-based inference in full likelihood methods has better properties in general than inference based on the maximum likelihood estimator and its estimated standard error, and this may well be the case for composite likelihood inference. A particularly important question is whether or not composite likelihood is more robust to model mis-specification than likelihood methods. It seems natural that it would be, since higher order dependencies are not modelled in, for example, pairwise likelihood, but it has proved difficult to formulate this more precisely. My other short term objectives relate to to further development of our work on higher order approximations: inference for vector parameters, the interface between Bayesian and likelihood approaches, the connection between default priors and the potential failure of flat priors to be well-calibrated, and semi-parametric models. The version of r^* from mean likelihood discussed in [10] opens up the possibility of developing higher order approximations for a wide range of pseudo-likelihoods, including partial likelihood and composite likelihood.

My long term objectives are to develop likelihood-based methods for inference in models with very complex structure, and to make likelihood-based inference as broadly accessible as possible. My exposure to models used in environmetrics, survey sampling and genomics, in large part through work as a scientific reviewer, has given me a sense of the practical issues faced when using complex models, as well as the need for some unifying perspectives on approaches to inference. In many respects the underlying problems in a wide range of applications are very similar, although each application has some unique details. Recurring themes include the use of multi-level models; the use of partially parametric models; and the potential for more accurate results by combining information from several sources and of several types, for example observational studies, geographic information systems, and experimental data. I believe it is important to abstract the unifying ideas behind these applications and develop them further, in part to provide a way of understanding the complexity of the models, and in part to provide a basis for approaching new applications. Making new theoretical results accessible to practitioners as quickly and as clearly as possible is also a long-term goal that informs all my research.

Methods and proposed approach

A range of problems associated with composite likelihood will be investigated. Professor Yun-Yi and I are studying the statistical properties of estimators obtained from biased estimating equations, using a new method of adjustment. We were motivated by the problem of longitudinal binary data with missing data and/or measurement error, where the estimating equations based on the observed complete data are biased, although much simpler to use. The first report on this work [2] has been accepted for publication, and our next goal is to extend this work to models with unknown nuisance parameters. The asymptotic theory of estimating equations can be related to that for composite likelihood, as the latter leads to unbiased, but not fully efficient, estimating functions, and we are also planning to study how the theory we developed for biased estimating equations relates to the robustness of estimators based on composite likelihood. PhD student Zi Jin investigated in her thesis the efficiency of composite likelihood estimators in models for discrete and continuous data. In some settings composite likelihood estimators are fully efficient, and we are working on an asymptotic theory to try to explain this. Mardia et al. (2009) have defined a class of closed exponential families, to try to explain the full efficiency of composite likelihood in special models. We plan to investigate extensions to this work, as their conditions for closure, and an additional condition needed called weak interaction, seem too strong to explain the high efficiency of composite likelihood inference in a wide range of applications. We will first study efficiency in binary data created by dichotomizing continuous data, and develop an extension of an argument outlined in Davison (2003, Ex. 10.17) for independent scalar responses.

An advantage of composite likelihood over GEE, is the existence of an objective function, which in principle has further information beyond that of the GEE, for example enabling choice between multiple roots. However likelihood ratio-type statistics based on composite likelihood have a complicated asymptotic distribution, a weighted sum of χ^2, with weights depending on the eigenvalues of a matrix related to the variance of the composite likelihood estimator. It should be possible to combine this with saddlepoint approximations for quadratic forms, discussed for example in Kuonen (1999), but this needs to be assessed numerically in a number of models: this will be a summer undergraduate project for 2010. If these results are promising the next step will be to develop asymptotic expansions under model mis-specification, building on work of Viraswami & Reid (1996). This work will be continued with a future PhD student. A quite different approach will be pursued by a research associate (L. Zeng) who just completed an M.Sc. in statistics. He is investigating higher order asymptotic theory for composite likelihood based on the results of [3], where the function $\varphi(\theta)$ is obtained using expected, rather than observed, log-likelihood, and can thus be computed from various types of pseudo-likelihood. Whether or not this can lead to higher order approximation is an open question. Zeng carrying out numerical work on the two proto-type examples studied in [11], where comparison with full likelihood models is analytically possible.

PhD student Xu began his research in June 2009 by seeking an appropriate formulation for studying the robustness of composite likelihood inference. His first challenge is to describe a family of models for binary data that have the same low dimensional marginal distributions, but different higher-dimensional margins, and to assess the performance of composite likelihood inference in this class of models. Tackling the robustness question of composite likelihood in general is quite challenging; it has proved difficult to come up with the right formulation of the problem. We are for now focussing attention on pairwise likelihood, where the robustness in question is against making strong assumptions about joint distributions of third and higher order. Xu is also studying the relationship of this to the development of joint distributions based on copulas. Another approach would be to study the results on optimal weighting of marginal distributions, discussed for example in Zhao & Joe (2005) and Kuk & Nott (2000) in the context of pairwise likelihood for clustered familial data, assuming that there is likely to be a trade-off between robustness and efficiency. Plante and I are just beginning work on adaptations of composite likelihood to meta-analysis, and as part of that he is investigating the connections between weighted likelihood, the topic of his thesis, and composite likelihood.

I have agreed to be a guest editor for a volume of *Statistica Sinica* on composite likelihood. Varin, Firth and I are just completing (October 09) a review paper giving an overview of research presented at the Warwick workshop of April (2008) and developments since then, updating and extending the review paper by Varin (2008) to include more emphasis on applications to Gaussian random fields and more general models in spatial analysis.

Methods to be used for new results in higher order approximation are a combination of analytical results based on asymptotic expansions and numerical work. In joint work with Sartori and Davison, Fraser and I are developing a version of directional tests based on Fraser's (2003) higher order approximation to likelihood. I also plan to use our new results on expected log-likelihood [3] in higher order inference for vector parameters of interest. Numerical work on the comparison of this approach with a vector-parameter version suggested in Skovgaard (2001) will be the topic of a summer undergraduate project for 2010. There is a great deal of scope for numerical work examining the agreement between Bayesian and frequentist inference in complex models through simulations and Laplace approximations. For likelihood-based and Bayesian approaches to agree beyond the first order of approximation, it is necessary for the prior to be data-dependent. The approach of

empirical Bayes inference also involves a data-dependent prior, although to my knowledge there is no current research that tries to make this connection. The natural place to start to investigate this is in the context of hierarchical models. Some very preliminary work is reported in [5], and the simplification of higher order approximations presented in [3] gives a way to extend this to much more complex hierarchical models. It would be very useful to have some understanding of when flat priors can be used without markedly affecting the posterior inference. Some progress on this is made in [15], but considering this in the context of applications, of the type treated for example in Gelman et al. (2007), would make this work more practically relevant. This will require substantial computing with large data sets, and would be suitable for an MSc project or for the first part of Lin's PhD research.

In [27] our goal was to make the use of higher order asymptotics easy for practitioners. My goal now is to provide a complete, but concise, account of the theory. This book will emphasize the methods Fraser and I have developed over the past several years, and their connection with the work outlined in the books by Barndorff-Nielsen & Cox (1994), Severini (2000) and Butler (2007). The aim is to provide an accessible systematic reference for students and researchers, to summarize the current state of research in this area, and to draw parallels between Bayesian and frequentist methods.

Anticipated significance of the work

Composite likelihood inference is becoming widely used, but there is as yet little work that provides a general understanding of its relation to the more widely used methods. There is also very little work on the asymptotic theory of composite likelihood when the dimension, p, increases, with fixed sample size, n, beyond that in [6]. The large p small n problem is important for the application of composite likelihood to problems in genetics and in geostatistics. Composite likelihood also has the potential to deepen our understanding of multivariate distributions. For example, it is relatively easy to introduce various types of correlation in discrete data by introducing latent random effects. It is much harder to see the structure of the resulting multivariate binary distribution, or even if indeed it is a real distribution. There is empirical evidence (e.g. Liang & Yu, 2003) that the composite likelihood surface can be not only computationally cheaper, but much smoother than the true likelihood surface. An understanding of these results would increase the utility of CL methods and potentially provide warnings about classes of problems that should not be treated in this way.

Improved approximations based on higher order likelihood theory are useful for applications, and are increasingly easier to implement. But a more important benefit of the detailed study of approximations is to deepen our understanding of the nature of model-based inference. For example, the fact that an $O(n^{-1})$ approximation to an arbitrary continuous density gives an $O(n^{-3/2})$ approximation to the p-value relies on a surprisingly simple property of the normal distribution (Andrews et al., 2005). Asymptotics also sheds light on when and why flat priors are highly informative for marginal inference about non-linear parameters, and shows explicitly that frequentist and Bayesian methods diverge at $O(n^{-1})$, unless data-dependent priors are used. Advances in the theory of statistical inference, which benefits both theoretical and applied statistical science.

Training to take place through the proposal

Training to take place is described throughout the proposal in the context of specific research problems. Students in my research group range from undergraduate to PDF, and are trained in computational, applied and theoretical areas of statistics. They have opportunities to contribute to all aspects of research, from background work through to publication, and regular group meetings help to develop their skills in collaboration and communication.

References

Andrews, D.F., Fraser, D.A.S. & Wong, A. (2005). Computation of distribution functions from likelihood information near observed data. *J. Statist. Plann. Inf.* **134**, 180–193.

Barndorff-Nielsen, O.E. & Cox, D.R. (1994). *Inference and Asymptotics.* Chapman & Hall, London.

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *J. R. Statist. Soc.* B, **34**, 192–236.

Butler, R. (2007). *Saddlepoint Approximations.* Cambridge University Press, Cambridge.

Crainiceanu, C., Ruppert D. & Claeskens, G. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika* **92**, 91–103.

Davison, A.C. (2003). *Statistical Models.* Cambridge University Press, Cambridge.

Fraser, D.A.S. & Massam, H. (1985). Conical tests: Observed levels of significance and confidence regions. *Statistische Hefte* **26**, 1-17.

Fraser, D.A.S. & Reid, N. (2002). Strong matching of frequentist and Bayesian parametric inference. *J. Statist. Plann. Inf.* **103**, 263–285.

Fraser, D.A.S. (2003). Likelihood for component parameters. *Biometrika* **90**, 327–339.

Gelman, A., Fagan, J. & Kiss, A. (2007). An analysis of the NYPD's stop-and-frisk policy in the context of claims of racial bias. *J. Amer. Statist. Assoc.* **102**, 813–823.

Kuonen, D. (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**, 929–935.

Kuk, A.Y.C. & Nott, D.J. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Stat. Prob. Lett.* **47**, 329–35.

Liang, G. & Yu, B. (2003). Maximum pseudo-likelihood estimation in network tomography. *IEEE Trans. Sig. Proc.* **51**, 2043–2053.

Lindsay, B.L. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes*, Ed. N.U. Prabhu, pp. 221–239. Providence: American Mathematical Society.

Lyons, B. & Peters, D. (2000). Applying Skovgaard's modified directed likelihood statistics to mixed linear models. *J. Statist. Comp. Sim.* **65**, 225–242.

Mardia, K.V., Kent, J.T., Hughes, G. and Taylor, C.C. (2009). Maximum likelihood estimation using composite likelihood for closed exponential families. *Biometrika*, to appear.

Severini, T.J. (2000). *Likelihood Methods in Statistics.* Oxford University Press, Oxford.

Skovgaard, I.M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2**, 145165.

Skovgaard, I.M. (2001). Likelihood asymptotics. *Scand. J. Statist.* **28**, 3–32.

Tibshirani, R.J. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**, 604–608.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis* **92**, 1-28.

Viraswami, K. & Reid, N. (1996). Higher order asymptotics under model misspecification. *Canad. J. Statist.* **24**, 263–278.

Zhao, Y. & Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canad. J. Statist.* **33**, 335–356.