Composite Likelihood Methods

Harry Joe (University of British Columbia),
Nancy Reid (University of Toronto),
Peter Xuekun Song (University of Michigan),
David Firth (University of Warwick),
Cristiano Varin (University Ca' Foscari)

April 22-27, 2012

1 Overview of the Field

Composite likelihood methods are extensions of the Fisherian likelihood theory, one of the most influential approaches in statistics. Such extensions are generally motivated by the issue of computational feasibility arising in the application of the likelihood method in high-dimensional data analysis. Complex dependence presents substantial challenges in statistical modelling and methods and in substantive applications. The idea of projecting high-dimensional complicated likelihood functions to low-dimensional computationally feasible likelihood objects is methodologically appealing. Composite likelihood inherits many of the good properties of inference based on the full likelihood function, but is more easily implemented with high-dimensional data sets. This methodology is, to some extent, an alternative to the Markov Chain Monte Carlo method, and its impact is unbounded.

The literature on both theoretical and practical issues for inference based on composite likelihood continues to expand quickly; the field of extremal processes for spatial data, of particular importance for climate modelling, is one of the most recent examples of an area where composite likelihood inference is both practical and efficient.

The first international workshop on composite likelihood methods was held at the University of Warwick in April 2008. It attracted participants from all over the world and was widely viewed as very successful. Following the workshop, a special issue of the journal *Statistica Sinica* devoted to composite likelihood was announced; it was published in January 2011. This issue includes two long overview papers, one of which is devoted to applications in statistical genetics; several papers developing new theory for inference based on composite likelihood; new results in the application of composite likelihood to time series, spatial processes, longitudinal data and missing data. The methodology has drawn considerable attention in a broad range of applied disciplines in which complex data structures arise. Some notable application areas include, statistical genetics, genetic epidemiology, finance, panel surveys, computer experiments, geostatistics and biostatistics.

2 Presentation Highlights

In the opening presentation, Varin described complex likelihoods where the ordinary likelihood function is difficult to evaluate or to specify. However, in many of these situations it is however possible to compute marginal or conditional densities for subsets of the data.

Terminology that he set for the workshop, and which are used for this report, are the following.

- pseudo-likelihood: any function of parameter and data that behaves in “some respect” as a likelihood;
- composite likelihood: one of many examples of pseudo-likelihoods based on terms that are logarithms of marginal and conditional densities;
- quasi-likelihood: different meanings, the two most common are (a) Wedderburn’s quasi-likelihood and variants (statisticians), (b) quasi-likelihoods as misspecified likelihoods (econometricians);
- limited information methods: used in psychometrics as inference procedures based on low-dimensional margins.

Let $\boldsymbol{\theta}$ be a parameter vector of a parametric model for an observation $\mathbf{y}$, a realization of random m-vector $\mathbf{Y}$. For independent and identically distributed replications $\mathbf{y}_1, \mathbf{y}_2, \ldots$, let n be the sample size. Then

- the composite log-density based on K different margins or conditional distributions has the form
$$cl(\boldsymbol{\theta}, \mathbf{y}) = \sum_{k=1}^{K} w_k l_{A_k}(\boldsymbol{\theta}, \mathbf{y}), \quad l_{A_k}(\boldsymbol{\theta}, \mathbf{y}) = \log f_{A_k}(y_j, j \in A_k; \boldsymbol{\theta}) \text{ for margin } A_k$$
- for a sample of size n, the maximum composite log-likelihood estimator $\hat{\boldsymbol{\theta}}$ maximizes $\sum_{i=1}^{n} cl(\boldsymbol{\theta}, \mathbf{y}_i)$.
- the composite score function is the partial derivative of the composite log-density with respect to the parameter vector:
$$\mathbf{u}(\boldsymbol{\theta}, \mathbf{y}) = \sum_{k=1}^{K} w_k \nabla_{\boldsymbol{\theta}} l_{A_k}(\boldsymbol{\theta}, \mathbf{y})$$
- sensitivity or Hessian matrix: $\mathbf{H}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\{-\nabla_{\boldsymbol{\theta}} \mathbf{u}(\boldsymbol{\theta}, \mathbf{Y})\}$
- variability matrix: $\mathbf{J}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}\{\mathbf{u}(\boldsymbol{\theta}, \mathbf{Y})\}$
- Godambe information: $\mathbf{G}(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta})\mathbf{J}^{-1}(\boldsymbol{\theta})\mathbf{H}(\boldsymbol{\theta})$. As $n \to \infty$, $\mathbf{G}^{-1}(\boldsymbol{\theta})$ is the asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ under some regularity conditions. In the case of observation of a single time series or random field, the asymptotics depend on ergodicity conditions and the above n is replaced by the observation length m.
- The composite likelihood version of AIC, as given in Varin and Vidoni (2005), is
$$\mathrm{CL} - \mathrm{AIC} = -2 \sum_{i} cl(\hat{\boldsymbol{\theta}}, \mathbf{y}_i) + 2\mathrm{tr}(\mathbf{J}(\hat{\boldsymbol{\theta}})\mathbf{H}^{-1}(\hat{\boldsymbol{\theta}}))$$
- The composite likelihood version of BIC, as given in Gao and Song (2010), is
$$\mathrm{CL} - \mathrm{BIC} = -2 \sum_{i} cl(\hat{\boldsymbol{\theta}}, \mathbf{y}_i) + (\log n)\mathrm{tr}(\mathbf{J}(\hat{\boldsymbol{\theta}})\mathbf{H}^{-1}(\hat{\boldsymbol{\theta}}))$$

Some challenges covered by the workshop and summarized in Varin’s talk included the following.

- Design issues: how do we select the set of marginal or conditional sets A_k, and how should they be combined through choice of weights w_k; this was discussed in presentations of Lindsay and others.
- Uncertainty estimation: $\hat{\boldsymbol{\theta}}$ is straightforward to obtain by inputting the negative of $\sum_i cl(\boldsymbol{\theta}, \mathbf{y}_i)$ into a numerical minimizer, and this can yield an estimate $\mathbf{H}$ at $\hat{\boldsymbol{\theta}}$, but estimation of the variability matrix $\mathbf{J}$ needed for $\mathbf{G}^{-1}$ and standard errors for components are $\hat{\boldsymbol{\theta}}$ can be computationally challenging.
- Calibration: this comes up in several contexts, including calibration of test statistics with nuisance parameters, as discussed in Salvan’s talk.

- Robustness to model misspecifications: examples of different robustness ideas were included in talks of Jordan and Xu.
- Prediction: how does composite likelihood for prediction compare with composite likelihood for estimation.
- Software development.

The most common form of composite marginal or conditional likelihood that has been used in applications is pairwise likelihood. Several presentations considered other sets of margins or conditional distributions; the talks on spatial extremes suggested that there can be gains in efficiencies with triple-wise or trivariate composite likelihood combined with increased (but feasible) computational time.

Lindsay compared various designs (choices of marginal or conditional distributions) such as pairwise, conditional pairs, and hybrids (combination of one-wise and pairwise likelihoods).

There are connections with the use of two-stage (and multi-stage) estimating equations in the copula modelling literature; the method is called inference functions for margins (IFM) in Joe (1997) — first univariate parameters are estimated from univariate likelihoods and then dependence parameters are estimated from higher-dimensional likelihoods.

In Vidoni's presentation, for predictive densities, another consideration is how to weight different components of cl; the best choices for estimation and prediction might not be the same. In Molenberghs' presentation, missing data were handled with inverse probability weighting. In family-based studies of genetic markers, Briollais and Choi had adjustments for ascertainment, for censoring, and for missing data.

Several notations of calibration came up in the workshop. In Salvan's presentation, this meant finding a constant c for the weights w_k in the composite likelihood so that for inference

$$2\sum_i [cl(\hat{\boldsymbol{\theta}}, \mathbf{y}_i) - cl(\boldsymbol{\theta}_0, \mathbf{y}_i)]$$

can be adjusted suitable to recover an approximate χ^2 distribution when the true parameter is $\boldsymbol{\theta}_0$. In Ribatet's presentation: a similar adjustment was mentioned for quasi-Bayes, with composite likelihood replacing likelihood, and the quasi-posterior or composite posterior distribution. Ng's presentation mentioned that if the weights are all multiplied by a common constant with $w_k \to cw_k$ for all k, then in the CL-AIC, $\mathrm{tr}(\mathbf{JH}^{-1}) \to c\mathrm{tr}(\mathbf{JH}^{-1})$ which implies that the penalty term $\mathrm{tr}(\mathbf{JH}^{-1}) = \mathrm{tr}(\mathbf{HG}^{-1})$ should not be interpreted as the effective number of parameters.

For estimation of $\mathbf{J}(\boldsymbol{\theta}) = \mathrm{E}[\mathbf{u}(\boldsymbol{\theta}, \mathbf{Y})\,\mathbf{u}^\top(\boldsymbol{\theta}, \mathbf{Y})]$, is some form of direct estimation better or is it better to estimate the inverse Godambe matrix $\mathbf{G}^{-1}$ of the maximum composite likelihood estimator via an appropriate jackknife or (parametric) bootstrap? Lindsay mentioned that approximately orthogonal pieces can make calculation of $\mathbf{J}$ or $\mathbf{G}$ simpler.

Over the workshop, in addition to those topics mentioned above, there was discussion of the theory of composite likelihood for incomplete data (Molenberghs), survey data and analysis of composite score equations from a design viewpoint (Yi) and model comparisons (Ng). A wide range of applications were covered within the special themed sessions on spatial statistics, multivariate extremes, psychometrics, genetics/genomics as well as other sessions. Applications included spatial-temporal data in air pollution and health (Bai); spatial-temporal data in fMRI (Kang); spatial extremes (Genton, Ribatet, Padoan); extreme rainfall events (Huser, Davison); ecology (Lele); Gaussian graphical models (Gao); random graph models for networks (Bellio); linkage disequilibrium, recombination rates, penetrance in genetics (Larribe, Choi, Briollais); genetic networks (Song); psychometrics and latent variable models (Moustaki, Vasdekis, Maydeu-Olivares); panel multinomial probit models for transportation choices (Bhat); multivariate times series of traffic accidents (Karlis).

There was also a session with demonstrations of software, followed by discussion. Moustaki gave a demonstration of some software for latent variable modeling used in psychometrics; these include implementations of limited information methods. Padoan gave a demonstration of an R package `CompRndFld`, Composite Likelihood for Random Fields, co-authored with Moreno Bevilacqua. Ribatet gave a demonstration of his R package `SpatialExtremes`, Modelling Spatial Extremes; in this package, max-stable processes are fit to spatial data using composite likelihood.

3 Recent Developments and Open Problems

For the study of efficiencies of various designs, it is helpful to have some models where composite likelihood with higher-dimensional margins are computationally feasible. There are models, such as random effects and structural equation models, for which some theoretical comparisons might be feasible.

There is still an open question about the sense in which composite likelihood estimation is more robust? For robustness against misspecification of the full distribution, this seems difficult to make precise. For studying the robustness of estimating equations based on low-dimensional log-likelihoods, it might be easier to make further study of IFM for copula models because one can more readily come up with different models where some set of margins are fixed and others can vary.

As mentioned in one talk, the case of $\mathbf{H}(\boldsymbol{\theta}) = \mathbf{J}(\boldsymbol{\theta})$ is called *information-unbiased* (a term from Lindsay 1982, Biometrika). Care must be taken in understanding this definition. More clearly, this is written as $\mathbf{H}(\boldsymbol{\theta}) \equiv \mathbf{J}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$ in a parameter space.

In the discussion on the Friday morning of the workshop, it was mentioned that for pairwise likelihood, $\mathbf{J}(\boldsymbol{\theta}) - \mathbf{H}(\boldsymbol{\theta})$ tends to be positive definite when the parameter $\boldsymbol{\theta}$ represents positive dependence; see the Appendix of Ribatet et al (2012) for the context of a Markov process. The intuitive explanation is that $\mathbf{J}(\boldsymbol{\theta}) - \mathbf{H}(\boldsymbol{\theta})$ involves covariances among different terms in the pairwise log-likelihood and $\mathbf{H}(\boldsymbol{\theta})$ involves the sum of the variance terms of the pairwise log-likelihood. Subsequent to the conference, it was checked that for the case of the parameter space with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ representing independence in the components of $\mathbf{Y}$, then $\mathbf{H}(\boldsymbol{\theta}_0) = \mathbf{J}(\boldsymbol{\theta}_0)$ under some assumptions, and furthermore, sometimes $\mathbf{G}^{-1}(\boldsymbol{\theta}_0) = \mathbf{H}^{-1}(\boldsymbol{\theta}_0)\mathbf{J}(\boldsymbol{\theta}_0)\mathbf{H}^{-1}(\boldsymbol{\theta}_0)$ matches inverse Fisher information and sometimes it doesn't.

An example was given by Xu (see also in Cox and Reid 2004; Joe and Lee 2009) for the 3-parameter exchangeable multivariate normal distribution where $\mathbf{H}(\boldsymbol{\theta}) \neq \mathbf{J}(\boldsymbol{\theta})$ but the maximum pairwise likelihood estimator is the same as the maximum likelihood estimator. More generally for structured models based on the multivariate normal distribution, the maximum pairwise likelihood estimator could be different from the maximum likelihood estimator depending on (a) the structural forms of the mean vector and covariance matrix, and (b) whether some parameters are assumed fixed or known. The relation of the maximum pairwise likelihood estimator and maximum likelihood estimator for these types of structured models was mentioned as a research problem in Maydeu's presentation.

The above example means that one cannot say that the composite likelihood estimate is not fully asymptotic efficient if $\mathbf{H}(\boldsymbol{\theta}) \neq \mathbf{J}(\boldsymbol{\theta})$.

Some other theoretical issues are the following.

- Does it matter if there is not a multivariate distribution compatible with, for example, bivariate margins? This depends on the inferences (e.g., joint tail probabilities) to be obtained from the model.
- How do we ensure identifiability of parameters?
- Can connections to weighted likelihoods provide additional insight?
- Is the composite likelihood ratio test preferable to Wald-type test?
- When is composite marginal likelihood preferred to composite conditional likelihood?
- For large p, small or moderate n asymptotics: is there consistency?

4 Outcome of the Meeting

With more opportunities for discussion in this second composite likelihood conference, there is a clearer picture of some of the challenges for composite likelihood. Some of these have been mentioned above.

Outside of the presentations, there were many opportunities for researchers to discuss their work in composite likelihood and give feedback to each other. Subgroups of the participants had other overlapping interests so there was also much discussion of other topics during meals etc. This workshop will lead to many future collaborative research efforts among the participants.

From the discussion of computing software for general use for composite likelihood, the conclusion seemed to be that creating software to cover the many existing applications of composite likelihood is premature until we have a clearer theoretical understanding of the construction of composite likelihood. However, a web page could be created to (a) collect the software packages related to composite likelihood, (b) suggest a common format for development of further R packages with composite likelihood estimation, providing a standardized user interface to enable easier application of composite likelihood methods.

Finally, invited sessions in other mainstream conferences of international statistical societies will continue the dissemination of research results on composite likelihood methods. An example is a session on composite likelihood at the World Congress in Probability and Statistics in July 2012.

References

[1] Cox, D.R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91, 729–737.

[2] Ribatet, M., Cooley, D. and Davison, A.C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 22, 813–845.

[3] Gao, X. and Song, P. X.-K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *J. Amer. Statist. Assoc*, 105, 1531–1540.

[4] Joe, H., (1997). *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

[5] Joe, H. and Lee, Y. (2009). On weighting of bivariate margins in pairwise likelihood. *J. Multivariate Analysis* 100, 670–685.

[6] Lindsay, B.G. (1982). Conditional score functions: some optimality results. *Biometrika*, 69, 503–512.

[7] Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* 92, 519–528.