Likelihood

Present position and future prospects

Nancy Reid

June 1, 2007
... Waterloo, 1970
A more philosophical approach

- Neyman: Foundations of Behavioristic Statistics
- Bartlett: When is Inference Statistical Inference?
- Fraser: Events, Information Processing, and the Structured Model
- Dempster: Model Searching and Estimation in the Logic of Inference
- Good, Hajèk, Barnard, Sprott, Rao, Basu, Godambe, Lindley, ...
- “philosophical” aspects of likelihood
- Likelihood principle, conditionality principle, sufficiency principle
- coherence of Bayesian inference
“... ubiquitous in applied work”
Estimating Genotypic Correlations and Their Standard Errors Using Multivariate Restricted Maximum Likelihood Estimation with SAS Proc MIXED

James B. Holland*

USDA-ARS Plant Science Research Unit, Dep. of Crop Science, Box 7620, North Carolina State University, Raleigh, NC 27605

*Corresponding author (James_Holland@ncsu.edu)

Plant breeders traditionally have estimated genotypic and phenotypic correlations between traits using the method of moments on the basis of a multivariate analysis of variance (MANOVA). Drawbacks of using the method of moments to estimate variance and covariance components include the possibility of obtaining estimates outside the admissible range and bias due to the use of small sample sizes. Restricted maximum likelihood (REML) is a widely used approach that provides unbiased estimators of variance and covariance components. However, the analysis of genotypic correlation matrices using this method is hindered by the covariance structures of the models. SAS Proc MIXED is a flexible procedure that can be used to fit a wide range of covariance structures. This paper illustrates how to use Proc MIXED to estimate genotypic correlations using REML.
Maximum Likelihood Estimation of Latent Affine Processes

David S. Bates
University of Iowa

This article develops a direct filtration-based maximum likelihood methodology for estimating the parameters and realizations of latent affine processes. Filtration is conducted in the transform space of characteristic functions, using a version of Bayes’ rule for recursively updating the joint characteristic function of latent variables and the data conditional upon past data. An application to daily stock market returns over 1953–1996 reveals substantial divergences from estimates based on the Efficient Methods of Moments (EMM) methodology; in particular, more substantial and time-varying jump risk. The implications for pricing stock index options are examined.
Abstract—Space–time block codes (STBCs) from orthogonal designs (ODs) and coordinate interleaved orthogonal designs (CIODs) have been attracting wider attention due to their amenability for fast (single-symbol) maximum-likelihood (ML) decoding, and full-rate with full-rank over quasi-static fading channels. However, these codes are instances of single-symbol decodable codes and it is natural to ask, if there exist codes other than STBCs form ODs and CIODs that allow single-symbol decoding? In this paper, the above question is answered in the affirmative by characterizing all linear STBCs, that allow single-symbol ML decoding (not necessarily full-diversity) over quasi-static fading channels calling them single-symbol decodable designs (SDD). The class SDD includes ODs and CIODs as proper subclasses. Further, among the SDD, a class of those that offer full-diversity, called Full-rank SDD (FSDD) are characterized and classified. We then concentrate on square designs and derive the maximal rate for square FSDDs using a constructional proof. It follows that 1) except for \( N = 2 \), square complex ODs are not maximal rate and 2) a rate one square FSDD exist only for two and four transmit antennas. For nonsquare designs, generalized coordinate-interleaved orthogonal designs (a superset of CIODs) are presented and analyzed. Finally, for rapid-fading channels an equivalent matrix channel representation is developed, which allows the results of quasi-static fading channels to be applied to rapid-fading channels. Using this representation we show that for rapid-fading channels the rate of single-symbol decodable STBCs are independent of the number of transmit antennas and inversely proportional to the block-length of the code. Significantly, the CIOD for two transmit antennas is the only STBC that is single-symbol decodable over difference between coded modulation [used for single-input single-output (SISO), single-input multiple-output (SIMO)] and space–time codes is that in coded modulation the coding is in time only while in space–time codes the coding is in both space and time and hence the name. STC can be thought of as a signal design problem at the transmitter to realize the capacity benefits of MIMO systems [1], [2], though, several developments toward STC were presented in [3]–[7] which combine transmit and receive diversity, much prior to the results on capacity. Formally, a thorough treatment of STCs was first presented in [8] in the form of trellis codes [space–time trellis codes (STTC)] along with appropriate design and performance criteria.

The decoding complexity of STTC is exponential in bandwidth efficiency and required diversity order. Starting from Alamouti [12], several authors have studied space–time block codes (STBCs) obtained from orthogonal designs (ODs) and their variations that offer fast decoding (single-symbol decoding or double-symbol decoding) over quasi-static fading channels [9]–[27]. But the STBCs from ODs are a class of codes that are amenable to single-symbol decoding. Due to the importance of single-symbol decodable codes, need was felt for rigorous characterization of single-symbol decodable linear STBCs.

Following the spirit of [11] by linear STBCs are characterized...
Cognitive Behavioral Therapy for Posttraumatic Stress Disorder in Women
A Randomized Controlled Trial

Paula P. Schnurr, PhD
Matthew J. Friedman, MD, PhD
Charles C. Engel, MD, MPH
Edna B. Foa, PhD
M. Tracie Shea, PhD
Bruce K. Chow, MS
Patricia A. Resick, PhD
Veronica Thurston, MBA
Susan M. Orsillo, PhD
Rodney Haug, PhD
Carole Turner, MN
Nancy Bernardy, PhD

Context The prevalence of posttraumatic stress disorder (PTSD) is elevated among women who have served in the military, but no prior study has evaluated treatment for PTSD in this population. Prior research suggests that cognitive behavioral therapy is a particularly effective treatment for PTSD.

Objective To compare prolonged exposure, a type of cognitive behavioral therapy, with present-centered therapy, a supportive intervention, for the treatment of PTSD.

Design, Setting, and Participants A randomized controlled trial of female veterans (n=277) and active-duty personnel (n=7) with PTSD recruited from 9 VA medical centers, 2 VA readjustment counseling centers, and 1 military hospital from August 2002 through October 2005.

Intervention Participants were randomly assigned to receive prolonged exposure (n=141) or present-centered therapy (n=143), delivered according to standard protocols in 10 weekly 90-minute sessions.

Main Outcome Measures Posttraumatic stress disorder symptom severity was the primary outcome. Comorbid symptoms, functioning, and quality of life were secondary outcomes. Blinded assessors collected data before and after treatment and at 3- and 6-month follow-up.
Accuracy of Coalescent Likelihood Estimates: Do We Need More Sites, More Sequences, or More Loci?

Joseph Felsenstein
Department of Genome Sciences and Department of Biology, University of Washington, Seattle

A computer simulation study has been made of the accuracy of estimates of $\Theta = 4N_e \mu$ from a sample from a single isolated population of finite size. The accuracies turn out to be well predicted by a formula developed by Fu and Li, who used optimistic assumptions. Their formulas are restated in terms of accuracy, defined here as the reciprocal of the squared coefficient of variation. This should be proportional to sample size when the entities sampled provide independent information. Using these formulas for accuracy, the sampling strategy for estimation of $\Theta$ can be investigated. Two models for cost have been used, a cost-per-base model and a cost-per-read model. The former would lead us to prefer to have a very large number of loci, each one base long. The latter, which is more realistic, causes us to prefer to have one read per locus and an optimum sample size which declines as costs of sampling organisms increase. For realistic values, the optimum sample size is 8 or fewer individuals. This is quite close to the results obtained by Pluzhnikov and Donnelly for a cost-per-base model, evaluating other estimators of $\Theta$. It can be understood by considering that the resources spent collecting larger samples prevent us from considering more loci. An examination of the efficiency of Watterson’s estimator of $\Theta$ was also made, and it was found to be reasonably efficient when the number of mutants per generation in the sequence in the whole population is less than 2.5.

Introduction

The availability of molecular sequencing at prices that even population biologists can afford has brought into existence new methods of estimation of population parameters. Sequence samples from populations enable one to make an estimate of the coalescent tree of genes connecting these sequences. I have argued (Felsenstein 1992a) that these enable a substantial increase in the accuracy of estimation of population parameters like $\Theta = 4N_e \mu$, the product of effective population size, and the neutral mutation rate per site. (This is usually expressed as $\theta$, the neutral mutation rate per locus but is perhaps better thought of in terms of the neutral mutation rate per site.)

Fu and Li (1993) analyzed my claim further. They developed some approximations to the accuracy of maximum likelihood estimation of $\Theta$. I will show below that these are remarkably good approximations, better than one might have expected and as assumed here. These allow one to explicitly check the effect of the number of loci, finding results which strongly back collecting more loci rather than more sites or more sequences, can be argued to be intuitively reasonable.

Fu (1994) developed a method which makes a UPGMA estimate of the coalescent tree and constructs a best linear unbiased estimate conditional on that being the correct tree. In his simulations using the infinite-sites model, his BLUE method achieved variances nearly as low as the Fu and Li lower bound. It is not obvious from this whether it would perform as well with data from an actual finite-sites DNA sequence model of evolution, where the tree is bound to be harder to infer. Nevertheless, the good behavior of BLUE suggests that a full likelihood method based on summing over all coalescent trees might do almost as well as the Fu-Li lower bound.

In the present paper, the results of a computer simulation of coalescent likelihood estimates of $\Theta$ will be described, demonstrating that one of Fu and Li’s optimistic approximation formulas does do a good job of cal-
Multidimensional mSUGRA likelihood maps

B. C. Allanach

DAMTP, CMS, Wilberforce Road, Cambridge, CB3 0WA, United Kingdom

C. G. Lester

Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, United Kingdom

(Received 18 November 2005; published 25 January 2006)

We calculate the likelihood map in the full 7-dimensional parameter space of the minimal symmetric standard model assuming universal boundary conditions on the supersymmetry breaking. Simultaneous variations of $m_0, A_0, M_{1/2}, \tan \beta, m_t, m_b$ and $\alpha_s(M_Z)$ are applied using a Markov Monte Carlo algorithm. We use measurements of $b \to s \gamma, (g - 2)_\mu$ and $\Omega_{DM} h^2$ in order to constrain the model. We present likelihood distributions for some of the sparticle masses, for the branching $B_s^0 \to \mu^+ \mu^-$ and for $m_{\tilde{t}} - m_{\chi_1^0}$. An upper limit of $2 \times 10^{-8}$ on this branching ratio might be achieved at the Tevatron, and would rule out 29% of the currently allowed likelihood. If one allows for non-three-neutralino components of dark matter, this fraction becomes 35%. The mass ordering allows the imprecise cascade decay $\tilde{q}_L \to \chi_2^0 \to \tilde{l}_R \to \chi_1^0$ with a likelihood of $24 \pm 4\%$. The stop-coannihilation region is highly disfavored, whereas the light Higgs region is marginally disfavored.

DOI: 10.1103/PhysRevD.73.015013

PACS numbers: 14.80.Ly
An apparatus for deriving amplitude values from an information signal, which amplitude values can be reference levels for the states of a finite state machine.
Some traditional aspects of likelihood inference

- plot the likelihood function, use it directly (Fisher, Edwards, Kalbfleisch, Royall, ...)
- Sprott’s approach to reparameterization to eliminate skewness
- calibrate the likelihood function
- as a probability distribution, using Bayes theorem
- asymptotic normality of the log-likelihood function (Fraser & McDunnough)
- likelihood as a means to summary statistics: maximum likelihood estimator, likelihood ratio statistic, estimating (score) function
Some more recent developments

- accommodation of nuisance parameters
- profile likelihood (maximize over nuisance parameters)
- “slice” through the likelihood function; easier to plot
- harder to calibrate; need an adjustment for nuisance parameter maximization
- leading to adjusted profile log-likelihoods
Calibrating the likelihood function

- gives a distribution for the parameter $\theta$, if combined with a prior
- gives an exact distribution for $\hat{\theta}$, in location models only
- gives an approximate distribution for $\hat{\theta}$ in general models
- leads to very accurate approximation to $p$-values and confidence limits
More accurate approximations

\[ f(y; \theta) = \exp\{- (y - \theta) - e^{-(y - \theta)}\} \]

\[ y^0 = 21.5 \]
... more accurate approximations
... more accurate approximations
“Likelihood-like” functions

- partial likelihood (Cox model), adjusted profile likelihood, quasi-likelihood, ...
- data has a complex structure; full joint distribution has complex dependence on the parameter and/or on the data
- ignore part of the data (partial likelihood) or part of the parameter (quasi-likelihood)
- try to retain “good” properties of likelihood (efficient, asymptotically normal, ...) 
- with a function that is simpler to deal with
Example: pairwise likelihood

- each observation $y_i = (y_{i1}, \ldots, y_{iq}) \sim f(y_i; \theta), \quad i = 1, \ldots, n$
- log-likelihood $\ell(\theta) = \sum_{i=1}^{n} \log f(y_i; \theta)$
- simpler:
  \[
  \ell_{\text{pair}}(\theta) = \sum_{i=1}^{n} \sum_{s<t} \log f_{st}(y_{is}, y_{it}; \theta)
  \]

- ignore information in full joint distribution
- easier to calculate, possibly more robust
- score function $U_{\text{pair}}(\theta)$ asymptotically normal, mean 0
- $\tilde{\theta}$ has variance $J^{-1}(\theta) I(\theta) J^{-1}(\theta)$
- $J(\theta) = E_{\theta}\{-\partial U_{\text{pair}}(\theta)/\partial \theta\}$
- $I(\theta) = E_{\theta}\{U_{\text{pair}}(\theta) U_{\text{pair}}(\theta)^T\}$
- generalized estimating equations: use 1st and 2nd moments
Example: symmetric normal

\[ Y_i \sim N(0, R), \quad \text{var} (Y_{ir}) = 1, \quad \text{corr} (Y_{ir}, Y_{is}) = \rho \]

\[
\text{a.var} (\hat{\rho}) = \frac{2}{nq(q-1)} \frac{(1 - \rho)^2}{(1 + \rho^2)^2} c(q^2, \rho^4)
\]
Example: dichotomized MV Normal

\[ Y_r = 1\{Z_r > 0\} \]
\[ Z \sim N(0, R) \]

\[
\begin{array}{cccccccc}
\rho & 0.02 & 0.05 & 0.12 & 0.20 & 0.40 & 0.50 \\
ARE & 0.998 & 0.995 & 0.992 & 0.968 & 0.953 & 0.968 \\
\rho & 0.60 & 0.70 & 0.80 & 0.90 & 0.95 & 0.98 \\
ARE & 0.953 & 0.903 & 0.900 & 0.874 & 0.869 & 0.850 \\
\end{array}
\]
Likelihood ratio statistics

- full log-likelihood

\[ w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi^2_q \]

- pairwise log-likelihood ($q = 1$)

\[ w_{\text{pair}}(\theta) \times \frac{J(\theta)}{I(\theta)} \xrightarrow{d} \chi^2_1 \]

- $J(\theta) = E_\theta\{-\partial U_{\text{pair}}(\theta)/\partial \theta\}$

- $I(\theta) = E_\theta\{U_{\text{pair}}(\theta)U_{\text{pair}}(\theta)^T\}$
... pairwise likelihood

- example of composite likelihood
- pseudo-likelihood for spatial data uses conditional densities instead of marginal
- several applications in current literature (Varin, 2006)
- nested random effects for binary data (Renard et al. 2004)
- generalized linear mixed models (Bellio & Varin 2005)
- familial data (Zhou & Joe 2005; Kuk & Nott 2000)
- network tomography (Liang & Yu 2003)
- efficient, easier to maximize
... in progress; Grace Yun-Yi, Zi Jin

- likelihood ratio inference for vector parameters, or eliminating nuisance parameters

\[ 2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \sum_{i=1}^{q} \lambda_i \chi^2_{1i} \]

- \( \lambda_i \) eigenvalues of \( I^{-1}(\theta)J(\theta) \)
- rescaling profile log-likelihood? (Stafford, 2000)
- asymptotic properties of rescaled log-likelihood?
- relation to generalized estimating equations?
- how to explain efficiency? (Mardia: \( Y \sim N(\mu, \sigma^2 R) \))
- large \( q \) asymptotics (genetics, spatial models, long time series, ...)
Important but omitted

- empirical likelihood (Qin)
- nonparametric and semiparametric likelihood (Murphy, Van der Vaart, Robins)
- maximization by parts (Song, Fan and Kalbfleisch, 2005; Song & Qu)
- weighting observations (Zidek, J.-F. Plante, Neal)
- ...
- Monte Carlo construction of full likelihood (Geyer & Thompson, 1995)
- Robins & Wasserman, 2000
... future prospects

- computational hurdles continue to diminish
- interpretation of results becomes more and more difficult
- likelihood provides a structure, a starting point
- Bayes/frequentist dialogue continues
- calibration of likelihood (computation of \( p \)-values) becomes more important
- new versions of likelihood are proposed