Likelihood based inference

1. Overview of classical asymptotics
2. Profile likelihood and nuisance parameters — NR 2013; 2010
3. *p* growing with *n* — Portnoy 1984, 5, 8
4. $p > n$: regularization — Buhlmann 2013; Taylor et al. 2014
5. approximate likelihoods — composite, quasi, empirical, ...

Models and likelihood

- Model for the probability distribution of y given x
- Density $f(y \mid x)$ with respect to, e.g., Lebesgue measure
- Parameters for the density $f(y \mid x; \theta)$, $\theta = (\theta_1, \ldots, \theta_p)$
- Data $y = (y_1, \ldots, y_n)$ often independent

- Likelihood function $L(\theta; y) \propto f(y; \theta)$ $(y_1, \ldots, y_n)$
- log-likelihood function $\ell(\theta; y) = \log L(\theta; y)$

- often $\theta = (\psi, \lambda)$

- θ could have very large dimension, $p > n$

- θ could have infinite dimension in principle

  $E(y \mid x) = \theta(x)$ 'smooth'

Examples

generalized linear mixed models

GLM: $$y_{ij} \mid u_i \sim \exp\{y_{ij}\eta_{ij} - b(\eta_{ij}) + c(y_{ij})\}$$

linear predictor: $$\eta_{ij} = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}}u_i \qquad j{=}1,\ldots n_i; \quad i{=}1,\ldots m$$

random effects: $$u_i \sim N_k(0, \Sigma)$$

log-likelihood:

$$\begin{aligned}
\ell(\beta,\Sigma) &= \sum_{i=1}^{m}\Biggl(y_i^{\mathrm{T}}X_i\beta - \frac{1}{2}\log|\Sigma| \\
&+ \log\int_{\mathbb{R}^k}\exp\{y_i^{\mathrm{T}}Z_iu_i - 1_i^{\mathrm{T}}b(X_i\beta + Z_iu_i) - \frac{1}{2}u_i^{\mathrm{T}}\Sigma^{-1}u_i\}du_i\Biggr)
\end{aligned}$$

Ormerod & Wand 2012

... complicated likelihoods

- example: clustered binary data
- latent variable:
  $z_{ir} = x_{ir}'\beta + b_i + \epsilon_{ir}, \quad b_i \sim N(0, \sigma_b^2), \quad \epsilon_{ir} \sim N(0, 1)$
- $r = 1, \dots, n_i$: observations in a cluster/family/school...
  $i = 1, \dots, n$ clusters
- random effect b_i introduces correlation between observations in a cluster
- observations: $y_{ir} = 1$ if $z_{ir} > 0$, else 0
- $Pr(y_{ir} = 1 \mid b_i) = \Phi(x_{ir}'\beta + b_i) = p_i \quad \Phi(z) = \int^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$
- likelihood $\theta = (\beta, \sigma_b)$
  $L(\theta; y) = \prod_{i=1}^n \log \int_{-\infty}^{\infty} \prod_{r=1}^{n_i} p_i^{y_{ir}} (1 - p_i)^{1 - y_{ir}} \phi(b_i, \sigma_b^2) db_i$
- more general: $z_{ir} = x_{ir}'\beta + w_{ir}'b_i + \epsilon_{ir}$

Renard et al. (2004)

... complicated likelihoods

Poisson $$f(y_t \mid \alpha_t; \theta) = \exp(y_t \log \mu_t - \mu_t)/y_t!$$

$$\log \mu_t = \beta + \alpha_t$$

autoregression

$$\alpha_t = \phi \alpha_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2), \quad |\phi| < 1, \quad \theta = (\beta, \phi, \sigma^2)$$

likelihood

$$L(\theta; y_1, \ldots, y_n) = \int \left(\prod_{t=1}^{n} f(y_t \mid \alpha_t; \theta) \right) f(\boldsymbol{\alpha}; \theta) d\boldsymbol{\alpha}$$

$L_{approx}(\theta; y)$ via Laplace with some refinements

Davis & Yau, 2011

... complicated likelihoods

multivariate extremes: example, wind speed at d locations

vector observations: $(X_{1i}, \ldots, X_{di}),\ i = 1, \ldots, n$

component-wise maxima: $Z_1, \ldots, Z_d; Z_j = \max(X_{j1}, \ldots, X_{jn})$

Z_j are transformed (centered and scaled)

joint distribution function:

$$\Pr(Z_1 \leq z_1, \ldots, Z_d \leq z_d) = \exp\{-V(z_1, \ldots, z_d)\}$$

$V(\cdot)$ can be parameterized via Gaussian process models

likelihood : need the joint derivatives of $V(\cdot)$

combinatorial explosion

Davison et al., 2012

... complicated likelihoods

Restricted Boltzmann machine:

$$f(v;\theta) = \frac{1}{Z(\theta)} \exp\{\sum_h (h^{\mathrm{T}} W v + \alpha^{\mathrm{T}} h + \beta^{\mathrm{T}} v)\}, \quad \theta = (W, \alpha, \beta)$$

observations: $v_1, \dots, v_n$, independent $\sim f(v;\theta)$; hidden units h

complete data likelihood

$$f(v,h;\theta) = \frac{1}{Z(\theta)} \exp\{f(v,h;\theta)\}$$

partition function: $Z(\theta) = \sum_{v,h} \exp\{f(v,h;\theta)\}$

Jan 30 MZ slides; GL slides

Why likelihood?

- makes probability modelling central
- emphasizes the inverse problem of reasoning from y to θ or $f(\cdot)$
- suggested by Fisher as a measure of plausibility

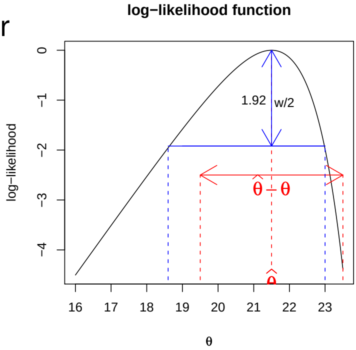
Royall, 1997

$$\begin{array}{ll} L(\hat{\theta})/L(\theta) \in (1,3) & \text{very plausible;} \\ L(\hat{\theta})/L(\theta) \in (3,10) & \text{implausible;} \\ L(\hat{\theta})/L(\theta) \in (10,\infty) & \text{very implausible} \end{array}$$

- converts a 'prior' probability $\pi(\theta)$ to a posterior $\pi(\theta \mid y)$ via Bayes' Theorem
- provides a conventional set of summary quantities: maximum likelihood estimator, score function, ...
- leading to approximate pivotal functions, based on normal distribution
- basis for comparison of models, using AIC or BIC

Derived quantities

- maximum likelihood estimator
  $\hat{\theta} = \arg\sup_\theta \log L(\theta; y)$
  $\quad = \arg\sup_\theta \ell(\theta; y)$



- observed Fisher information
  $j(\hat{\theta}) = -\,\partial^2\ell(\theta)/\partial\theta^2\big|_{\hat{\theta}}$

- efficient score function
  $\ell'(\theta) = \partial\ell(\theta; y)/\partial\theta$

  $\ell'(\hat{\theta}) = 0$

  assuming enough regularity

- $\ell'(\theta; y) = \sum_{i=1}^n \partial \log f_{Y_i}(y_i; \theta)/\partial\theta$

  $y_1, \ldots, y_n$ independent

Limiting results

no nuisance parameters

$$\ell'(\theta)^{\mathrm{T}} j^{-1}(\hat{\theta})\ell'(\theta) \xrightarrow{\mathcal{L}} \chi^2_p,$$

$$q(\theta) \equiv (\hat{\theta} - \theta)^{\mathrm{T}} j(\hat{\theta})(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \chi^2_p,$$

$$w(\theta) \equiv 2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{\mathcal{L}} \chi^2_p$$

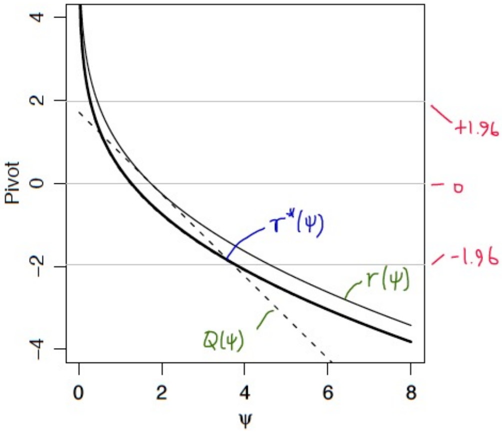
Approximate pivots $p = 1$

$$s(\theta) \equiv \ell'(\theta) j^{-1/2}(\hat{\theta}) \quad \dot{\sim} \quad N(0,1)$$

$$q(\theta) \equiv (\hat{\theta} - \theta) j^{1/2}(\hat{\theta}) \quad \dot{\sim} \quad N(0,1)$$

$$r(\theta) \equiv \pm\surd[2\{\ell(\hat{\theta}) - \ell(\theta)\}^{1/2}] \quad \dot{\sim} \quad N(0,1)$$

... approximate pivots scalar parameter of interest



Nuisance parameters: $\theta = (\psi, \lambda)$

- $\hat{\lambda}_\psi$ constrained maximum likelihood estimator
- profile log-likelihood $\ell_{\mathrm{p}}(\psi) = \ell(\psi, \hat{\lambda}_\psi)$

$$r_e(\psi; y) = (\hat{\psi} - \psi) j_{\mathrm{p}}^{1/2}(\hat{\psi}) \quad \dot{\sim} \quad N(0,1)$$

$$r(\psi; y) = \pm\sqrt{[2\{\ell_{\mathrm{p}}(\hat{\psi}) - \ell_{\mathrm{p}}(\psi)\}]} \quad \dot{\sim} \quad N(0,1)$$

$$\pi_{\mathrm{m}}(\psi \mid y) \quad \dot{\sim} \quad N\{\hat{\psi}, j_{\mathrm{p}}^{-1/2}(\hat{\psi})\}$$

$j_{\mathrm{p}}(\psi) = -\ell_{\mathrm{p}}''(\psi)$; profile information

- treat profile log-likelihood as a one-parameter log-likelihood

The problem with profiling

- $\ell_{\mathsf{p}}(\psi) = \ell(\psi, \hat{\lambda}_\psi)$ used as a 'regular' likelihood, with the usual asymptotics
- neglects errors in the estimation of the nuisance parameter
- can be very large when there are many nuisance parameters

- example: $Y \sim N(X\beta, \sigma^2 I), \quad \hat{\sigma}^2 = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})/n$
- badly biased if $\dim(\beta)$ large relative to n
- easy fix: $\tilde{\sigma}^2 = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})/(n - p)$

- example: $Y_{ij} \sim N(\mu_i, \sigma^2), j = 1, \ldots, n; i = 1, \ldots, p$
- $\hat{\sigma}^2 \xrightarrow{p} \dfrac{n-1}{n}\sigma^2$ as $p \to \infty, n$ fixed

Neyman & Scott, 1948

Reminder: deriving limit results

- $\ell'(\hat\theta; y) = 0 = \ell'(\theta; y) + (\hat\theta - \theta)\ell''(\theta; y)$
- $\ell'(\theta; y)\{-\ell''(\theta; y)\}^{-1} \doteq \hat\theta - \theta$
- $\underbrace{\ell'(\theta; y)}_{\xrightarrow{\mathcal{L}} N(0, i(\theta))} \quad \underbrace{\{-\ell''(\theta; y)\}^{-1}}_{\xrightarrow{p} i^{-1}(\theta)} \quad \Longrightarrow (\hat\theta - \theta) \xrightarrow{\mathcal{L}} N(0, i^{-1}(\theta))$

  $i(\theta) = \mathrm{E}\{j(\theta)\} = \mathrm{E}\{-\ell''(\theta)\} = \mathrm{cov}\{\ell'(\theta)\}$
- M estimator: $\tilde\theta_\rho = \mathrm{argmin}_\theta \Sigma\rho(y_i; \theta)$
- Solution $\Sigma\psi(y_i; \tilde\theta_\rho) = 0, \quad \psi(y; \theta) = \partial\rho(y_i; \theta)/\partial\theta$
- $\tilde\theta_\rho \xrightarrow{\mathcal{L}} N\{0, G^{-1}(\theta)\}$ $\qquad \mathrm{E}\{\psi(Y; \theta)\} = 0$
- $G(\theta) =$
  $\mathrm{E}\{-\partial\psi(Y; \theta)/\partial\theta\}[\mathrm{cov}\{\psi(Y; \theta)\}]^{-1}\mathrm{E}\{-\partial\psi(Y; \theta)/\partial\theta\}$

big data asymptotics

- Neyman-Scott problems: n fixed, $p \to \infty$
- Donoho $n, p \to \infty, \quad p/n \to \beta < \infty$
- likelihood results $n, p \to \infty, \quad p^2/n \to \beta < \infty$

  Portnoy, 1984, 5, 8
- Laplace approx $n, p \to \infty, \quad p = o(n^{1/3})$

  Shun & McCullagh, 1995
- $p > n$: regularize
- lasso

$$\text{argmin}_\beta \, (y - X\beta)^{\mathrm{T}}(y - X\beta) - \lambda\Sigma_j|\beta_j|$$

no intercept

- ridge regression

$$\text{argmin}_\beta \, (y - X\beta)^{\mathrm{T}}(y - X\beta) - \lambda\Sigma_j\beta_j^2$$

$n, p \to \infty$

Portnoy, 1984,5,8

- Model: $y_i = x_i^{\mathrm{T}}\beta + Z_i, \quad i = 1, \ldots, n$ independent
- M-estimation:

$$\sum_{i=1}^{n} x_i \psi(y_i - x_i^{\mathrm{T}}\hat{\beta}) = 0 \tag{1}$$

- result: if ψ is monotone, and $p\log(p)/n \to 0$, and conditions on X, then

  there is a solution of (1) satisfying $||\hat{\beta} - \beta||^2 = O(p/n)$

- "rows of X behave like a sample from a distribution in $\mathbb{R}^p$"

- if $p^{3/2}\log n/n \to 0$, then

$$\max|x_i^{\mathrm{T}}(\hat{\beta} - \beta)| \xrightarrow{p} 0$$

- and

$$a_n^{\mathrm{T}}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, \sigma^2)$$

$\sigma^2 = a_n^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}a_n \mathrm{E}\psi^2(Z)/\{E\psi'(Z)\}^2$

$n, p \to \infty$

Portnoy, 1984,5,8

- Model: $y_i \sim \exp\{\theta^{\mathrm{T}} y - \psi(\theta)\}, i = 1, \ldots, n$ independent; $p = p_n$
- maximum likelihood estimate $\psi'(\hat{\theta}_n) = \bar{y}_n$
- under conditions on the eigenvalues of $\psi''(\theta)$ and moment conditions on y, Fisher information matrix

$$||\hat{\theta}_n - \theta_n||^2 \leq c\frac{p}{n}, \text{ in probability },$$

-

$$||\hat{\theta} - \theta - \bar{y}|| = O_p(p/n) \text{ if } p/n \to 0,$$

- $p^{3/2}/n \to 0$:

$$\sqrt{n} a_n^{\mathrm{T}}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N(0, 1),$$

  likelihood ratio test of simple hypothesis asymptotically χ^2_p
- "asymptotic approximations are trustworthy if $p^{3/2}/n$ is small, but may be very wrong if p^2/n is not small"
- MLE 'will tend to be' consistent if $p/n \to 0$

cf. also El Karoui et al., 2013, PNAS

$n < p$

Bühlmann et al., Lockhart et al.

$$\hat{\beta}_{\text{Lasso}} = \text{argmin}_\beta(||y - X\beta||_2^2 + \lambda||\beta||_1)$$

- prediction: $||X(\hat{\beta}_{\text{Lasso}} - \beta^0||_2^2/n$ 'small'
- estimation: $||\hat{\beta}_{\text{Lasso}} - \beta^0||_q, \quad q \in 1, 2$ 'small'
- selection: $\mathbb{P}(\hat{S} = S_0)$ 'large' S_0 is the 'active set': $\{j : \beta_j^0 \neq 0\}$
- under restricted eigenvalue conditions on X, can get results like

$$||\hat{\beta}_{\text{Lasso}} - \beta^0||_1 = O_p(s_0\sqrt{\log(p)/n}), \quad \lambda \approx \sqrt{\log(p)/n}$$

- what about estimated standard errors for $\hat{\beta}_{\text{Lasso}}$?
- Bühlmann, 2013: the ridge regression estimate

$$\hat{\beta}_R = \text{argmin}_\beta(||y - X\beta||_2^2 + \lambda||\beta||_2),$$

can be bias-corrected

$n < p$

Bühlmann et al., Lockhart et al.

$$\hat{\beta}_{\text{Lasso}} = \text{argmin}_{\beta}(||y - X\beta||_2^2 + \lambda||\beta||_1)$$

First, we define some needed quantities. Let A be the active set just before λ_k, and suppose that predictor j enters at λ_k. Denote by $\hat{\beta}(\lambda_{k+1})$ the solution at the next knot in the path λ_{k+1}, using predictors $A \cup \{j\}$. Finally, let $\tilde{\beta}_A(\lambda_{k+1})$ be the solution of the lasso problem using only the active predictors X_A, at $\lambda = \lambda_{k+1}$. To be perfectly explicit,

$$\tilde{\beta}_A(\lambda_{k+1}) = \underset{\beta_A \in \mathbb{R}^{|A|}}{\text{argmin}} \frac{1}{2}\|y - X_A\beta_A\|_2^2 + \lambda_{k+1}\|\beta_A\|_1. \tag{4}$$

We propose the *covariance test statistic* defined by

$$T_k = (\langle y, X\hat{\beta}(\lambda_{k+1})\rangle - \langle y, X_A\tilde{\beta}_A(\lambda_{k+1})\rangle)/\sigma^2. \tag{5}$$

$$T_k \xrightarrow{\mathcal{L}} \text{Exp}(1)$$

Taylor et al. 2014

Likelihood in complex models

- simplify the likelihood
  - composite likelihood
  - variational approximation
  - Laplace approximation to integrals
- change the mode of inference
  - quasi-likelihood
  - indirect inference
- simulate
  - approximate Bayesian computation
  - MCMC

Composite likelihood

- also called pseudo-likelihood
- reduce high-dimensional dependencies by ignoring them

- for example, replace $f(y_{i1}, \ldots, y_{ik}; \theta)$ by

pairwise marginal $\displaystyle\prod_{j<j'} f_2(y_{ij}, y_{ij'}; \theta)$, or

conditional $\displaystyle\prod_{j} f_c(y_{ij} \mid y_{\mathcal{N}(ij)}; \theta)$

- Composite likelihood function

$$CL(\theta; y) \propto \prod_{i=1}^{n} \prod_{j<j'} f_2(y_{ij}, y_{ij'}; \theta)$$

- Composite ML estimates are consistent, asymptotically normal, not fully efficient   Besag, 1975; Lindsay, 1988

Example: AR Poisson

Davis & Yau, 2011

- Likelihood

$$L(\theta; y_1, \ldots, y_n) = \int \left(\prod_{t=1}^{n} f(y_t \mid \alpha_t; \theta) \right) f(\boldsymbol{\alpha}; \theta) d\boldsymbol{\alpha}$$

- Composite likelihood

$$CL(\theta; y_1, \ldots, y_n) = \prod_{t=1}^{n-1} \int \int f(y_t \mid \alpha_t; \theta) f(y_{t+1} \mid \alpha_{t+1}; \theta) f(\alpha_t, \alpha_{t+1}; \theta) d\alpha_t d\alpha_{t+1}$$

- consecutive pairs
- Time-series asymptotic regime one vector y of increasing length
- Composite ML estimator still consistent, asymptotically normal, estimable asymptotic variance
- Efficient, relative to a Laplace-type approximation
- Surprises: AR(1), fully efficient; MA(1), poor; ARFIMA(0,d,0), ok

Example: spatial extremes

Davison et al., 2012; & Huser, 2015

$$\Pr(Z_1 \le z_1, \dots, Z_d \le z_d) = \exp\{-V(z_1, \dots, z_d; \theta)\}$$

- pairwise composite likelihood used to compare the fits of several competing models
- model choice using "CLIC", an analogue of AIC

  $-2\log(\widehat{CL}) + \mathrm{tr}(J^{-1}K)$
- Davison et al. 2012 applied this to annual maximum rainfall at several stations near Zurich
- "fitting max-stable processes to spatial or spatio-temporal block maxima is awkward ... the use of composite likelihoods ... has become widely used" Davison & Huser

Example: Ising model

Ising model:

$$f(y;\theta) = \exp\Big(\sum_{(j,k)\in E} \theta_{jk} y_j y_k\Big)\frac{1}{Z(\theta)}$$

neighbourhood contributions

$$f(y_j \mid y_{(-j)};\theta) = \frac{\exp(2y_j \sum_{k\neq j}\theta_{jk}y_k)}{\exp(2y_j \sum_{k\neq j}\theta_{jk}y_k)+1}$$

penalized CL estimation based on sample $y^{(1)},\dots,y^{(n)}$

$$\max_{\theta}\left\{\sum_{i=1}^{n}\ell_j(\theta;y^{(i)}) - \sum_j\sum_k P_\lambda(|\theta_{jk}|)\right\}$$

Xue et al., 2012

Ravikumar et al., 2010

Quasi-likelihood

- simplify the model
- $$E(y_i;\theta) = \mu_i(\theta); \qquad \mathrm{Var}(y_i;\theta) = \phi\nu_i(\theta)$$
- consistent with generalized linear models
- example: over-dispersed Poisson responses
- PQL uses this construction, but with random effects

  Molenberghs & Verbeke, Ch. 14
- why does it work?
- score equations are the same as for a 'real' likelihood

  hence unbiased
- derivative of score function equal to variance function

  special to GLMs

Indirect inference

- composite likelihood estimators are consistent

  under conditions ...

- because $\log CL(\theta; y) = \sum_{i=1}^{n} \sum_{j<j'} \log f(y_j, y_{j'}; \theta)$
- derivative w.r.t. θ has expected value 0

- what happens if an estimating equation $g(y; \theta)$ is biased?
- $g(y_1, \ldots, y_n; \tilde{\theta}_n) = 0; \qquad \tilde{\theta}_n \to \theta^*$ $\qquad$ $\mathrm{E}g(Y; \theta^*) = 0$

- $\theta^* = \tilde{k}(\theta)$; invertible? $\theta = k(\theta^*)$ $\qquad$ $\tilde{k}^{-1} \equiv k$

- new estimator $\hat{\theta}_n = k(\tilde{\theta}_n)$
- $k(\cdot)$ is a bridge function, connecting wrong value of θ to the right one $\qquad$ Yi & R, 2010; Jiang & Turnbull, 2004

... indirect inference

Smith, 2008

- model of interest

$$y_t = G_t(y_{t-1}, x_t, \epsilon_t; \theta), \quad \theta \in \mathbb{R}^d$$

- likelihood is not-computable, but can simulate from the model

- simple (wrong) model

$$y_t \sim f(y_t \mid y_{t-1}, x_t; \theta^*), \quad \theta^* \in \mathbb{R}^p$$

- find the MLE in the simple model, $\hat{\theta}^* = \hat{\theta}^*(y_1, \dots, y_n)$, say

- use simulated samples from model of interest
  to find the 'best' β
- 'best' θ gives data that reproduces $\hat{\theta}^*$

Shalizi, 2013

... indirect inference

Smith, 2008

- simulate samples $y_t^m, \quad m = 1, \ldots, M$ at some value θ

- compute $\hat{\theta}^*(\theta)$ from the simulated data

$$\hat{\theta}^*(\theta) = \arg\max_{\theta^*} \sum_m \sum_t \log f(y_t^m \mid y_{t-1}^m, x_t; \theta^*)$$

- choose θ so that $\hat{\theta}^*(\theta)$ is as close as possible to $\hat{\theta}^*$

- if $p = d$ simply invert the 'bridge function'
- usually $p > d$
  - $\hat{\theta}_1 = \arg\min_\theta \{\hat{\theta}^*(\theta) - \hat{\theta}\}^{\mathrm{T}} W \{\hat{\theta}^*(\theta) - \hat{\theta}\}$
  - $\hat{\beta}_2 = \arg\min_\theta (\sum_t \log f(y_t \mid y_{t-1}, x_t, \hat{\theta}^*(\theta)) - \sum_t \log f(y_t \mid y_{t-1}, x_t, \hat{\theta}))$
- estimates of θ are consistent, asymptotically normal, but not efficient

Approximate Bayesian Computation

Marin et al., 2010

- simulate θ' from $\pi(\theta)$
- simulate data z from $f(\cdot;\theta')$
- if $z = y$ then θ' is an observation from posterior $\pi(\cdot \mid y)$
- actually $s(z) = s(y)$ for some set of statistics
- actually $\rho\{s(z), s(y)\} < \epsilon$ for some distance function $\rho(\cdot)$

Fearnhead & Prangle, 2011

- many variations, using different MCMC methods to select candidate values θ'

... approximate Bayesian computation

$M/G/1$ queue: exponential arrival times, general service times, single server

observations y_i: times between departures from the queue

unobserved variables V_i: arrival time of customer i

model:

- $V_1 \sim \text{Exp}(\theta_3)$
- $V_i \mid V_{i-1} \sim V_{i-1} + \text{Exp}(\theta_3)$
- $Y_i \mid X_{i-1}, V_i \sim \text{Uniform}\{\theta_1 + max(0, V_i - X_{i-1}),$
  $\theta_2 + max(0, V_i - X_{i-1})\} \qquad X_i = \sum_{j=1}^{i} Y_j$
- service time $\sim U(\theta_1, \theta_2)$

ABC: use quantiles of departure times as summary statistics

Indirect Inference: use $\bar{y}$, $y_{(1)}$, $\hat{\theta}_2$ from steady-state model

Table 7. Mean quadratic losses for various analyses of 50 $M/G/1$ data sets†

Method	θ_1	θ_2	θ_3
Comparison	1.1	2.2	0.0013
Comparison + regression	*0.020*	1.1	*0.0013*
Semi-automatic ABC	*0.022*	1.0	*0.0013*
Semi-automatic predictors	0.024	1.2	0.0017
Indirect inference	0.18	*0.42*	0.0033

†Losses within 10% of the smallest values for that parameter are italicized.

Fearnhead & Prangle, 2011

ABC and Indirect Inference

Cox & Kartsonaki, 2012

- both methods need a set of parameter values from which to simulate: θ' or θ
- both methods need a set of auxiliary functions of the data $s(y)$ or $\hat{\theta}^*(y)$
- in indirect inference, $\hat{\theta}^*$ is the 'bridge' to the parameters of real interest, θ
- C & K use orthogonal designs based on Hadamard matrices to chose θ'
- and calculate summary statistics focussed on individual components of θ
- MCMC estimation of log-likelihood function

Geyer & Thompson, 1992

cond. comp. likelihood poor for Ising model Okabayashi et al., 2011

Variational methods

Ormerod & Wand, 2010

- in a Bayesian context, want $f(\beta \mid y)$
  use an approximation $q(\beta)$
- dependence of q on y suppressed

- choose $q(\beta)$ to be
  - simple to calculate
  - close to posterior

- simple to calculate
  - $q(\beta) = \prod q_j(\beta_j)$
  - simple parametric family

- close to posterior: miminize Kullback-Leibler divergence

$$KL(q \,||\, f_{post}) = \int q(\beta) \log\{q(\beta)/f(\beta \mid y)\} d\beta$$

... variational methods

Titterington, 2006

- close to posterior:

$$\min_q \int q(\beta)\log\{q(\beta)/f(\beta \mid y)\}d\beta = \min_q KL(q \,||\, f_{post})$$

- equivalent to

best LB for marginal $f(y)$

$$\max_q \int q(\beta)\log\{f(y,\beta)/q(\beta)\}d\beta$$

- in a likelihood context $\log f(y;\theta) = \log \int f(y \mid \beta;\theta)f(\beta)d\beta$

$$= \int q(\beta)\log\{f(y,\beta;\theta)/q(\beta)\}d\beta + KL(q \,||\, f_{post})$$

-

$$\log f(y;\theta) \geq \int q(\beta)\log\{f(y,\beta;\theta)/q(\beta)\}d\beta$$

here β represent random effects u, or b, or ...

Example: GLMM

Ormerod & Wand, 2012

log-likelihood:

$$
\begin{aligned}
\ell(\beta,\Sigma) &= \sum_{i=1}^{m}\Bigg(y_i^{\mathrm{T}}X_i\beta - \frac{1}{2}\log|\Sigma| \\
&+ \log\int_{\mathbb{R}^k}\exp\{y_i^{\mathrm{T}}Z_iu_i - 1_i^{\mathrm{T}}b(X_i\beta+Z_iu_i) - \frac{1}{2}u_i^{\mathrm{T}}\Sigma^{-1}u_i\}du_i\Bigg) \\
&= \sum_{i=1}^{m}\Bigg(y_i^{\mathrm{T}}X_i\beta - \frac{1}{2}\log|\Sigma| \\
&+ \log\int_{\mathbb{R}^k}\exp\{y_i^{\mathrm{T}}Z_iu_i - 1_i^{\mathrm{T}}b(X_i\beta+Z_iu_i) - \frac{1}{2}u_i^{\mathrm{T}}\Sigma^{-1}u_i\}\frac{\phi_{\Lambda_i}(u-\mu_i)}{\phi_{\Lambda_i}(u-\mu_i)}du_i\Bigg)
\end{aligned}
$$

variational approx:

$$
\begin{aligned}
\ell(\beta,\Sigma) &\geq \sum_{i=1}^{m}\Bigg(y_i^{\mathrm{T}}X_i\beta - \frac{1}{2}\log|\Sigma|\Bigg) \\
&+\sum_{i=1}^{m}E_{u\sim N(\mu_i,\Lambda_i)}\Bigg(y_i^{\mathrm{T}}Z_iu - 1_i^{\mathrm{T}}b(X_i\beta+Z_iu) - \frac{1}{2}u^{\mathrm{T}}\Sigma^{-1}u - \log\{\phi_{\Lambda_i}(u-\mu_i)\}\Bigg) \\
&\equiv \ell(\beta,\Sigma,\mu,\Lambda)
\end{aligned}
$$

simplifies to k one-dim. integrals

... variational approximations

Ormerod & Wand, 2012

- $$\ell(\beta, \Sigma) \geq \ell(\beta, \Sigma, \mu, \Lambda)$$
- variational estimate:

$$\ell(\tilde\beta, \tilde\Sigma, \tilde\mu, \tilde\Lambda) = \arg\max_{\beta,\Sigma,\mu,\Lambda} \ell(\tilde\beta, \tilde\Sigma, \tilde\mu, \tilde\Lambda)$$

- inference for $\tilde\beta, \tilde\Sigma$? consistency? asymptotic normality?

  Hall, Ormerod, Wand, 2011; Hall et al. 2011
- emphasis on algorithms and model selection

  e.g. Tan & Nott, 2013, 2014
- VL: approx $L(\theta; y)$ by a simpler function of θ, e.g. $\prod q_j(\theta)$
- CL: approx $f(y; \theta)$ by a simpler function of y, e.g. $\prod f(y_j; \theta)$

Laplace approximation

$$\ell(\theta; y) = \log \int f(y \mid b; \theta) g(b) db = \log \int \exp\{Q(b, y, \theta)\} db, \text{say}$$

$$\ell_{Lap}(\theta; y) = Q(\tilde{b}, y, \theta) - \frac{1}{2} \log |Q''(\tilde{b}, y, \theta)| + c$$

using Taylor series expansion of $Q(\cdot, y, \theta)$ about $\tilde{b}$

simplification of the Laplace approximation leads to PQL:

$$\ell_{PQL}(\theta, b; y) = \log f(y \mid b; \theta) - \frac{1}{2} b^{\mathrm{T}} \Sigma^{-1} b$$

Breslow & Clayton, 1993

to be jointly maximized over b and θ — and parameters in Σ

PQL can be viewed as linearizing $E(y)$ and then using results for linear mixed models — Molenberghs & Verbeke, 2006

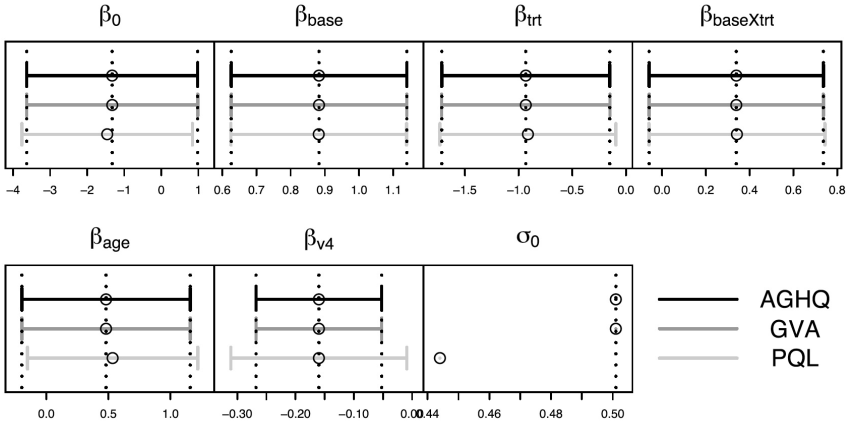


Figure 2. Point estimates and approximate 95% confidence intervals based on each of AGHQ, GVA, and PQL for the *Epilepsy* data random intercept model. The vertical dotted lines correspond to the AGHQ values.

implemented in `lme4` as `glmer`, in `MASS` as `glmmPQL`

Ormerod & Wand, 2012

References

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician* **24**, 179–195.

Breslow, N.E. & Clayton, D. G. (1993). Approximate inference in generalizsed linear mixed models. *J. Am. Statist. Assoc.* **88**, 9–25.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19**, 1212 – 1242.

Bühlmann, P., Kalisch, M. and Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and its Applications* **1**, 255–278.

Cox, D.R. & Kartsonaki, C. (2012). The fitting of complex parametric models. *Biometrika* **99**, 741–747.

Davis, R. & Yau, C.Y. (2011). Comments on pairwise likelihood in time series. *Statistica Sinica* **21**, 255–277.

Davison, A.C., (2012). Statistical modeling of spatial extremes. *Statistical Science* **27**, 161–186.

Davison, A.C. & Huser, R.(2015). Statistics of Extremes *Annual Reviews* **2**, to appear.

El Karoui, N., Bean, D., Bickel, P.J., Lim, C. and Yu, B. (2013). On robust regression with hig-dimensional predictors. *PNAS* **110**, 14557 – 14562.

Fearnhead, P. & Prangle, (2012). Approximate lieklihood methods for estimating local recombination rates *J. R. Statist. Soc.* B **64**, 657–680.

Geyer, C. & Thompson, E.A. (1992). Constrained MC maximum likelihood... *J. R. Statist. Soc.* B **54**, 657–699.

Jiang, W. & Turnbull, B. (2004). The indirect methods ... *Statistical Science* **19**, 239–263 .

Lindsay, B. (1988). Composite likelihood methods. *Contemp. Math.* **80**, 220–239.

Lockhart, R., Taylor, J., Tibshirani, R.J. and Tibshirani, R. (2014). A signficance test for the lasso. *Ann. Statist.* **42**, 413 – 468.

Marin, J.-M. et al. (2010). Approximate Bayesian computational methods. *Stat. & Computing* **22**, 1167–1180.

Molenberghs, G. & Verbeke, G. (2006). *Discrete Longitudinal Data* Springer, New York.

... references

Okabayashi, X. Johnson, L.& Geyer, C.J. (2011). Extending pseudo-likelihood *Statistica Sinica* **21**, 331–347.
Ormerod, & Wand, M. (2012). Gaussian variational approximate inference... *J Comp Graph Statist* **21**, 2–17.
Ormerod, & Wand, M. (2010). Explaining variational approximations. *Am. Stat.* **64**, 140–153.

Portnoy, S. (1984). Asymptotic behaviour of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.* **12**, 1298 – 1309.

Portnoy, S. (1985). Asymptotic behaviour of M-estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13**, 1403 – 1417.

Portnoy, S. (1988). Asymptotic behaviour of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16**, 356–366.

Ravikumar et al. (2010). High-dimensional Ising model selection... *Ann. Statist.* **38**, 1287–1319.

Reid, N. (2013). Aspects of likelihood inference. *Bernoulli* **19**, 1404–1418.

Reid, N. (2010). Likelihood inference. *Wiley Interdisciplinary Reviews in Computational Statistics*, **5**, 517–525.

Renard, D. Molenberghs, G. and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Comp. Stat. Data. Anal.* **44**, 649–667.

Royall, R.J. (1997). *Statistical Evidence...*. Chapman & Hall, London.

Shalizi, C. (2013). Notebooks. indirect inference

Shun, Z. & McCullagh, P. (1995). Laplace approximation ... *J. R. Statist. Soc.* B *57*, 749–760.

Smith, A.A. (2008). Indirect inference. in *New Palgrave Dictionary of Economics* 2nd ed.

Taylor, J., Lockhart, R., Tibshirani, R.J. and Tibshirani, R. (2014). Exact post-selection inference for forward stepwise and least angle regression. http://arxiv.org/pdf/1401.3889v4.pdf

Titterington, D.M. (2006). Bayesian methods for neural networks ... *Statistical Science* **19**, 128–139.

Xue, L., Zou, H. & Cai, T. (2012). Nonconcave penalized composite conditional likelihood... *Ann. Statist.* **40**, 1403–1429.

Yi, G. & Reid, N. (2010). A note on misspecified estimating equations. *Statistica Sinica* **20**, 1749–1769.