Likelihood based inference

1. Overview of classical asymptotics

2. Profile likelihood and nuisance parameters

3. \( p \) growing with \( n \)

4. \( p > n \): regularization

5. approximate likelihoods
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 \( \theta = (\psi, \lambda) \)

- \( \theta \) could have very large dimension, \( p > n \)

- \( \theta \) could have infinite dimension in principle

\[
E(y \mid x) = \theta(x) \quad \text{‘smooth’}
\]
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: \[ \eta_{ij} = x_{ij}^T \beta + z_{ij}^T u_i \quad j=1,...,n_i; \quad i=1,...,m \]

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

log-likelihood:

\[
\ell(\beta, \Sigma) = \sum_{i=1}^{m} \left( y_i^T X_i \beta - \frac{1}{2} \log |\Sigma| \right) \\
+ \log \int_{\mathbb{R}^k} \exp\{y_i^T Z_i u_i - 1_i^T b(X_i \beta + Z_i u_i) - \frac{1}{2} u_i^T \Sigma^{-1} u_i\} du_i
\]

Ormerod & Wand 2012
... complicated likelihoods

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

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

\[ L_{\text{approx}}(\theta; y) \text{ via Laplace with some refinements} \]

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
Restricted Boltzmann machine:

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

observations: \( v_1, \ldots, v_n \), 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 $\theta$ or $f(\cdot)$
- suggested by Fisher as a measure of plausibility

\[
\frac{L(\hat{\theta})}{L(\theta)} \in (1, 3) \quad \text{very plausible;}
\]
\[
\frac{L(\hat{\theta})}{L(\theta)} \in (3, 10) \quad \text{implausible;}
\]
\[
\frac{L(\hat{\theta})}{L(\theta)} \in (10, \infty) \quad \text{very implausible}
\]

- 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

Royall, 1997
Derived quantities

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

- observed Fisher information
  \[j(\hat{\theta}) = -\frac{\partial^2 \ell(\theta)}{\partial \theta^2} \bigg|_{\hat{\theta}}\]

- efficient score function
  \[\ell'(\theta) = \frac{\partial \ell(\theta; y)}{\partial \theta}\]
  \[\ell'(\hat{\theta}) = 0\]

- \[\ell'(\theta; y) = \sum_{i=1}^{n} \frac{\partial \log f_Y(y_i; \theta)}{\partial \theta}\]

assuming enough regularity

\[y_1, \ldots, y_n\] independent
Limiting results

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

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

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

Approximate pivots \( p = 1 \)

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

\[ q(\theta) \equiv (\hat{\theta} - \theta) j^{1/2}(\hat{\theta}) \sim N(0, 1) \]

\[ r(\theta) \equiv \pm \sqrt{2\{\ell(\hat{\theta}) - \ell(\theta)\}^{1/2}} \sim N(0, 1) \]
... approximate pivots

scalar parameter of interest
Nuisance parameters: $\theta = (\psi, \lambda)$

- $\hat{\lambda}_\psi$ constrained maximum likelihood estimator

- profile log-likelihood $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$

\[
re(\psi; y) = (\hat{\psi} - \psi)j_p^{1/2}(\hat{\psi}) \sim N(0, 1)
\]

\[
r(\psi; y) = \pm \sqrt{2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}} \sim N(0, 1)
\]

\[
\pi_m(\psi | y) \sim N\{\hat{\psi}, j_p^{-1/2}(\hat{\psi})\}
\]

\[
\pi_m(\psi | y) \sim N\{\hat{\psi}, j_p^{-1/2}(\hat{\psi})\}
\]

\[
\ell_p(\psi) = -\ell_p''(\psi); \text{ profile information}
\]

- treat profile log-likelihood as a one-parameter log-likelihood
The problem with profiling

- $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$ 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:** $Y \sim N(X\beta, \sigma^2 I)$, $\hat{\sigma}^2 = (y - X\hat{\beta})^T(y - X\hat{\beta})/n$
- badly biased if $\text{dim}(\beta)$ large relative to $n$
- easy fix: $\tilde{\sigma}^2 = (y - X\hat{\beta})^T(y - X\hat{\beta})/(n - p)$

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

Neyman & Scott, 1948
Reminder: deriving limit results

\[ \ell'(\hat\theta; y) = 0 = \ell'(\theta; y) + (\hat\theta - \theta)\ell''(\theta; y) \]

\[ \ell'(\theta; y)\{−\ell''(\theta; y)\}^{-1} \overset{\text{L}}{\rightarrow} \hat\theta - \theta \]

\[ \ell'(\theta; y) \{−\ell''(\theta; y)\}^{-1} \Rightarrow (\hat\theta - \theta) \overset{\text{L}}{\rightarrow} N(0, i^{-1}(\theta)) \]

\[ i(\theta) = E\{j(\theta)\} = E\{−\ell''(\theta)\} = \text{cov}\{\ell'(\theta)\} \]

\[ \textbf{M estimator: } \tilde{\theta}_{\rho} = \text{argmin}_{\theta} \Sigma \rho(y_i; \theta) \]

\[ \Sigma \psi(y_i; \tilde{\theta}_{\rho}) = 0, \quad \psi(y; \theta) = \partial \rho(y_i; \theta) / \partial \theta \]

\[ \tilde{\theta}_{\rho} \overset{\text{L}}{\rightarrow} N\{0, G^{-1}(\theta)\} \]

\[ G(\theta) = \text{E}\{-\partial \psi(Y; \theta) / \partial \theta\} \text{[cov}\{\psi(Y; \theta)\}\text{]}^{-1} \text{E}\{-\partial \psi(Y; \theta) / \partial \theta\} \]

\[ \text{E}\{\psi(Y; \theta)\} = 0 \]

Topics in Inference  Fields Institute, 2015
big data asymptotics

- Neyman-Scott problems: $n$ fixed, $p \to \infty$

- Donoho $n, p \to \infty$, $p/n \to \beta < \infty$

- likelihood results $n, p \to \infty$, $p^2/n \to \beta < \infty$

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

$\beta > n$: regularize

- lasso

$$\argmin_{\beta} (y - X\beta)^T(y - X\beta) - \lambda \sum_j |\beta_j|$$

no intercept

- ridge regression

$$\argmin_{\beta} (y - X\beta)^T(y - X\beta) - \lambda \sum_j \beta_j^2$$

Portnoy, 1984, 5, 8

Shun & McCullagh, 1995
\[ n, p \to \infty \] Portnoy, 1984,5,8

- **Model:** \[ y_i = x_i^T \beta + Z_i, \quad i = 1, \ldots, n \]
- **\( M \)-estimation:**
  \[
  \sum_{i=1}^{n} x_i \psi(y_i - x_i^T \hat{\beta}) = 0 \quad (1)
  \]
- **result:** if \( \psi \) 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
  \[
  \max |x_i^T(\hat{\beta} - \beta)| \xrightarrow{p} 0
  \]
- and
  \[
  a_n^T(\hat{\beta} - \beta) \overset{\mathcal{L}}{\to} N(0, \sigma^2)
  \]
  \[
  \sigma^2 = a_n^T(X^T X)^{-1} a_n E\psi^2(Z) / \{E\psi'(Z)\}^2
  \]
$n, p \to \infty$

- Model: $y_i \sim \exp\{\theta^T y - \psi(\theta)\}, i = 1, \ldots, n$ 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 $y$,

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

- $||\hat{\theta} - \theta - \bar{y}|| = O_p(p/n)$ if $p/n \to 0$,
- $p^{3/2}/n \to 0$:

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

likelihood ratio test of simple hypothesis asymptotically $\chi^2_p$

- “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
\[ \hat{\beta}_{\text{Lasso}} = \arg\min_\beta (\| y - X\beta \|^2_2 + \lambda \| \beta \|^1) \]

- 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': \( \{ j : \beta^0_j \neq 0 \} \)

- 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

\[ \hat{\beta}_R = \arg\min_\beta (\| y - X\beta \|^2_2 + \lambda \| \beta \|^2), \]

\text{can be bias-corrected}
\[ \hat{\beta}_{\text{Lasso}} = \arg\min_{\beta} \left( \| y - X\beta \|^2 + \lambda \| \beta \|_1 \right) \]

First, we define some needed quantities. Let \( A \) be the active set just before \( \lambda_k \), and suppose that predictor \( j \) enters at \( \lambda_k \). Denote by \( \hat{\beta}(\lambda_{k+1}) \) the solution at the next knot in the path \( \lambda_{k+1} \), 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,

\[
(4) \quad \tilde{\beta}_A(\lambda_{k+1}) = \arg\min_{\beta_A \in \mathbb{R}^{|A|}} \frac{1}{2} \| y - X_A\beta_A \|^2 + \lambda_{k+1}\| \beta_A \|_1.
\]

We propose the covariance test statistic defined by

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

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

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 $\prod_{j<j'} f_2(y_{ij}, y_{ij'}; \theta)$, or
  
  conditional $\prod_{j} f_c(y_{ij} | y_{N(ij)}; \theta)$

- 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

David & Yau, 2011

- **Likelihood**

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

- **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

\[
Pr(Z_1 \leq z_1, \ldots, Z_d \leq z_d) = \exp\{ -V(z_1, \ldots, z_d; \theta) \}
\]

▶ pairwise composite likelihood used to compare the fits of several competing models

▶ model choice using “CLIC”, an analogue of AIC

\[-2 \log(\hat{CL}) + \text{tr}(J^{-1}K)\]

▶ 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:

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

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 \( y^{(1)}, \ldots, y^{(n)} \)

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

Xue et al., 2012
Ravikumar et al., 2010
Quasi-likelihood

- simplify the model

\[
E(y_i; \theta) = \mu_i(\theta); \quad \text{Var}(y_i; \theta) = \phi \nu_i(\theta)
\]

- 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. \( \theta \) has expected value 0

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

\[ E g(Y; \theta^*) = 0 \]

- \( \theta^* = \tilde{k}(\theta) \); invertible? \( \theta = k(\theta^*) \)

\[ \tilde{k}^{-1} \equiv k \]

- new estimator \( \hat{\theta}_n = k(\tilde{\theta}_n) \)
- \( k(\cdot) \) is a bridge function, connecting wrong value of \( \theta \) to the right one

Yi & R, 2010; Jiang & Turnbull, 2004
... indirect inference

- 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, \( \hat{\theta}^* = \hat{\theta}^*(y_1, \ldots, y_n) \), say

- use simulated samples from model of interest to find the ‘best’ \( \beta \)

- ‘best’ \( \theta \) gives data that reproduces \( \hat{\theta}^* \)
... indirect inference

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

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

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

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

- if $p = d$ simply invert the ‘bridge function’

- usually $p > d$
  - $\hat{\theta}_1 = \arg \min_{\theta} \left\{ \hat{\theta}^*(\theta) - \hat{\theta} \right\}^T W \left\{ \hat{\theta}^*(\theta) - \hat{\theta} \right\}$
  - $\hat{\beta}_2 = \arg \min_{\theta} \left( \sum_t \log f(y_t | y_{t-1}, x_t, \hat{\theta}^*(\theta)) - \sum_t \log f(y_t | y_{t-1}, x_t, \hat{\theta}) \right)$

- estimates of $\theta$ are consistent, asymptotically normal, but not efficient
Approximate Bayesian Computation

- simulate $\theta'$ from $\pi(\theta)$

- simulate data $z$ from $f(\cdot; \theta')$

- if $z = y$ then $\theta'$ 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 $\rho(\cdot)$

- many variations, using different MCMC methods to select candidate values $\theta'$

Fearnhead & Prangle, 2011
... approximate Bayesian computation

\textbf{M/G/1 queue}: exponential arrival times, general service times, single server

\textbf{observations} \( y_i \): times between departures from the queue

\textbf{unobserved variables} \( V_i \): arrival time of customer \( i \)

\textbf{model}:

\begin{itemize}
  \item \( V_1 \sim \text{Exp}(\theta_3) \)
  \item \( V_i | V_{i-1} \sim V_{i-1} + \text{Exp}(\theta_3) \)
  \item \( Y_i | X_{i-1}, V_i \sim \text{Uniform}\{\theta_1 + \max(0, V_i - X_{i-1}), \theta_2 + \max(0, V_i - X_{i-1})\} \)
  \item service time \( X_i = \sum_{j=1}^{i} Y_j \)
\end{itemize}

\textbf{ABC}: use quantiles of departure times as summary statistics

\textbf{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†

<table>
<thead>
<tr>
<th>Method</th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\theta_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>1.1</td>
<td>2.2</td>
<td>0.0013</td>
</tr>
<tr>
<td>Comparison + regression</td>
<td>0.020</td>
<td>1.1</td>
<td>0.0013</td>
</tr>
<tr>
<td>Semi-automatic ABC</td>
<td>0.022</td>
<td>1.0</td>
<td>0.0013</td>
</tr>
<tr>
<td>Semi-automatic predictors</td>
<td>0.024</td>
<td>1.2</td>
<td>0.0017</td>
</tr>
<tr>
<td>Indirect inference</td>
<td>0.18</td>
<td>0.42</td>
<td>0.0033</td>
</tr>
</tbody>
</table>

†Losses within 10% of the smallest values for that parameter are italicized.
ABC and Indirect Inference

- both methods need a set of parameter values from which to simulate: $\theta'$ or $\theta$
- 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, $\theta$
- C & K use orthogonal designs based on Hadamard matrices to chose $\theta'$
- and calculate summary statistics focussed on individual components of $\theta$
- MCMC estimation of log-likelihood function

Geyer & Thompson, 1992
cond. comp. likelihood poor for Ising model
Okabayashi et al., 2011
Variational methods

- 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: minimize Kullback-Leibler divergence

$$KL(q \parallel f_{post}) = \int q(\beta) \log \{q(\beta)/f(\beta \mid y)\} d\beta$$
... variational methods

- close to posterior:

\[
\min_q \int q(\beta) \log \left\{ \frac{q(\beta)}{f(\beta \mid y)} \right\} d\beta = \min_q KL(q \parallel f_{post})
\]

- equivalent to

\[
\max_q \int q(\beta) \log \left\{ \frac{f(y, \beta)}{q(\beta)} \right\} d\beta
\]

- in a likelihood context

\[
\log f(y; \theta) = \log \int f(y \mid \beta; \theta)f(\beta)d\beta
\]

\[
= \int q(\beta) \log \left\{ \frac{f(y, \beta; \theta)}{q(\beta)} \right\} d\beta + KL(q \parallel f_{post})
\]

- \[
\log f(y; \theta) \geq \int q(\beta) \log \left\{ \frac{f(y, \beta; \theta)}{q(\beta)} \right\} d\beta
\]

here $\beta$ represent random effects $u$, or $b$, or...
Example: GLMM

log-likelihood:

$$\ell(\beta, \Sigma) = \sum_{i=1}^{m} \left( y_i^T X_i \beta - \frac{1}{2} \log |\Sigma| \right)$$

$$+ \log \int_{\mathbb{R}^k} \exp\{y_i^T Z_i u_i - 1_i^T b(X_i \beta + Z_i u_i) - \frac{1}{2} u_i^T \Sigma^{-1} u_i\} du_i$$

$$= \sum_{i=1}^{m} \left( y_i^T X_i \beta - \frac{1}{2} \log |\Sigma| \right)$$

$$+ \log \int_{\mathbb{R}^k} \exp\{y_i^T Z_i u_i - 1_i^T b(X_i \beta + Z_i u_i) - \frac{1}{2} u_i^T \Sigma^{-1} u_i\} \phi_{\Lambda_i}(u - \mu_i) \phi_{\Lambda_i}(u - \mu_i) du_i$$

variational approx:

$$\ell(\beta, \Sigma) \geq \sum_{i=1}^{m} \left( y_i^T X_i \beta - \frac{1}{2} \log |\Sigma| \right)$$

$$+ \sum_{i=1}^{m} E_{u \sim N(\mu_i, \Lambda_i)} \left( y_i^T Z_i u - 1_i^T b(X_i \beta + Z_i u) - \frac{1}{2} u^T \Sigma^{-1} u - \log\phi_{\Lambda_i}(u - \mu_i) \right)$$

$$\equiv \ell(\beta, \Sigma, \mu, \Lambda)$$

simplifies to $k$ one-dim. integrals
... variational approximations

\[ \ell(\beta, \Sigma) \geq \ell(\beta, \Sigma, \mu, \Lambda) \]

▶ 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 \( \theta \), 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

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

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

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

simplification of the Laplace approximation leads to PQL:

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

Breslow & Clayton, 1993

to be jointly maximized over \(b\) and \(\theta\)

PQL can be viewed as linearizing \(E(y)\) and then using results
for linear mixed models

Molenberghs & Verbeke, 2006
implemented in lme4 as glmer, in MASS as glmmPQL

Ormerod & Wand, 2012

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.
References

