Inference without likelihoods

Richard E. Chandler

Department of Statistical Science, University College London
(r.chandler@.ucl.ac.uk)
Joint work with Joao Jesus

Maurice Priestley commemoration day, 18th December 2013

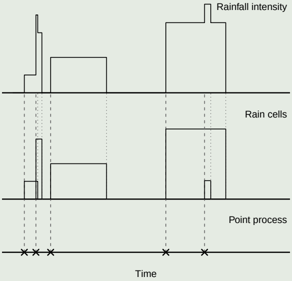
Motivation

- Likelihood function fundamental to most statistical inference
  - Measures relative fidelity of model to data under different parameter values
- But may be unable or unwilling to formulate likelihood in some settings, e.g.:
  - Dependent non-Gaussian processes: relative scarcity of tractable multivariate distributions
  - Where data do not correspond directly to model structure (e.g. models in continuous time, data aggregated or sampled at discrete time points)
  - Where likelihood would encourage fidelity to features of the data that (simplified) models were not designed to reproduce
  - Where models are non-probabilistic

Example: point process models for rainfall

- Hydrologists need models to simulate rainfall time series for use when designing dams, reservoirs, sewage systems etc.
- Popular class of models based on point processes
  - Used in 'weather generator' provided in official UK climate projections (http://ukclimateprojections.defra.gov.uk)

- Simplified representation of rainfall mechanism: superposition of rain cells, each attached to event of a point process
- Each cell has random duration & constant random intensity.
- Rainfall intensity at any time is sum of intensities over all active cells.



Inference for point process rainfall models

- Model parameters are (e.g.) cell arrival rate, mean cell duration, mean cell intensity etc.
- Rainfall data typically totals over (e.g.) hourly intervals
- Likelihood-based inference infeasible: joint density of data unavailable
- Likelihood-based inference also undesirable because of rectangular temporal profiles of cells:
  - Observed rainfall series rarely contain same value in successive wet intervals, need new cell at each time point to achieve this using model ('fidelity to data')
- Models usually fitted using generalised method of moments: match observed and modelled values of selected properties for which analytical expressions are available

Beyond likelihood: estimating functions

- Many problem-specific techniques available to overcome difficulties with likelihood-based inference (EM algorithm, Bayesian methods, composite likelihood, ...)
- Focus here on estimating functions (EFs) as unifying theory
- EFs widely known as 'folklore' in statistical community — but most literature focused on optimality in specific settings
- Aim here to summarise theory in accessible & generally applicable terms, & look at some applications

Reference

Jesus, J. and R.E. Chandler (2011). Estimating functions and the generalized method of moments. *Interface Focus*, **1(6)**, 871-885.

Remainder of talk

1. Review of EF theory
   (a) Main definitions & properties
   (b) Example 1: the generalised method of moments
   (c) Example 2: Whittle likelihood
2. Application to rainfall models
3. Summary

Estimating functions: overview of theory

Definition (estimating function / equation)

Given a model with $p \times 1$ parameter vector θ, and a $n \times 1$ vector $\mathbf{y}$ of data values, suppose that θ is estimated by solving an equation of the form

$$\mathbf{g}(\theta; \mathbf{y}) = \mathbf{0} \tag{1}$$

where $\mathbf{g}(\cdot;\cdot)$ is a vector-valued function containing p elements. Such a function $\mathbf{g}(\cdot;\cdot)$ is an *estimating function* (EF), and an equation of the form (1) is an *estimating equation*.

- Often $\mathbf{g}(\theta;\cdot)$ is gradient vector (e.g. of log-likelihood or error sum of squares) — but framework doesn't require this

Asymptotics: target of estimation

- Extend notation: let $\mathbf{Y}_n$ be $n \times 1$ vector of random variables; $\mathbf{g}_n(\cdot;\cdot)$ be corresponding EF; $\hat{\theta}_n$ be root of equation

$$\mathbf{g}_n(\theta; \mathbf{Y}_n) = \mathbf{0}\ . \tag{2}$$

- Implicit assumption: (2) has at least one root.

Definition (target of estimation)

Assume existence of sequence (η_n) of $p \times p$ matrices, independent of θ and such that as $n \to \infty$, $\eta_n \mathbf{g}_n(\theta; \mathbf{Y}_n)$ converges uniformly in probability to a non-random function, $\mathbf{g}_\ell(\theta)$ say, with the following properties:

1. The equation $\mathbf{g}_\ell(\theta) = \mathbf{0}$ has a unique root at $\theta = \theta_0$.
2. $g_\ell(\cdot)$ is bounded away from zero except in neighbourhood of θ_0.

Then θ_0 is *target of estimation* or *object of inference*.

Asymptotics: convergence of the estimator

Result

Under conditions given above, as $n \to \infty$ the EF defines a unique estimator $\hat{\theta}_n$ that converges in probability to θ_0.

Comments on conditions:

- Often easy to establish pointwise convergence of $\eta_n \mathbf{g}_n(\theta; \mathbf{Y}_n)$ but uniform convergence can be technically challenging
- Some approaches to ensure uniform convergence:
  - Assume parameter space is compact.
  - Impose conditions of smoothness on EFs $\{\mathbf{g}_n(\theta; \cdot)\}$.
  - Write EF as continuous function of finite vector $\mathbf{T}_n(\mathbf{Y}_n)$ of statistics, which itself converges in probability to some limiting value.
- More details: van der Vaart (1998) *Asymptotic statistics*, Ch. 5.

Asymptotics: limiting distribution

Result

- Assume existence of sequences (γ_n) and (δ_n) of invertible $p \times p$ matrices that do not depend on θ and are such that:
  1. $\lim_{n\to\infty} \mathrm{Var}\left(\tilde{\mathbf{g}}_n\left(\theta_0; \mathbf{Y}_n\right)\right) = \tilde{\Sigma}$ where $\tilde{\mathbf{g}}_n\left(\theta; \mathbf{Y}_n\right) = \gamma_n \mathbf{g}_n\left(\theta; \mathbf{Y}_n\right)$ and $\tilde{\Sigma}$ is a positive definite matrix.
  2. Defining $\tilde{\mathbf{G}}_n(\theta) = \partial \tilde{\mathbf{g}}_n / \partial\theta$, within a neighbourhood of θ_0 the matrix $\tilde{\mathbf{G}}_n(\theta)\delta_n$ converges uniformly in probability to an invertible matrix $\mathbf{M}(\theta)$ with elements that are continuous functions of θ.
- Then $\lim_{n\to\infty} \mathrm{E}\left(\hat{\theta}_n\right) = \theta_0$ & $\lim_{n\to\infty} \mathrm{Var}\left(\delta_n^{-1}\hat{\theta}_n\right) = \mathbf{M}_0^{-1}\tilde{\Sigma}\left(\mathbf{M}_0^{-1}\right)'$ where $\mathbf{M}_0 = \mathbf{M}\left(\theta_0\right)$.
- If, in addition, $\tilde{\mathbf{g}}_n\left(\theta; \mathbf{Y}_n\right)$ has limiting multivariate normal (MVN) distribution then so does $\delta_n^{-1}\hat{\theta}_n$.

Comments on limiting distribution

- Conditions are easy to check & hold in wide variety of settings
- Can often set $\eta_n = n^{-1}\mathbf{I}_{p\times p}$, $\gamma_n = \delta_n = n^{-1/2}\mathbf{I}_{p\times p}$ but different choices needed for (e.g.) long-memory processes, combinations of stationary and non-stationary elements of $\mathbf{g}_n(\cdot;\cdot)$ etc.
- Limiting result more usefully restated for operational use:

Operational statement of limiting result

Let Σ_n denote covariance matrix of $\mathbf{g}_n(\theta_0;\mathbf{Y}_n)$. Then under previous assumptions, and if $\mathbf{G}_0 = \mathrm{E}\left[\partial\mathbf{g}_n/\partial\theta|_{\theta=\theta_0}\right]$ exists, $\hat{\theta}_n \sim \mathrm{MVN}\left(\theta_0, \mathbf{G}_0^{-1}\Sigma_n\left[\mathbf{G}_0^{-1}\right]'\right)$ approximately in large samples.

Extensions of result

- Generalisation available without requiring existence of expectations or covariance matrices (Sweeting, 1980, *Ann. Stat.* **8**, 1375-1381).
- Extension to processes for which sequence $\left(\tilde{\mathbf{G}}_n(\theta)\delta_n\right)$ converges in distribution to random matrix $\mathbf{M}_0$; then inference about θ_0 is conditional upon realised value of $\mathbf{M}_0$ (Sweeting, 1992, *Ann. Stat.*, **20**, 580-589).
  - Needed, e.g., when regressing time series upon random walk covariate

Model comparison

- Limiting result forms basis for testing hypotheses of form $H_0 : \Xi\theta = \xi_0$ where Ξ is $q \times p$ matrix of rank q.
- Let $\Gamma_n = \mathbf{G}_0^{-1}\Sigma_n\left[\mathbf{G}_0^{-1}\right]'$ be approximate covariance matrix of $\hat{\theta}$ from operational version of limiting result; then

$$\hat{\xi}_n = \Xi\hat{\theta}_n \sim \mathrm{MVN}\left(\Xi\theta_0, \Xi\Gamma_n\Xi'\right) \tag{3}$$

- Suggests quasi-Wald test statistic

$$\left(\hat{\xi}_n - \xi_0\right)\left[\Xi\Gamma_n\Xi'\right]^{-1}\left(\hat{\xi}_n - \xi_0\right)' \tag{4}$$

with approximate χ^2_q distribution under H_0.
- Alternative: quasi-score test based on value of EF itself (easiest when EF is gradient vector so that value under H_0 is defined)

Model comparison continued

- Final option when EF is gradient vector: $\mathbf{g}_n(\theta; \mathbf{Y}_n) = \partial Q_n / \partial \theta$ say
- Let $\tilde{\theta}_n$ be optimiser of Q_n under restriction $\Xi\theta = \xi_0$; then test can be based on statistic

$$2\left| Q_n\left(\tilde{\theta}_n; \mathbf{Y}_n\right) - Q_n\left(\hat{\theta}_n; \mathbf{Y}_n\right)\right| \tag{5}$$

- Null distribution is that of $\mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z}$ where $\mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{I}_{q\times q})$ and $\mathbf{A} = \Xi\mathbf{G}_0^{-1}\Xi$ — can approximate with scaled and shifted χ^2 dbn.
- **NB** results yield standard χ^2 asymptotics when $\mathbf{g}(\cdot;\cdot)$ is gradient of log-likelihood.
- Some practical and theoretical benefits from adjusting $Q_n(\cdot;\cdot)$ before calculating (5) — see Chandler & Bate, 2007, *Biometrika*, **94**, 167-183 in context of mis-specified log-likelihoods.

Practical issues

Recap: limiting result

$\hat{\theta}_n \sim \mathrm{MVN}\left(\theta_0, \mathbf{G}_0^{-1}\Sigma_n\left[\mathbf{G}_0^{-1}\right]'\right)$ approx., where $\Sigma_n = \mathrm{Var}\left[\mathbf{g}_n\left(\theta_0; \mathbf{Y}_n\right)\right]$ & $\mathbf{G}_0 = \mathrm{E}\left[\partial \mathbf{g}_n / \partial\theta|_{\theta=\theta_0}\right]$.

- Need consistent estimators of Σ_n & $\mathbf{G}_0$
  - Can use any estimator for which estimation error is asymptotically negligible compared with quantity being estimated.
- Some options:
  - Plug estimate $\hat{\theta}_n$ into expressions for $\mathbf{G}_0$ and Σ_n, if available.
  - For $\mathbf{M}_0$, numerical differentiation of $\mathbf{g}_n(\cdot;\cdot)$ at $\hat{\theta}_n$.
  - Use empirical estimator for Σ_n — needs replication e.g. by splitting data into (quasi-)independent subsets

Example 1: the generalised method of moments (GMM)

- Consider vector $\mathbf{T}_n = \mathbf{T}_n(\mathbf{Y}_n)$ of $k \geq p$ summary statistics with:
  - $\mathrm{E}[\mathbf{T}_n] = \tau(\theta)$
  - $\lim_{n\to\infty} \mathrm{Var}[\gamma_n \mathbf{h}_n(\theta; \mathbf{Y}_n)] = \mathbf{S}$ for some sequence (γ_n) of $k \times k$ matrices that do not depend on θ, where $\mathbf{h}_n(\theta; \mathbf{Y}_n) = \mathbf{T}_n - \tau(\theta)$.
- Estimate θ by minimising

$$Q_n(\theta; \mathbf{Y}_n) = \left[\tilde{\mathbf{h}}_n(\theta; \mathbf{Y}_n)\right]' \mathbf{W}_n \tilde{\mathbf{h}}_n(\theta; \mathbf{Y}_n) \ . \tag{6}$$

  where
  - $\tilde{\mathbf{h}}_n(\theta; \mathbf{Y}_n) = \gamma_n \mathbf{h}_n(\theta; \mathbf{Y}_n)$
  - $\mathbf{W}_n$ is $k \times k$ matrix with $\mathrm{p}\lim_{n\to\infty} \mathbf{W}_n = \mathbf{W}$ (+ve definite)
- Resulting EF is

$$\mathbf{g}_n(\theta; \mathbf{Y}_n) = \tilde{\mathbf{H}}_n'(\theta) \mathbf{W}_n \tilde{\mathbf{h}}_n(\theta; \mathbf{Y}_n) \tag{7}$$

  where $\tilde{\mathbf{H}}_n(\theta) = \partial\tilde{\mathbf{h}}_n/\partial\theta = -\gamma_n \partial\tau/\partial\theta$.

GMM: comments

- Requirements for EF asymptotics translate into convergence and continuity requirements for $\mathbf{T}_n$ and $\tau(\theta)$, their properties & derivatives
- Large-sample covariance matrix suggests optimal choice of **W** is $\mathbf{W} = \mathbf{S}^{-1}$
  - Recap: **S** is limiting covariance matrix of normalised summary statistics
  - **NB** however: **S** must be estimated — sampling errors here can affect inference particularly if $k^2 \gg p$ & elements of **S** are estimated separately
  - Alternative ('2-step procedure'): use preliminary estimate of θ to obtain improved estimate of **S**, then re-estimate θ

Example 2: the Whittle likelihood

- Often want to study stationary processes for which likelihood function is analytically / computationally intractable
- 1950s: Whittle formulated frequency-domain approximation to full likelihood for zero-mean Gaussian processes
- Subsequent work extended approach to linear, long-memory, ARCH, locally stationary . . . processes
- Alternative justification (REC & TSR, Athens Conference, 1996): treat sample Fourier coefficients as observations and use standard large-sample properties (approx. independent & normal with variance proportional to spectral density):
  - Justifies use of Whittle estimator in non-Gaussian settings
  - Accommodates processes with non-zero mean by incorporating Fourier coefficient at zero frequency

Whittle likelihood from Fourier coefficients

Definition (Whittle log-likelihood for a stationary process)

$$\log L(\theta) = -\sum_{j=0}^{\lfloor n/2 \rfloor} \left(1 - \frac{1}{2}\delta_{j,n/2}\right) \left[\frac{I(\omega_j)}{h(\omega_j;\theta)} + \log h(\omega_j;\theta)\right] - \frac{1}{2}\left[\log h(0;\theta) + \frac{(A_0 - n\mu(\theta))^2}{h(0;\theta)}\right] \text{, where} \quad (8)$$

- $\delta_{\cdot,\cdot}$ is Kronecker delta
- $I(\omega_j)$ is periodogram at frequency $\omega_j = 2\pi j/n$
- $h(\omega;\theta)$ is theoretical spectral density at frequency ω
- $A_0 = \sum_{t=1}^{n} Y_t$ is sample Fourier coefficient at zero frequency
- $\mu(\theta)$ is theoretical mean of process

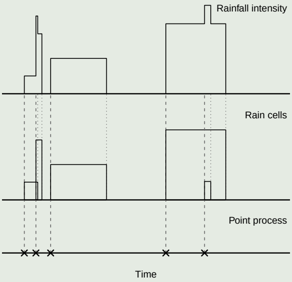
Inference using the Whittle likelihood

- Usual approach to inference / uncertainty of Whittle estimator requires fourth-order spectral density — limits practical application
- EF treatment with empirical covariance matrix estimation circumvents this:
  - First noted for zero-mean processes in Heyde, 1997, *Quasi-Likelihood and its applications*.
- Inclusion of zero-frequency term requires careful treatment (& many results from Priestley, Robinson etc.)
- EF treatment with previous assumptions also requires $0 < h(\omega;\theta) < \infty$; first & second θ-derivatives of $h(\omega;\theta)$ finite & continuous; first & second ω-derivatives of $h(\omega;\theta)$ finite.
  - Finite spectral density rules out long-memory processes for this treatment

Application to rainfall models

The Neyman-Scott rectangular pulses model

- 'Storm origins': homogeneous Poisson process, rate λ
- Each storm has random number of cells, $C \sim Poi(\mu_C)$
- Within storm, cell origins displaced from storm origin independently according to $Exp(\beta)$
- Cell durations: independent $Exp(\eta)$



- Cell intensities: independent with mean μ_X and variance σ_X^2
- This is model used in official UK climate projections

GMM for Neyman-Scott model

- Simulation study to assess performance
- Work with $\theta = (\log\lambda \ \log\mu_X \ \log(\sigma_X/\mu_X) \ \log\mu_C \ \log\beta \ \log\eta)'$ (more computationally stable than original parameterisation)
- Generate 1000 simulated datasets
  - Each represents 20 years' worth of hourly data for one calendar month (30 days) — typical of availability in applications
  - Parameters representative of UK winter rainfall
- GMM properties $\mathbf{T}_n$: mean; variance of 1-, 6- & 24-hour totals; ACF(1) for 1- & 24-hour totals; proportion of dry hours & days
  - Typical of hydrological practice
  - Calculated separately for each month — 20 replicates per simulation allows empirical covariance matrix estimation
  - Quenouille estimator used for ACF to ensure $E(\mathbf{T}_n) = \tau$

UCL

GMM simulation study continued

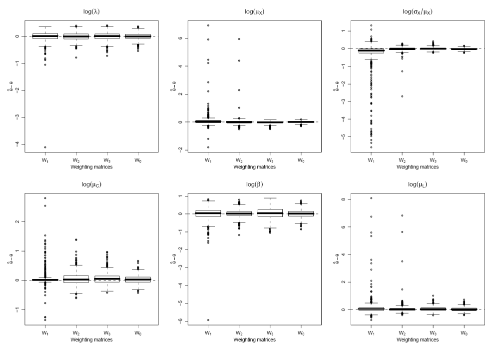
- Recap: GMM estimator minimises $[\mathbf{T}_n - \tau(\theta)]' \mathbf{W}_n [\mathbf{T}_n - \tau(\theta)]$.
- Different options for $\mathbf{W}_n$ compared:
  - $\mathbf{W}_1$: diagonal, equal weights
  - $\mathbf{W}_2$: diagonal, increased weight to 1-hour mean, variance & proportion dry (common hydrological practice)
  - $\mathbf{W}_3$: diagonal, inverses of variances of elements of $\mathbf{T}_n$, obtained by simulation from initial fit using inverses of empirical variances.
  - $\mathbf{W}_0$: inverse of covariance matrix of $\mathbf{T}_n$, obtained by simulation from initial fits used for $\mathbf{W}_3$.

  NB $\mathbf{W}_3$ & $\mathbf{W}_0$ yield two-step estimators
- Performance assessment:
  - Bias & variability of estimators
  - Coverages of quasi-Wald confidence intervals for each parameter
  - Coverages of confidence regions for θ based on values of GMM objective function

GMM simulations: bias & variability

Simulated distributions of estimation errors

- All weighting schemes deliver approx. unbiased estimators
- $\mathbf{W}_1$ & $\mathbf{W}_2$ prone to outliers
- Distributions ≈ normal for $\mathbf{W}_3$ & $\mathbf{W}_0$



- $\mathbf{W}_0$ most efficient as expected
- $\mathbf{W}_3$ close to $\mathbf{W}_0$ (& to first stage in two-step estimator)

GMM simulations: coverages

Level		$\log\lambda$	$\log\mu_X$	$\log\sigma_X/\mu_X$	$\log\mu_C$	$\log\beta$	$\log\eta$	θ
95%	$\mathbf{W}_1$	0.94	0.97	0.99	0.99	0.98	0.97	0.89
	$\mathbf{W}_2$	0.92	0.90	0.90	0.95	0.93	0.98	0.89
	$\mathbf{W}_3$	0.92	0.95	0.93	0.96	0.94	0.96	0.92
	$\mathbf{W}_0$	0.94	0.94	0.92	0.94	0.92	0.94	0.94
99%	$\mathbf{W}_1$	0.98	0.99	0.99	0.99	1.00	0.98	0.96
	$\mathbf{W}_2$	0.98	0.97	0.96	0.99	0.97	0.99	0.96
	$\mathbf{W}_3$	0.98	0.98	0.98	0.99	0.98	0.99	0.97
	$\mathbf{W}_0$	0.99	0.98	0.97	0.99	0.97	0.99	0.98

- Coverages reasonable for $\mathbf{W}_3$ & $\mathbf{W}_0$; less accurate for $\mathbf{W}_1$ & $\mathbf{W}_2$
- Slight undercoverage of all confidence regions for θ based on values of objective function

Whittle likelihood for Neyman-Scott model

- Similar simulation experiment carried out
- For this model, derivative matrix $\mathbf{G}_0 = \partial \mathbf{g} / \partial \theta$ ill-conditioned for Whittle EFs: simplify so that cell intensities $\sim Exp(1/\mu_X)$ & $\sigma_X = \mu_X$.
- Results indicate that estimators are approx. unbiased but asymptotic theory can overestimate sampling variability
  - Possibly due to use of empirical covariance matrix of Whittle EFs
  - But Wald-based confidence intervals have reasonable coverage
- Poor coverage of confidence regions for θ based on values of log-likelihood itself (e.g. 77% instead of 95%)
- Whittle estimates more variable than GMM ones for this model

Summary

- Estimating functions provide general framework for studying many inference methods
- Consistency, asymptotic distributions etc. verified using (fairly) easy-to-check conditions
- Empirical / two-step covariance matrix estimation is useful alternative to (e.g.) use of fourth-order properties in spectral estimation
- Optimal GMM estimation preferable to spectral likelihoods in inference for (challenging) stochastic rainfall models