Parametric inference for missing data problems

Clive Newstead[1]

Supervised by Giorgos Sermaidis[2] and Paul Fearnhead[2]

Setting

The scenario we consider is as follows:

- $X_1, \dots, X_n$ latent variables
- $Y_1, \dots, Y_n$ observations
- $(X_{1:n}, Y_{1:n}) \sim \pi(\,\cdot\,;\theta)$
- θ^* true parameter value
- $L_n(\theta)$ observed (marginal) likelihood

We would like to find maximum likelihood estimator

$$\theta_n \doteq \mathrm{argmax}_\theta L_n(\theta)$$

However, it is not possible to do this directly: the likelihood is not available in closed form because it is a high-dimensional integral

$$L_n(\theta) = \int \pi(X_{1:n}, Y_{1:n};\theta)\,\mathrm{d}X_{1:n}$$

We must therefore resort to estimation.

Simulated likelihood

A current method for overcoming this problem, known as simulated likelihood, can be used to estimate the likelihood and the MLE using Monte Carlo sampling techniques. The process is as follows:

1. Choose a proposal density g for $X_{1:n}$
2. Simulate $r(n)$ draws $X^1_{1:n}, \dots, X^r_{1:n}$ from g
3. Define the unbiased estimator

$$\widehat{L}^r_n(\theta) \doteq \frac{1}{r(n)} \sum_{i=1}^{r} \frac{\pi(X^i_{1:n}, Y_{1:n};\theta)}{g(X^i_{1:n};\theta)}$$

We define an estimator of the MLE by

$$\widehat{\theta}^r_n \doteq \mathrm{argmax}_\theta \widehat{L}^r_n(\theta)$$

Geyer (1994) showed that that

$$\widehat{\theta}^r_n \xrightarrow{p} \theta_n \text{ as } r \to \infty$$

We would also like

$$\widehat{\theta}^r_n \xrightarrow{p} \theta^* \text{ as } n \to \infty$$

However, this occurs only when $r(n)$ increases at least exponentially in n, so computation is infeasible for large data sets (Cappé et al. 2002).

Motivation and method

If we estimate the **log-likelihood** instead of the likelihood, we will achieve consistency even for constant r. In order to do this we first need to find a way to unbiasedly estimate logarithms of expectations. Then we will apply this to the simulated likelihood estimator.

Estimating a logarithm

If Z is a random variable and g is a function, it is not true in general that $g(\mathbb{E}[Z]) = \mathbb{E}[g(Z)]$ unless g is linear.

Consider the case of estimating $\log\mu$ where $\mu > 0$ is unknown. Fearnhead et al. (2008) made use of Taylor expansions to estimate an exponential. We follow a similar method for a logarithm:

$$\log\mu = \log c + \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j}\left(\frac{\mu - c}{c}\right)^j = \log c + \mathbb{E}\left[\frac{(-1)^{\kappa+1}}{\kappa p(\kappa)}\left(\frac{\mu-c}{c}\right)^{\kappa}\right]$$

where $c > 0$ is a user-defined constant and κ is a random variable distributed according to the probability mass function p with support $\mathbb{N}$.

Thus if $Z_1, \dots, Z_\kappa$ are iid random variables with mean $\mu > 0$ and variance $\sigma^2 < \infty$ then $\log\mu$ can be estimated unbiasedly by

$$\hat{\ell} \doteq \log c + \frac{(-1)^{\kappa+1}}{\kappa p(\kappa)} \prod_{i=1}^{\kappa} \frac{Z_i - c}{c}$$

What is the 'best' estimator?

In order to optimise our estimator $\hat{\ell}$ of $\log\mu$, we would like to minimise its variance. Using the method of Lagrange multipliers, it can be shown that the variance is minimised when $\kappa \sim \mathrm{Log}(q)$, the logarithmic (series) distribution with parameter

$$q = \frac{1}{c}\sqrt{\sigma^2 + (\mu - c)^2}$$

which is valid for all $c > \dfrac{\sigma^2 + \mu^2}{2\mu}$.

If we now minimise $\mathbb{V}[\hat{\ell}]$ by choice of c, we find that the optimal value of c is given by

$$c = \frac{\sigma^2 + \mu^2}{\mu}, \quad q = \frac{\sigma}{\sqrt{\sigma^2 + \mu^2}}$$

Potential drawbacks

This is promising since when $\sigma^2 = 0$ we get $\hat{\ell} = \log\mu$ with probability 1. It also minimises $\mathbb{E}[\kappa]$, the expected number of required simulations. However, in general we do not know the value of μ, which we can think of as being the true likelihood, or σ^2. We must therefore resort to making an educated guess for the value of c. This may be possible if we have an upper bound for μ and σ^2 is adequately small.

Estimating the log-likelihood

We consider an iid case where X_i are drawn from the probability density function $f_X(\,\cdot\,;\theta^*)$ and Y_i are drawn from the density $f_{Y|X}(\,\cdot\,;\theta^*)$. If g denotes a proposal density for X, then we can simulate r draws of $X_{1:n}$ from the g. We can then estimate the likelihood as follows.

- Estimate the density f_Y of Y_i by

$$\widehat{f}_Y(Y_i;\theta) \doteq \frac{1}{r}\sum_{j=1}^{r} \frac{f_{Y|X}(Y_i|X^j_i;\theta) f_X(X^j_i;\theta)}{g(X^j_i|Y_i;\theta)}$$

- Let $\widehat{L}^r_n(\theta) = \widehat{f}_Y(Y_1;\theta)\widehat{f}_Y(Y_2;\theta)\cdots\widehat{f}_Y(Y_n;\theta)$

Then $\widehat{L}^r_n(\theta)$ is an unbiased estimator of $L_n(\theta)$.

Next we simulate $\kappa \sim \mathrm{Log}(q)$ and let $Z_1, \dots, Z_\kappa$ be independent realisations of $\widehat{L}^r_n(\theta)$. Thus $\ell_n(\theta) \doteq \log L_n(\theta)$ is estimated unbiasedly by

$$\widehat{\ell}^r_n(\theta) \doteq \log c + \log(1-q)\prod_{1}^{\kappa}\frac{c - Z_i}{cq}$$

where c and q are chosen appropriately.

By the above theory, if $\widehat{\theta}^r_n \doteq \mathrm{argmax}_\theta \widehat{\ell}^r_n(\theta)$ then $\widehat{\theta}^r_n \xrightarrow{p} \theta^*$ as $n \to \infty$ for fixed r.

Example

Consider the case where

$$X_i \sim \mathcal{N}(\theta, \sigma^2_X) \text{ and } Y_i|X_i \sim X_i + \sigma_Y t_d$$

where the true value of the parameter θ is θ^*. The optimal choice of proposal distribution g would be the conditional distribution of $X_i|Y_i$ with parameter θ. This is not easy to compute, so we instead suppose that $Y_i|X_i$ is normally distributed with mean X_i and variance σ^2_Y. If this were the case then we would have

$$X_i|Y_i \sim \mathcal{N}\left(\theta + \frac{\sigma^2_X(Y_i - \theta)}{\sigma^2_X + \sigma^2_Y}, \frac{\sigma^2_X\sigma^2_Y}{\sigma^2_X + \sigma^2_Y}\right)$$

so we use this as our choice of proposal density g.

We will take:

- $\theta^* = 1$ (true parameter value)
- $r = 10$ (Monte Carlo sample size)
- $\sigma_X = \sigma_Y = 1$
- $d = 5$ (number of degrees of freedom in t-distribution)

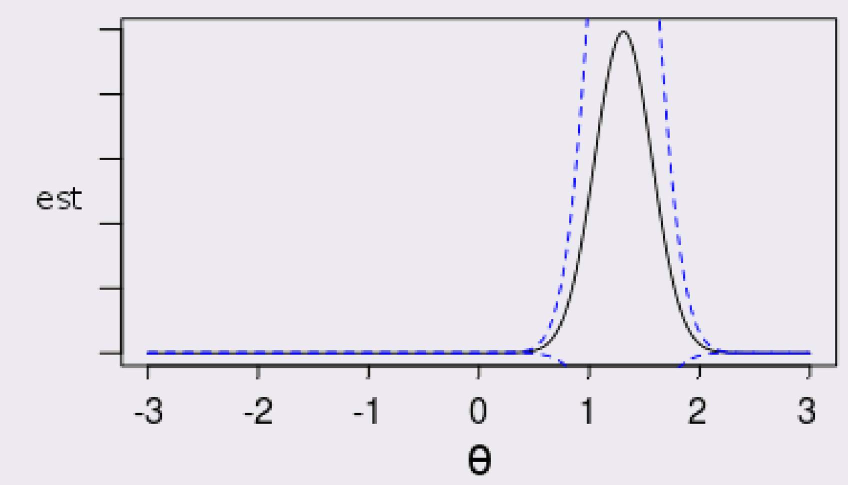
and look at the estimated parameter values $\hat{\theta}^{10}_n$ for $n = 1$, 10, 50 and 100.

Results

The estimated parameter values were given by

n	$\hat{\theta}^{10}_n$	Absolute error
1	−0.78	1.22
10	1.30	0.30
50	1.08	0.08
100	0.96	0.04

As expected, the parameter estimates $\hat{\theta}^{10}_n$ got closer to the true parameter value 1 as n grew. A sample plot (for $n = 10$) of the estimated likelihood for $\theta \in [-3, 3]$ is shown below.



The blue dashed lines are 95% confidence intervals.

Applications and future research

Theoretically the logarithmic estimator can be applied to any problem where the simulated likelihood approach is appropriate. This includes discretely observed diffusions, mixture problems and more general missing data problem.

Ideas for future research include applying the logarithmic estimator to a wider variety of situations, investigating an adaptive method for estimating $L_n(\theta)$, reducing sensitivity to the choice of proposal density g and finding better ways of choosing c and q given that we do not know the true likelihood or the exact value of the variance of the estimator.

References

- A. Beskos, O. Papaspilipoulos, G. O. Roberts, P. Fearnhead (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *Journal of the Royal Statistical Society, Series B: Statistical Methodology 68(3), 333-382.*
- O. Cappé, R. Douc, E. Moulines, C. Robert (2002). On the convergence of the Monte Carlo maximum likelihood method for latent variable models. *Scandinavian Journal of Statistics 29(4), 615-635*
- P. Fearnhead, O. Papaspiliopoulos, G. O. Roberts (2008). Particle filters for partially observed diffusions. *Journal of the Royal Statistical Society, Series B: Statistical Methology 70(4), 755-777.*
- G. Geyer (1994). On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society, Series B: Statistical methodology 56, 261-274.*

[1]University of Cambridge (Robinson College) [2]University of Lancaster (Department of Mathematics and Statistics)