Log-Density Estimation with Application to Approximate Likelihood Inference

Martin Hazelton[1]

Institute of Fundamental Sciences
Massey University

19 November 2015

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

[1] Email: m.hazelton@massey.ac.nz

Motivation for Studying Log-Density Estimation

- Probability density function is fundamental to huge swathe of statistical theory and methods.
- Often the density occurs naturally on the log-scale.
- Critical example is likelihood theory, where log-likelihood rather than likelihood plays predominant role.
- Other instances of use of log-density include as elements of information criteria, and log-relative risk function in spatial epidemiology.

General Approach

(log-density) estimation $\neq$ log-(density estimation)

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

General Approach

(log-density) estimation $\neq$ log-(density estimation)

- There are some methods that target log-density directly.
  - Maximum penalized likelihood method (Silverman, 1982)
  - Local likelihood density estimation (Loader,1996)
  - Spline-based methods (O'Sullivan, 1988)
- We prefer transformation of *appropriately modified* kernel density estimates.

Loader, C. (1996). *Ann. Statist.* **24**, 1602–1618.

O'Sullivan, F. (1988). *SIAM J. Sci. Stat. Comp.* **9**, 363–379.

Silverman, B. W. (1982). *Ann. Statist.*, **10**, 795–810.

Kernel Density Estimation

- $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ a d-dimensional random sample, common density function f.
- Kernel estimate of f defined by

$$\hat{f}_h(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{x}_i)$$

  - $K_h(\boldsymbol{x}) = h^{-d} K(\boldsymbol{x}/h)$ is scaled kernel
  - Unscaled kernel function K is radially symmetric density function with $\int K(\boldsymbol{u}) ||\boldsymbol{u}||^2 \, d\boldsymbol{u} = d$
  - h is the bandwidth.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Smoothing Regimens

- Will be interested in optimal smoothing for given estimation point $\boldsymbol{x}$.
- We employ isotropic smoothing.
  - Requires selection of only scalar $h = h(\boldsymbol{x})$.
  - Variables must either be on comparable scale to begin with, or are pre-scaled.
- More general smoothing regimens are possible; e.g. full bandwidth matrix (Wand & Jones, 1993).
  - Local selection of full bandwidth matrices seems challenging.

Wand, M. P. & Jones, M. C. (1993). *JASA* **88**, 520–528.

Naive Kernel Log-Density Estimation

- Define $\psi(\boldsymbol{x}) = \log(f(\boldsymbol{x}))$.
- Naive kernel estimator is $\log(\hat{f}_h(\boldsymbol{x}))$.
- Note that if K has compact support then $P(\hat{f}_h(\boldsymbol{x}) = 0) > 0$.
- Hence $\log(\hat{f}_h(\boldsymbol{x}))$ has no finite moments.
    - Introduces problems with (standard) theoretical analysis.
        - E.g. globally optimal (MISE minimizing) bandwidth does not exist.
    - In practice could always choose $h(\boldsymbol{x})$ sufficiently large conditional on data.
    - Nevertheless, unboundedness of log function at zero has major ramifications for selection of bandwidths.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Modified Kernel Log-Density Estimator

- Consider henceforth the modified log-density estimator $\hat{\psi}_h(\boldsymbol{x}) = \log(\hat{f}_h(\boldsymbol{x}) + e^{-n})$.

Theorem

Assume that

(A1) *K is a radially symmetric probability density function with $\int K(\boldsymbol{u})||\boldsymbol{u}||^2\,d\boldsymbol{u} = d$ and $\int K(\boldsymbol{u})||\boldsymbol{u}||^4\,d\boldsymbol{u} < \infty$.*

(A2) *All partial derivatives of f up to and including order 2 are continuous in a neighbourhood of $\boldsymbol{x}$.*

(A3) $0 < f(\boldsymbol{x})$.

(A4) $h = O(n^{-1/(d+4)})$.

Then $|\hat{\psi}_h(\boldsymbol{x})|^M$ is uniformly integrable for any positive integer M, and hence $\hat{\psi}_h(\boldsymbol{x})$ has finite moments of all orders for all $n > 0$.

What's the Fuss?

Asymptotically errors in $\hat{\psi}_h(\boldsymbol{x}) = \log(\hat{f}_h(\boldsymbol{x}) + e^{-n})$ look like relative errors in $\hat{f}_h(\boldsymbol{x})$.

$$\begin{aligned}\mathrm{MSE}\left(\hat{\psi}_h(\boldsymbol{x})\right) &= \mathrm{E}\left[\left(\hat{\psi}_h(\boldsymbol{x}) - \psi(\boldsymbol{x})\right)^2\right] \\ &= \mathrm{AMSE}\left(\hat{\psi}_h(\boldsymbol{x})\right) + o(h^4 + n^{-1}h^{-d})\end{aligned}$$

with

$$\begin{aligned}\mathrm{AMSE}\left(\hat{\psi}_h(\boldsymbol{x})\right) &= \frac{h^4}{4}\frac{(\nabla f(\boldsymbol{x}))^2}{f(\boldsymbol{x})^2} + \frac{R(K)}{nh^d f(\boldsymbol{x})} \\ &= \frac{1}{f(\boldsymbol{x})^2}\mathrm{AMSE}\left(\hat{f}_h(\boldsymbol{x})\right).\end{aligned}$$

where $R(g) = \int g(\boldsymbol{x})^2\, d\boldsymbol{x}$ for any square integrable function g.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

What's the Fuss

continued

- Asymptotically optimal bandwidth

$$h_{as} = \left(\frac{dR(K)f(\boldsymbol{x})}{(\nabla^2 f(\boldsymbol{x}))^2} \right)^{1/(d+4)} n^{-1/(d+4)}.$$

- In principle this applies for both estimation of $f(\boldsymbol{x})$ and $\psi(\boldsymbol{x})$.
- While asymptotics provide a (surprisingly) reliable guide to finite sample behaviour for estimation of f, not the case when estimating ψ.
  - Particularly when the estimation point $\boldsymbol{x}$ lies in an area of low density.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Asymptotics Versus Real Life

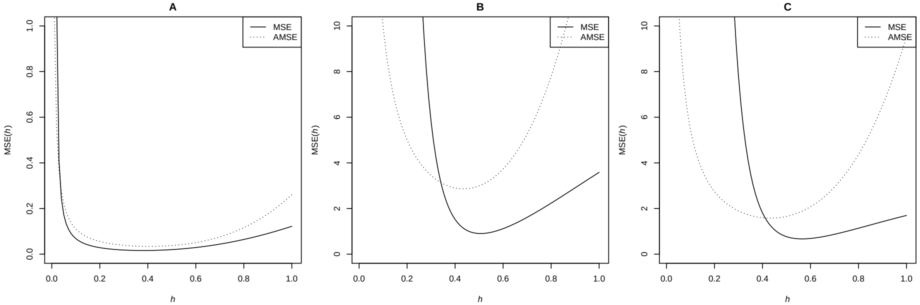


Figure: Comparison of exact and asymptotic versions of MSE($\hat{\psi}_h(\boldsymbol{x})$). For panels (A) and (B), target density is standard normal estimated at points $x = 0$ and $x = 3$ respectively. For (C) target density is bivariate standard normal, estimated at $\boldsymbol{x} = (2, 2)^{\mathsf{T}}$. In each case $n = 100$.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Approaches to Bandwidth Selection

- Majority of methods for bandwidth selection in density estimation target asymptotic form of bandwidth.
- That will not work well here.
- Bootstrap and smoothed cross-validation (SCV) methods target exact form of MSE.
- Such methods seem to work well for local density estimation, and for density estimation in ‘high’ (for kernel methods) dimensions.

Smoothed Cross-Validation Bandwidth Selection

For density estimation on the raw scale, SCV estimate of $\mathrm{MSE}(\hat{f}(\boldsymbol{x}))$ is

$$\begin{aligned}\mathrm{SCV}_f(h) &= \left(\mathrm{E}^\dagger[f_h^\dagger(\boldsymbol{x})] - \hat{f}_\lambda(\boldsymbol{x})\right)^2 + \frac{\hat{f}_\lambda(\boldsymbol{x})R(K)}{nh^d} \\ &= \left(\hat{f}_\lambda * K_h(\boldsymbol{x}) - \hat{f}_\lambda(\boldsymbol{x})\right)^2 + \frac{\hat{f}_\lambda(\boldsymbol{x})R(K)}{nh^d}.\end{aligned}$$

- $f_h^\dagger(\boldsymbol{x})$ denotes kernel density estimate constructed using random sample of size n drawn from pilot density $\hat{f}_\lambda$.
- $\mathrm{E}^\dagger$ indicates expectation with respect to $f_h^\dagger(\boldsymbol{x})$, conditional on the original data.
    - Note that expectation evaluated analytically.
- Symbol $*$ denotes a convolution.
- SCV bandwidth selector is minimizer of $\mathrm{SCV}_f(h)$.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Application of SCV to Log-Density Estimation

Direct adaptation of SCV to the estimator $\hat{\psi}_h(\boldsymbol{x})$ gives

$$\begin{aligned}\mathrm{SCV}(h) &= \left(\mathrm{E}^\dagger[\hat{\psi}_h^\dagger(\boldsymbol{x})] - \log(\hat{\psi}_\lambda(\boldsymbol{x}))\right)^2 + \frac{R(K)}{\hat{f}_\lambda(\boldsymbol{x})nh^d} \\ &= \left(\mathrm{E}^\dagger[\log((f_h^\dagger(\boldsymbol{x}) + e^{-n})/\hat{f}_\lambda(\boldsymbol{x}))]\right)^2 + \frac{R(K)}{\hat{f}_\lambda(\boldsymbol{x})nh^d}.\end{aligned}$$

where $\hat{\psi}_h^\dagger(\boldsymbol{x}) = \log(f_h^\dagger(\boldsymbol{x}) + e^{-n})$.

Application of SCV to Log-Density Estimation

Direct adaptation of SCV to the estimator $\hat{\psi}_h(\boldsymbol{x})$ gives

$$\begin{aligned}\mathrm{SCV}(h) &= \left(\mathrm{E}^\dagger[\hat{\psi}_h^\dagger(\boldsymbol{x})] - \log(\hat{\psi}_\lambda(\boldsymbol{x}))\right)^2 + \frac{R(K)}{\hat{f}_\lambda(\boldsymbol{x})nh^d} \\ &= \left(\mathrm{E}^\dagger[\log((f_h^\dagger(\boldsymbol{x}) + e^{-n})/\hat{f}_\lambda(\boldsymbol{x}))]\right)^2 + \frac{R(K)}{\hat{f}_\lambda(\boldsymbol{x})nh^d}.\end{aligned}$$

where $\hat{\psi}_h^\dagger(\boldsymbol{x}) = \log(f_h^\dagger(\boldsymbol{x}) + e^{-n})$.

Problem: the squared bias term cannot be evaluated in closed form in this case.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Approximate Smoothed Cross-Validation (ASCV)

We propose approximate smoothed cross-validation (ASCV) criterion,

$$\text{ASCV}(h) = \left(\log((\hat{f}_\lambda * K_h(\boldsymbol{x}) + e^{-n})/\hat{f}_\lambda(\boldsymbol{x}))\right)^2 + \frac{R(K)}{\hat{f}_\lambda(\boldsymbol{x})nh^d}.$$

- Obtained by making school boy mistake of interchanging order of expectation and transformation.
- Result is ASCV(h) only approximates SCV(h)...
- ... but straightforward to show that the approximation error is order $O_p(n^{-1}h^{2-d})$ which is asymptotically negligible.
- Idea is that ASCV(h) will capture important characteristics of MSE(h).
  - In particular will reflect large errors associated with overly small bandwidths.
- ASCV bandwidth selector, $\hat{h}$, is minimizer of ASCV(h).

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Tuning SCV

Properties of $\hat{h}$ depend on choice of λ, the bandwidth used to construct the pilot (bootstrap) density $\hat{f}_\lambda$.

Theorem

Under suitable regularity conditions,

$$\mathrm{E}\left[\left(\frac{\hat{h}-h_{as}}{h_{as}}\right)^2\right] = \frac{1}{(d+4)^2}\left[\lambda^4\left(\frac{\Theta(\boldsymbol{x})}{\nabla^2 f(\boldsymbol{x})} - \frac{\nabla^2 f(\boldsymbol{x})}{2f(\boldsymbol{x})}\right)^2 + 4\frac{f(\boldsymbol{x})}{(\nabla^2 f(\boldsymbol{x}))^2}\frac{R(\nabla^2 K)}{n\lambda^{d+4}}\right] + o(\lambda^4 n^{-1}\lambda^{-d-4})$$

where

$$\Theta(\boldsymbol{x}) = \sum_{i=1}^{d}\frac{\partial^4}{\partial x_i^4}f(\boldsymbol{x}) + \sum_{i=1}^{d}\sum_{i\neq j\ j=1}^{d}\frac{\partial^2}{\partial x_i^2}\frac{\partial^2}{\partial x_j^2}f(\boldsymbol{x}).$$

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Tuning SCV (continued)

Corollary

Under the assumption that $2\Theta(\boldsymbol{x})f(\boldsymbol{x}) \neq (\nabla^2 f(\boldsymbol{x})$*, the value of the pilot bandwidth to minimize* $\mathrm{E}[((\hat{h} - h_{as})/h_{as})^2]$ *is*

$$\lambda_0 = \left[\frac{4(d+4)f(\boldsymbol{x})^3 R(\nabla^2 K)}{(2\Theta(\boldsymbol{x})f(\boldsymbol{x}) - (\nabla^2 f(\boldsymbol{x}))^2)^2}\right]^{1/(d+8)} n^{-1/(d+8)}.$$

In practice replace functionals by pilot estimates thereof.

Tuning SCV (continued)

Corollary

Under the assumption that $2\Theta(\boldsymbol{x})f(\boldsymbol{x}) \neq (\nabla^2 f(\boldsymbol{x})$*, the value of the pilot bandwidth to minimize* $\mathrm{E}[((\hat{h} - h_{as})/h_{as})^2]$ *is*

$$\lambda_0 = \left[\frac{4(d+4)f(\boldsymbol{x})^3 R(\nabla^2 K)}{(2\Theta(\boldsymbol{x})f(\boldsymbol{x}) - (\nabla^2 f(\boldsymbol{x}))^2)^2} \right]^{1/(d+8)} n^{-1/(d+8)}.$$

In practice replace functionals by pilot estimates thereof.

Corollary

Using a pilot bandwidth $\lambda \propto n^{-1/(d+8)}$ *gives the optimal rate of convergence for the bandwidth selector:*
$(\hat{h} - h_{as})/h_{as}) = O_p(n^{-2/(d+8)})$.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Numerical Results

Standard normal target density

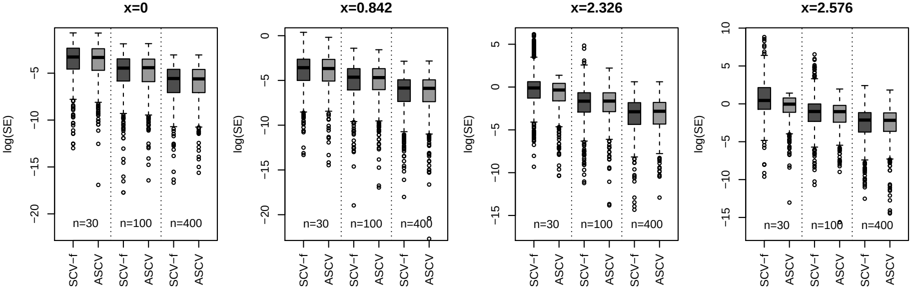


Figure: Log-squared error for log-density estimates using smooth cross-validation for density estimation (SCV-f) and approximate smooth cross-validation for log-density estimation (ASCV).

Numerical Results

Standard exponential target density

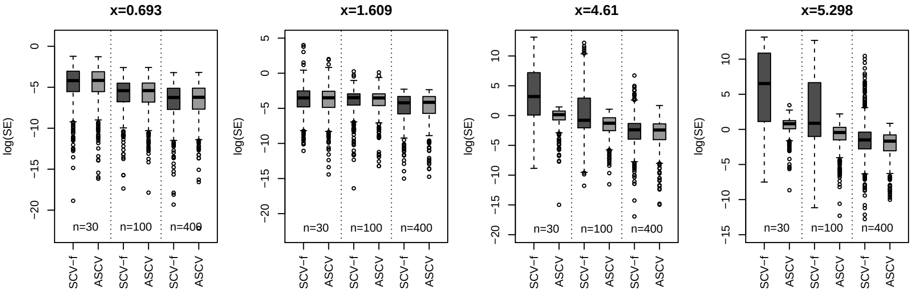


Figure: Log-squared error for log-density estimates using smooth cross-validation for density estimation (SCV-f) and approximate smooth cross-validation for log-density estimation (ASCV).

Numerical Results

Standard t target density

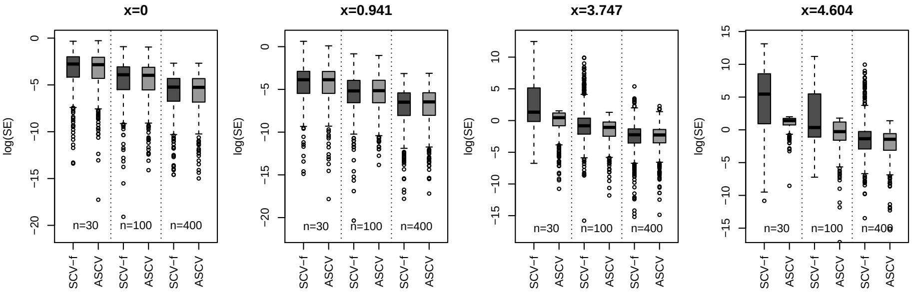


Figure: Log-squared error for log-density estimates using smooth cross-validation for density estimation (SCV-f) and approximate smooth cross-validation for log-density estimation (ASCV).

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Numerical Results

Uncorrelated bivariate normal target density

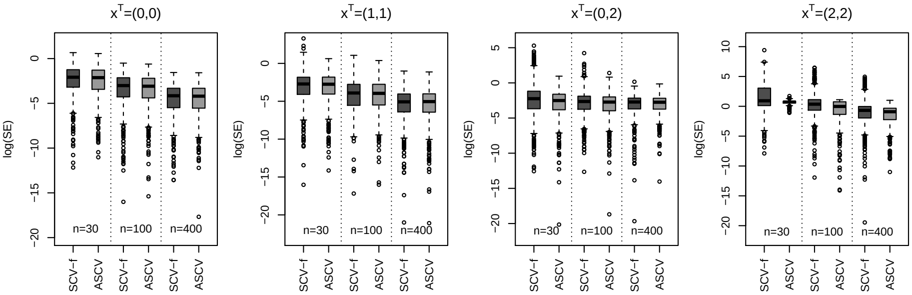


Figure: Log-squared error for log-density estimates using smooth cross-validation for density estimation (SCV-f) and approximate smooth cross-validation for log-density estimation (ASCV).

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Numerical Results

Correlated bivariate normal target density

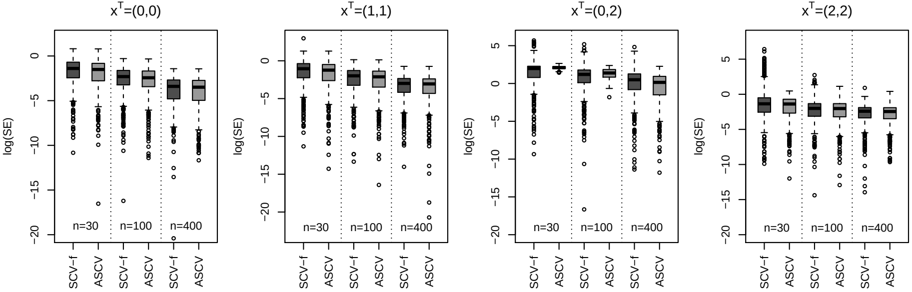


Figure: Log-squared error for log-density estimates using smooth cross-validation for density estimation (SCV-f) and approximate smooth cross-validation for log-density estimation (ASCV).

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Numerical Results

Bivariate t target density

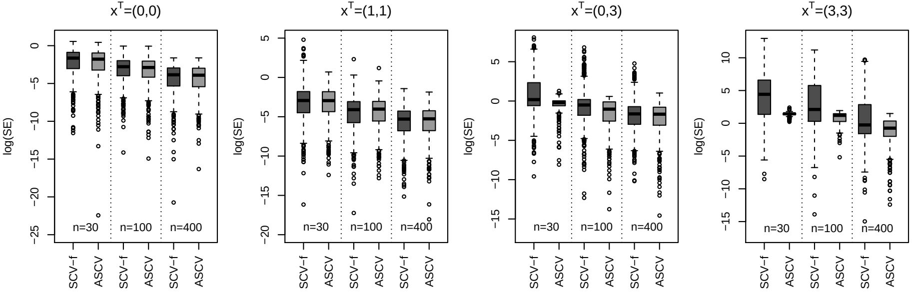


Figure: Log-squared error for log-density estimates using smooth cross-validation for density estimation (SCV-f) and approximate smooth cross-validation for log-density estimation (ASCV).

Introduction to Approximate Likelihood Inference (ALI)

- Consider statistical model dependent upon parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\mathsf{T}$.
- Observe d-variate random sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$.
  - In the spirit of approximate Bayesian computation, $\boldsymbol{y}_i$ may be summary statistics from ith observation.
- Wish to conduct inference about $\boldsymbol{\theta}$.
- Suppose likelihood function is intractable...
- ... but we can simulate realizations of data for any given $\boldsymbol{\theta}$.
  - Denote by $\boldsymbol{x}_1^{\boldsymbol{\theta}}, \ldots, \boldsymbol{x}_n^{\boldsymbol{\theta}}$ a set of n independent simulated realizations.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Approximate Likelihood Inference Methodology

- Can use simulations $\boldsymbol{x}_1^{\boldsymbol{\theta}}, \ldots, \boldsymbol{x}_n^{\boldsymbol{\theta}}$ to construct estimate $\log(\hat{f}_h(\cdot|\boldsymbol{\theta}))$ of log-density at $\boldsymbol{\theta}$.
- Then compute estimate of log-likelihood:

$$\hat{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \hat{\psi}_{h_i}(\boldsymbol{y}_i|\boldsymbol{\theta}).$$

- Entire log-likelihood function (near the maximum) obtained by applying smoother to estimates of $\ell(\boldsymbol{\theta})$ over parameter grid.
- Calculation of (approximate) maximum likelihood estimate proceed by maximizing the fitted smoother.

Approximate Likelihood Inference Methodology

Practical Implementation

- Idea is surprisingly old (Diggle & Gratton, 1984) but arguably received less attention than is due.
  - Everyone wants to do Approximate Bayesian Computation.
  - That also involves smoothing process (needs work).
- Methodology depends critically on log-density estimation.
- Previous implementations used ad hoc bandwidth selection.
- We will implement using ASCV bandwidths.

Diggle, P. J. & Gratton, R. J. (1984). *JRSSB* **46**, 193–227.

Example

Inference for a Model of Migration in Sumba Using Genetic Data

- Genome-wide data collected from groups of individuals in several villages on island of Sumba in eastern Indonesia.
- Five language clusters on Sumbda.
- Focus here is two pairs of villages and the rates of migration between them.
  - First pair is Mamboro and Wanokaka, from same language cluster.
  - Second pair is Loli and Kodi, from two different language clusters.
- Apart from language, two village pairs are very comparable.
- Interest is in long-term rates of within-pair migration taking into account the geographical distances between the villages.
- Are migration rates affected by language differences?
- Joint work with Murray Cox (Massey University).

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Example

continued

- We have a complex stochastic model (based on a structured coalescent) to describe genetic profiles of communities.
- In model, migration rate for pair i is described by parameter θ_i.
  - Represents proportion of genes that move between the two populations in one generation.
- Observed genetic data for community pair i condensed to a single summary statistic, y_i, the fixation index F_{ST}.
  - This measures genetic variation within each village as proportion relative to total genetic variation.
  - Can be expected to be informative about the parameter θ_i.
  - Observed data is $\boldsymbol{y} = (y_1, y_2)^{\mathsf{T}} = (0.01412, 0.01259)^{\mathsf{T}}$.

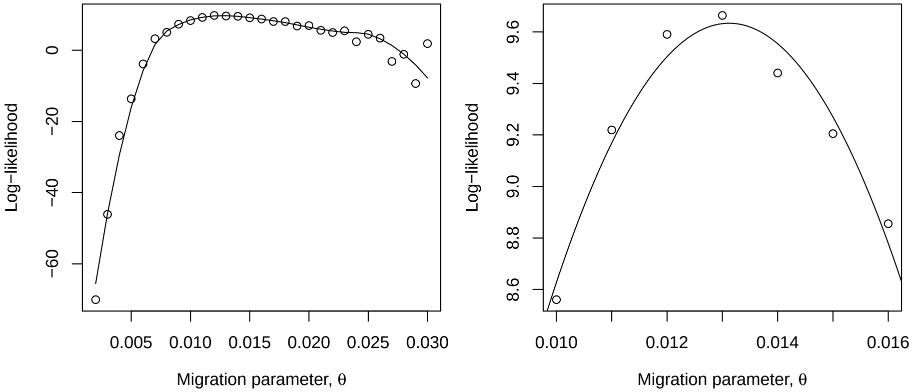
Example

continued

- Model log-likelihood $\ell(\boldsymbol{\theta}) = \log(f(\boldsymbol{y}|\boldsymbol{\theta}))$ is intractable.
- Can produce simulations $(x_1^{\boldsymbol{\theta}}, x_2^{\boldsymbol{\theta}})^{\mathsf{T}}$ of fixation index for any $\boldsymbol{\theta}$.
- Consider two variants of model:
  - Small model: $\theta_1 = \theta_2$, so migration rate not affected by language differences.
  - Large model: $\theta_1 \neq \theta_2$.
- Estimate log-likelihood for both models.
  - Get approximate MLEs for both models.
  - Do approximate likelihood ratio test to compare models.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

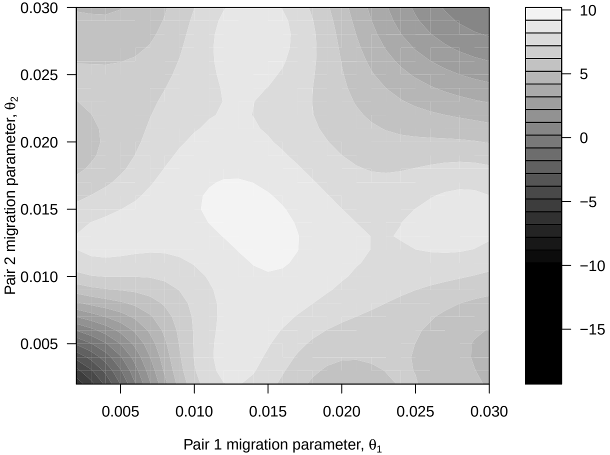
Approximate Log-Likelihood for Small Model



Approximate MLE: $\hat{\theta} = 0.0131$

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Approximate Log-Likelihood for Large Model



Approximate MLE: $(\hat{\theta}_1, \hat{\theta}_2)^\mathsf{T} = (0.0129, 0.0132)^\mathsf{T}$

Approximate Log-Likelihood Ratio Test

- Test hypothesis $H_0 : \theta_1 = \theta_2$ using approximate likelihood ratio test.
- Implemented using $n = 5000$ simulated realizations for both models evaluated at their respective maximum likelihood estimates.
- Approximate log-likelihood ratio test statistic was $\hat{D} = -0.135$.
- Estimated Monte Carlo standard error of $\hat{\sigma}(\hat{D}) = 0.130$.
    - So impossible negativity of $\hat{D}$ explicable by simulation induced noise.
- Conclusion is that language differences have no affect on long term migration rates.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

For a Copy of these Slides...

www.massey.ac.nz/~mhazelto/talks/wwpms2015.pdf