Profile likelihood, smoothed likelihood, and Monte Carlo likelihood

Edward Ionides
University of Michigan, Department of Statistics

Lecture 5 at Wharton Statistics Department
Wednesday 10th May, 2017

Slides are online at
`http://dept.stat.lsa.umich.edu/~ionides/talks/upenn`

Overview

Our agenda is to collect together some approaches to likelihood based inference:

- Profile likelihood.
- Fisher information (observed and expected).
- Local asymptotic normality.
- Smoothed likelihood.
- Monte Carlo likelihood.

We then see how these ideas combine to facilitate inference for complex dynamic systems.

Profile likelihood: some definitions

- The log likelihood function for model $f_Y(y\,;\theta)$ and data y^* is $\lambda(\theta\,;y^*) = \log f_Y(y^*\,;\theta)$,
- A maximum likelihood estimate (MLE) is $\hat{\theta}^* = \hat{\theta}(y^*) = \arg\max_\theta \lambda(\theta\,;y^*)$.
- We suppose $\theta = (\phi, \psi)$ with $\phi \in \mathbb{R}^1$ and $\psi \in \mathbb{R}^{p-1}$. Here, ϕ is a focal parameter for which we are interested in obtaining a confidence interval.
- The **profile log likelihood function** for ϕ is defined as $\lambda^P(\phi\,;y^*) = \max_\psi \lambda\big((\phi,\psi)\,;y^*\big)$.
- The profile log likelihood is maximized at a marginal MLE, $\hat{\phi}^* = \hat{\phi}(y^*) = \arg\max_\phi \lambda^P(\phi\,;y^*)$.
- A profile likelihood confidence interval with cutoff δ is defined as $\big\{\phi : \lambda^P(\phi\,;y^*) > \lambda^P\big(\hat{\phi}^*\,;y^*\big) - \delta\big\}$.

Profile likelihood: some history

- Profile likelihood confidence intervals are equivalent to likelihood ratio tests, which have a long history.
- Box and Cox (1964) graphed the profile likelihood under the name of maximized likelihood, and constructed confidence intervals using the χ^2 cutoff.
- Cox and Snell (1970) made an early use of the name "profile likelihood." Use of "maximized likelihood" continued through 1970's but is now antiquated.
- Much work in the 1980's focused on how to modify profile likelihood for improved higher-order asymptotic behavior (Barndorff-Nielsen, 1983).
- Profile likelihood has uses in semiparametric inference (Murphy and van der Vaart, 2000). The proportional hazard "partial likelihood" (Cox, 1972) is a semiparametric profile likelihood.

Fisher information and observed Fisher information

- Fisher information, evaluated at the MLE, is
  $I_{ij} = \mathbb{E}\left[-\frac{\partial}{\partial\theta_i \partial\theta_j} \lambda(\hat{\theta}^* \,; Y) \right]$
- Observed Fisher information is
  $I^*_{ij} = -\frac{\partial}{\partial\theta_i \partial\theta_j} \lambda(\hat{\theta}^* \,; y^*)$
- The asterisk denoting observed Fisher information indicates the additional data dependence.
- Corresponding standard errors and 95% confidence intervals for $\phi = \theta_i$ are
  $$\begin{array}{rclcrcl} \mathrm{SE}_F &=& \sqrt{[I^{-1}]_{ii}} & \quad & \mathrm{CI}_F &=& \left[\hat{\theta}^* - 1.96\,\mathrm{SE}_F, \hat{\theta}^* + 1.96\,\mathrm{SE}_F\right] \\ \mathrm{SE}^*_F &=& \sqrt{[I^{*\,-1}]_{ii}} & & \mathrm{CI}^*_F &=& \left[\hat{\theta}^* - 1.96\,\mathrm{SE}^*_F, \hat{\theta}^* + 1.96\,\mathrm{SE}^*_F\right] \end{array}$$
- An identity: $\frac{d^2}{d\phi^2} \lambda_P(\phi \,; y^*) = -\left[{I^*}^{-1}_{ii} \right]^{-1}$ [Exercise].
- Observed Fisher information together with the MLE fully escribes a quadratic log likelihood function.
- For a quadratic likelihood function, CI^*_F is equal to the profile likelihood confidence interval [Exercise],
  $\mathrm{CI}_P = \left\{ \phi : \lambda_P(\phi \,; y^*) > \lambda_P(\hat{\phi}^* \,; y^*) - 1.92 \right\}$.

In favor of observed Fisher information and profile likelihood

- Heuristically, the error on an estimator depends on the amount of information observed in the actual experiment.
- Efron and Hinkley (1978) argued for observed Fisher information, SE_F^*, over SE_F. Formal reasoning was limited to special cases, with arguments based on ancillarity.
- Lindsay and Li (1997) used a risk framework to show SE_F^* generally gives asymptotically optimal estimation of $(\hat{\phi} - \phi)^2$.
- CI_P transforms naturally if ϕ is reparameterized by $h(\phi)$. If there is an unknown h such that the log likelihood is [approximately] quadratic for any y^*, CI_P [approximately] corresponds to CI_F^* computed on this scale. Thus, heuristically, we may expect CI_P to have comparable asymptotic optimality to CI_F^* but better finite sample behavior.

Local asymptotic normality

- LAN (Le Cam, 1986) concerns a sequence of statistical models, $f_{Y,n}(y_n\,;\theta)$, and the behavior of the log likelihood ratio, $\Lambda_n(\theta) = \log f_{Y_n}(Y_n\,;\theta) - \log f_{Y_n}(Y_n\,;\theta_0)$ when $Y_n \sim F_{Y,n}(y_n\,;\theta_0)$.
- $f_{Y,n}(y_n\,;\theta)$ has LAN with information matrix K if there is a sequence of random variables $\Delta_n \xrightarrow{d} N(0,K)$ such that, for all bounded $\{t_n\}$,

$$\Lambda_n\big(\theta_0 + t_n n^{-1/2}\big) = \; t_n^T \Delta_n - \tfrac{1}{2} t_n^T K t_n + o_p(1;\theta_0).$$

- K is the asymptotic information rate concerning θ; it coincides with the Fisher information under regularity conditions. From Hájek's convolution theorem, an estimator $\hat{\theta}_n$ is asymptotically efficient if LAN holds and $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0,K^{-1})$
- Bickel et al. (1993) discuss LAN and demonstrate the utility of the LAN framework for semiparametrics

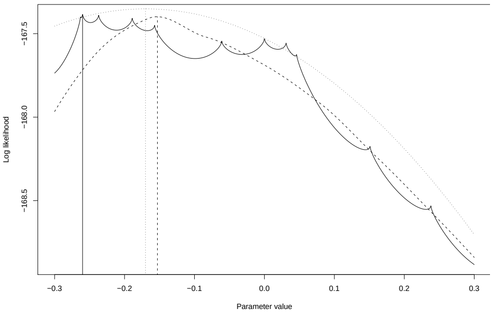
Maximum quadratic likelihood estimator (MQLE)

- LAN justifies the following estimation procedure:
  1. Evaluate the log likelihood at a grid of points in the neighborhood of a $\sqrt{n}$-consistent estimator.
  2. Fit a quadratic through these points.
  3. Obtain the maximum of this quadratic.
- This is called Le Cam's one-step estimator. We will say **maximum quadratic likelihood estimator (MQLE)**.
- MQLE is efficient under LAN and more generally when the likelihood is locally asymptotic quadratic (Le Cam, 1986).
- Under regularity, LAN is equivalent to asymptotic normality of MLE.
- MQLE can succeed when LAN holds but the MLE behaves badly.
- LAN can be easier to prove than asymptotic normality of the MLE.
- The one-step estimator is, in some sense, better than the MLE.

Maximum quadratic likelihood estimator (MQLE)

- LAN justifies the following estimation procedure:
  1. Evaluate the log likelihood at a grid of points in the neighborhood of a $\sqrt{n}$-consistent estimator.
  2. Fit a quadratic through these points.
  3. Obtain the maximum of this quadratic.
- This is called Le Cam's one-step estimator. We will say **maximum quadratic likelihood estimator (MQLE)**.
- MQLE is efficient under LAN and more generally when the likelihood is locally asymptotic quadratic (Le Cam, 1986).
- Under regularity, LAN is equivalent to asymptotic normality of MLE.
- MQLE can succeed when LAN holds but the MLE behaves badly.
- LAN can be easier to prove than asymptotic normality of the MLE.
- The one-step estimator is, in some sense, better than the MLE.
- **But, who would use LAN for data analysis when the log likelihood is grossly non-quadratic**

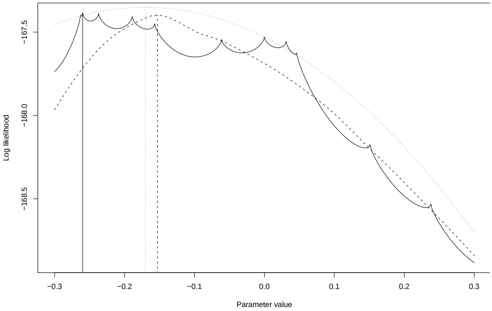
A non-smooth likelihood: $f_Y(y|\theta) \propto \exp(-|y-\theta|^\alpha)$



- Dotted: MQLE, initialized at the median.
- Dashed: maximum smoothed likelihood estimator (MSLE).
- Solid: likelihood and MLE.

- This model does not satisfy the usual Cramér conditions for the MLE. MQLE and MSLE are 15% more efficient than MLE (Ionides, 2005).
- Perhaps more importantly, they are not worse!

A non-smooth likelihood: $f_Y(y|\theta) \propto \exp(-|y-\theta|^\alpha)$



- Dotted: MQLE, initialized at the median.
- Dashed: maximum smoothed likelihood estimator (MSLE).
- Solid: likelihood and MLE.

- This model does not satisfy the usual Cramér conditions for the MLE. MQLE and MSLE are 15% more efficient than MLE (Ionides, 2005).
- Perhaps more importantly, they are not worse!
- **In higher dimensions, or when we must rely on Monte Carlo approximation of the likelihood, MQLE and MSLE may be easier to implement.**

More on maximum smoothed likelihood estimation (MSLE)

- MSLE (Ionides, 2005) involves the following steps:

  1. Evaluate the log likelihood at a grid of points in the neighborhood of a $\sqrt{n}$-consistent estimator.
  2. Fit a smooth curve through these points.
  3. Obtain the maximum of this smooth curve.

- MSLE replaces the quadratic of MQLE with a smoother.
- The smoothed likelihood can be used to construct profile confidence intervals.
- As long as the smoother fits a quadratic through points on a quadratic, MSLE inherits asymptotic optimality from MQLE.
- The `loess` smoother in R is a 2nd order local polynomial smoother with this property.

Monte Carlo profile confidence intervals for dynamic systems

- Monte Carlo methods to evaluate and maximize the likelihood function enable the construction of confidence intervals and hypothesis tests, facilitating scientific investigation using models for which the likelihood function is intractable.
- When Monte Carlo error can be made small, by sufficiently exhaustive computation, then the standard theory and practice of likelihood-based inference applies. One may still want to use MSLE to enable reliable inference at reduced computational cost.
- As datasets become larger, and models more complex, situations arise where no reasonable amount of computation can render Monte Carlo error negligible.
- We seek profile likelihood methodology enabling frequentist inferences accounting for Monte Carlo error.
- This methodology facilitates inference for computationally challenging dynamic latent variable models.

A metamodel for a Monte Carlo profile

- A **Monte Carlo metamodel** is a statistical model fitted to output of a Monte Carlo algorithm.
- We have independent Monte Carlo profile likelihood evaluations $\big(\breve{\lambda}^P_k(y^*), k \in 1\!:\!K\big)$ at points $\phi_{1:K} = (\phi_1, \dots, \phi_K)$.
- Without loss of generality we can write

  $$[M1] \qquad \breve{\lambda}^P_k(y^*) = \lambda^P(\phi_k\,; y^*) + \beta_k(y^*) + \epsilon_k(y^*), \quad k \in 1\!:\!K,$$

  where Monte Carlo errors $\epsilon_{1:K}(Y)$ are, **by construction, mean zero and independent** conditional on Y. In $[M1]$, $\beta_k(y^*)$ is Monte Carlo bias.
- Local to the MLE, we may make additional metamodel assumptions:
  $[M2] \qquad \beta_k(y^*) = \beta(y^*)$: constant bias.
  $[M3] \qquad \mathrm{Var}\big[\epsilon_k(y^*)\big] = \sigma^2(y^*) < \infty$: constant variance.
- We can complete the metamodel by proposing parametric or nonparametric specifications of $\lambda^P(\phi\,; y^*)$.

A quadratic metamodel for the profile likelihood

- LAN suggests a quadratic metamodel, $\breve{\lambda}^P_k(y) = -\hat{a}(y)\phi_k^2 + \hat{b}(y)\phi_k + \hat{c}(y) + \epsilon_k, \quad \mathrm{Var}(\epsilon_k) = \sigma^2(y).$
- The unknown coefficients $\hat{a}^* = \hat{a}(y^*)$, $\hat{b}^* = \hat{b}(y^*)$ and $\hat{c}^* = \hat{c}(y^*)$ make a quadratic approximation to the intractable likelihood.
- We fit the metamodel to the Monte Carlo profile evaluations, using linear regression to estimate $(\hat{a}^*, \hat{b}^*, \hat{c}^*)$ by $(\breve{a}^*, \breve{b}^*, \breve{c}^*)$.
- The marginal MLE $\hat{\phi}^*$ can be approximated by the maximum of $\breve{\lambda}^Q(\phi\,; y^*)$, which is given by $\breve{\phi}^Q(y^*, \epsilon) = \breve{b}(y^*, \epsilon)/2\breve{a}(y^*, \epsilon)$
- We can separate the variability of $\breve{\phi}^Q(Y, \epsilon)$ into two components:
  1. **Statistical error** is uncertainty from randomness in the data, viewed as a draw from the statistical model. This is the usual statistical error of $\hat{b}(y^*)/2\hat{a}(y^*)$ as an estimate of ϕ_0.
  2. **Monte Carlo** error is the uncertainty from implementing a Monte Carlo estimator. This is the error in $\breve{b}(y^*, \epsilon)/2\breve{a}(y^*, \epsilon)$ as a Monte Carlo estimate of $\hat{b}(y^*)/2\hat{a}(y^*)$.

Monte Carlo error and statistical error

- Routine application of the delta method gives a central limit approximation for the Monte Carlo error on the maximum, conditional on $Y = y^*$,

$$\frac{\breve{b}^*}{2\breve{a}^*} \approx N\left[\left(\frac{\hat{b}^*}{2\hat{a}^*}\right), \mathrm{SE}^2_{\mathrm{mc}}\right],$$

  where

$$\mathrm{SE}^2_{\mathrm{mc}} = \frac{1}{4\breve{a}^{*2}}\left\{\mathrm{Var}\left[\breve{b}^*\right] - \frac{2\breve{b}^*}{\breve{a}^*}\mathrm{Cov}\left[\breve{a}^*, \breve{b}^*\right] + \frac{\breve{b}^{*2}}{\breve{a}^{*2}}\mathrm{Var}\left[\breve{a}^*\right]\right\}.$$

- The usual statistical standard error, $1/\sqrt{2\hat{a}^*}$, is not available to us. Its Monte Carlo estimate is

$$\mathrm{SE}_{\mathrm{stat}} = \frac{1}{\sqrt{2\breve{a}^*}}.$$

- Under suitable regularity, these two error sources are additive, and so

$$\mathrm{SE}_{\mathrm{total}} = \sqrt{\mathrm{SE}^2_{\mathrm{mc}} + \mathrm{SE}^2_{\mathrm{stat}}}.$$

Using $\mathrm{SE}_{\mathrm{total}}$ for a Monte Carlo adjusted profile (MCAP)

- The usual χ^2 cutoff for profile confidence intervals is based on quadratic asymptotics. It is robust to reparameterization, and can be applied to either the actual profile or a smoothed version.
- Exactly the same argument can be applied to give a cutoff for a smoothed Monte Carlo profile based on a quadratic approximation:

$$\delta = \breve{a}^* \times \left(z_\alpha \times \mathrm{SE}_{\mathrm{total}}\right)^2 = z_\alpha^2 \left(\breve{a}^* \times \mathrm{SE}_{\mathrm{mc}}^2 + \frac{1}{2}\right),$$

  where z_α is the $1 - \alpha_2$ normal quantile.
- if $\mathrm{SE}_{\mathrm{mc}} = 0$, the cutoff for $\alpha = 0.05$ reduces to $\delta = 1.96^2/2 = 1.92$.
- We apply this cutoff after estimating the profile via a locally weighted quadratic smoother. $\mathrm{SE}_{\mathrm{mc}}$ can be computed using the local weights at the maximum.

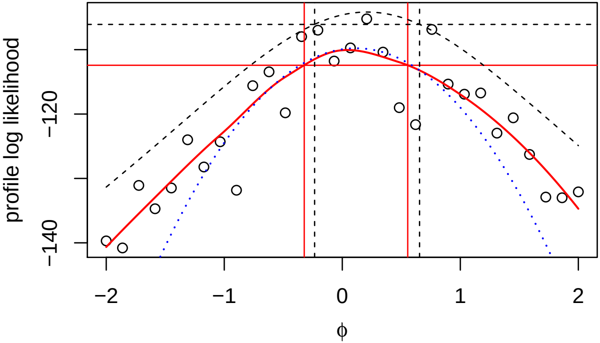
Using $\mathrm{SE}_{\mathrm{total}}$ for a Monte Carlo adjusted profile (MCAP)

- The usual χ^2 cutoff for profile confidence intervals is based on quadratic asymptotics. It is robust to reparameterization, and can be applied to either the actual profile or a smoothed version.
- Exactly the same argument can be applied to give a cutoff for a smoothed Monte Carlo profile based on a quadratic approximation:

$$\delta = \breve{a}^* \times \big(z_\alpha \times \mathrm{SE}_{\mathrm{total}}\big)^2 = z_\alpha^2 \left(\breve{a}^* \times \mathrm{SE}^2_{\mathrm{mc}} + \frac{1}{2} \right),$$

  where z_α is the $1-\alpha_2$ normal quantile.
- if $\mathrm{SE}_{\mathrm{mc}} = 0$, the cutoff for $\alpha = 0.05$ reduces to $\delta = 1.96^2/2 = 1.92$.
- We apply this cutoff after estimating the profile via a locally weighted quadratic smoother. $\mathrm{SE}_{\mathrm{mc}}$ can be computed using the local weights at the maximum.
- We call this procedure a **Monte Carlo adjusted profile (MCAP)**.

A toy: importance sampling for a log normal model



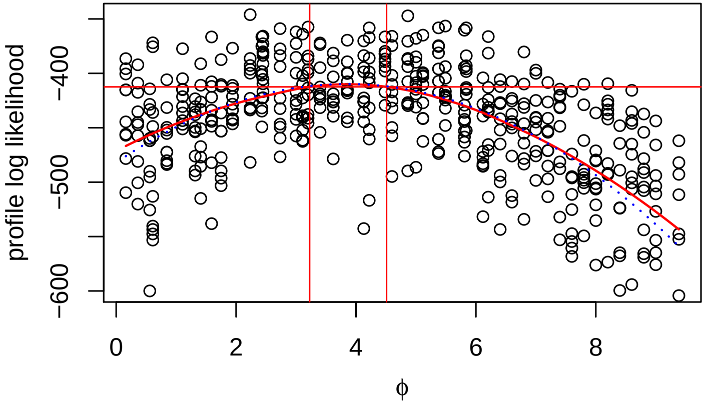
Points show Monte Carlo profile evaluations. Black dashed lines: exact profile and 95% confidence interval. Solid red lines: MCAP confidence interval. Dotted blue line: quadratic approximation.

	Exact profile	MCAP profile	Bootstrap	Quadratic
Coverage %	94.3	93.4	93.3	93.3
Mean width	0.78	0.88	0.94	0.92

Inferring population dynamics from genetic sequence data

- Genetic sequence data on a sample of individuals in an ecological system has potential to reveal population dynamics.
- Extraction information on population dynamics from genetic data has been termed *phylodynamics* (Grenfell et al., 2004).
- Inference via the full likelihood stretches modern computational capabilities, but can be done using the `genPomp` algorithm of Smith et al. (2017).
- The `genPomp` algorithm is an application of iterated filtering methodology (Ionides et al., 2015) to phylodynamic models and data.
- However, the `genPomp` algorithm leads to estimators with high Monte Carlo variance, indeed, too high for reasonable amounts of computation resources to reduce Monte Carlo variability to negligibility.
- This situation provides a useful scenario to demonstrate our methodology.

Monte Carlo profile for genetic data on HIV dynamics



- ϕ models HIV transmitted by recently infected, diagnosed individuals.
- The MCAP cutoff is 2.35, compared to the unadjusted cutoff of 1.92.
- The computation of this figure took approximately 10 days using 500 cores on a Linux cluster.
- The standard error of the profile evaluations is around 25 log units.

Comparison with methods based on summary statistics

- We have focused on likelihood-based confidence intervals.
- An alternative to likelihood-based inference is to compare the data with simulations using some summary statistic.
- Various plug-and-play methodologies of this kind have been proposed, such as synthetic likelihood (Wood, 2010) and nonlinear forecasting (Ellner et al., 1998).
- For large nonlinear systems, it can be hard to find low-dimensional summary statistics that capture a good fraction of the information in the data.
- Even summary statistics derived by careful scientific or statistical reasoning have been found surprisingly uninformative compared to the whole data likelihood in both scientific investigations (Shrestha et al., 2011) and simulation experiments (Fasiolo et al., 2016).

Comparison with Bayesian computation

- Much attention has been given to scaling Bayesian computation to complex models and large data. Latent process models are closely related computationally to Bayesian inference: Bayesian parameters are latent random variables.
- Bayesian Numerical methods such as expectation propagation (EP), variational Bayes, and posterior interval estimation (PIE) are effective for some model classes. They emphasize hierarchical models, where the joint density of the data and latent variables can be conveniently factorized. The `genPomp` example doesn't have this structure; and the MCAP methodology has no such requirement.
- Some simulation-based Bayesian methods use unbiased Monte Carlo likelihood evaluations inside an MCMC algorithm (Andrieu and Roberts, 2009). Error in likelihood evaluation slows MCMC convergence. Optimal trade-off between number of MCMC iterations and time spent on each likelihood evaluation occurs at a Monte Carlo likelihood std. deviation of one log unit (Doucet et al., 2015). For the `genPomp` example, Monte Carlo errors that small are infeasible.

Conclusions

- MCAP provides a simple and general approach to inference when the signal-to-noise ratio in the Monte Carlo profile log likelihood is sufficient to uncover the main features of this function, up to an unimportant vertical shift.
- For large datasets in which the signal (quantified as the curvature of the log likelihood) is large, the methodology can be effective even when the Monte Carlo noise is far too big to carry out standard Bayesian MCMC techniques.
- Although the frequentist motivation for likelihood-based inference differs from the goal of Bayesian posterior inference, both approaches can be used for deductive scientific reasoning (Gelman and Shalizi, 2013; Ionides et al., 2017).

References I

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient computation. *Annals of Statistics*, 37:697–725.

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70(2):343–365.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, pages 211–252.

Cox, D. and Snell, E. (1970). *The analysis of binary data*. Methuen and Co, London.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 34:187–220.

References II

Doucet, A., Pitt, M. K., Deligiannidis, G., and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102:295–313.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, 65:457–483.

Ellner, S. P., Bailey, B. A., Bobashev, G. V., Gallant, A. R., Grenfell, B. T., and Nychka, D. W. (1998). Noise and nonlinearity in measles epidemics: Combining mechanistic and statistical approaches to population modeling. *American Naturalist*, 151:425–440.

Fasiolo, M., Pya, N., and Wood, S. N. (2016). A comparison of inferential methods for highly nonlinear state space models in ecology and epidemiology. *Statistical Science*, 31:96–118.

Gelman, A. and Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66:8–38.

References III

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303:327–332.

Ionides, E. L. (2005). Maximum smoothed likelihood estimation. *Statistica Sinica*, 15:1003–1014.

Ionides, E. L., Giessing, A., Ritov, Y., and Page, S. E. (2017). Response to the ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 71:88–89.

Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S., and King, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences of the USA*, 112:719–724.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

Lindsay, B. G. and Li, B. (1997). On second-order optimality of the observed Fisher information. *Annals of Statistics*, 25(5):2172–2199.

References IV

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95:449–465.

Shrestha, S., King, A. A., and Rohani, P. (2011). Statistical inference for multi-pathogen systems. *PLoS Computational Biology*, 7:e1002135.

Smith, R. A., Ionides, E. L., and King, A. A. (2017). Infectious disease dynamics inferred from genetic data via sequential Monte Carlo. *Molecular Biology and Evolution*, pre-published online, doi:10.1093/molbev/msx124.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466:1102–1104.