Saddlepoint approximation for likelihoods

Jesse Goodman

University of Auckland

Based on joint work with Godrick Oketch, Rachel Fewster, Alice Hankin, Sixiang Shan, Sarah Marais, and George Fan

Part of the Marsden funded project “Fast statistical methods for enigmatic sensor data” with Rachel Fewster, Martin Hazelton, and Ben Stevenson

Biometrics in the Bush Capital, November 2025

Outline

Outline

Density, MGF, CGF, saddlepoint

Ingredients

- Random variable Y
- Density function $f(y)$
- Moment generating function $M(s) = \mathbb{E}(e^{sY})$
- Cumulant generating function $K(s) = \log M(s)$
- Saddlepoint $\hat{s}$ solving

$$K'(\hat{s}) = y \qquad \text{(SE)}$$

Saddlepoint approximation – continuous RV

$$\hat{f}(y) = \frac{\exp\left(K(\hat{s}) - \hat{s}y\right)}{\sqrt{2\pi K''(\hat{s})}} \qquad \text{(SPA)}$$

Integer and multivariate versions

Saddlepoint Approximation – continuous RV, density $f(y)$

$$\hat{f}(y) = \frac{\exp\left(K(\hat{s}) - \hat{s}y\right)}{\sqrt{2\pi K''(\hat{s})}} \quad \text{where } K'(\hat{s}) = y$$

Saddlepoint Approximation – integer-valued, PMF $f(y)$

$$\hat{f}(y) = \frac{\exp\left(K(\hat{s}) - \hat{s}y\right)}{\sqrt{2\pi K''(\hat{s})}} \quad \text{where } K'(\hat{s}) = y$$

Saddlepoint Approximation – multivariate

$$\hat{f}(y) = \frac{\exp\left(K(\hat{s}) - \hat{s}y\right)}{\sqrt{\det(2\pi K''(\hat{s}))}} \quad \text{where } K'(\hat{s}) = y \quad \text{and}$$

- $M(s) = \mathbb{E}(e^{sY}) = \mathbb{E}(e^{s_1 Y_1 + \cdots + s_d Y_d})$ is the multivariate MGF
- K' and K'' are the gradient and Hessian of the multivariate CGF $K(s) = \log M(s)$

Saddlepoint approximation – summary

The saddlepoint approximation is a systematic method for converting a known MGF into an approximate probability density/mass function.

- Edgeworth expansions
- Laplace approximations
- contour integration
- Watson's lemma
- exponential families
- tilting
- ...

Outline

Saddlepoint approximations – uses

First strand – classical

Understand sampling distributions

Setup and aims:

- The distribution Y is fixed, either a sampling distribution or a related statistic
- We seek theoretical understanding of the density function $f_Y(y)$ as a function of y, including tail behaviour as $|y| \to \infty$
- . . . particularly in the limit where Y is a sum of n i.i.d. terms as $n \to \infty$

Features

The saddlepoint approximation gives approximate densities $\hat{f}_Y(y)$ with good uniformity in y, a fairly simple functional form as a function of y and n, and easily interpretable error estimates.

Saddlepoint approximation – likelihood

Second strand – recent

Use saddlepoint to approximate the likelihood.

Saddlepoint approximation – summary

The saddlepoint approximation is a systematic method for converting a known MGF into an approximate probability likelihood function.

Saddlepoint approximation to the likelihood

$$\hat{L}(\theta; y) = \hat{f}(y; \theta) = \frac{\exp\left(K'(\hat{s}; \theta) - \hat{s}y\right)}{\sqrt{\det(2\pi K''(\hat{s}; \theta)}} \quad \text{where } K'(\hat{s}; \theta) = y$$

as an approximation to $L(\theta; y) = f(y; \theta)$.

Outline

Example – wildlife abundance estimation

Two-source capture-recapture models

At every Biometrics conference ("capture occasion"), I ask everyone to make two fingerprints (right thumb, left index finger) and send them to me anonymously. Not everyone listens, and some people send only one fingerprint, or send nothing. After several conferences, I can get some extra information about the population by cross-matching fingerprints.

- If Louise sends me both fingerprints on the same piece of paper, I can match all her fingerprints from all seminars.
- If James sends me some right-thumbprints and some left-index-fingerprints, but never on the same piece of paper, I cannot match his fingerprints. James's papers will contribute to two piles, "unmatched left" and "unmatched right."

Example – wildlife abundance estimation

Two-source capture-recapture models

At every Biometrics conference ("capture occasion"), I ask everyone to make two fingerprints (right thumb, left index finger) and send them to me anonymously. Not everyone listens, and some people send only one fingerprint, or send nothing. After several conferences, I can get some extra information about the population by cross-matching fingerprints.

- Using pencil and paper, we can determine what we would have recorded from one individual, if we knew their responses across all capture occasions
- For wildlife, we can formulate sensible individual-based parametric models for how each animal responds at each capture occasion
  - Tigers and stripe patterns, recorded by camera traps
  - Whales and barnacle patterns, recorded during photo surveys
  - Mixed photo-genotype studies
  - Other effects: misidentification, open populations...

Outline

When else?

The simplest setting for the saddlepoint approximation is for i.i.d. sums:

$$Y_\theta = \sum_{i=1}^{n} X_\theta^{(i)}, \quad X_\theta^{(i)} \sim X_\theta \text{ i.i.d.}$$

Then

$$M_Y(s;\theta) = M_X(s;\theta)^n, \quad K_Y(s;\theta) = nK_X(s;\theta),$$

and the saddlepoint approximation has nice n-dependence:

$$\hat{L}(\theta; y) = \frac{\exp\left(n\left[K_X(\hat{s}) - \hat{s}x\right]\right)}{\sqrt{\det(2\pi n K_X''(\hat{s}))}} \quad \text{where } x = y/n.$$

Individual-based models

Definition

An individual-based model for population totals Y_θ means a model

$$Y_\theta = \sum_{i=1}^{n} X_\theta^{(i)}, \quad X_\theta^{(i)} \sim X_\theta \text{ i.i.d.}$$

- The contribution of a single individual is modelled by a parametric distribution $\theta \mapsto X_\theta$ for which we know the (multivariate) MGF
- Contributions are independent across different individuals

- Specifying X_θ and n fully determines the parametric model $\theta \mapsto Y_\theta$, and $K_Y(s;\theta) = nK_X(s;\theta)$ is easily computed...
- ...but it may be complicated to compute $L_Y(\theta; y)$

Note: we do not observe the individual contributions $X_\theta^{(i)}$, only the population total Y_θ.

Example: branching process for population sizes

Example – Davison, Hautphenne & Kraus

Once every year, count the number of birds on an island. Model this time series by a Galton-Watson branching process, where at each step

$$Y(t) \stackrel{d}{=} \sum_{i=1}^{Y(t-1)} X^{(i)}$$

Individual-based model:

- The individuals are birds
- The $X^{(i)}$'s are i.i.d. copies of the offspring distribution for a single individual, which we model by a parametric distribution $\theta \mapsto X_\theta$
- θ contains per-individual parameters, eg. birth and death rates
- The observed data are the population totals across all individuals at the previous generation

Other kinds of model features

The saddlepoint likelihood method can be applied to any model for which the CGF $K_Y(s;\theta)$ is available.

- For several common model building-block operations, the model CGF K_Y is available in terms of CGFs K_X for the "ingredient" distributions.

Randomly stopped sums

$$Y_\theta = \sum_{i=1}^{N_\theta} X_\theta^{(i)}, \qquad X_\theta^{(i)}\text{'s i.i.d. and independent of } N_\theta.$$

Compound distributions

$$Y_\theta \sim \text{Poisson}(X_\theta)$$

and other additive families such as Gamma(shape $= X$, rate $= r(\theta)$), Negative Binomial, Normal.

Other kinds of model features

Inhomogeneous sums

$$Y_\theta = X_{\theta,1} + \cdots + X_{\theta,r}$$

where $X_{\theta,i}$ are independent but not identically distributed.

Example – INAR(p), Pedeli, Davison & Fokianos

An integer-valued autoregressive model of order p for a time series of count data, where at each step

$$Y(t) \stackrel{d}{=} \text{Binomial}(Y(t-1), q_1) + \cdots + \text{Binomial}(Y(t-p), q_p) + \xi_\theta,$$

with all terms independent. The parameters of interest are the probabilities $q_1, \ldots, q_p$ (the autoregressive parameters) and any parameters in the innovation distribution ξ_θ.

Other kinds of model features

Thinning and splitting

- Each individual in a population of random size N_θ is kept with probability p and discarded otherwise, and $Y_{\theta,p}$ counts the number of kept individuals.
  - $Y_{\theta,p} \sim \text{Binomial}(N_\theta, p)$
- Each individual in a population of random size N_θ is assigned to one of r categories with probabilities $p_1, \ldots, p_r$, and $Y_{\theta,\vec{p}}$ is the vector of counts in each category.
  - $Y_{\theta,p} \sim \text{Multinomial}(N_\theta, \vec{p})$

Correlated count variables, partial summaries

Example:

$$Y_{\theta,1} = X_\theta^{(1)} + Z_\theta, \quad \ldots, \quad Y_{\theta,k} = X_\theta^{(k)} + Z_\theta,$$

with $X_\theta^{(i)}$ i.i.d. and independent of a single shared random variable Z_θ.

Other kinds of model features

Linear mapping

For a random vector X_θ and a deterministic matrix A, set

$$Y_\theta = AX_\theta$$

Common use case:

Each individual has a latent category that, if known, determines how they are counted in vector Y of population totals.

- A is a deterministic matrix, with one column for each possible latent category and one row for each measured total in Y
- the entries of X_θ count the number of individuals with each latent category

The two-source capture recapture model has this form.

Outline

Why does the saddlepoint approximation work?

Tilting

Define the tilted random variable $X^{(s)}$ by

$$f_{X^{(s)}}(x) = \frac{e^{sx}}{M(s)} f(x)$$

Then

$$M_{X^{(s_0)}}(s) = M(s_0 + s)/M(s_0), \quad K_{X^{(s_0)}}(s) = K(s_0 + s) - K(s_0)$$

and we can recover $f(x)$ from $f_{X^{(s)}}(x)$:

$$f(x) = \exp\left(K(s) - sx\right) f_{X^{(s)}}(x)$$

Why does the saddlepoint approximation work?

Tilting

Define the tilted random variable $X^{(s)}$ by

$$f_{X^{(s)}}(x) = \frac{e^{sx}}{M(s)} f(x), \qquad f(x) = \exp\left(K(s) - sx\right) f_{X^{(s)}}(x)$$

- The family of tilted distributions is precisely that exponential family for which X is the sufficient statistic
- The saddlepoint equation $K'(\hat{s}) = x$ is the constraint that the tilted distribution $X^{(\hat{s})}$ should have mean equal to the observed value x

Why does the saddlepoint approximation work?

Tilting

Define the tilted random variable $X^{(s)}$ by

$$f_{X^{(s)}}(x) = \frac{e^{sx}}{M(s)} f(x), \qquad f(x) = \exp\left(K(s) - sx\right) f_{X^{(s)}}(x)$$

Saddlepoint approximation via tilting

1. Tilt X so that its tilted mean is $\mathbb{E}(X^{(\hat{s})}) = K'(\hat{s}) = x$
   - i.e. find within its exponential family that distribution having mean x
2. Relate $f(x)$ to $f_{X^{(\hat{s})}}(x)$ in terms of the relative entropy $K(\hat{s}) - \hat{s}K'(\hat{s})$
3. Approximate the tilted density $f_{X^{(\hat{s})}}(x)$ by the normal approximation to $X^{(\hat{s})}$ at its mean.

Why does the saddlepoint approximation work?

Visualization – tilting and saddlepoint approximation

Outline

Saddlepoint approximations for likelihoods

Idea

Use saddlepoint to approximate the likelihood.

Setup and aims:

- The distribution Y_θ is a parametric model for the data vector to be obtained from our experiment
- MGFs, CGFs, densities and saddlepoint approximations depend on the parameter vector θ

Saddlepoint approximation to the likelihood

$$\hat{L}(\theta; y) = \hat{f}(y; \theta) = \frac{\exp\left(K'(\hat{s}; \theta) - \hat{s}y\right)}{\sqrt{\det(2\pi K''(\hat{s}; \theta)}} \quad \text{where } K'(\hat{s}; \theta) = y$$

$$L(\theta; y) = f(y; \theta).$$

Saddlepoint approximations and MLEs

Idea

Use saddlepoint to approximate the likelihood.

$$\hat{L}(\theta; y) = \hat{f}(y; \theta) = \frac{\exp\left(K'(\hat{s}; \theta) - \hat{s}y\right)}{\sqrt{\det(2\pi K''(\hat{s}; \theta)}} \quad \text{where } K'(\hat{s}; \theta) = y$$

- We want to maximise L to obtain the MLE

$$\theta_{\text{MLE}}(y) = \operatorname*{argmax}_{\theta} L(\theta; y)$$

- . . . but if we cannot compute L, we must settle for the

Saddlepoint MLE

$$\hat{\theta}_{\text{MLE}}(y) = \operatorname*{argmax}_{\theta} \hat{L}(\theta; y)$$

The Saddlepoint Likelihood Method

Idea

Use the saddlepoint approximation to compute approximate MLEs.

$$\hat{\theta}_{\text{MLE}}(y) = \underset{\theta}{\text{argmax}}\, \hat{L}(\theta; y)$$

Key Question: How well does the saddlepoint MLE work?

How big is the discrepancy $\delta = \theta_{\text{MLE}}(y) - \hat{\theta}_{\text{MLE}}(y)$ between the true MLE and saddlepoint MLE?

Context for comparison

We should compare the discrepancy δ to the scale of sampling variability $\theta_{\text{MLE}}(y) - \theta_0$, asymptotically in a relevant limit.

Approximations and MLEs

General Key Question

How big of a discrepancy δ is introduced by using an approximation instead of the true log-likelihood function?

Related Key Question

What if we use the normal approximation instead of the saddlepoint approximation?

- Let $\tilde{Y}_\theta$ be Normal with the same mean and variance as Y_θ (as functions of θ)
- Let $\tilde{L}(\theta; y)$ be the likelihood function for $\tilde{Y}_\theta$.
- Maximise to obtain $\tilde{\theta}_{\text{MLE}}(y) = \text{argmax}_\theta \, \tilde{L}(\theta; y)$.
- How big is $\theta_{\text{MLE}}(y) - \widetilde{\theta}_{\text{MLE}}(y)$?

Limiting setup

$$Y_\theta = \sum_{i=1}^{n} X_\theta^{(i)}, \quad X_\theta^{(i)} \sim X_\theta \text{ i.i.d.}$$

Limiting setup

We consider the limit $n \to \infty$, everything else fixed.

Small-sample asymptotics

In practice we are thinking of n being large but not overwhelmingly large, and we will compare inverse powers of n.

- Note: we do not observe the individual contributions $X_\theta^{(i)}$, only the population total Y_θ with large population size n.
  - Thus n is not quite a sample size, the data vector Y_θ has fixed dimension, and there is no guarantee that $n \to \infty$ lets us identify the data-generating distribution.

Limiting setup

$$Y_\theta = \sum_{i=1}^{n} X_\theta^{(i)}, \quad X_\theta^{(i)} \sim X_\theta \text{ i.i.d.}$$

Limiting setup

We consider the limit $n \to \infty$, everything else fixed.

Small-sample asymptotics

In practice we are thinking of n being large but not overwhelmingly large, and we will compare inverse powers of n.

- Note: we do not observe the individual contributions $X_\theta^{(i)}$, only the population total Y_θ with large population size n.
  - Thus n is not quite a sample size, the data vector Y_θ has fixed dimension, and there is no guarantee that $n \to \infty$ lets us identify the data-generating distribution.

“Mean-like” identifiability condition

The error in the saddlepoint approximation has largely universal scaling behaviour. However, the effect on the MLE is model-dependent and depends on the overall shape of the log-likelihood function.

“Fully-identifiable” condition

The Jacobian matrix $(\frac{\partial}{\partial\theta_j}\mathbb{E}(X_{\theta,i}))_{i,j}$ has rank $=$ #parameters

- This implies $\theta \mapsto \mathbb{E}(X_\theta)$ is (locally) one-to-one
- Informally, all the parameters are “mean-like”
- More formally, the model is “fully identifiable at the level of the sample mean”

MLE discrepancy – fully identifiable case

Assume:

- appropriate regularity conditions hold
- the identifiability condition holds – "all parameters are mean-like"
- $y = y_n \sim Y_{\theta_0,n}$ is drawn according to the model with true parameter θ_0

Then:

- the saddlepoint discrepancy $\delta_n = \theta_{\mathrm{MLE}}(y_n) - \hat{\theta}_{\mathrm{MLE}}(y_n)$ is $O_{\mathbb{P}}(1/n^2)$
- the normal approximation discrepancy $\theta_{\mathrm{MLE}}(y_n) - \widetilde{\theta}_{\mathrm{MLE}}(y_n)$ is $O_{\mathbb{P}}(1/n)$
- the sampling variability is $O(1/\sqrt{n})$
- all three MLEs are consistent and asymptotically normal estimators of θ_0, and

$$\sqrt{n}\left(\theta_{\mathrm{MLE}} - \theta_0, \hat{\theta}_{\mathrm{MLE}} - \theta_0, \widetilde{\theta}_{\mathrm{MLE}} - \theta_0\right) \to (Z, Z, Z)$$

where Z is normal with an explicit covariance matrix.

Estimating the discrepancy

Estimated discrepancy

The (unknown) discrepancy $\delta = \theta_{\text{MLE}}(y) - \hat{\theta}_{\text{MLE}}(y)$ can be estimated by an explicit quantity $\hat{\delta}$ computed only in terms of K_Y and the saddlepoint log-likelihood.

With Godrick Oketch & Rachel Fewster

- Code to quickly and painlessly generate saddlepoint approximations and MLEs
- An estimate of the discrepancy in the MLE...
- ...and code to generate this estimate painlessly
  - devtools::install_github("godrick/saddlepoint")

Theorem (Godrick Oketch, Rachel Fewster, J.G.)

The estimated discrepancy $\hat{\delta}$ closely approximates the true discrepancy δ in the sense that

$$\delta, \hat{\delta} = O(n^{-2}), \quad \delta - \hat{\delta} = O(n^{-3})$$

in the fully identifiable case.

Outline

How is this implemented in code?

Example (Two-source capture-recapture model, continued)

The observed data vector is given by

$$Y_\theta = AX_\theta$$
$$X_\theta \sim \text{Multinomial}(N, h(\theta))$$
$$\theta = (N, p_L, p_R, p_B)$$

The vectors Y_θ and X_θ hold the counts of various observed and true capture histories, and A is the 0-1 matrix where $A_{ij} = 1$ if true capture history j contributes to the count for observed capture history i. The parameters have constraints $p_B < p_L, p_B < p_R$. Formulas for $h(\theta)$ and A_{ij} are reasonably straightforward but are omitted.

Two-source model in code

Listing 1: Pseudocode for two-source capture-recapture model

```
devtools::install\_github("godrick/saddlepoint")
# theta = (N, pL, pR, pB)
# Define adaptor function h with theta as the argument
h <- function(theta) {...}

# Define the constraints on the model parameters as a
# function of theta: pB < pL and pB < pR
# The constraints are set to be non-positive: (pB - pL < 0, pB - pR < 0)
constraints.on.theta <- function(theta){
  list(constraints = c(theta[4] - theta[2], theta[4] - theta[3]),
  jacobian = rbind(c(0, -1, 0, 1), c(0, 0, -1, 1)))
}
# Build the CGF for Y as a function of model parameter vector theta
K.X <- MultinomialModelCGF(n = adaptor(indices = 1), prob.vec = h)
K.Y <- linearlyMappedCGF(cgf = K.X, matrix_A = ...)
# Find the estimate of theta
find.saddlepoint.MLE(observed.data = Y, cgf = K.Y,
          starting.theta = ... ,
          user.ineq.constraint.function = constraints.on.theta,
          discrepancy = TRUE)
```

Thank you.

Example – traffic models

Traffic models (from Martin Hazelton, lightly adapted)

Place traffic monitoring devices at various points in a city's road network. At the end of a monitoring period, each monitoring device sends us the number of cars that drove past the device. The observed data y is the vector of all car counts, one entry for each device. A sensible model Y_θ must account for cars driving past several nearby devices.

Individual-based model:

- The individuals are drivers
- Each individual chooses a route according to probabilities obtained from θ, independently across individuals
- Using a city map, we can determine which routes go past which traffic monitoring devices
- The individual contribution X_θ is the 0-1 vector recording which devices the individual drives past on their chosen route

MLE discrepancy – partially identifiable case

Assumption

We assume that the parameter vector can be partitioned as

$$\theta = (\omega, \tau) \quad \text{where} \quad \mathbb{E}(Y_\theta) \text{ depends only on } \omega.$$

Assume:

- $y_n = n\mathbb{E}(Y_{\theta_0}) + \sqrt{n}z_0$ is consistent with true mean-controlling parameter ω_0
  - Thus: the z-score vector for y_n is constant as a function of n
- The Jacobian of $\omega \mapsto \mathbb{E}(X_\theta)$ has rank $=$ #parameters in ω
- There is a local MLE for τ with $\omega = \omega_0$ fixed
  - τ must be "variance-like", i.e., has a non-trivial effect on $\mathrm{Var}(X_\theta)$ while keeping $\mathbb{E}(X_\theta)$ fixed

Then: mean-like discrepancy $\omega_{\mathrm{MLE}}(y_n) - \hat{\omega}_{\mathrm{MLE}}(y_n) = O(1/n^{3/2})$
variance-like discrepancy $\tau_{\mathrm{MLE}}(y_n) - \hat{\tau}_{\mathrm{MLE}}(y_n) = O(1/n)$
Under a further explicit condition
$\omega_{\mathrm{MLE}}(y_n) - \hat{\omega}_{\mathrm{MLE}}(y_n) = O(1/n^2)$

Other limiting frameworks

The saddlepoint approximation is well-adapted to the limit
$n = (\text{number of i.i.d. summands}) \to \infty$

- $n \to \infty$ makes the approximation error smaller
- $n \to \infty$ is also the region where direct calculation is harder

However, (number of i.i.d. summands) $\to \infty$ does not always apply. Other relevant limits:

- The classical statistical paradigm is (number of i.i.d. observations) $\to \infty$.
  - The saddlepoint approximation does not benefit from this limit.
- (length of time series) $\to \infty$.
- Point process data with (number of spatial observations) $\to \infty$.

Approximating the discrepancy

Relative error in the saddlepoint approximation is

$$\frac{L(\theta; y)}{\hat{L}(\theta; y)} =$$

$$\mathbb{E}\left(\exp\left(\varepsilon^{-2}\left[K(s + i\varepsilon Z; \theta) - K(s; \theta) - i\varepsilon Z K'(s; \theta) + \tfrac{1}{2}\varepsilon^2 Z K''(s; \theta) Z^T\right]\right.\right.$$

evaluated with

$$Z \sim \mathcal{N}(0, K''(s; \theta)^{-1}), \qquad s = \hat{s}, \quad \varepsilon = 1.$$

The limiting setup $Y_\theta = \sum_{i=1}^{n} X_\theta^{(i)}$ corresponds to setting $\varepsilon = 1/\sqrt{n}$. Higher-order approximations are obtained by expanding as a Taylor series around $\varepsilon = 0$.

Approximating the discrepancy – open question

$$\mathbb{E}\left(\exp\left(\varepsilon^{-2}\left[K(s+i\varepsilon Z;\theta)-K(s;\theta)-i\varepsilon Z K'(s;\theta)+\tfrac{1}{2}\varepsilon^2 Z K''(s;\theta)Z\right.\right.\right.$$

We can expand the exponential as

$$1+\varepsilon\cdot[\text{odd powers of } Z]+\frac{\varepsilon^2}{24}\sum_{i_1,i_2,i_3,i_4}\frac{\partial^4 K}{\partial s_{i_1}\partial s_{i_2}\partial s_{i_3}\partial s_{i_4}}Z_{i_1}Z_{i_2}Z_{i_3}Z_{i_4}$$
$$-\frac{\varepsilon^2}{72}\left(\sum_{i_1,i_2,i_3}\frac{\partial^3 K}{\partial s_{i_1}\partial s_{i_2}\partial s_{i_3}}Z_{i_1}Z_{i_2}Z_{i_3}\right)^2+\cdots$$

For each fixed value of Z, computing the sums has a fixed complexity not depending on dimension – this is the Cheap Gradient Principle.

Open question

Can the Gaussian expectations be computed systematically with a fixed complexity as dimension grows?

Thank you.