A probabilistic diagnostic tool to assess Laplace approximations

Proof of concept and non-asymptotic experimentation

Shaun McDonald[1], Dave Campbell[2], Haoxuan Zhou[3]

[1]Department of Statistics and Actuarial Science
Simon Fraser University

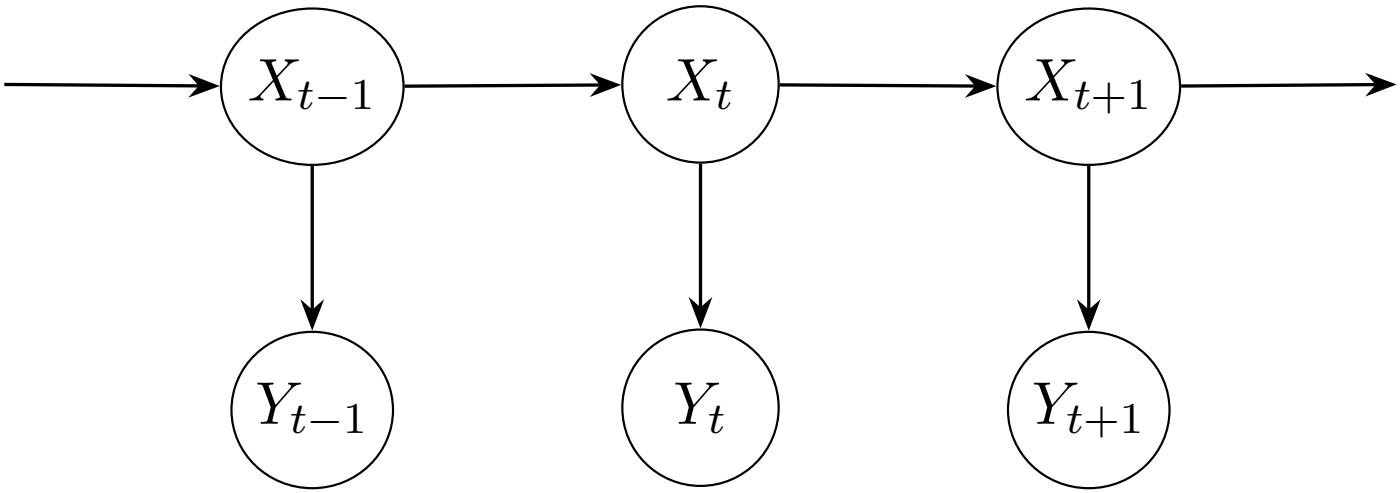
[2]School of Mathematics and Statistics
Carleton University

U of E Statistics Seminar, 15 November 2021

Outline

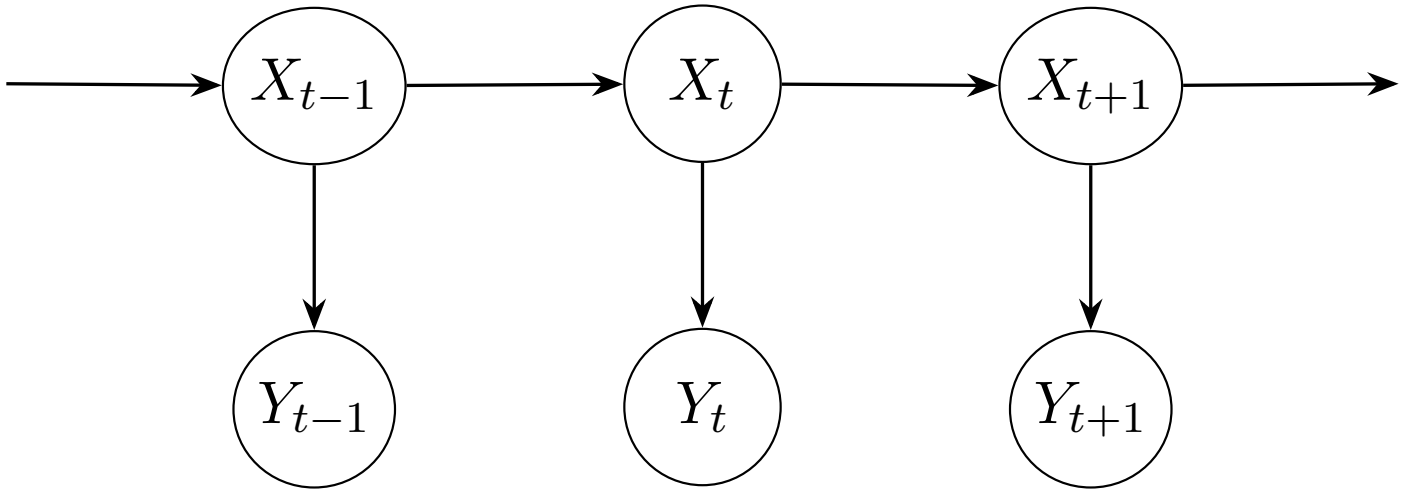
Table of Contents

Motivation: state-space models (SSM's)



- y_t: observations at time t
- $x_t \in \mathbb{R}^q$: hidden state at time t

Motivation: state-space models (SSM's)



Joint likelihood

$$p_{x,y}(\boldsymbol{x}, \boldsymbol{y} \mid \theta) = p(x_1 \mid \theta) \left[\prod_{t=2}^{T} p(x_t \mid x_{t-1}, \theta)\right] \left[\prod_{t=1}^{T} p(y_t \mid x_t, \theta)\right]$$

- $\boldsymbol{x} = (x_1, \ldots, x_T), x_t \in \mathbb{R}^q$: hidden states at times t
- $\boldsymbol{y} = (y_1, \ldots, y_T)$: observations at times t
- θ: model parameters

Example: stock assessment model

- *Stock assessment model* (SAM): nonlinear SSM for fish populations
- y_t: observed fish abundances for year t
- x_t: true abundances, fishing mortality rates
- θ: correlation, variance, scaling parameters
- Aeberhard et al. [2], Nielsen and Berg [15]

The process equation describes the dynamics in the unobserved states and is based on the conditional expectation of the current states given the previous states:

$$\mathrm{E}[\boldsymbol{X}_t|\boldsymbol{X}_{t-1}] = \begin{cases} \log N_{1,t} &= \log N_{1,t-1} \\ \log N_{a,t} &= \log N_{a-1,t-1} - F_{a-1,t-1} - M_{a-1,t-1}, \qquad 2 \le a < A \\ \log N_{A,t} &= \log[N_{A-1,t-1}\exp(-F_{A-1,t-1} - M_{A-1,t-1}) \\ &\quad + N_{A,t-1}\exp(-F_{A,t-1} - M_{A,t-1})] \\ \log F_{a,t} &= \log F_{a,t-1}, \qquad 1 \le a \le A, \end{cases}$$

where A denotes the largest age class. These equations assume a random walk for $\log N_{1,t}$ and for the whole vector $(\log F_{1,t}, \ldots, \log F_{A,t})^\top$, a survival process for $\log N_{a,t}$ where the combination of F and M represents total mortality, and a modified survival process for the plus group in $\log N_{A,t}$. The corresponding distribution $P_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ is a multivariate Gaussian with zero mean vector. The first A Gaussian error components are independent, while we enforce a first-order autoregressive correlation structure for the others:

$$\mathrm{Cor}[\log(F_{a,t}), \log(F_{\tilde{a},t})] = \rho^{|a-\tilde{a}|},$$

where the between-age correlation ρ is an element of $\boldsymbol{\theta}$. Other fixed parameters include four separate variances: one for recruitment ($\sigma^2_{N_{a=1}}$), one for survival ($\sigma^2_{N_{a>1}}$), one for fishing mortality at age 1 ($\sigma^2_{F_{a=1}}$), and one for fishing mortality at older ages ($\sigma^2_{F_{a\ge 2}}$).

The observation equation relates the unobserved states to the observed response variables through a conditional expectation:

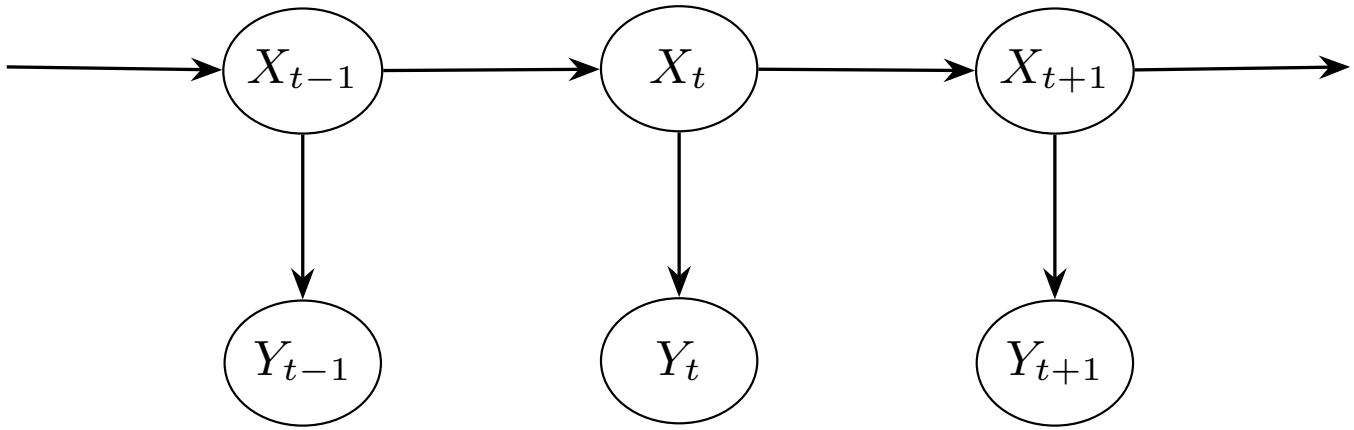
$$\mathrm{E}[\boldsymbol{Y}_t|\boldsymbol{X}_t] = \begin{cases} \log C_{a,t} = \log\left[\frac{F_{a,t}}{Z_{a,t}}(1-\exp(-Z_{a,t}))N_{a,t}\right] \\ \log I^{(s)}_{a,t} = \log\left[Q^{(s)}_a \exp(-Z_{a,t}\frac{D^{(s)}}{365})N_{a,t}\right], \end{cases} \qquad 1 \le a \le A,$$

where $s = 1, 2$ identifies the surveys, the largest age class A is 5 for $s = 1$ and 4 for $s = 2$, $Z_{a,t} = M_{a,t} + F_{a,t}$ is the total mortality rate, $D^{(s)}$ is the number of days into the year when survey (s) was conducted, and $Q^{(s)}_a$ are so-called catchability coefficients that scale the survey relative indices to the stock abundance. The catchabilities are unknown parameters that need to be estimated, there are nine of them, as they are distinct for each age class and each survey. Auxiliary information and expertise from fisheries scientists cast doubt on the reliability of the absolute level of the catches between 1993 and 2005, hence extra catch scaling parameters τ_t are added (and estimated) for these years:

$$\log C_{a,t} = \log\left[\frac{1}{\tau_t}\frac{F_{a,t}}{Z_{a,t}}(1-\exp(-Z_{a,t}))N_{a,t}\right], \quad t \in \{1993, \ldots, 2005\}.$$

Aeberhard et al. [2]

Motivation: SSM's



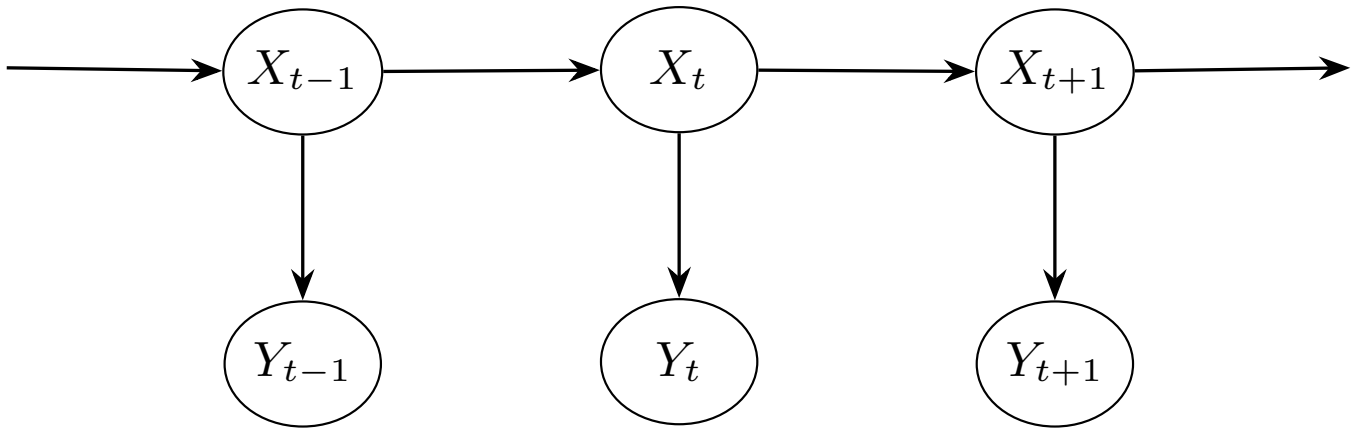
Joint likelihood

$$p_{x,y}\left(\boldsymbol{x}, \boldsymbol{y} \mid \theta\right) = p\left(x_1 \mid \theta\right) \left[\prod_{t=2}^{T} p\left(x_t \mid x_{t-1}, \theta\right)\right] \left[\prod_{t=1}^{T} p\left(y_t \mid x_t, \theta\right)\right]$$

Ideally: estimate θ by maximizing *marginal likelihood*

$$p_y\left(\boldsymbol{y} \mid \theta\right) = \int_{\mathbb{R}^d} p_{x,y}\left(\boldsymbol{x}, \boldsymbol{y} \mid \theta\right) \mathrm{d}\boldsymbol{x}$$

Motivation: SSM's



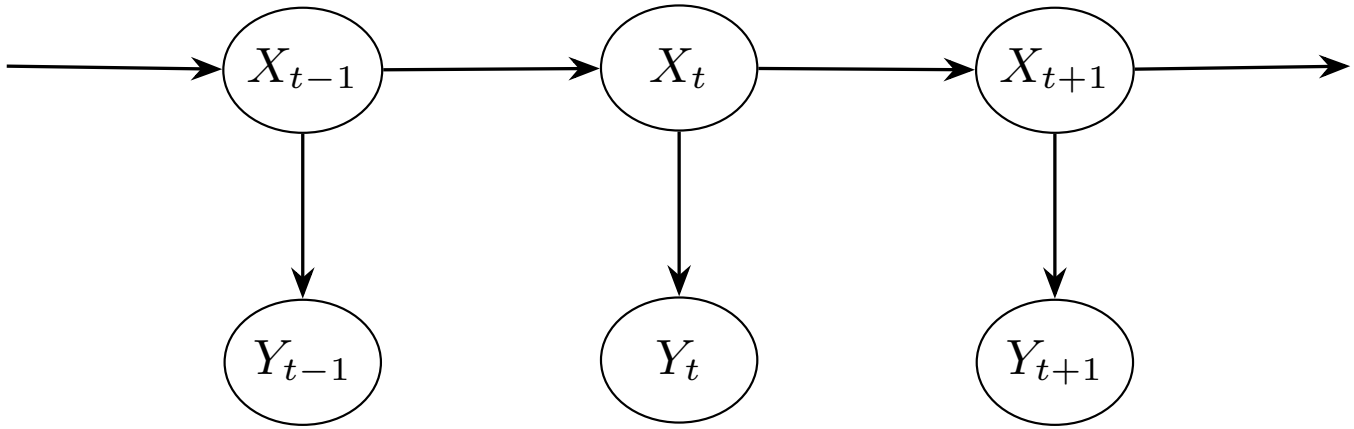
Joint likelihood

$$p_{x,y}\left(\boldsymbol{x}, \boldsymbol{y} \mid \theta\right) = p\left(x_1 \mid \theta\right) \left[\prod_{t=2}^{T} p\left(x_t \mid x_{t-1}, \theta\right)\right] \left[\prod_{t=1}^{T} p\left(y_t \mid x_t, \theta\right)\right]$$

In practice: p_y intractable (integral over $d = qT$ dimensions); estimate θ by maximizing *Laplace approximation* (LA) [e.g. 13]

$$L_\theta\left(p_{x,y}\right) \approx \int_{\mathbb{R}^d} p_{x,y}\left(\boldsymbol{x}, \boldsymbol{y} \mid \theta\right) \mathrm{d}\boldsymbol{x}$$

Motivation: SSM's



Joint likelihood

$$p_{x,y}\left(\boldsymbol{x}, \boldsymbol{y} \mid \theta\right) = p\left(x_1 \mid \theta\right) \left[\prod_{t=2}^{T} p\left(x_t \mid x_{t-1}, \theta\right)\right] \left[\prod_{t=1}^{T} p\left(y_t \mid x_t, \theta\right)\right]$$

In practice: p_y intractable (integral over $d = qT$ dimensions); estimate θ by maximizing *Laplace approximation* (LA) [e.g. 13]

$$L_\theta\left(p_{x,y}\right) \approx \int_{\mathbb{R}^d} p_{x,y}\left(\boldsymbol{x}, \boldsymbol{y} \mid \theta\right) \mathrm{d}\boldsymbol{x}$$

But is this always a good idea?

Laplace approximation

In general terms:

- Given: function $f : \mathbb{R}^d \to \mathbb{R}_{>0}$
- Want: $F = \int_{\mathbb{R}^d} f(x)\mathrm{d}x$
- Assumptions on f:
  - Global maximum at *mode* $\hat{x} \in \mathbb{R}^d$
  - Hessian of $\log f$ at $\hat{x}$, denoted H, is negative definite

Laplace approximation

- Given: function $f : \mathbb{R}^d \to \mathbb{R}_{>0}$
- Want: $F = \int_{\mathbb{R}^d} f(x)\mathrm{d}x$
- Assumptions on f:
  - Global maximum at *mode* $\hat{x} \in \mathbb{R}^d$
  - Hessian of $\log f$ at $\hat{x}$, denoted H, is negative definite
- Taylor expansion of $\log f$ gives *Gaussian approximation to* f:

$$\phi(x) := f\left(\hat{x}\right) \exp\left[\frac{1}{2}\left(x - \hat{x}\right)^\top H \left(x - \hat{x}\right)\right]$$

Laplace approximation

- Given: function $f : \mathbb{R}^d \to \mathbb{R}_{>0}$
- Want: $F = \int_{\mathbb{R}^d} f(x)\mathrm{d}x$
- Assumptions on f:
  - Global maximum at *mode* $\hat{x} \in \mathbb{R}^d$
  - Hessian of $\log f$ at $\hat{x}$, denoted H, is negative definite

$$\phi(x) := f\left(\hat{x}\right) \exp\left[\frac{1}{2}\left(x - \hat{x}\right)^\top H \left(x - \hat{x}\right)\right]$$

- Integrate Gaussian approximation to get *Laplace approximation* [14]:

$$L\left(f\right) := \int_{\mathbb{R}^d} \phi(x)\mathrm{d}x = f\left(\hat{x}\right)\sqrt{\left(2\pi\right)^d \det\left(-H^{-1}\right)} \approx F$$

Laplace approximation

- LA is much faster than other numerical methods, Monte Carlo, importance sampling
- **But** it may be less accurate: 2nd-order Taylor approximation to $\log f$ assumes f is roughly Gaussian in shape
- Find a way to assess this assumption — "middle ground"
- With moderate computation, answer the question

Laplace approximation

- LA is much faster than other numerical methods, Monte Carlo, importance sampling
- **But** it may be less accurate: 2nd-order Taylor approximation to $\log f$ assumes f is roughly Gaussian in shape
- Find a way to assess this assumption — "middle ground"
- With moderate computation, answer the question

Is f "Gaussian enough" to justify the LA?

Laplace approximation

- LA is much faster than other numerical methods, Monte Carlo, importance sampling
- **But** it may be less accurate: 2nd-order Taylor approximation to $\log f$ assumes f is roughly Gaussian in shape
- Find a way to assess this assumption — "middle ground"
- With moderate computation, answer the question

  Is f "Gaussian enough" to justify the LA?
- If yes: use LA to estimate integral
- If no: use other method(s)

Laplace approximation

- LA is much faster than other numerical methods, Monte Carlo, importance sampling
- **But** it may be less accurate: 2nd-order Taylor approximation to $\log f$ assumes f is roughly Gaussian in shape
- Find a way to assess this assumption — "middle ground"
- With moderate computation, answer the question

  Is f "Gaussian enough" to justify the LA?
- If yes: use LA to estimate integral
- If no: use other method(s)

Develop a **diagnostic for the LA** using *probabilistic numerics*

Table of Contents

Intro to probabilistic numerics/BQ

- *Probabilistic numerics*: use probability theory for estimation & uncertainty quantification in standard numerical problems (e.g. quadrature) [8]

Intro to probabilistic numerics/BQ

- *Probabilistic numerics*: use probability theory for estimation & uncertainty quantification in standard numerical problems (e.g. quadrature) [8]

Conventional quadrature: given f, estimate integral F using limited information (weighted sum of f-values)

Intro to probabilistic numerics/BQ

- *Probabilistic numerics*: use probability theory for estimation & uncertainty quantification in standard numerical problems (e.g. quadrature) [8]

Conventional quadrature: given f, estimate integral F using limited information (weighted sum of f-values)

Bayesian quadrature (BQ): put prior on f, use f-values to induce posterior on F [5]

- Posterior mean $\Rightarrow$ estimate ($=$ quadrature)
- Posterior variance $\Rightarrow$ uncertainty quantification

BQ details

- Use "importance re-weighting trick" to re-express F [12, 17]:

$$F = \int_{\mathbb{R}^d} f(x)\mathrm{d}x = \int_{\mathbb{R}^d} r(x)g(x)\mathrm{d}x = \int_{\mathbb{R}^d} r(x)\mathrm{d}G(x)$$

  where G is a Gaussian measure with density g and $r = f/g$

BQ details

$$F = \int_{\mathbb{R}^d} f(x)\mathrm{d}x = \int_{\mathbb{R}^d} r(x)g(x)\mathrm{d}x \tag{1}$$

- Gaussian process prior $r \sim \mathcal{GP}\left(m_0^x, C_0^x\right)$
  - Prior mean is re-weighted Gaussian approximation to f:

$$m_0^x(x) := \frac{f\left(\hat{x}\right)\exp\left[\frac{1}{2}\left(x-\hat{x}\right)^\top H\left(x-\hat{x}\right)\right]}{g(x)} = \frac{\phi(x)}{g(x)}$$

  - Covariance C_0^x to be defined later

BQ details

$$F = \int_{\mathbb{R}^d} f(x)\mathrm{d}x = \int_{\mathbb{R}^d} r(x)g(x)\mathrm{d}x$$
$$r \sim \mathcal{GP}\left(m_0^x, C_0^x\right)$$
$$m_0^x = \frac{\phi}{g}$$

- Evaluate r at *interrogation points* $\boldsymbol{s} := \{s_1, \ldots, s_n\}, s_i \in \mathbb{R}^d$
- Condition on $r(\boldsymbol{s}) = \left(r(s_1), \ldots, r(s_n)\right)^\top \in \mathbb{R}^n$ to get posterior [18]

$$r \mid r(\boldsymbol{s}) \sim \mathcal{GP}\left(m_1^x, C_1^x\right)$$
$$m_1^x(x) = m_0^x(x) + C_0^x(x, \boldsymbol{s})^\top \left[C_0^x(\boldsymbol{s}, \boldsymbol{s})\right]^{-1} \left(r(\boldsymbol{s}) - m_0^x(\boldsymbol{s})\right)$$
$$C_1^x(x, z) = C_0^x(x, z) - C_0^x(x, \boldsymbol{s})^\top \left[C_0^x(\boldsymbol{s}, \boldsymbol{s})\right]^{-1} C_0^x(z, \boldsymbol{s})$$

BQ details

$$
\begin{aligned}
r \mid r(\boldsymbol{s}) &\sim \mathcal{GP}\left(m_1^x, C_1^x\right) \\
m_1^x(x) &= m_0^x(x) + C_0^x(x, \boldsymbol{s})^\top \left[C_0^x(\boldsymbol{s}, \boldsymbol{s})\right]^{-1} \left(r(\boldsymbol{s}) - m_0^x(\boldsymbol{s})\right) \\
C_1^x(x, z) &= C_0^x(x, z) - C_0^x(x, \boldsymbol{s})^\top \left[C_0^x(\boldsymbol{s}, \boldsymbol{s})\right]^{-1} C_0^x(z, \boldsymbol{s})
\end{aligned}
$$

- Induces Normal posterior on F [18]:

$$
\begin{aligned}
F \mid r(\boldsymbol{s}) &\sim \mathcal{N}\left(m_1, C_1\right) \\
m_1 &= L(f) + \left[\int_{\mathbb{R}^d} C_0^x(z, \boldsymbol{s}) \mathrm{d}G(z)\right]^\top \left[C_0^x(\boldsymbol{s}, \boldsymbol{s})\right]^{-1} \left(r(\boldsymbol{s}) - m_0^x(\boldsymbol{s})\right) \\
C_1 &= C_0 - \left[\int_{\mathbb{R}^d} C_0^x(x, \boldsymbol{s}) \mathrm{d}G(x)\right]^\top \left[C_0^x(\boldsymbol{s}, \boldsymbol{s})\right]^{-1} \left[\int_{\mathbb{R}^d} C_0^x(x, \boldsymbol{s}) \mathrm{d}G(x)\right]
\end{aligned}
$$

BQ details

$$F \mid r(\boldsymbol{s}) \sim \mathcal{N}(m_1, C_1)$$

$$m_1 = L(f) + \left[\int_{\mathbb{R}^d} C_0^x(z, \boldsymbol{s}) \mathrm{d}G(z)\right]^\top [C_0^x(\boldsymbol{s}, \boldsymbol{s})]^{-1} (r(\boldsymbol{s}) - m_0^x(\boldsymbol{s}))$$

$$C_1 = C_0 - \left[\int_{\mathbb{R}^d} C_0^x(x, \boldsymbol{s}) \mathrm{d}G(x)\right]^\top [C_0^x(\boldsymbol{s}, \boldsymbol{s})]^{-1} \left[\int_{\mathbb{R}^d} C_0^x(x, \boldsymbol{s}) \mathrm{d}G(x)\right]$$

- Note: posterior mean m_1 is LA plus "correction term" (weighted sum of function values)
- Diagnostic:

  Reject LA iff $L(f) \notin \left(m_1 - 1.96\sqrt{C_1}, m_1 + 1.96\sqrt{C_1}\right)$

BQ details

$$F \mid r(\boldsymbol{s}) \sim \mathcal{N}(m_1, C_1)$$

$$m_1 = L(f) + \left[\int_{\mathbb{R}^d} C_0^x(z, \boldsymbol{s}) \mathrm{d}G(z)\right]^\top [C_0^x(\boldsymbol{s}, \boldsymbol{s})]^{-1} (r(\boldsymbol{s}) - m_0^x(\boldsymbol{s}))$$

$$C_1 = C_0 - \left[\int_{\mathbb{R}^d} C_0^x(x, \boldsymbol{s}) \mathrm{d}G(x)\right]^\top [C_0^x(\boldsymbol{s}, \boldsymbol{s})]^{-1} \left[\int_{\mathbb{R}^d} C_0^x(x, \boldsymbol{s}) \mathrm{d}G(x)\right]$$

- Note: posterior mean m_1 is LA plus "correction term" (weighted sum of function values)
- Diagnostic:

Reject LA iff $L(f) \notin \left(m_1 - 1.96\sqrt{C_1}, m_1 + 1.96\sqrt{C_1}\right)$
(i.e. iff diagnostic p-value < 0.05)

Table of Contents

Design

3 design choices to make:

1. Interrogation grid
2. Covariance kernel
3. Integrating measure

Design

3 design choices to make:

1. Interrogation grid
2. Covariance kernel
3. Integrating measure

- Traditionally, BQ seeks high accuracy & low uncertainty
  - Design choices may be optimized for specific integrand
- Different goals here:

Design

3 design choices to make:

1. Interrogation grid
2. Covariance kernel
3. Integrating measure

- Traditionally, BQ seeks high accuracy & low uncertainty
  - Design choices may be optimized for specific integrand
- Different goals here:
  - Quick, "one-size-fits-all"
  - Don't want to reject *every* function — "*good*-**enough**-*ness-of-fit*"

Interrogation grid

- Start with *preliminary grid* $\boldsymbol{s}^* = \{s_1^*, \ldots, s_n^*\}$

Assume $\boldsymbol{s}^*$:

1. Is a *fully symmetric set* [11]
2. Contains multiples of standard basis vectors ("points along the axes")
3. Is centered at origin

Interrogation grid

- Start with *preliminary grid* $\boldsymbol{s}^* = \{s_1^*, \ldots, s_n^*\}$
- Recall: H is Hessian of $\log f$ at mode
- Take eigendecomposition $-H^{-1} = VDV^\top$ and let $T := V\sqrt{D}$

Interrogation grid

- Start with *preliminary grid* $\boldsymbol{s}^* = \{s_1^*, \ldots, s_n^*\}$
- Recall: H is Hessian of $\log f$ at mode
- Take eigendecomposition $-H^{-1} = VDV^\top$ and let $T := V\sqrt{D}$
- Interrogation points: $s_i = Ts_i^* + \hat{x}$

Interrogation grid

- Start with *preliminary grid* $\boldsymbol{s}^* = \{s_1^*, \ldots, s_n^*\}$
- Recall: H is Hessian of $\log f$ at mode
- Take eigendecomposition $-H^{-1} = VDV^\top$ and let $T := V\sqrt{D}$
- Interrogation points: $s_i = Ts_i^* + \hat{x}$

Grid $\boldsymbol{s}$ is:

1. Centered at mode
2. Aligned w/ "principal axes" of ϕ
3. Scaled in each direction based on H

Interrogation grid

- Start with *preliminary grid* $\boldsymbol{s}^* = \{s_1^*, \ldots, s_n^*\}$
- Recall: H is Hessian of $\log f$ at mode
- Take eigendecomposition $-H^{-1} = VDV^\top$ and let $T := V\sqrt{D}$
- Interrogation points: $s_i = Ts_i^* + \hat{x}$

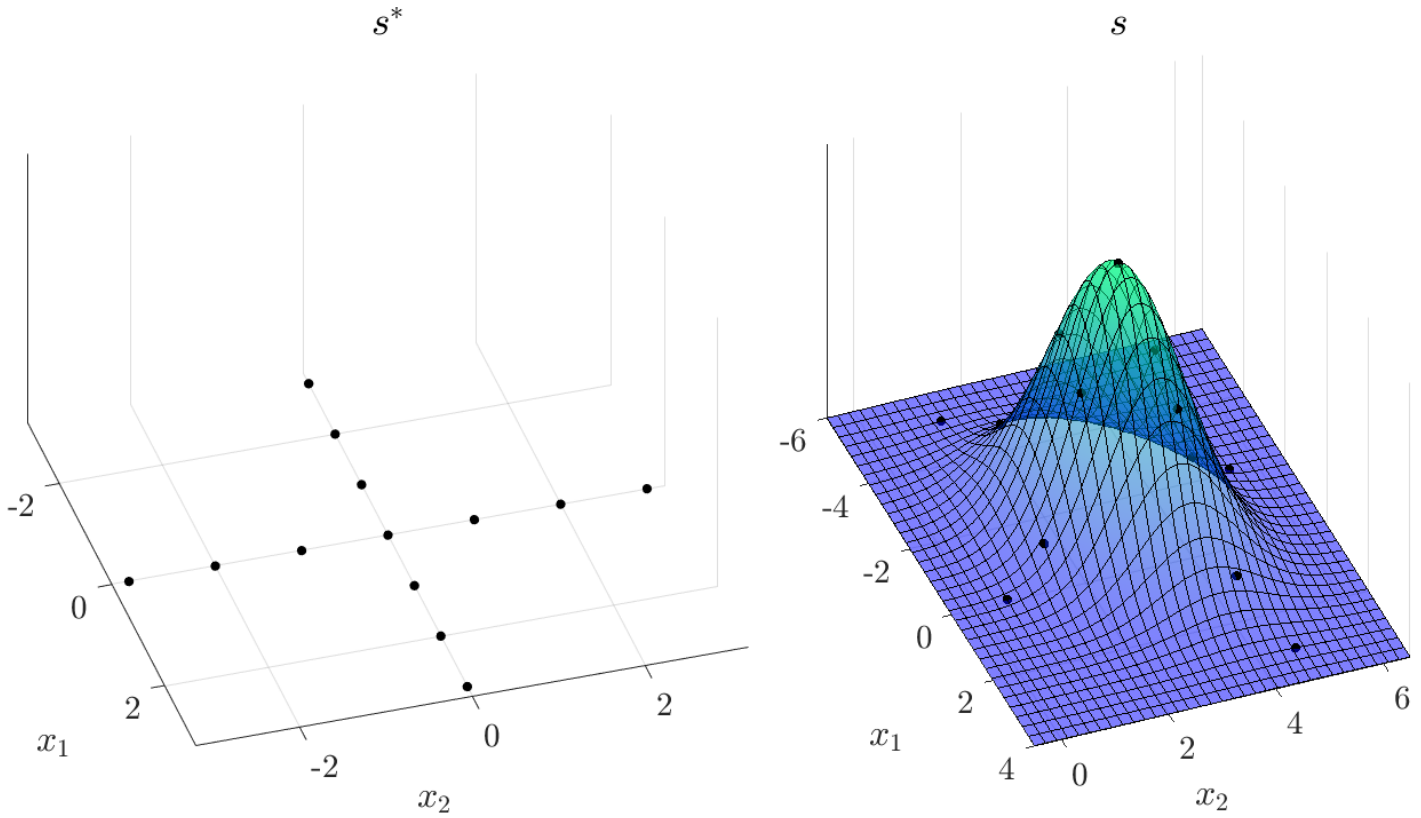
Example:

- Let $X \sim \mathcal{N}\left(\hat{x}, -H^{-1}\right)$, $Y =$ 1st principal component of X
- If $s_i^* = (m, 0, \ldots, 0)$, then s_i is "m standard deviations" (of Y) away from mode (in direction of Y) [10]

Interrogation grid

Example:

- Let $X \sim \mathcal{N}\left(\hat{x}, -H^{-1}\right)$, $Y =$ 1st principal component of X
- If $s_i^* = (m, 0, \ldots, 0)$, then s_i is "m standard deviations" (of Y) away from mode (in direction of Y)

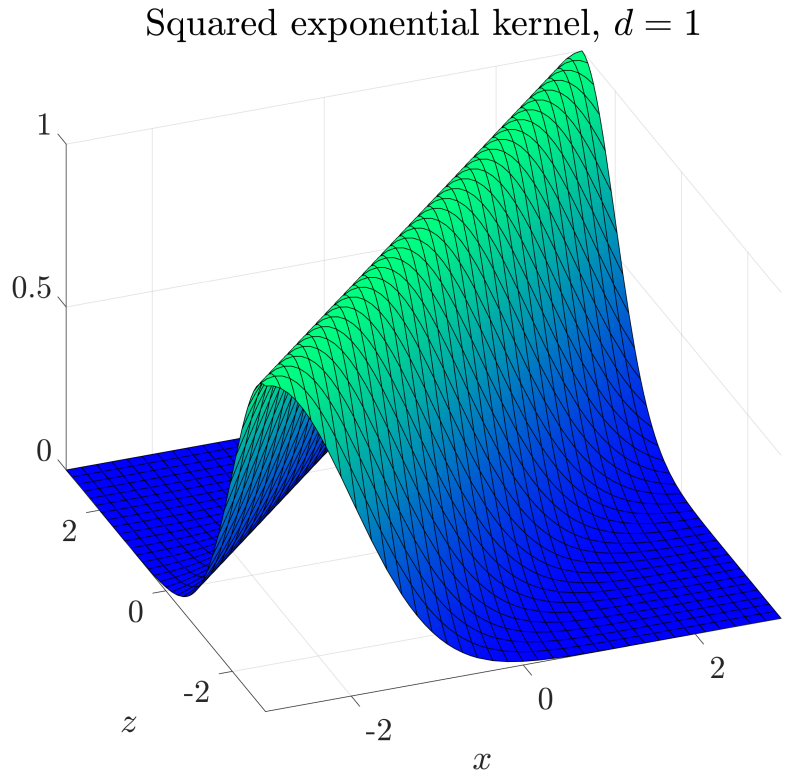


Covariance kernel

Squared exponential kernel [16, 18]:

$$\kappa(x, z) := \alpha^{-d} \exp\left[-\frac{\|x - z\|^2}{2\lambda^2}\right]$$

- $\lambda =$ length-scale (shape)
- $\alpha =$ precision



Covariance kernel

Squared exponential kernel [16, 18]:

$$\kappa(x, z) := \alpha^{-d} \exp\left[-\frac{\|x - z\|^2}{2\lambda^2}\right]$$

- $\lambda =$ length-scale (shape)
- $\alpha =$ precision

Modify based on function of interest f:

$$C_0^x(x, z) = f(\hat{x})^2 \det\left(-H^{-1}\right) \kappa\left(T^{-1}x, T^{-1}z\right)$$

Covariance kernel

Squared exponential kernel [16, 18]:

$$\kappa(x, z) := \alpha^{-d} \exp\left[-\frac{\|x - z\|^2}{2\lambda^2}\right]$$

- $\lambda =$ length-scale (shape)
- $\alpha =$ precision

Modify based on function of interest f:

$$C_0^x(x, z) = f(\hat{x})^2 \det\left(-H^{-1}\right) \kappa\left(T^{-1}x, T^{-1}z\right)$$

Covariance between two points essentially based on *Mahalanobis distance* [1]

Integrating measure

- Recall:

$$F = \int_{\mathbb{R}^d} f(x)\mathrm{d}x = \int_{\mathbb{R}^d} r(x)g(x)\mathrm{d}x = \int_{\mathbb{R}^d} r(x)\mathrm{d}G(x)$$

where G is a Gaussian measure with density g and $r = f/g$ [16, 12].

Integrating measure

- Recall:

$$F = \int_{\mathbb{R}^d} f(x)\mathrm{d}x = \int_{\mathbb{R}^d} r(x)g(x)\mathrm{d}x = \int_{\mathbb{R}^d} r(x)\mathrm{d}G(x)$$

  where G is a Gaussian measure with density g and $r = f/g$ [16, 12].
- Take G s.t. g is slightly wider than ϕ:
- $G = \mathcal{N}\left(\hat{x}, -\gamma^2 H^{-1}\right)$, where $\gamma > 1$ is a hyperparameter (to be discussed later)

Invariance

- These design choices ensure *invariance*:
- For fixed $\boldsymbol{s}^*, \lambda, \alpha, \gamma$, can be shown that diagnostic outcome is unchanged by scaling of f or affine transformation of domain
  - "Standardized" design, like in sigma-point methods (see Särkkä et al. [19])
- All that matters in determining rejection/nonrejection are *relative differences* between f and ϕ — "how Gaussian f is"

Hyperparameter calibration

- Recall diagnostic goals:
  - Quick, "one-size-fits-all"
  - Don't want to reject *every* function — "*good*-**enough**-*ness-of-fit*"
- Thus, given dimension d & preliminary grid $\boldsymbol{s}^*$, want *one* set of hyperparameters $(\lambda, \alpha, \gamma)$ to use for *every* function
- Pick **calibration function** to set hyperparameters

Hyperparameter calibration

- Let $\tau_{\nu,d} = d$-dimensional Student's t density w/ν degrees of freedom
- To calibrate d-dimensional diagnostic (given $\boldsymbol{s}^*$), use $\tau_{\nu_d,d}$, where ν_d is s.t. $L\left(\tau_{\nu_d,d}\right) = 0.95$:

Hyperparameter calibration

- Let $\tau_{\nu,d} = d$-dimensional Student's t density w/ν degrees of freedom
- To calibrate d-dimensional diagnostic (given $\boldsymbol{s}^*$), use $\tau_{\nu_d,d}$, where ν_d is s.t. $L\left(\tau_{\nu_d,d}\right) = 0.95$:
  1. Set $\gamma = \sqrt{1.5\left(\nu_d + d\right) / \left(\nu_d + d - 3\right)}$ (heuristic: ensures g is a bit wider than $\tau_{\nu_d,d}$)

Hyperparameter calibration

- Let $\tau_{\nu,d} = d$-dimensional Student's t density w/ν degrees of freedom
- To calibrate d-dimensional diagnostic (given $\boldsymbol{s}^*$), use $\tau_{\nu_d,d}$, where ν_d is s.t. $L\left(\tau_{\nu_d,d}\right) = 0.95$:
  1. Set $\gamma = \sqrt{1.5\left(\nu_d + d\right) / \left(\nu_d + d - 3\right)}$ (heuristic: ensures g is a bit wider than $\tau_{\nu_d,d}$)
  2. Set λ s.t. posterior diagnostic mean m_1 is close to true integral $\int \tau_{\nu_d,d} = 1$
     - If possible, set λ s.t. ("unweighted") posterior mean function $m_1^x \cdot g$ is close to $\tau_{\nu_d,d}$ throughout $\mathbb{R}^d$

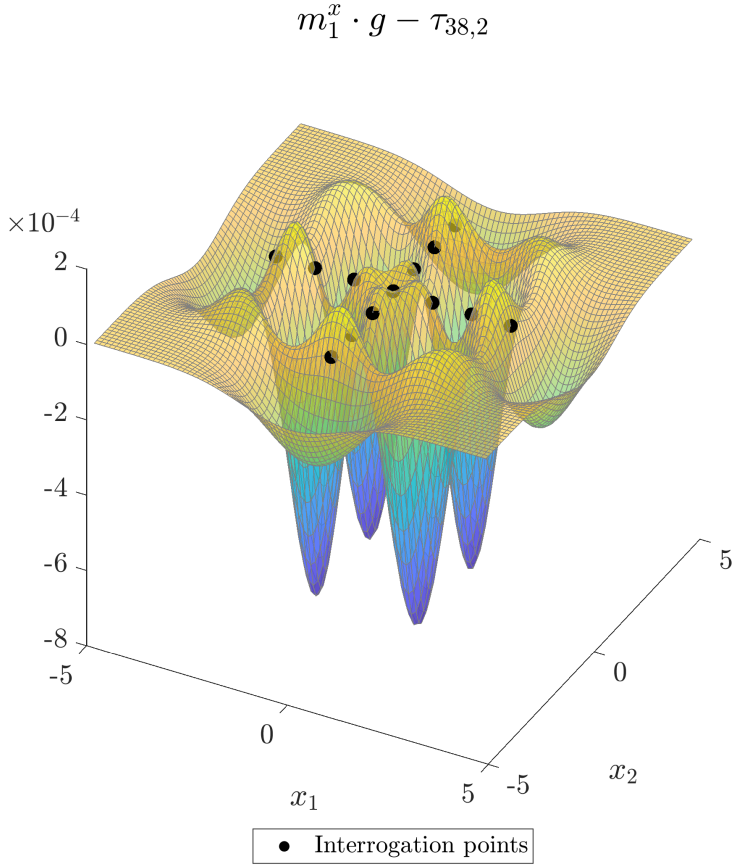
Hyperparameter calibration

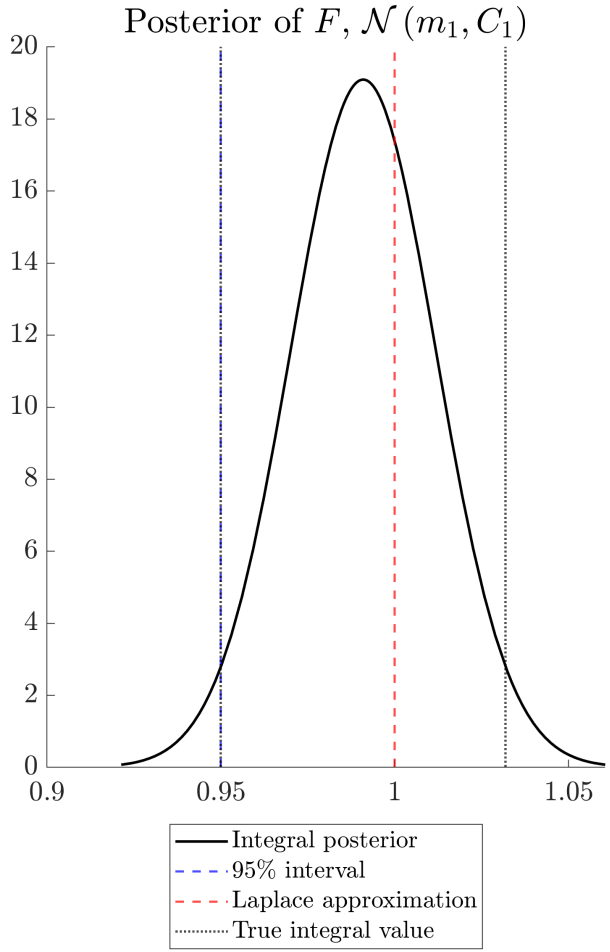
- Let $\tau_{\nu,d} = d$-dimensional Student's t density w/ν degrees of freedom
- To calibrate d-dimensional diagnostic (given $\boldsymbol{s}^*$), use $\tau_{\nu_d,d}$, where ν_d is s.t. $L\left(\tau_{\nu_d,d}\right) = 0.95$:
  1. Set $\gamma = \sqrt{1.5\left(\nu_d + d\right) / \left(\nu_d + d - 3\right)}$ (heuristic: ensures g is a bit wider than $\tau_{\nu_d,d}$)
  2. Set λ s.t. posterior diagnostic mean m_1 is close to true integral $\int \tau_{\nu_d,d} = 1$
     - If possible, set λ s.t. ("unweighted") posterior mean function $m_1^x \cdot g$ is close to $\tau_{\nu_d,d}$ throughout $\mathbb{R}^d$
  3. Set α s.t.$L\left(\tau_{\nu_d,d}\right)$ is on boundary of rejection region

Example: two-dimensional calibration

In low dimensions, can numerically approximate "L^2 error" $\int_{\mathbb{R}^2} \left(m_1^x(x)g(x) - \tau_{38,2}(x)\right)^2 \mathrm{d}x$ and minimize w.r.t. λ

$$\lambda = 4.2241,\ \gamma = 1.2734,\ \alpha = 0.023142$$

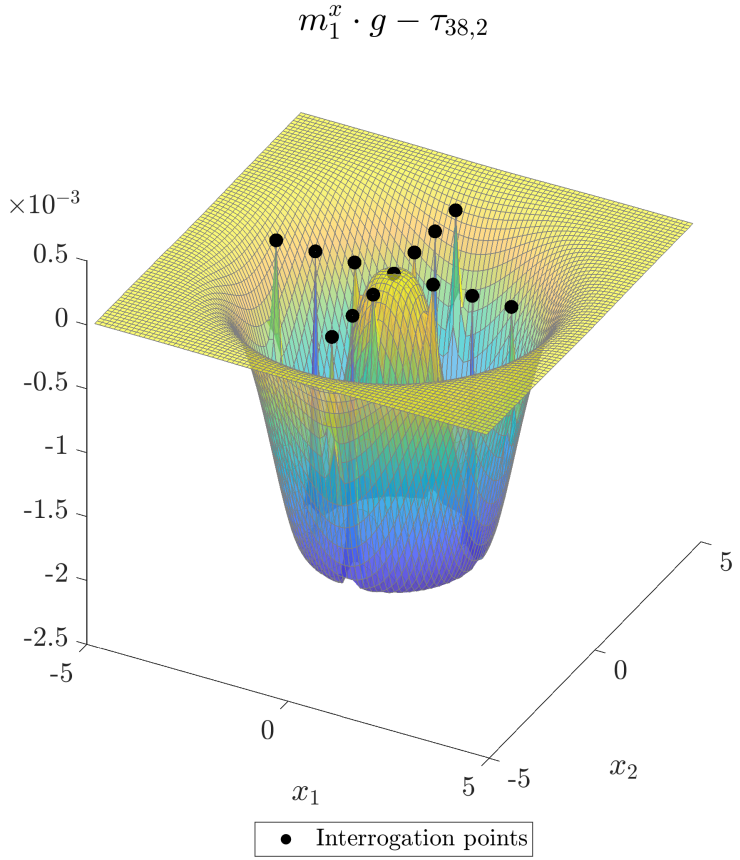


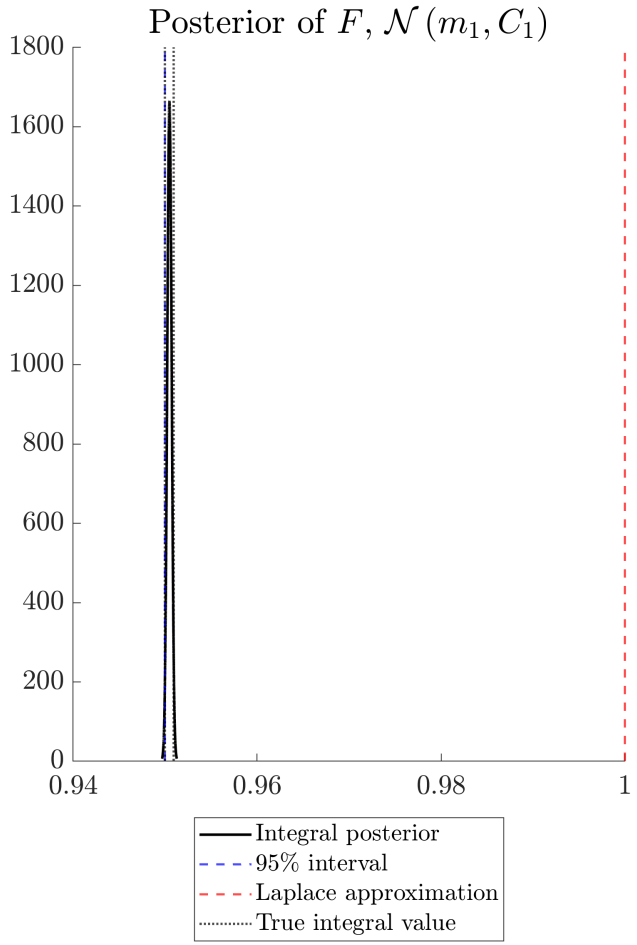


Example: two-dimensional calibration

Low λ results in *undersmoothing* (no interpolation b/t interrogation points)

$$\lambda = 0.0729,\ \gamma = 1.2734,\ \alpha = 25.2372$$

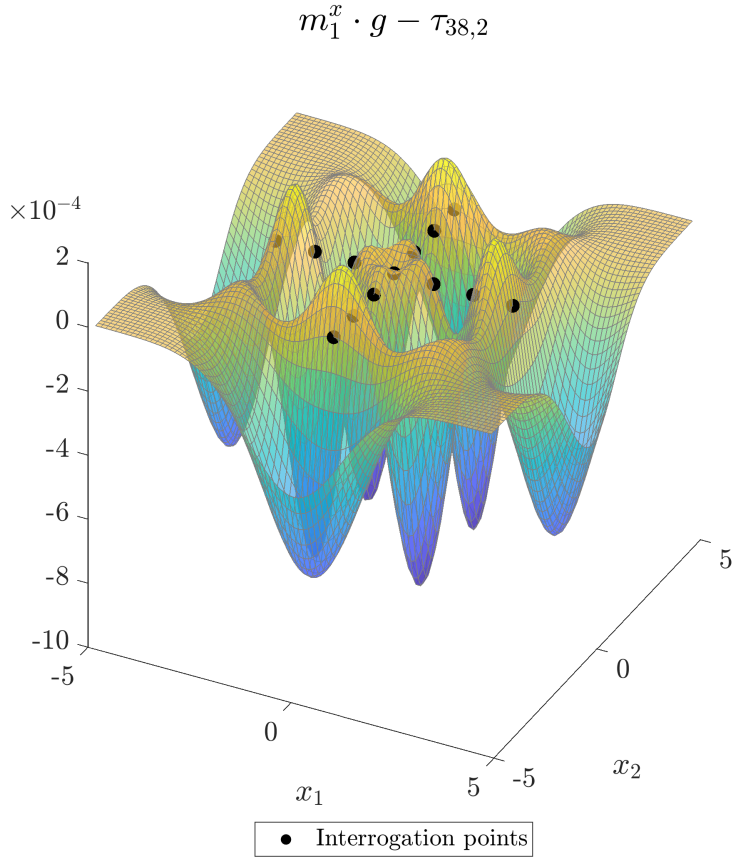


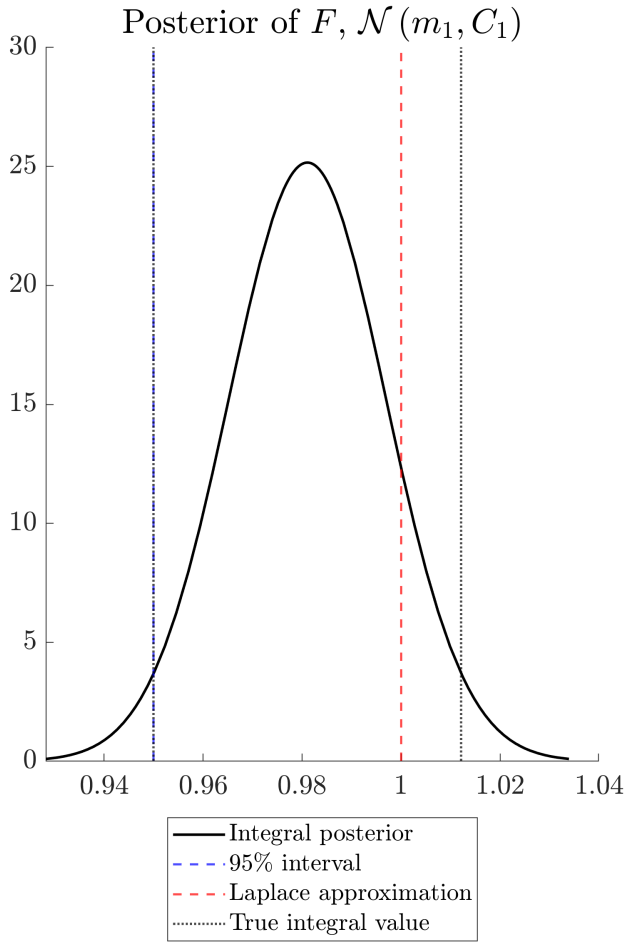


Example: two-dimensional calibration

High γ increases sensitivity w.r.t. λ, which causes *oversmoothing* (extrapolation problems) when λ slightly too large

$$\lambda = 1.3,\ \gamma = 3,\ \alpha = 1.39$$





Rationale

- Calibration function $\tau_{\nu_d,d}$ is "borderline" — just Gaussian enough not to reject L.A.

Rationale

- Calibration function $\tau_{\nu_d,d}$ is "borderline" — just Gaussian enough not to reject L.A.
- Recall: posterior integral mean = LA + correction
- Consider "normalized correction term"

$$\Delta(f) := \frac{\sqrt{\det(-H)}}{f(\hat{x})} \left[\int_{\mathbb{R}^d} C_0^x(z, \boldsymbol{s}) \mathrm{d}G(z)\right]^\top [C_0^x(\boldsymbol{s}, \boldsymbol{s})]^{-1} (r(\boldsymbol{s}) - m_0^x(\boldsymbol{s}))$$

Rationale

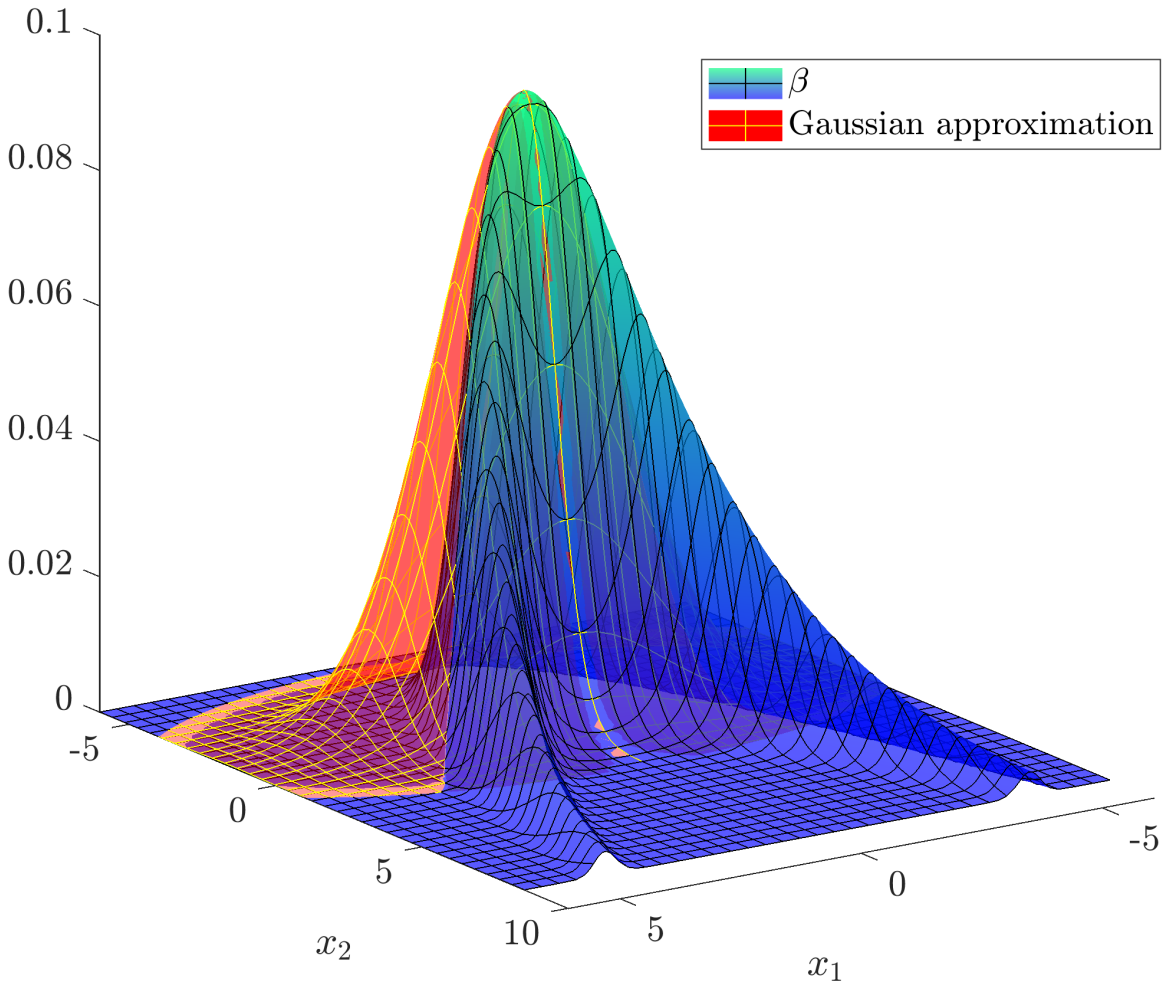
- Calibration function $\tau_{\nu_d,d}$ is "borderline" — just Gaussian enough not to reject L.A.
- Recall: posterior integral mean = LA + correction
- Consider "normalized correction term"

$$\Delta(f) := \frac{\sqrt{\det(-H)}}{f(\hat{x})} \left[\int_{\mathbb{R}^d} C_0^x(z, \boldsymbol{s}) \mathrm{d}G(z)\right]^\top [C_0^x(\boldsymbol{s}, \boldsymbol{s})]^{-1} (r(\boldsymbol{s}) - m_0^x(\boldsymbol{s}))$$

- Can be shown that, given $\boldsymbol{s}^*, \lambda, \alpha, \gamma$,
  - LA rejected for $f \Leftrightarrow |\Delta(f)| > |\Delta(\tau_{\nu_d,d})|$ ("f not Gaussian enough")
  - LA not rejected for $f \Leftrightarrow |\Delta(f)| \leq |\Delta(\tau_{\nu_d,d})|$ ("f sufficiently Gaussian")

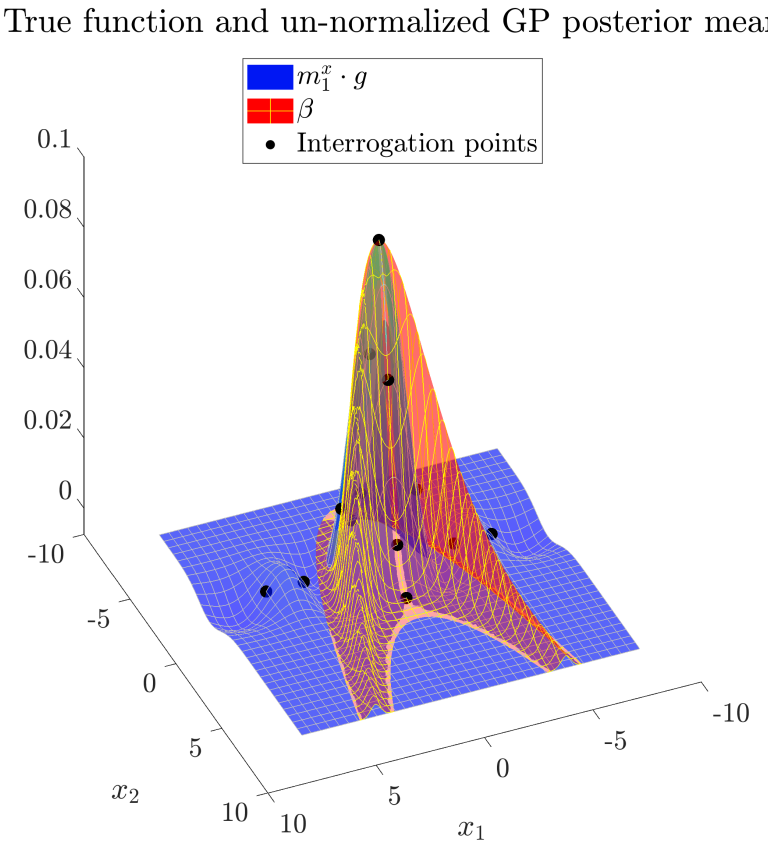
Example: two-dimensional banana

- "Twist" one coordinate of bivariate Gaussian to get banana-shaped β [7]
- Turns out that $L(\beta) = \int \beta = 1$ (LA is true)
- Just "coincidence" — β clearly not well-approximated by Gaussian



Example: two-dimensional banana

- Calibrated diagnostic rejects due to non-Gaussian shape
- This is fine — main focus is *assumptions underlying* LA



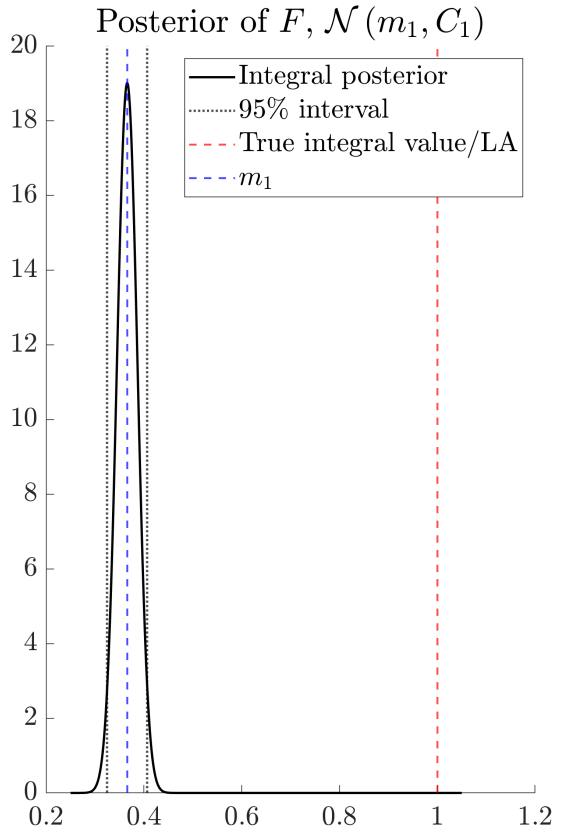


Table of Contents

The curse of dimensionality

Is f "Gaussian enough" to justify the LA?

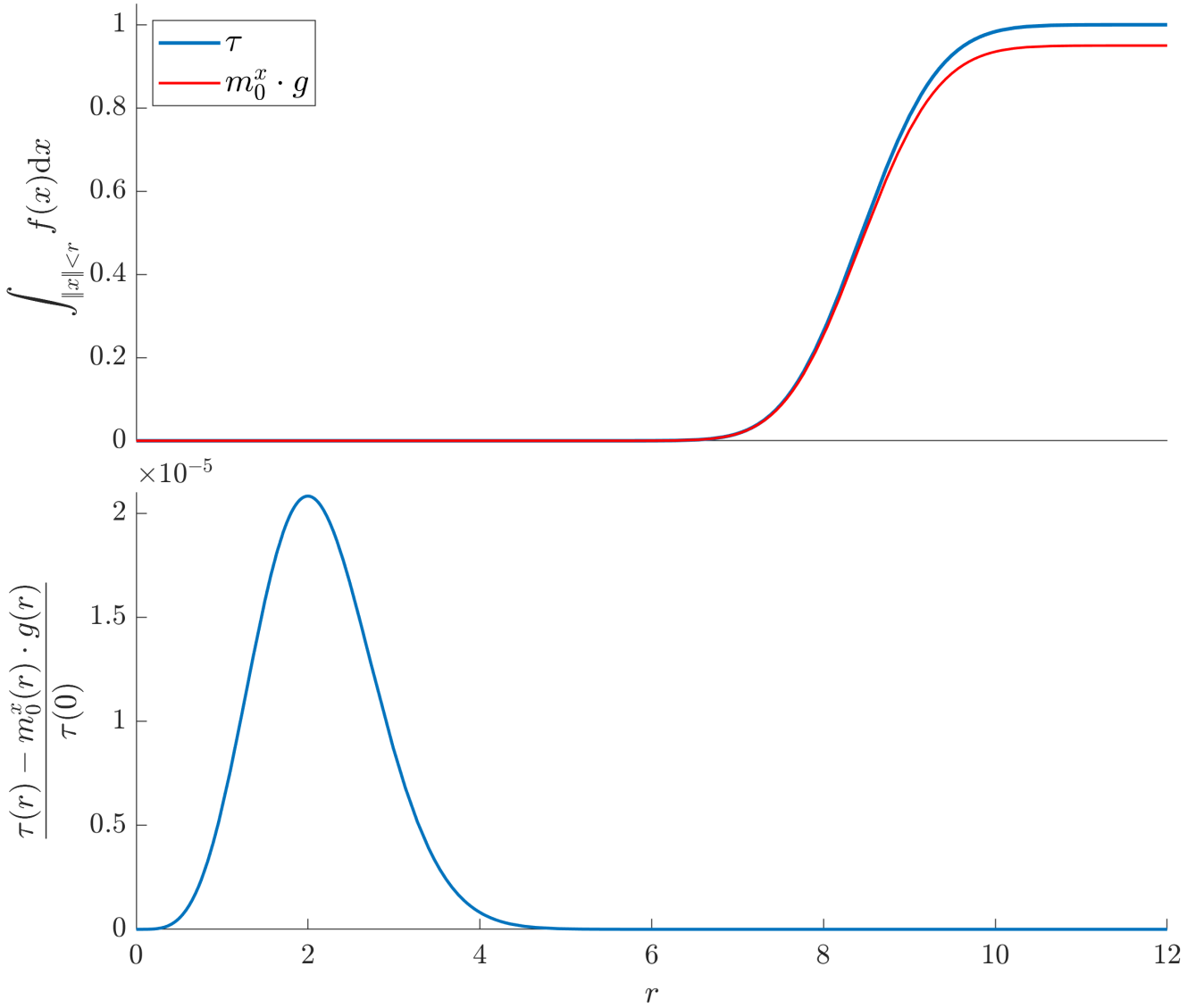
The curse of dimensionality

Is f "Gaussian enough" to justify the LA?

- In high dimensions, trickier to answer this question — info about a function's shape is more divorced from the value of its integral
- Most obvious "shape" information is in high-density region near mode, but for high d *most mass is in tails* [3]
- e.g. for standard Gaussian, most mass is $\mathcal{O}\left(\sqrt{d}\right)$ from origin in shell of width $\mathcal{O}(1)$ [4]

Example: 72-dimensional t density w/25921 d.f.

Compounded over large volumes, tiny shape differences make all the difference

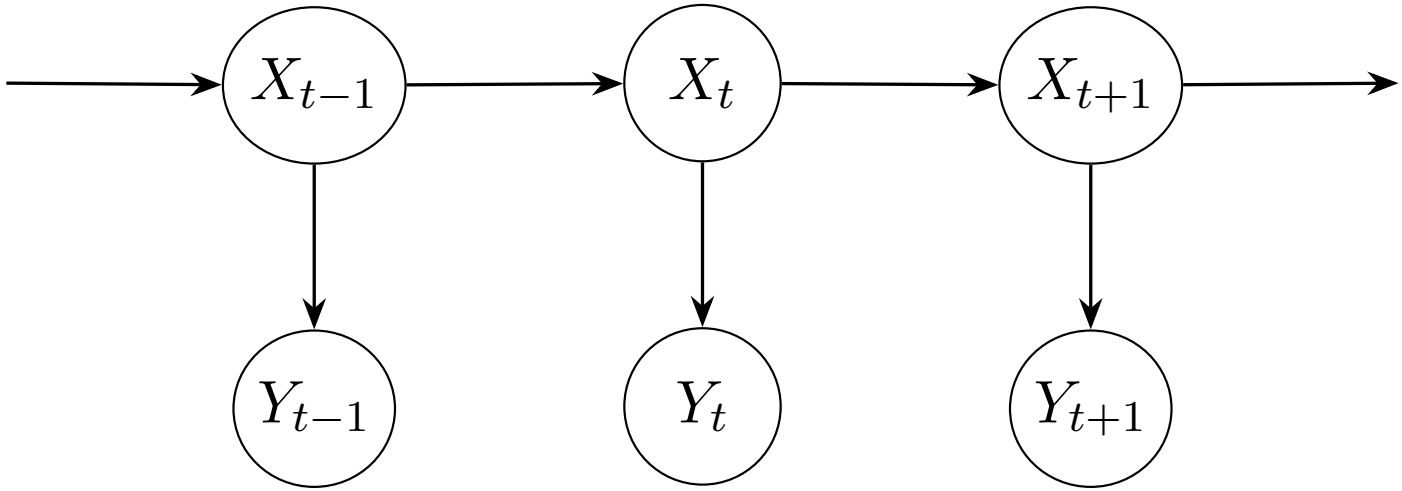


The curse of dimensionality

- Interrogation points should be in "typical set" where most mass lies [3]
  - Assessing if f is Gaussian enough *in the tails* to justify LA
- **Idea**: put preliminary points $\boldsymbol{s}^*$ at origin and $\pm\sqrt{d}e_i$, where $e_i = i^{\text{th}}$ standard basis vector
- $2d+1$ points in "cross-shaped" grid; similar point set as *cubature Kalman filter* [9]
- Calibration is harder than low dimensions:
  - d too high to minimize "L^2 error"
  - Shape differences too tiny to see
- Thus, can only ensure integral estimate m_1 is close to true value for calibration function

Example: stock assessment model

Recall SSM:



$$p_{x,y}\left(\boldsymbol{x}, \boldsymbol{y} \mid \theta\right) = p\left(x_1 \mid \theta\right) \left[\prod_{t=2}^{T} p\left(x_t \mid x_{t-1}, \theta\right)\right] \left[\prod_{t=1}^{T} p\left(y_t \mid x_t, \theta\right)\right]$$

Given $\boldsymbol{y}$, maximize $L_\theta\left(p_{x,y}\right) \approx p_y\left(\boldsymbol{y} \mid \theta\right)$ w.r.t. θ to get $\hat{\theta}$

Example: stock assessment model

- *Stock assessment model* (SAM): nonlinear SSM for fish populations
- y_t: observed fish abundances for year t
- x_t: true abundances, fishing mortality rates
- θ: correlation, variance, scaling parameters
- Aeberhard et al. [2], Nielsen and Berg [15]

The process equation describes the dynamics in the unobserved states and is based on the conditional expectation of the current states given the previous states:

$$\mathrm{E}[\boldsymbol{X}_t|\boldsymbol{X}_{t-1}] = \begin{cases} \log N_{1,t} &= \log N_{1,t-1} \\ \log N_{a,t} &= \log N_{a-1,t-1} - F_{a-1,t-1} - M_{a-1,t-1}, \qquad 2 \le a < A \\ \log N_{A,t} &= \log[N_{A-1,t-1}\exp(-F_{A-1,t-1} - M_{A-1,t-1}) \\ &\quad + N_{A,t-1}\exp(-F_{A,t-1} - M_{A,t-1})] \\ \log F_{a,t} &= \log F_{a,t-1}, \qquad 1 \le a \le A, \end{cases}$$

where A denotes the largest age class. These equations assume a random walk for $\log N_{1,t}$ and for the whole vector $(\log F_{1,t}, \ldots, \log F_{A,t})^\top$, a survival process for $\log N_{a,t}$ where the combination of F and M represents total mortality, and a modified survival process for the plus group in $\log N_{A,t}$. The corresponding distribution $P_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ is a multivariate Gaussian with zero mean vector. The first A Gaussian error components are independent, while we enforce a first-order autoregressive correlation structure for the others:

$$\mathrm{Cor}[\log(F_{a,t}), \log(F_{\tilde{a},t})] = \rho^{|a-\tilde{a}|},$$

where the between-age correlation ρ is an element of $\boldsymbol{\theta}$. Other fixed parameters include four separate variances: one for recruitment ($\sigma^2_{N_{a=1}}$), one for survival ($\sigma^2_{N_{a>1}}$), one for fishing mortality at age 1 ($\sigma^2_{F_{a=1}}$), and one for fishing mortality at older ages ($\sigma^2_{F_{a\ge 2}}$).

The observation equation relates the unobserved states to the observed response variables through a conditional expectation:

$$\mathrm{E}[\boldsymbol{Y}_t|\boldsymbol{X}_t] = \begin{cases} \log C_{a,t} = \log\left[\frac{F_{a,t}}{Z_{a,t}}(1-\exp(-Z_{a,t}))N_{a,t}\right] \\ \log I^{(s)}_{a,t} = \log\left[Q^{(s)}_a \exp(-Z_{a,t}\frac{D^{(s)}}{365})N_{a,t}\right], \end{cases} \qquad 1 \le a \le A,$$

where $s = 1, 2$ identifies the surveys, the largest age class A is 5 for $s = 1$ and 4 for $s = 2$, $Z_{a,t} = M_{a,t} + F_{a,t}$ is the total mortality rate, $D^{(s)}$ is the number of days into the year when survey (s) was conducted, and $Q^{(s)}_a$ are so-called catchability coefficients that scale the survey relative indices to the stock abundance. The catchabilities are unknown parameters that need to be estimated, there are nine of them, as they are distinct for each age class and each survey. Auxiliary information and expertise from fisheries scientists cast doubt on the reliability of the absolute level of the catches between 1993 and 2005, hence extra catch scaling parameters τ_t are added (and estimated) for these years:

$$\log C_{a,t} = \log\left[\frac{1}{\tau_t}\frac{F_{a,t}}{Z_{a,t}}(1-\exp(-Z_{a,t}))N_{a,t}\right], \quad t \in \{1993, \ldots, 2005\}.$$

Aeberhard et al. [2]

Example: stock assessment model

- Fit SAM's to North Sea cod data [2]
- Model 1: data from $t = 1970, \ldots, 1975$
- Model 2: data from $t = 2005, \ldots, 2011$
- $x_t \in \mathbb{R}^{12} \Rightarrow d = 12 \times 6 = 72$
- For each model: use TMB package [13] to estimate $\hat{\theta}$ w/LA
- Use diagnostic to check if $p_{x,y}\left(\boldsymbol{x}, \boldsymbol{y} \mid \hat{\theta}\right)$ is Gaussian enough to justify $p_y\left(\boldsymbol{y} \mid \hat{\theta}\right) \approx L_{\hat{\theta}}\left(p_{x,y}\right)$

The process equation describes the dynamics in the unobserved states and is based on the conditional expectation of the current states given the previous states:

$$\mathrm{E}[\boldsymbol{X}_t|\boldsymbol{X}_{t-1}] = \begin{cases} \log N_{1,t} &= \log N_{1,t-1} \\ \log N_{a,t} &= \log N_{a-1,t-1} - F_{a-1,t-1} - M_{a-1,t-1}, \qquad 2 \le a < A \\ \log N_{A,t} &= \log[N_{A-1,t-1}\exp(-F_{A-1,t-1} - M_{A-1,t-1}) \\ &\quad + N_{A,t-1}\exp(-F_{A,t-1} - M_{A,t-1})] \\ \log F_{a,t} &= \log F_{a,t-1}, \qquad 1 \le a \le A, \end{cases}$$

where A denotes the largest age class. These equations assume a random walk for $\log N_{1,t}$ and for the whole vector $(\log F_{1,t}, \ldots, \log F_{A,t})^\top$, a survival process for $\log N_{a,t}$ where the combination of F and M represents total mortality, and a modified survival process for the plus group in $\log N_{A,t}$. The corresponding distribution $P_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ is a multivariate Gaussian with zero mean vector. The first A Gaussian error components are independent, while we enforce a first-order autoregressive correlation structure for the others:

$$\mathrm{Cor}[\log(F_{a,t}), \log(F_{\bar{a},t})] = \rho^{|a-\bar{a}|},$$

where the between-age correlation ρ is an element of $\boldsymbol{\theta}$. Other fixed parameters include four separate variances: one for recruitment ($\sigma^2_{N_{a=1}}$), one for survival ($\sigma^2_{N_{a>1}}$), one for fishing mortality at age 1 ($\sigma^2_{F_{a=1}}$), and one for fishing mortality at older ages ($\sigma^2_{F_{a\ge 2}}$).

The observation equation relates the unobserved states to the observed response variables through a conditional expectation:

$$\mathrm{E}[\boldsymbol{Y}_t|\boldsymbol{X}_t] = \begin{cases} \log C_{a,t} = \log\left[\frac{F_{a,t}}{Z_{a,t}}(1 - \exp(-Z_{a,t}))N_{a,t}\right] \\ \log I^{(s)}_{a,t} = \log\left[Q^{(s)}_a \exp(-Z_{a,t}\frac{D^{(s)}}{365})N_{a,t}\right], \end{cases} \qquad 1 \le a \le A,$$

where $s = 1, 2$ identifies the surveys, the largest age class A is 5 for $s = 1$ and 4 for $s = 2$, $Z_{a,t} = M_{a,t} + F_{a,t}$ is the total mortality rate, $D^{(s)}$ is the number of days into the year when survey (s) was conducted, and $Q^{(s)}_a$ are so-called catchability coefficients that scale the survey relative indices to the stock abundance. The catchabilities are unknown parameters that need to be estimated, there are nine of them, as they are distinct for each age class and each survey. Auxiliary information and expertise from fisheries scientists cast doubt on the reliability of the absolute level of the catches between 1993 and 2005, hence extra catch scaling parameters τ_t are added (and estimated) for these years:

$$\log C_{a,t} = \log\left[\frac{1}{\tau_t}\frac{F_{a,t}}{Z_{a,t}}(1 - \exp(-Z_{a,t}))N_{a,t}\right], \quad t \in \{1993, \ldots, 2005\}.$$

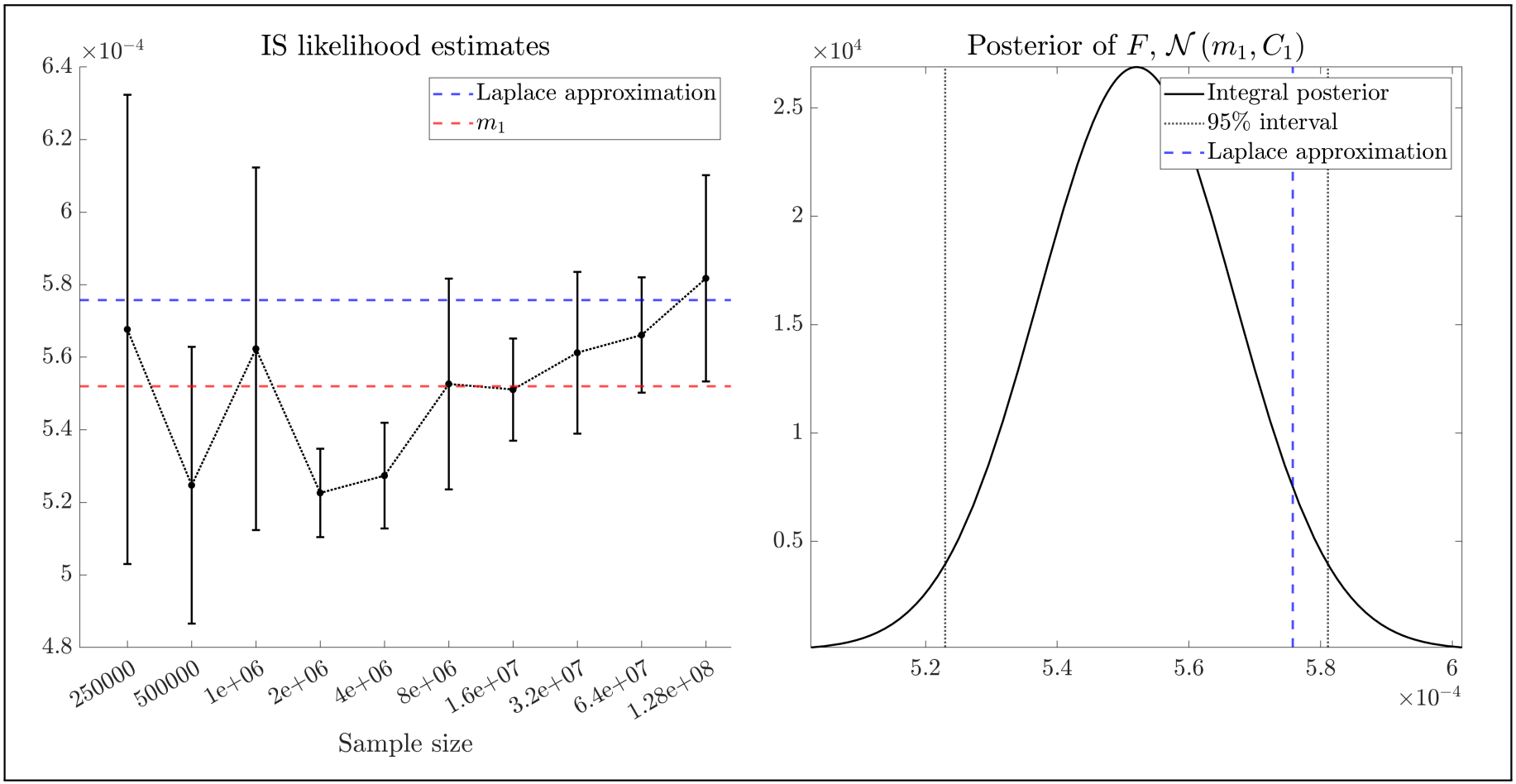
Aeberhard et al. [2]

Another way: `checkConsistency`

- TMB package contains `checkConsistency` function to validate LA [13]
- Generate n datasets $\boldsymbol{y}^* \sim p_y\left(\cdot \mid \hat{\theta}\right)$, approximate score test for $\mathbb{E}_y\left[\nabla_\theta \log L_{\hat{\theta}}\left(p_{x,y}\right)\right] = 0$
- Checks if p_y and $L\left(p_{x,y}\right)$ are similar *as functions of* $\boldsymbol{y}$
- Diagnostic: checks if $p_{x,y}$ is Gaussian enough *as function of* $\boldsymbol{x}$ to justify LA for *observed* $\boldsymbol{y}$
- Still useful to compare

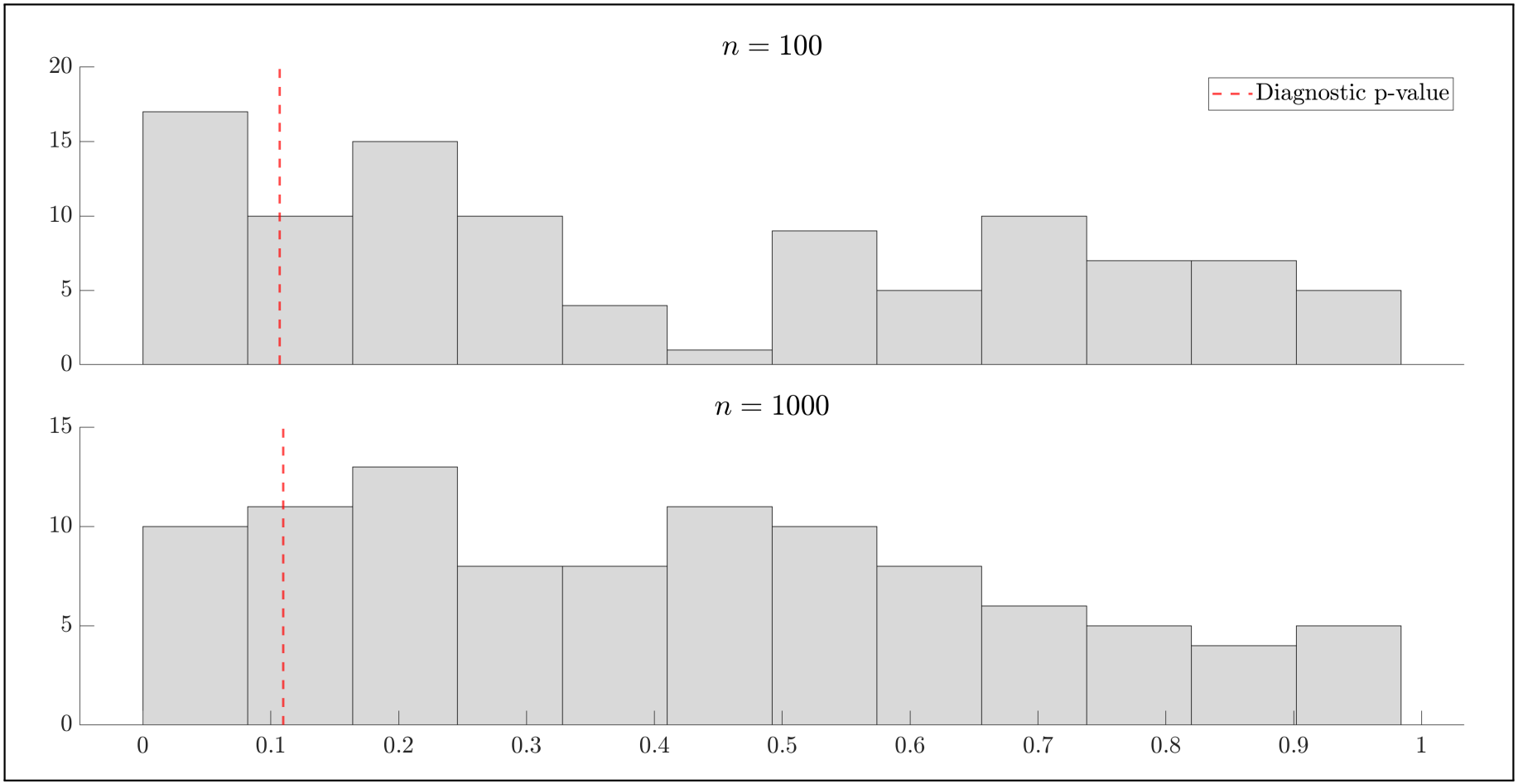
Results: 1970–1975 data

Importance samplers used for rough idea of "ground truth"

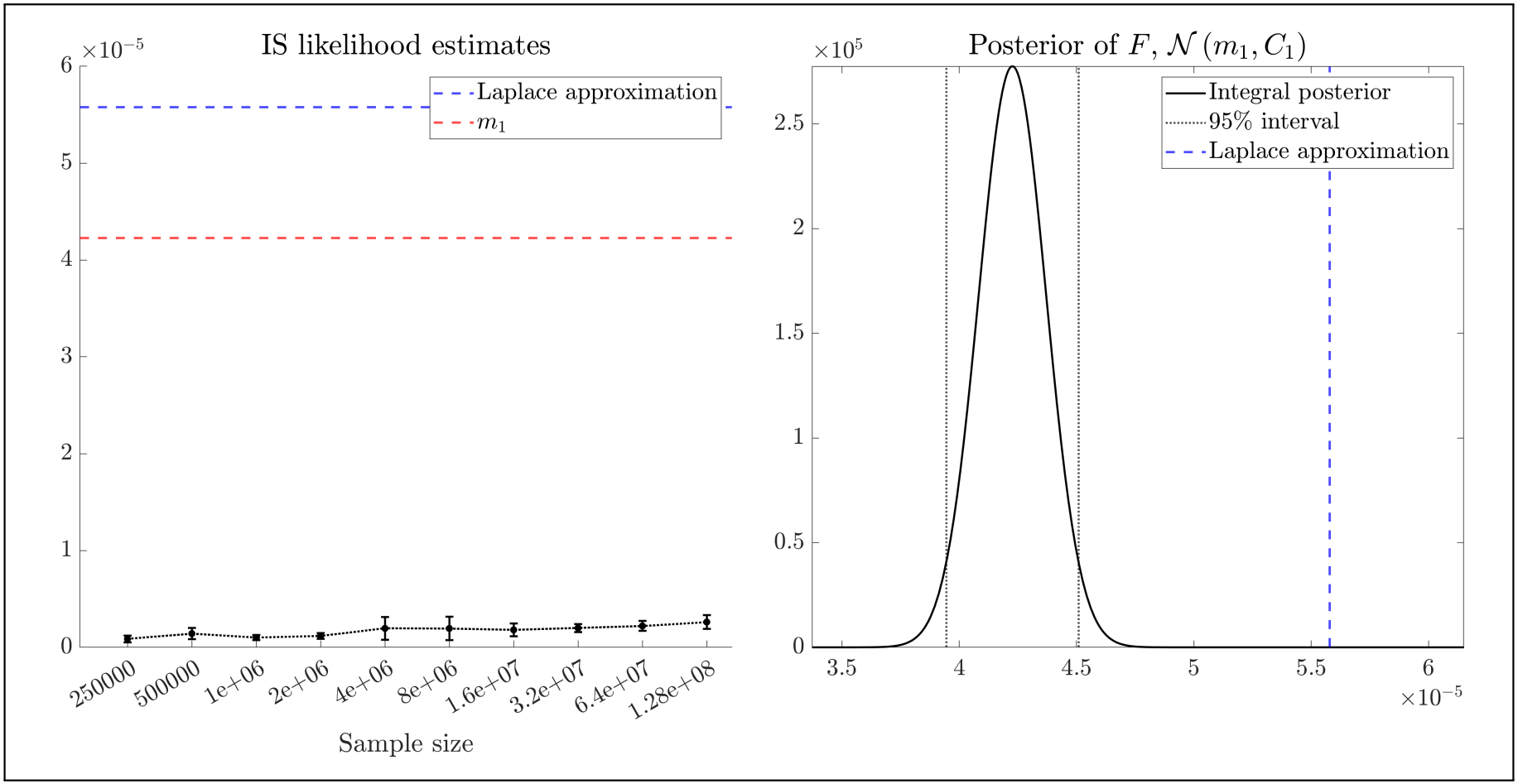


Results: 1970–1975 data

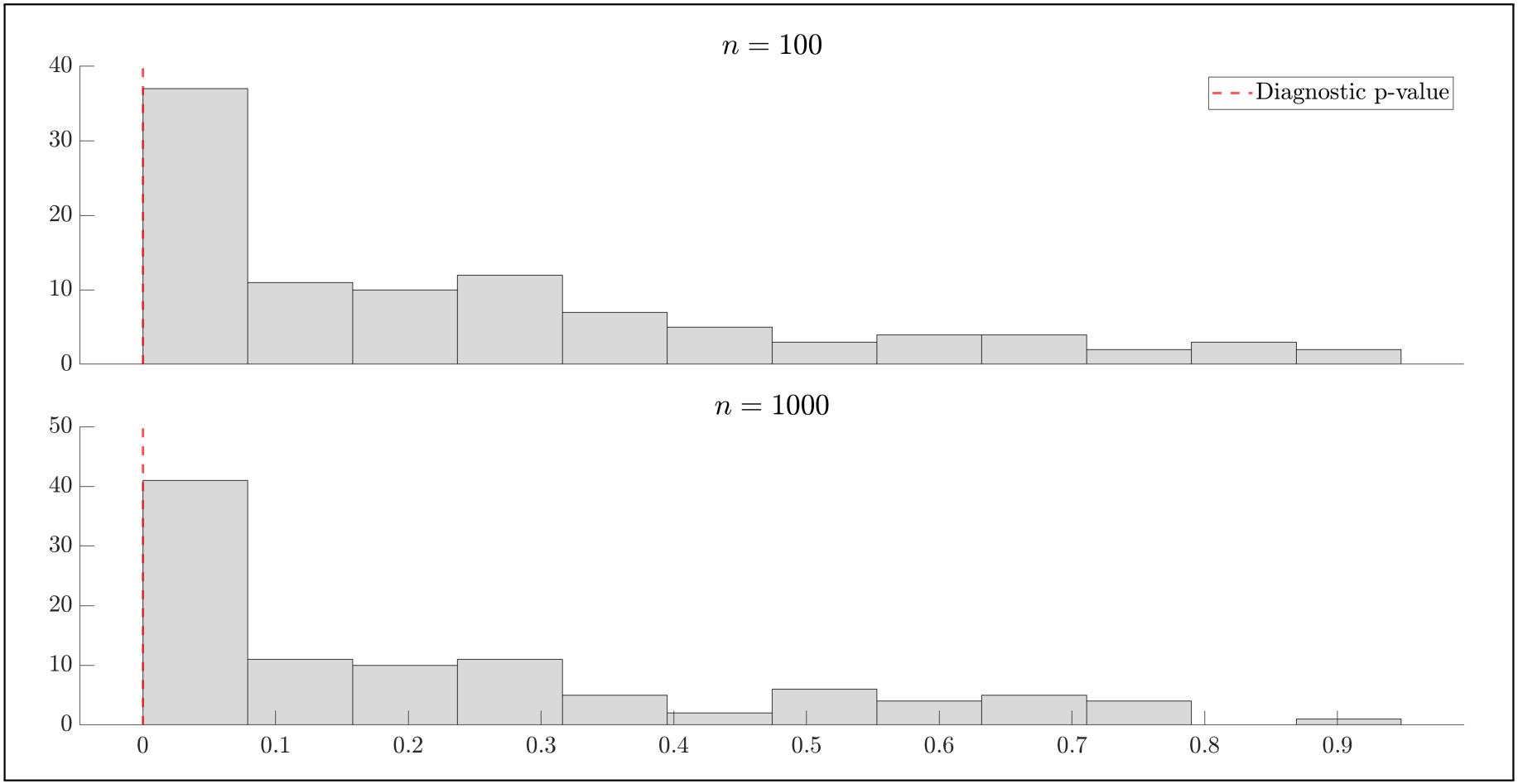
p-values from repeated `checkConsistency` runs with $n = 100, 1000$ simulated datasets



Results: 2005–2011 data



Results: 2005–2011 data



Computation time

Time (seconds)	1970–1975	2005–2011
`checkConsistency`, $n = 100$	2.511 ± 0.035	7.367 ± 0.136
`checkConsistency`, $n = 1000$	25.115 ± 0.152	73.584 ± 0.489
Diagnostic	0.009 ± 0.007	0.012 ± 0.0003

...and even longer for importance samplers

Table of Contents

Discussion/conclusions

- "Medium-effort", one-size-fits-all tool to check if function shape justifies use of LA
  - *Diagnostic*: accurate integral estimation is secondary concern
- BQ provides natural, probabilistic way to build tool
- "Good-enough-ness-of-fit"
  - Don't want to reject every slight deviation from Gaussian shape, esp. in high dimensions
- High dimensions require more care — harder to determine integral w/limited shape info
- Future work:
  1. Fold into model fitting: check $L_\theta\left(p_{x,y}\right)$ at every iteration, rather than last one
  2. Different interrogation grid structures (e.g. higher-order sparse grids w/fully symmetric methods of Karvonen and Särkkä [11])
  3. Different prior structure: e.g. GP prior on $\log f$ instead of f [6]

References I

[1] Shigeo Abe. Training of support vector machines with Mahalanobis kernels. In *Proc. International Conference on Artificial Neural Networks (ICANN 2005)*, pages 571–576. Springer, Berlin, Heidelberg, 2005. ISBN 3540287558. doi: 10.1007/11550907_90. URL http://www2.eedept.kobe-u.ac.jp/$\sim$abe.

[2] William H Aeberhard, Joanna Mills Flemming, and Anders Nielsen. Review of State-Space Models for Fisheries Science. *Annual Review of Statistics and Its Application*, 5:215–235, 2018. doi: 10.1146/annurev-statistics. URL https://doi.org/10.1146/annurev-statistics-.

[3] Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. Technical report, 2018.

[4] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, jan 2020. ISBN 9781108755528. doi: 10.1017/9781108755528. URL https://www-cambridge-org.proxy.lib.sfu.ca/core/books/foundations-of-data-science/6A43CE830DE83BED6CC5171E62B0AA9E.

[5] François-Xavier Briol, Chris J. Oates, Mark Girolami, Michael A. Osborne, and Dino Sejdinovic. Probabilistic Integration: A Role in Statistical Computation? *Statistical Science*, 34(1):1–22, 2019. URL http://www.

References II

[6] Henry Chai and Roman Garnett. Improving Quadrature for Constrained Integrands. Technical report, 2019.

[7] Heikki Haario, Eero Saksman, and Johanna Tamminen. Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14: 375–395, 1999.

[8] Philipp Hennig, Michael A. Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150142, jul 2015. ISSN 1364-5021. doi: 10.1098/rspa.2015.0142. URL `https://royalsocietypublishing.org/doi/10.1098/rspa.2015.0142`.

[9] Ienkaran and Simon Haykin. Cubature kalman filters. In *IEEE Transactions on Automatic Control*, volume 54, pages 1254–1269, 2009. doi: 10.1109/TAC.2009.2019800.

[10] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag New York, New York, 2 edition, 2002. ISBN 0-387-95442-2. doi: 10.1007/b98835. URL `http://link.springer.com/10.1007/b98835`.

References III

[11] Toni Karvonen and Simo Särkkä. Fully symmetric kernel quadrature. *SIAM Journal on Scientific Computing*, 40(2):A697–A720, mar 2018. ISSN 10957197. doi: 10.1137/17M1121779.

[12] Marc Kennedy. Bayesian quadrature with non-normal approximating functions. *Statistics and Computing*, 8:365–375, 1998.

[13] Kasper Kristensen, Anders Nielsen, Casper W. Berg, Hans Skaug, and Bradley M. Bell. TMB: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, 70(1):1–21, apr 2016. ISSN 15487660. doi: 10.18637/jss.v070.i05. URL `https://www.jstatsoft.org/index.php/jss/article/view/v070i05/v70i05.pdfhttps://www.jstatsoft.org/index.php/jss/article/view/v070i05`.

[14] D. V. Lindley. The Use of Prior Probability Distributions in Statistical Inference and Decisions. In *Proc. 4th Berkeley Symp. on Math. Stat. and Prob.*, volume 4, pages 453–468. University of California Press, jan 1961.

[15] Anders Nielsen and Casper W. Berg. Estimation of time-varying selectivity in stock assessments using state-space models. *Fisheries Research*, 158:96–101, 2014. doi: 10.1016/j.fishres.2014.01.014.

References IV

[16] A. O'Hagan. Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference*, 29:245–260, 1991. doi: 10.1016/0378-3758(91)90002-V.

[17] Michael Osborne. *Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature*. PhD thesis, University of Oxford, 2010.

[18] Carl Edward Rasmussen and Zoubin Ghahramani. Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 489–496, 2003. URL `http://www.gatsby.ucl.ac.uk`.

[19] Simo Särkkä, Jouni Hartikainen, Lennart Svensson, and Fredrik Sandblom. On the relation between Gaussian process quadratures and sigma-point methods. Technical report, 2015.

End

Acknowledgements: thanks to Richard Lockhart, Michael Osborne, and Anders Nielsen for their valuable insights.

Thanks for listening!