Assessing Sensitivity to the Stick-Breaking Prior in Bayesian Nonparametrics

May 5th, 2021

University of California, Berkeley

Collaborators

Ryan Giordano
MIT

Runjing (Bryan) Liu
UC Berkeley

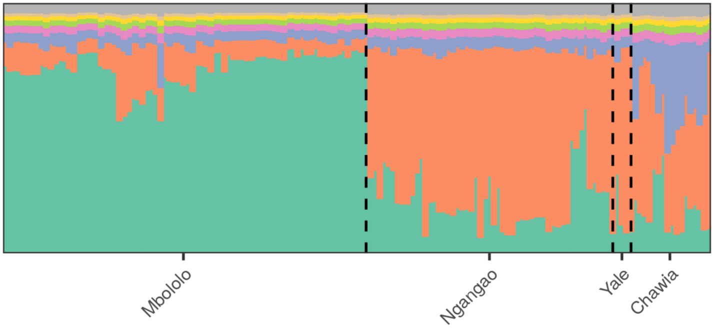
Michael I. Jordan
UC Berkeley

Tamara Broderick
MIT

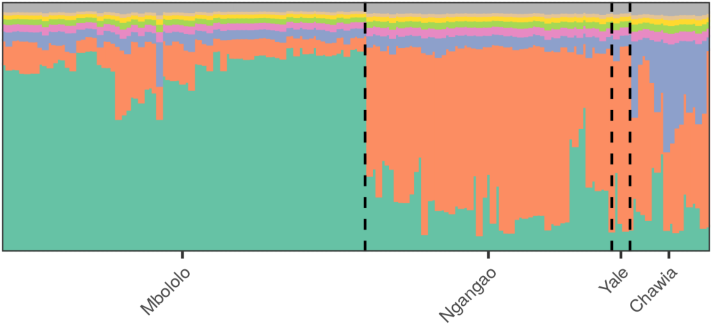
Motivating Example

Inferring population structure from genomic sequences.

- Genetic data from Taita thrush, an endangered bird species native to Kenya [Galbusera et al., 2000]
- Microsatellites sequences of 155 individuals at 7 loci.



Motivating Example

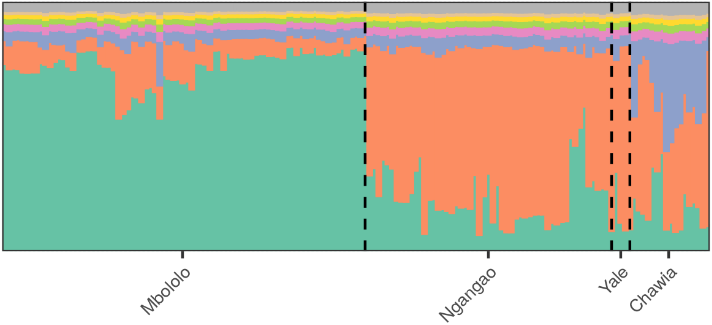


- Three primary populations .
- Many small, rare populations .

Question: How many distinct populations (clusters) are there...

- ...in this dataset?
- ...with more than N loci?
- ...in a future dataset of the same size?

Motivating Example



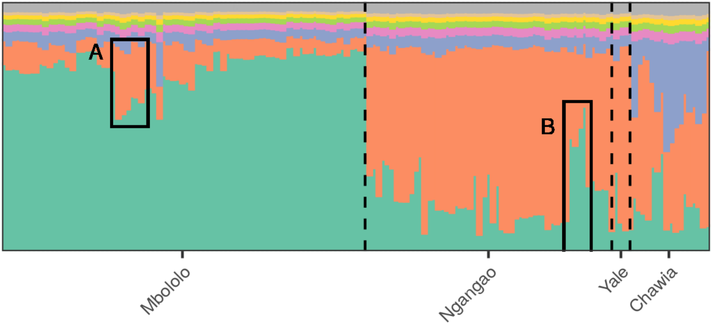
Individuals are generally clustered by geographic locations:

Mbololo ≈ ■ Ngangao ≈ ■ Chawia ≈ ■ + ■ + ■

Question: Which individuals cluster together?

Exceptions to the clustering give evidence of historical migrations.

Motivating Example



Individuals are generally clustered by geographic locations:

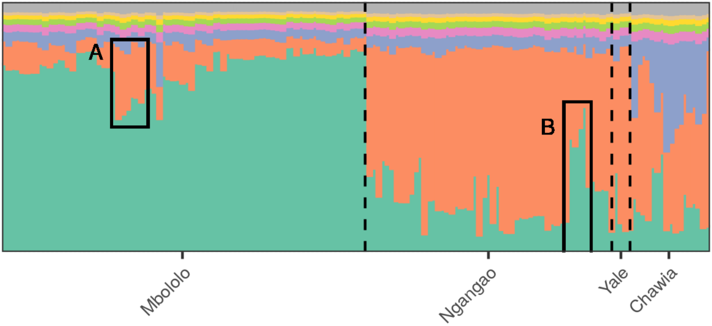
Mbololo ≈ ■ Ngangao ≈ ■ Chawia ≈ ■ + ■ + ■

Question: Which individuals cluster together?

Exceptions to the clustering give evidence of historical migrations.

For example, the groups of individuals in A and B suggest migration between the Mbololo and Ngangao locations.

Motivating Example

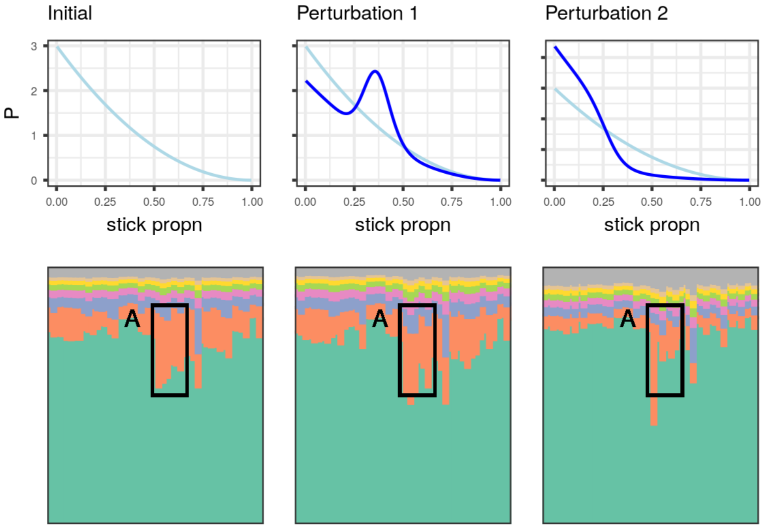


How many distinct clusters are there? Which individuals cluster together?

A **discrete Bayesian nonparametric (BNP)** model makes these questions amenable to Bayesian inference...

...but the answer may depend on the **prior you choose.**

Motivating Example



Research Problem

A discrete Bayesian nonparametric (BNP) model makes scientific questions amenable to Bayesian inference.

Research Problem

A discrete Bayesian nonparametric (BNP) model makes scientific questions amenable to Bayesian inference.

We approximate the exact posterior using variational Bayes (VB).

Research Problem

A discrete Bayesian nonparametric (BNP) model makes scientific questions amenable to Bayesian inference.

We approximate the exact posterior using variational Bayes (VB).

Question: How sensitive is the VB approximation, and the resulting inferences, to BNP model choices?

Research Problem

A discrete Bayesian nonparametric (BNP) model makes scientific questions amenable to Bayesian inference.

We approximate the exact posterior using variational Bayes (VB).

Question: How sensitive is the VB approximation, and the resulting inferences, to BNP model choices?

Problem: Re-running VB for multiple model choices is expensive.

Research Problem

A discrete Bayesian nonparametric (BNP) model makes scientific questions amenable to Bayesian inference.

We approximate the exact posterior using variational Bayes (VB).

Question: How sensitive is the VB approximation, and the resulting inferences, to BNP model choices?

Problem: Re-running VB for multiple model choices is expensive.

We propose: A linear approximation to efficiently estimate BNP sensitivity from a single run of VB. The linear approximation can both:

- Provide approximate sensitivity with no refitting, or
- Guide the choice of priors for refitting.

Outline

- The BNP model
- The variational approximation
- Hyperparameter sensitivity
- Functional sensitivity and influence functions
- Results on population genetics modeling of the Taita thrush

The BNP Model [Sethuraman, 1994]

A **Dirichlet process prior** allows for an infinite number of components.

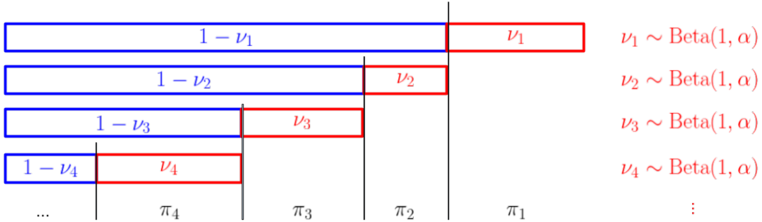


Figure 2: A schematic of the Dirichlet process prior

While there are an infinite number of **components**, there are a finite number of **clusters** in a given dataset.

Posterior quantities depend on the BNP prior, which is defined by the density of the stick-breaking process $\nu_k \sim \mathcal{P}(\nu_k)$.

If $\nu_k \sim \mathrm{Beta}(1, \alpha)$ what should α be?
Why should $\mathcal{P}(\nu_k)$ even be in the Beta family?

Variational Inference [Jordan et al., 1999]

Variational inference is an expansion-based methodology

- *Example*: algebraic vs. variational definition of the maximum eigenvalue

$$Ax = \lambda x \quad \textit{vs.} \quad \lambda = \max_x \left\{ \frac{x^T A x}{x^T x} \right\}$$

In general, we define an object (e.g., an integral) via an optimization problem, using test functions to obtain necessary conditions for optimality

E.g., likelihood-based objects naturally lend themselves to optimization problems involving the KL divergence, with the test functions being exponential-family densities

Here we go further, using test functions to probe sensitivities in function spaces of interest

Variational Stick-Breaking [Blei and Jordan, 2006]

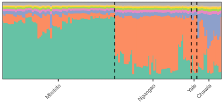
Let ζ denote all model variable, including stick lengths $\nu = (\nu_1, \nu_2, ...)$. Let x denote the observed data. The posterior $\mathcal{P}(\zeta|x)$ is intractable.

We approximate $\mathcal{P}(\zeta|z)$ using distributions $\mathcal{Q}(\zeta|\eta)$, parameterized by a finite-dimensional $\eta \in \Omega_\eta \subseteq \mathbb{R}^{D_\eta}$. We solve

$$\hat{\eta} := \underset{\eta \in \Omega_\eta}{\operatorname{argmin}} \mathrm{KL}\left(\eta\right) \quad \text{where} \quad \mathrm{KL}\left(\eta\right) := \mathrm{KL}\left(\mathcal{Q}(\zeta|\eta) || \mathcal{P}(\zeta|x)\right)$$

Note:

- The optimal variational parameters $\hat{\eta}$ depend on the prior through optimizing the KL objective.
- The approximate posterior quantities are then functions of $\hat{\eta}$, e.g.

$\hat{\eta} \mapsto \underset{\mathcal{Q}(\zeta|\hat{\eta})}{\mathbb{E}}$ [#clusters] or $\hat{\eta} \mapsto$ 

How do these approximate posterior quantities depend on the stick-breaking prior?

Hyperparameter Sensitivity

Let t be some real-valued hyperparameter for the stick-breaking density.

Hyperparameter Sensitivity

Let t be some real-valued hyperparameter for the stick-breaking density.

Write $\hat{\eta}(t) := \mathrm{argmin}_{\eta} \mathrm{KL}\left(\eta, t\right) = \mathrm{argmin}_{\eta} \mathrm{KL}\left(\mathcal{Q}(\zeta|\eta) || \mathcal{P}(\zeta|x, t)\right)$.

Hyperparameter Sensitivity

Let t be some real-valued hyperparameter for the stick-breaking density.

Write $\hat{\eta}(t) := \mathrm{argmin}_\eta \, \mathrm{KL}\left(\eta, t\right) = \mathrm{argmin}_\eta \, \mathrm{KL}\left(\mathcal{Q}(\zeta|\eta)||\mathcal{P}(\zeta|x, t)\right)$.

Problem: Evaluating $\hat{\eta}(t)$ requires solving a new optimization problem.

Hyperparameter Sensitivity

Let t be some real-valued hyperparameter for the stick-breaking density.

Write $\hat{\eta}(t) := \mathrm{argmin}_{\eta}\,\mathrm{KL}\left(\eta, t\right) = \mathrm{argmin}_{\eta}\,\mathrm{KL}\left(\mathcal{Q}(\zeta|\eta)||\mathcal{P}(\zeta|x,t)\right)$.

Problem: Evaluating $\hat{\eta}(t)$ requires solving a new optimization problem.

We propose: Approximate $\hat{\eta}(t)$ with a first-order Taylor expansion:

$$\hat{\eta}(t) \approx \hat{\eta}(0) + \left.\frac{d\hat{\eta}(t)}{dt}\right|_{t=0} t.$$

Hyperparameter Sensitivity

Let t be some real-valued hyperparameter for the stick-breaking density.

Write $\hat{\eta}(t) := \mathrm{argmin}_{\eta} \mathrm{KL}(\eta, t) = \mathrm{argmin}_{\eta} \mathrm{KL}(\mathcal{Q}(\zeta|\eta)||\mathcal{P}(\zeta|x, t))$.

Problem: Evaluating $\hat{\eta}(t)$ requires solving a new optimization problem.

We propose: Approximate $\hat{\eta}(t)$ with a first-order Taylor expansion:

$$\hat{\eta}(t) \approx \hat{\eta}(0) + \left.\frac{d\hat{\eta}(t)}{dt}\right|_{t=0} t.$$

- We need only use a linear approximation for the map $t \mapsto \hat{\eta}(t)$. We can retain nonlinearities in the map $\hat{\eta} \mapsto \underset{\mathcal{Q}(\zeta|\hat{\eta})}{\mathbb{E}}[\#\text{clusters}]$, etc.

Hyperparameter Sensitivity

Let t be some real-valued hyperparameter for the stick-breaking density.

Write $\hat{\eta}(t) := \mathrm{argmin}_{\eta} \mathrm{KL}\left(\eta, t\right) = \mathrm{argmin}_{\eta} \mathrm{KL}\left(\mathcal{Q}(\zeta|\eta) || \mathcal{P}(\zeta|x, t)\right)$.

Problem: Evaluating $\hat{\eta}(t)$ requires solving a new optimization problem.

We propose: Approximate $\hat{\eta}(t)$ with a first-order Taylor expansion:

$$\hat{\eta}(t) \approx \hat{\eta}(0) + \left.\frac{d\hat{\eta}(t)}{dt}\right|_{t=0} t.$$

- We need only use a linear approximation for the map $t \mapsto \hat{\eta}(t)$. We can retain nonlinearities in the map $\hat{\eta} \mapsto \underset{\mathcal{Q}(\zeta|\hat{\eta})}{\mathbb{E}}$ [#clusters], etc.
- This is "Bayesian local robustness" for VB [cf. Gustafson, 1996]

Hyperparameter Sensitivity

Let t be some real-valued hyperparameter for the stick-breaking density.

Write $\hat{\eta}(t) := \mathrm{argmin}_\eta \mathrm{KL}(\eta, t) = \mathrm{argmin}_\eta \mathrm{KL}(\mathcal{Q}(\zeta|\eta)||\mathcal{P}(\zeta|x, t))$.

Problem: Evaluating $\hat{\eta}(t)$ requires solving a new optimization problem.

We propose: Approximate $\hat{\eta}(t)$ with a first-order Taylor expansion:

$$\hat{\eta}(t) \approx \hat{\eta}(0) + \left.\frac{d\hat{\eta}(t)}{dt}\right|_{t=0} t.$$

- We need only use a linear approximation for the map $t \mapsto \hat{\eta}(t)$. We can retain nonlinearities in the map $\hat{\eta} \mapsto \underset{\mathcal{Q}(\zeta|\hat{\eta})}{\mathbb{E}}[\#\text{clusters}]$, etc.
- This is "Bayesian local robustness" for VB [cf. Gustafson, 1996]
- The derivative can be evaluated using the implicit function theorem and modern automatic differentiation.

Computing the Derivative [Giordano et al., 2018]

Theorem 1. (The derivative $d\hat{\eta}(t)/dt$.)

Computing the Derivative [Giordano et al., 2018]

Theorem 1. (The derivative $d\hat{\eta}(t)/dt$.)

Define $\hat{\eta} = \hat{\eta}(0)$, $\hat{H} := \left.\frac{\partial^2 \mathrm{KL}(\eta)}{\partial\eta\partial\eta^T}\right|_{\hat{\eta}}$ and $\nabla_\eta \log \mathcal{Q}\left(\nu|\hat{\eta}\right) := \left.\frac{\log \mathcal{Q}(\nu|\hat{\eta})}{\partial\eta}\right|_{\hat{\eta}}$.

Computing the Derivative [Giordano et al., 2018]

Theorem 1. (The derivative $d\hat{\eta}(t)/dt$.)

Define $\hat{\eta} = \hat{\eta}(0)$, $\hat{H} := \left.\frac{\partial^2 \mathrm{KL}(\eta)}{\partial \eta \partial \eta^T}\right|_{\hat{\eta}}$ and $\nabla_\eta \log \mathcal{Q}\left(\nu|\hat{\eta}\right) := \left.\frac{\log \mathcal{Q}(\nu|\hat{\eta})}{\partial \eta}\right|_{\hat{\eta}}$.

Assume:

- The Hessian at the optimum, $\hat{H}$, is non-singular.

Computing the Derivative [Giordano et al., 2018]

Theorem 1. (The derivative $d\hat{\eta}(t)/dt$.)

Define $\hat{\eta} = \hat{\eta}(0)$, $\hat{H} := \left.\frac{\partial^2 \mathrm{KL}(\eta)}{\partial \eta \partial \eta^T}\right|_{\hat{\eta}}$ and $\nabla_\eta \log \mathcal{Q}\left(\nu|\hat{\eta}\right) := \left.\frac{\log \mathcal{Q}(\nu|\hat{\eta})}{\partial \eta}\right|_{\hat{\eta}}$.

Assume:

- The Hessian at the optimum, $\hat{H}$, is non-singular.
- The optimal VB parameters, $\hat{\eta}$, are interior.

Computing the Derivative [Giordano et al., 2018]

Theorem 1. (The derivative $d\hat{\eta}(t)/dt$.)

Define $\hat{\eta} = \hat{\eta}(0)$, $\hat{H} := \left.\frac{\partial^2 \mathrm{KL}(\eta)}{\partial \eta \partial \eta^T}\right|_{\hat{\eta}}$ and $\nabla_\eta \log \mathcal{Q}\left(\nu|\hat{\eta}\right) := \left.\frac{\log \mathcal{Q}(\nu|\hat{\eta})}{\partial \eta}\right|_{\hat{\eta}}$.

Assume:

- The Hessian at the optimum, $\hat{H}$, is non-singular.
- The optimal VB parameters, $\hat{\eta}$, are interior.
- We can exchange limits and $\mathcal{Q}$ expectations as needed in a neighborhood of $\hat{\eta}$ and $t = 0$.
    - This imposes some regularity conditions on the prior $\mathcal{P}(\nu|t)$.

Computing the Derivative [Giordano et al., 2018]

Theorem 1. (The derivative $d\hat{\eta}(t)/dt$.)

Define $\hat{\eta} = \hat{\eta}(0)$, $\hat{H} := \left.\frac{\partial^2 \mathrm{KL}(\eta)}{\partial\eta\partial\eta^T}\right|_{\hat{\eta}}$ and $\nabla_\eta \log \mathcal{Q}\left(\nu|\hat{\eta}\right) := \left.\frac{\log \mathcal{Q}(\nu|\hat{\eta})}{\partial\eta}\right|_{\hat{\eta}}$.

Assume:

- The Hessian at the optimum, $\hat{H}$, is non-singular.
- The optimal VB parameters, $\hat{\eta}$, are interior.
- We can exchange limits and $\mathcal{Q}$ expectations as needed in a neighborhood of $\hat{\eta}$ and $t = 0$.
    - This imposes some regularity conditions on the prior $\mathcal{P}(\nu|t)$.

Then the map $t \mapsto \hat{\eta}(t)$ is continuously differentiable at $t = 0$ with

$$\left.\frac{d\hat{\eta}(t)}{dt}\right|_0 = -\hat{H}^{-1} \underset{\mathcal{Q}_{\hat{\eta}}}{\mathbb{E}}\left[\nabla_\eta \log \mathcal{Q}\left(\nu|\hat{\eta}\right) \left.\frac{\partial \log \mathcal{P}(\nu|t)}{\partial t}\right|_{t=0}\right].$$

□

Computing the Derivative [Giordano et al., 2018]

Theorem 1. (The derivative $d\hat{\eta}(t)/dt$.)

Define $\hat{\eta} = \hat{\eta}(0)$, $\hat{H} := \left.\frac{\partial^2 \mathrm{KL}(\eta)}{\partial\eta\partial\eta^T}\right|_{\hat{\eta}}$ and $\nabla_\eta \log \mathcal{Q}(\nu|\hat{\eta}) := \left.\frac{\log \mathcal{Q}(\nu|\hat{\eta})}{\partial\eta}\right|_{\hat{\eta}}$.

Assume:

- The Hessian at the optimum, $\hat{H}$, is non-singular.
- The optimal VB parameters, $\hat{\eta}$, are interior.
- We can exchange limits and $\mathcal{Q}$ expectations as needed in a neighborhood of $\hat{\eta}$ and $t = 0$.
    - This imposes some regularity conditions on the prior $\mathcal{P}(\nu|t)$.

Then the map $t \mapsto \hat{\eta}(t)$ is continuously differentiable at $t = 0$ with

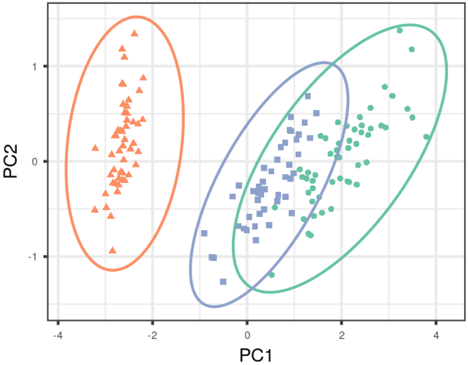
$$\left.\frac{d\hat{\eta}(t)}{dt}\right|_0 = -\hat{H}^{-1} \underset{\mathcal{Q}_{\hat{\eta}}}{\mathbb{E}} \left[\nabla_\eta \log \mathcal{Q}(\nu|\hat{\eta}) \left.\frac{\partial \log \mathcal{P}(\nu|t)}{\partial t}\right|_{t=0}\right].$$

□

Note: The computation of $\hat{H}^{-1}$ is the computationally difficult part. For our BNP problem, $\hat{H}$ is sparse.

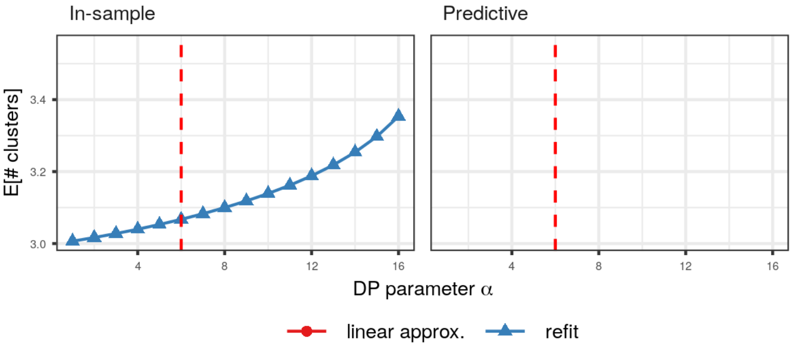
A Simple Example: Iris Data

We fit a Gaussian mixture model with a DP prior to the iris data.



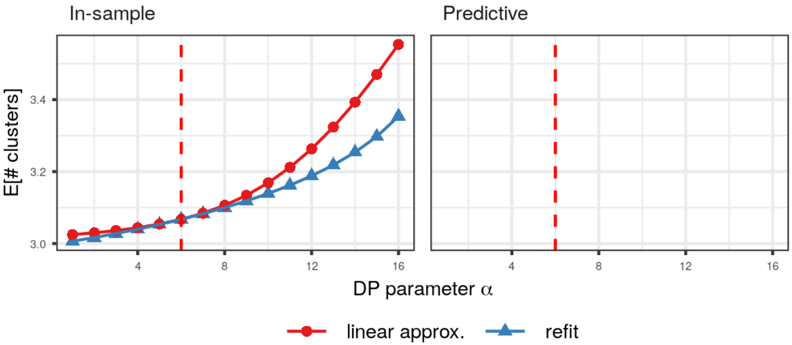
The iris data in principal component space and GMM fit at $\alpha = 6$.

Iris Data: Parametric Sensitivity



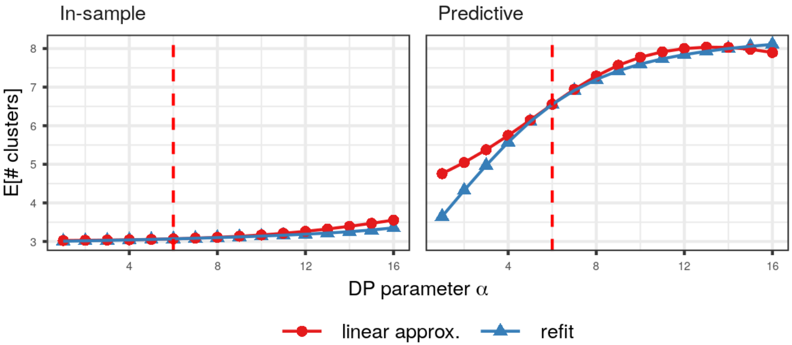
The expected number of posterior clusters in the iris data as α varies.

Iris Data: Parametric Sensitivity



The expected number of posterior clusters in the iris data as α varies.

Iris Data: Parametric Sensitivity



The expected number of posterior clusters in the iris data as α varies.

Functional Sensitivity [Gustafson, 1996]

What about stick-breaking priors not in the Beta family?

Functional Sensitivity [Gustafson, 1996]

What about stick-breaking priors not in the Beta family?

Let $\mathcal{P}_0(\nu)$ be the stick-breaking prior used to compute $\hat{\eta}$. Suppose we wish to replace $\mathcal{P}_0(\nu)$ with another density, $\mathcal{P}_1(\nu)$.

Functional Sensitivity [Gustafson, 1996]

What about stick-breaking priors not in the Beta family?

Let $\mathcal{P}_0(\nu)$ be the stick-breaking prior used to compute $\hat{\eta}$. Suppose we wish to replace $\mathcal{P}_0(\nu)$ with another density, $\mathcal{P}_1(\nu)$.

Define the "perturbed" prior as:

$$\mathcal{P}(\nu|\phi) \propto \mathcal{P}_0(\nu) \exp(\phi(\nu)) \quad \text{with} \quad \phi(\nu) = \log \mathcal{P}_1(\nu) - \log \mathcal{P}_0(\nu)$$

Functional Sensitivity [Gustafson, 1996]

What about stick-breaking priors not in the Beta family?

Let $\mathcal{P}_0(\nu)$ be the stick-breaking prior used to compute $\hat{\eta}$. Suppose we wish to replace $\mathcal{P}_0(\nu)$ with another density, $\mathcal{P}_1(\nu)$.

Define the "perturbed" prior as:

$$\mathcal{P}(\nu|\phi) \propto \mathcal{P}_0(\nu) \exp(\phi(\nu)) \quad \text{with} \quad \phi(\nu) = \log \mathcal{P}_1(\nu) - \log \mathcal{P}_0(\nu)$$

Then $t \mapsto \mathcal{P}(\nu|t\phi)$ parameterizes a path from $\mathcal{P}_0$ to $\mathcal{P}_1$ for $t \in [0, 1]$.

Functional Sensitivity [Gustafson, 1996]

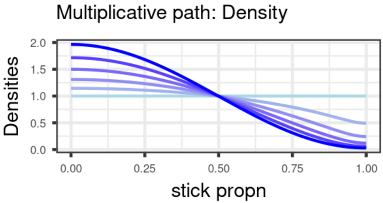
What about stick-breaking priors not in the Beta family?

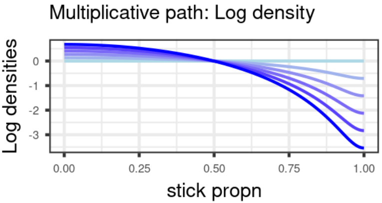
Let $\mathcal{P}_0(\nu)$ be the stick-breaking prior used to compute $\hat{\eta}$. Suppose we wish to replace $\mathcal{P}_0(\nu)$ with another density, $\mathcal{P}_1(\nu)$.

Define the "perturbed" prior as:

$$\mathcal{P}(\nu|\phi) \propto \mathcal{P}_0(\nu)\exp(\phi(\nu)) \quad \text{with} \quad \phi(\nu) = \log \mathcal{P}_1(\nu) - \log \mathcal{P}_0(\nu)$$

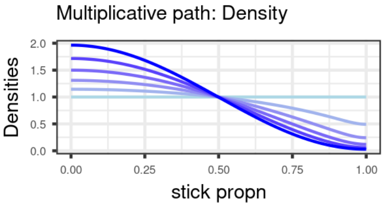
Then $t \mapsto \mathcal{P}(\nu|t\phi)$ parameterizes a path from $\mathcal{P}_0$ to $\mathcal{P}_1$ for $t \in [0, 1]$.

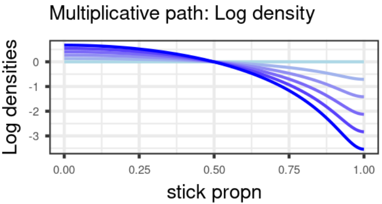




Functional Sensitivity [Gustafson, 1996]

For any particular ϕ, we can try to apply Theorem 1 to $t \mapsto \mathcal{P}(\nu|t\phi)$.

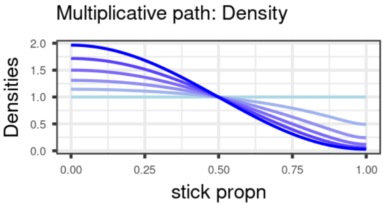


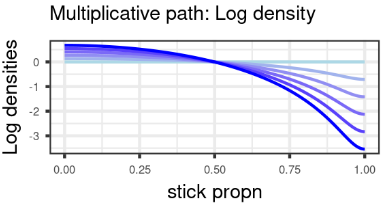


Functional Sensitivity [Gustafson, 1996]

For any particular ϕ, we can try to apply Theorem 1 to $t \mapsto \mathcal{P}(\nu|t\phi)$.

But it would be nice to safely search the space of functions ϕ.





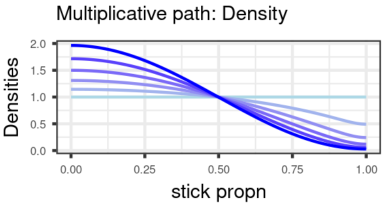
Functional Sensitivity [Gustafson, 1996]

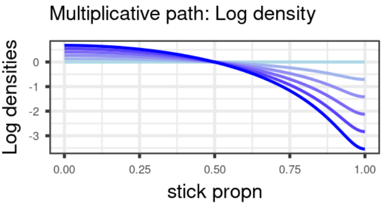
For any particular ϕ, we can try to apply Theorem 1 to $t \mapsto \mathcal{P}(\nu|t\phi)$.

But it would be nice to safely search the space of functions ϕ.

Questions:

- Can we specify a general condition on ϕ for Theorem 1 to apply?





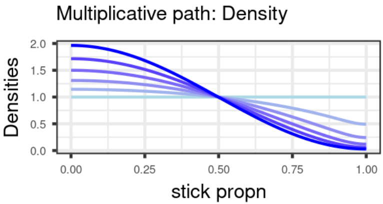
Functional Sensitivity [Gustafson, 1996]

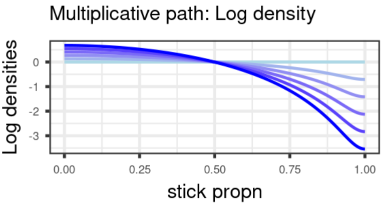
For any particular ϕ, we can try to apply Theorem 1 to $t \mapsto \mathcal{P}(\nu|t\phi)$.

But it would be nice to safely search the space of functions ϕ.

Questions:

- Can we specify a general condition on ϕ for Theorem 1 to apply?
- Is the derivative a good linear approximation for all such functions?





Functional Sensitivity: Differentiability

Let L_∞ denote the vector space of bounded, Lebesgue-measurable functions with norm $\|\phi\|_\infty := \operatorname{esssup}_\nu |\phi(\nu)|$.

Functional Sensitivity: Differentiability

Let L_∞ denote the vector space of bounded, Lebesgue-measurable functions with norm $\|\phi\|_\infty := \operatorname{esssup}_\nu |\phi(\nu)|$.

Proposition.

If $\phi \in L_\infty$, then $\mathcal{P}(\nu|\phi)$ is a valid density (positive and normalizable).

Functional Sensitivity: Differentiability

Let L_∞ denote the vector space of bounded, Lebesgue-measurable functions with norm $\|\phi\|_\infty := \operatorname{esssup}_\nu |\phi(\nu)|$.

Proposition.

If $\phi \in L_\infty$, then $\mathcal{P}(\nu|\phi)$ is a valid density (positive and normalizable).

Theorem 2. (Validity of the derivative in L_∞.)

Functional Sensitivity: Differentiability

Let L_∞ denote the vector space of bounded, Lebesgue-measurable functions with norm $\|\phi\|_\infty := \mathrm{esssup}_\nu |\phi(\nu)|$.

Proposition.

If $\phi \in L_\infty$, then $\mathcal{P}(\nu|\phi)$ is a valid density (positive and normalizable).

Theorem 2. (Validity of the derivative in L_∞.)

If $\phi \in L_\infty$, then the map $t \mapsto \mathcal{P}(\nu|t\phi)$ satisfies the conditions of Theorem 1, so $t \mapsto \hat{\eta}(t\phi)$ is continuously differentiable.

Functional Sensitivity: Differentiability

Let L_∞ denote the vector space of bounded, Lebesgue-measurable functions with norm $\|\phi\|_\infty := \operatorname{esssup}_\nu |\phi(\nu)|$.

Proposition.

If $\phi \in L_\infty$, then $\mathcal{P}(\nu|\phi)$ is a valid density (positive and normalizable).

Theorem 2. (Validity of the derivative in L_∞.)

If $\phi \in L_\infty$, then the map $t \mapsto \mathcal{P}(\nu|t\phi)$ satisfies the conditions of Theorem 1, so $t \mapsto \hat{\eta}(t\phi)$ is continuously differentiable.

Further, the derivatives provides a uniformly good linear approximation in an $\|\cdot\|_\infty$-neighborhood of the zero function. In other words, the map $\phi \mapsto \hat{\eta}(\phi)$ from $L_\infty \mapsto \mathbb{R}^D$ is *Fréchet differentiable* at zero. $\square$

Functional Sensitivity: Differentiability

Let L_∞ denote the vector space of bounded, Lebesgue-measurable functions with norm $\|\phi\|_\infty := \operatorname{esssup}_\nu |\phi(\nu)|$.

Proposition.

If $\phi \in L_\infty$, then $\mathcal{P}(\nu|\phi)$ is a valid density (positive and normalizable).

Theorem 2. (Validity of the derivative in L_∞.)

If $\phi \in L_\infty$, then the map $t \mapsto \mathcal{P}(\nu|t\phi)$ satisfies the conditions of Theorem 1, so $t \mapsto \hat{\eta}(t\phi)$ is continuously differentiable.

Further, the derivatives provides a uniformly good linear approximation in an $\|\cdot\|_\infty$-neighborhood of the zero function. In other words, the map $\phi \mapsto \hat{\eta}(\phi)$ from $L_\infty \mapsto \mathbb{R}^D$ is *Fréchet differentiable* at zero. $\square$

Note: Arguably, Fréchet differentiability is a minimal requirement for using the linear approximation to safely search the space of functions.

Functional Sensitivity: Influence Functions

Corollary of Theorem 2. (Influence functions.)

Take a continuously differentiable quantity of interest $g(\eta)$, e.g.

$$g_{\mathrm{cl}}(\eta) = \mathop{\mathbb{E}}_{\mathcal{Q}_\eta}[\#\text{clusters}]$$

Functional Sensitivity: Influence Functions

Corollary of Theorem 2. (Influence functions.)

Take a continuously differentiable quantity of interest $g(\eta)$, e.g.

$$g_{\mathrm{cl}}(\eta) = \mathop{\mathbb{E}}_{\mathcal{Q}_\eta}[\#\text{clusters}]$$

Let $S_g(\phi)$ be the *local sensitivity* of g in the direction ϕ:

$$S_g(\phi) := \left.\frac{dg(\hat{\eta}(t\phi))}{dt}\right|_{t=0}.$$

Functional Sensitivity: Influence Functions

Corollary of Theorem 2. (Influence functions.)

Take a continuously differentiable quantity of interest $g(\eta)$, e.g.

$$g_{\mathrm{cl}}(\eta) = \mathop{\mathbb{E}}_{\mathcal{Q}_\eta} [\#\text{clusters}]$$

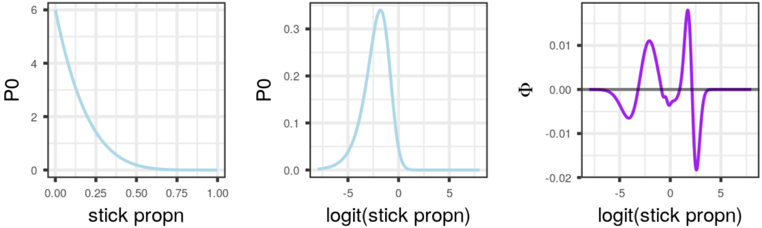
Let $S_g(\phi)$ be the *local sensitivity* of g in the direction ϕ:

$$S_g(\phi) := \left. \frac{dg(\hat{\eta}(t\phi))}{dt} \right|_{t=0}.$$

If $\|\phi\|_\infty < \infty$, the local sensitivity can be expressed as an inner product between an *influence function* Ψ and the functional perturbation ϕ:

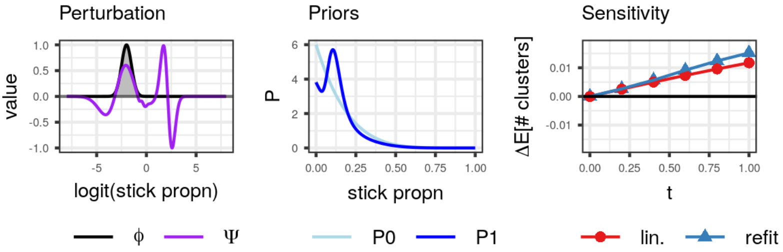
$$\begin{aligned} S_g(\phi) &= - \left. \frac{dg(\eta)}{d\eta^T} \right|_{\hat{\eta}} \hat{H}^{-1} \mathop{\mathbb{E}}_{\mathcal{Q}_{\hat{\eta}}} \left[\nabla_\eta \log \mathcal{Q}\left(\nu | \hat{\eta}\right) \phi(\nu) \right] \\ &= \int \Psi(\nu) \phi(\nu) \, d\nu. \end{aligned}$$

Iris Data: Influence Functions

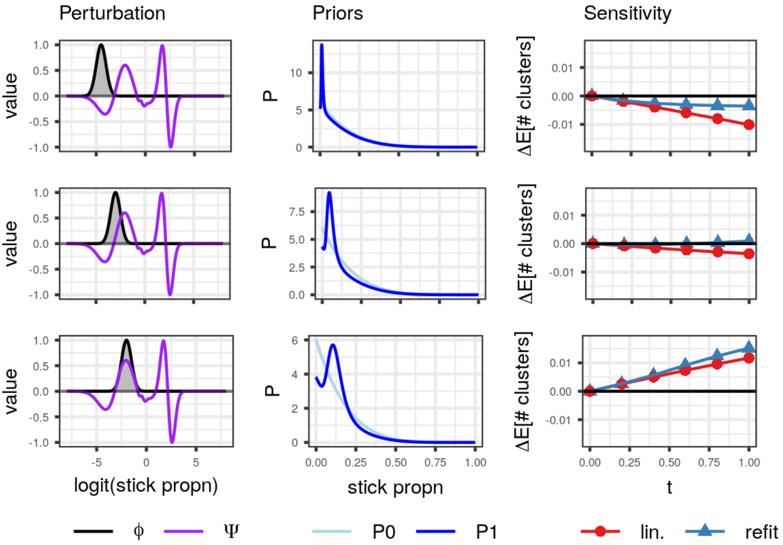


The influence function for the number of clusters, g_{cl}.

Iris Data: Functional Perturbations



Iris Data: Functional Perturbations



Functional Perturbations: Worst Case [Gustafson, 1996]

Which perturbation ϕ maximizes the sensitivity $S_g(\phi)$?

Functional Perturbations: Worst Case [Gustafson, 1996]

Which perturbation ϕ maximizes the sensitivity $S_g(\phi)$?

That is, can we find the **worst-case** ϕ in the L-infinity ball of radius δ,

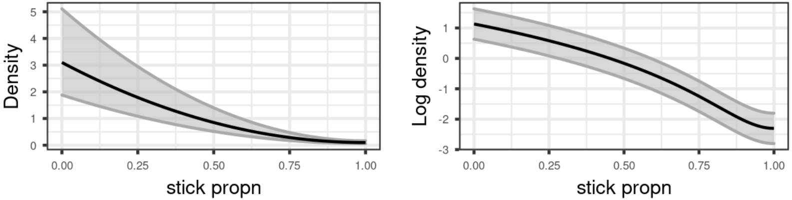
$$B_\delta := \{\phi : \|\phi\|_\infty < \delta\}?$$

Functional Perturbations: Worst Case [Gustafson, 1996]

Which perturbation ϕ maximizes the sensitivity $S_g(\phi)$?

That is, can we find the **worst-case** ϕ in the L-infinity ball of radius δ,

$$B_\delta := \{\phi : \|\phi\|_\infty < \delta\}?$$

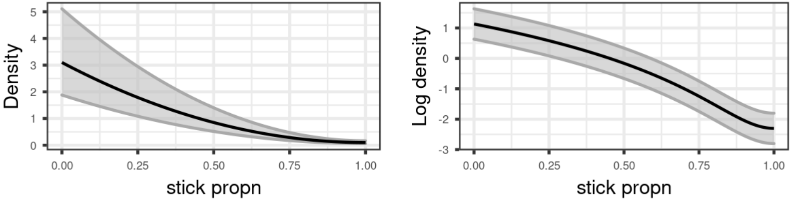


Functional Perturbations: Worst Case [Gustafson, 1996]

Which perturbation ϕ maximizes the sensitivity $S_g(\phi)$?

That is, can we find the **worst-case** ϕ in the L-infinity ball of radius δ,

$$B_\delta := \{\phi : \|\phi\|_\infty < \delta\}?$$

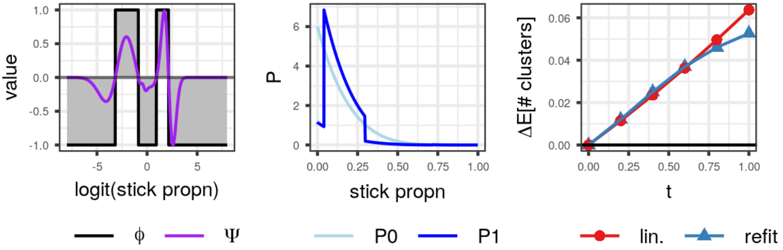


Using the influence function and Hölder's inequality,

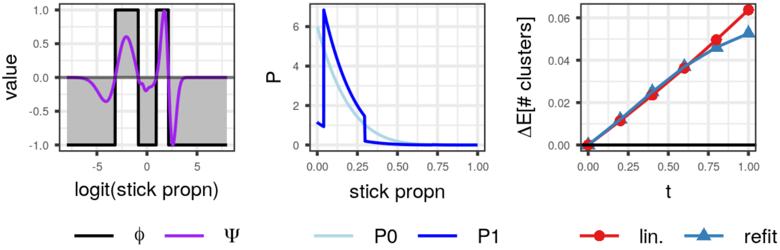
$$\sup_{\phi \in \mathcal{B}_\delta} S_g(\phi) = \sup_{\phi \in \mathcal{B}_\delta} \int \Psi(\nu)\phi(\nu) d\nu = \delta \int |\Psi(\nu)| \, d\nu, \text{ achieved at}$$

$$\phi^*(\nu) = \delta \, \mathrm{sign}(\Psi(\nu)).$$

Iris Data: Worst-Case Perturbation



Iris Data: Worst-Case Perturbation



The worst-case prior may look unreasonable.

But if the worst-case sensitivity is small, it is evidence of robustness.

Functional sensitivity: other paths [Gustafson, 1996]

For $\mathcal{P}(\nu_k|\phi)$, we used a multiplicative perturbation.

Could we have used other paths through function space?

Functional sensitivity: other paths [Gustafson, 1996]

For $\mathcal{P}(\nu_k|\phi)$, we used a multiplicative perturbation.

Could we have used other paths through function space?

Consider, for example, "mixture distributions":

$$\mathcal{P}(\nu|\phi_{mix}) \propto \mathcal{P}_0(\nu) + \phi_{mix}(\nu) \quad \text{and} \quad \phi_{mix}(\nu) = \mathcal{P}_1(\nu) - \mathcal{P}_0(\nu)$$

Then $t \mapsto \mathcal{P}(\nu|t\phi_{mix})$ also parameterizes a path from $\mathcal{P}_0$ to $\mathcal{P}_1$.

Functional sensitivity: other paths [Gustafson, 1996]

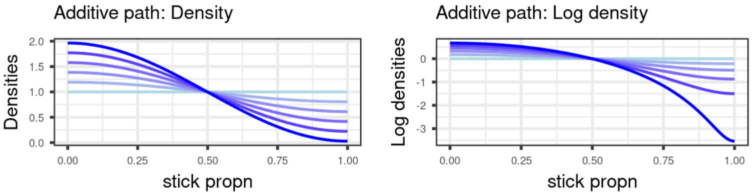
For $\mathcal{P}(\nu_k|\phi)$, we used a multiplicative perturbation.

Could we have used other paths through function space?

Consider, for example, "mixture distributions":

$$\mathcal{P}(\nu|\phi_{mix}) \propto \mathcal{P}_0(\nu) + \phi_{mix}(\nu) \quad \text{and} \quad \phi_{mix}(\nu) = \mathcal{P}_1(\nu) - \mathcal{P}_0(\nu)$$

Then $t \mapsto \mathcal{P}(\nu|t\phi_{mix})$ also parameterizes a path from $\mathcal{P}_0$ to $\mathcal{P}_1$.



Functional sensitivity: other paths [Gustafson, 1996]

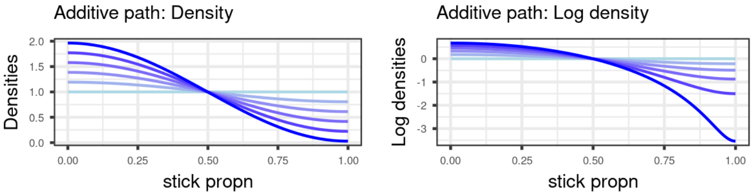
For $\mathcal{P}(\nu_k|\phi)$, we used a multiplicative perturbation.

Could we have used other paths through function space?

Consider, for example, "mixture distributions":

$$\mathcal{P}(\nu|\phi_{mix}) \propto \mathcal{P}_0(\nu) + \phi_{mix}(\nu) \quad \text{and} \quad \phi_{mix}(\nu) = \mathcal{P}_1(\nu) - \mathcal{P}_0(\nu)$$

Then $t \mapsto \mathcal{P}(\nu|t\phi_{mix})$ also parameterizes a path from $\mathcal{P}_0$ to $\mathcal{P}_1$.



Question: Is there anything wrong with using ϕ_{mix} with our VB approximation?

Functional Sensitivity: Other Paths

Functional Sensitivity: Other Paths

Theorem 3. (Differentiability of other paths.)

Functional Sensitivity: Other Paths

Theorem 3. (Differentiability of other paths.)

Let $S_{mix} := \{\phi_{mix} : \phi_{mix} = \mathcal{P}_1 - \mathcal{P}_0 \text{ for some density } \mathcal{P}_1 \ll \mathcal{P}_0\}$.

Functional Sensitivity: Other Paths

Theorem 3. (Differentiability of other paths.)

Let $S_{mix} := \{\phi_{mix} : \phi_{mix} = \mathcal{P}_1 - \mathcal{P}_0 \text{ for some density } \mathcal{P}_1 \ll \mathcal{P}_0\}$.

For any $\phi_{mix} \in S_{mix}$, the conditions of Theorem 1 are satisfied under some additional mild integrability assumptions on $\mathcal{Q}_\eta$. So the map $t \mapsto \hat{\eta}(t\phi_{mix})$ is continuously differentiable.

Functional Sensitivity: Other Paths

Theorem 3. (Differentiability of other paths.)

Let $S_{mix} := \{\phi_{mix} : \phi_{mix} = \mathcal{P}_1 - \mathcal{P}_0 \text{ for some density } \mathcal{P}_1 \ll \mathcal{P}_0\}$.

For any $\phi_{mix} \in S_{mix}$, the conditions of Theorem 1 are satisfied under some additional mild integrability assumptions on $\mathcal{Q}_\eta$. So the map $t \mapsto \hat{\eta}(t\phi_{mix})$ is continuously differentiable.

But normalizability of $\mathcal{P}(\nu|\phi_{mix})$ is determined by $\|\phi_{mix}\|_1$, and the error of the derivative is arbitrarily large in any $\|\cdot\|_1$-neighborhood of the zero function.

Functional Sensitivity: Other Paths

Theorem 3. (Differentiability of other paths.)

Let $S_{mix} := \{\phi_{mix} : \phi_{mix} = \mathcal{P}_1 - \mathcal{P}_0 \text{ for some density } \mathcal{P}_1 \ll \mathcal{P}_0\}$.

For any $\phi_{mix} \in S_{mix}$, the conditions of Theorem 1 are satisfied under some additional mild integrability assumptions on $\mathcal{Q}_\eta$. So the map $t \mapsto \hat{\eta}(t\phi_{mix})$ is continuously differentiable.

But normalizability of $\mathcal{P}(\nu|\phi_{mix})$ is determined by $\|\phi_{mix}\|_1$, and the error of the derivative is arbitrarily large in any $\|\cdot\|_1$-neighborhood of the zero function.

$\Rightarrow$ No extension of S_{mix} to L_1 of the map $\phi_{mix} \mapsto \hat{\eta}(\phi_{mix})$ can be Fréchet differentiable.

$\square$

Functional Sensitivity: Other Paths

Theorem 3. (Differentiability of other paths.)

Let $S_{mix} := \{\phi_{mix} : \phi_{mix} = \mathcal{P}_1 - \mathcal{P}_0 \text{ for some density } \mathcal{P}_1 \ll \mathcal{P}_0\}$.

For any $\phi_{mix} \in S_{mix}$, the conditions of Theorem 1 are satisfied under some additional mild integrability assumptions on $\mathcal{Q}_\eta$. So the map $t \mapsto \hat{\eta}(t\phi_{mix})$ is continuously differentiable.

But normalizability of $\mathcal{P}(\nu|\phi_{mix})$ is determined by $\|\phi_{mix}\|_1$, and the error of the derivative is arbitrarily large in any $\|\cdot\|_1$-neighborhood of the zero function.

$\Rightarrow$ No extension of S_{mix} to L_1 of the map $\phi_{mix} \mapsto \hat{\eta}(\phi_{mix})$ can be Fréchet differentiable.

$\square$

Note: An analogous result holds for all L_p spaces with $p < \infty$.

Functional Sensitivity: Other Paths

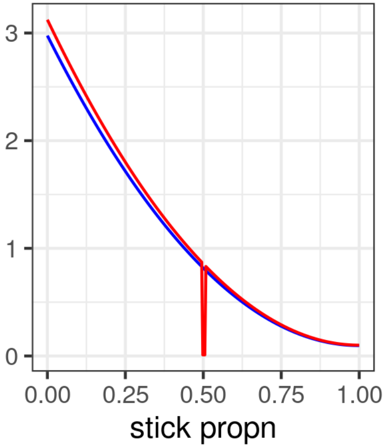
What went wrong with the mixture distribution?

Functional Sensitivity: Other Paths

What went wrong with the mixture distribution?

These red and blue densities are

- Distant in KL and $\|\cdot\|_\infty$, but
- Close in $\|\cdot\|_p$ when $p < \infty$.



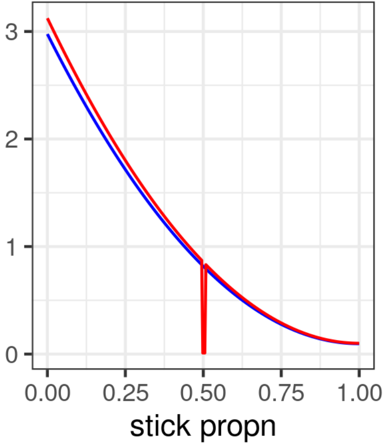
Functional Sensitivity: Other Paths

What went wrong with the mixture distribution?

These red and blue densities are

- Distant in KL and $\|\cdot\|_\infty$, but
- Close in $\|\cdot\|_p$ when $p < \infty$.

A parameterization + prior normalizability dictates a norm.



Functional Sensitivity: Other Paths

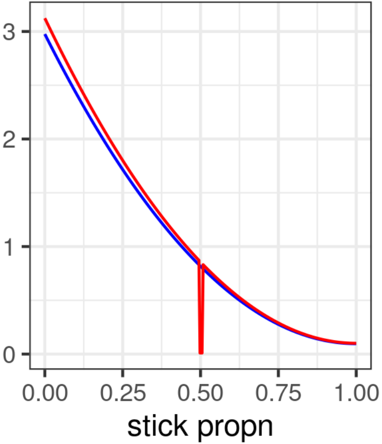
What went wrong with the mixture distribution?

These red and blue densities are

- Distant in KL and $\|\cdot\|_\infty$, but
- Close in $\|\cdot\|_p$ when $p < \infty$.

A parameterization + prior normalizability dictates a norm.

For differentiability of $\hat{\eta}$, the norm's topology must match that of KL.



Functional Sensitivity: Other Paths

What went wrong with the mixture distribution?

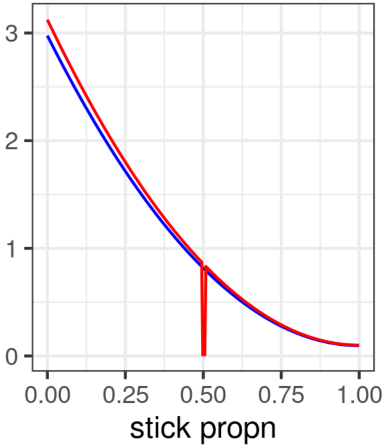
These red and blue densities are

- Distant in KL and $\|\cdot\|_\infty$, but
- Close in $\|\cdot\|_p$ when $p < \infty$.

A parameterization + prior normalizability dictates a norm.

For differentiability of $\hat{\eta}$, the norm's topology must match that of KL.

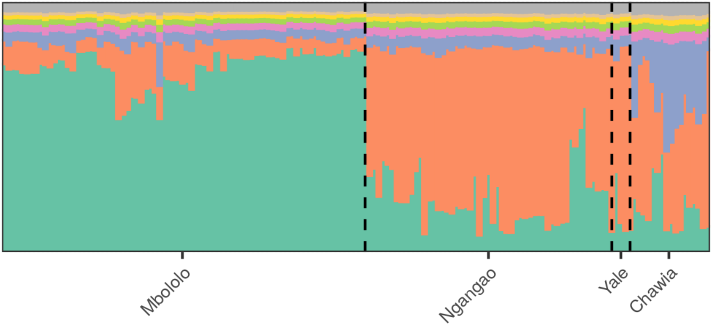
$\Rightarrow$ **We consider only multiplicative perturbations for VB.**



Results on fastSTRUCTURE [Raj et al., 2014]

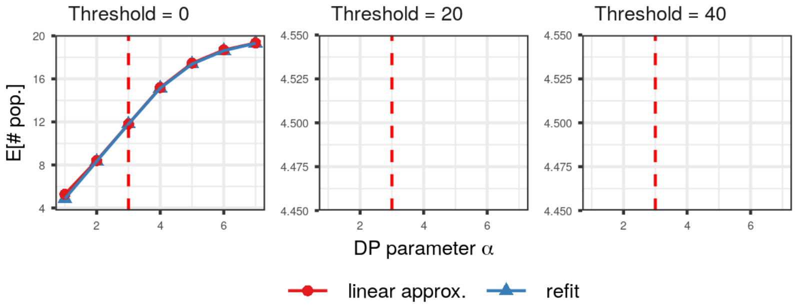
We adapt fastSTRUCTURE a Bayesian model for population genetics, to include a BNP prior.

We study genetic data from the Taita thrush, an endangered bird species. The data consists of microsatellites sequences of 155 individuals at 7 loci.



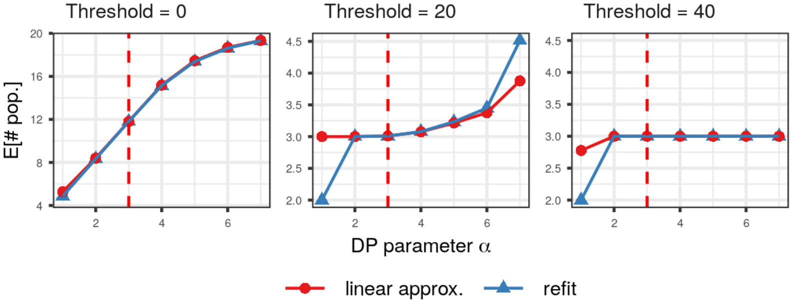
The intitial fit at $\alpha = 3$.

fastSTRUCTURE: Parametric Sensitivity



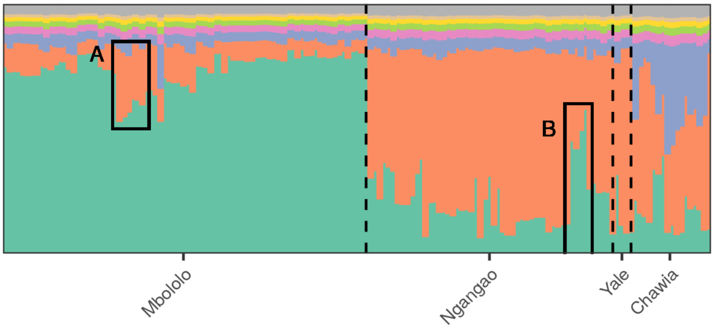
Expected number of posterior in-sample clusters in the thrush data as α varies.

fastSTRUCTURE: Parametric Sensitivity

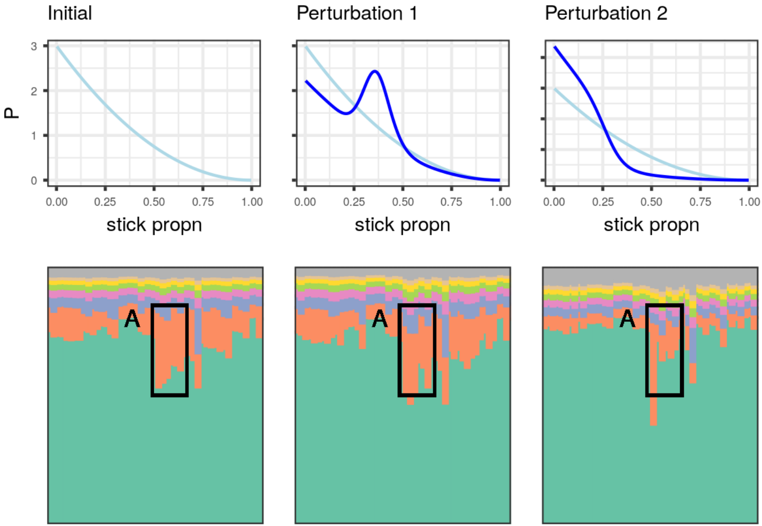


Expected number of posterior in-sample clusters in the thrush data as α varies.

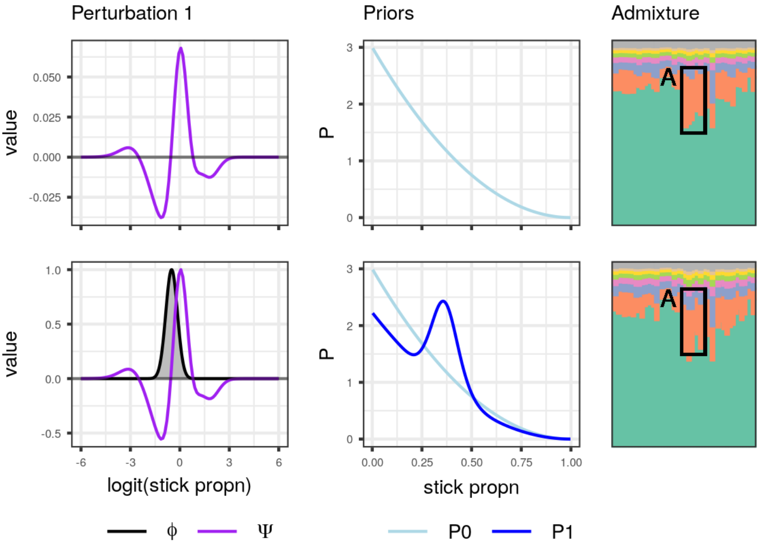
fastSTRUCTURE: Evidence of Migration?



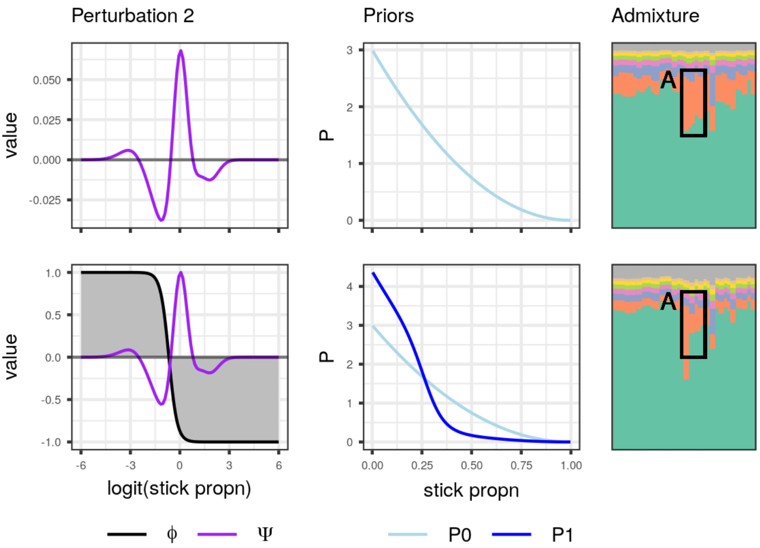
fastSTRUCTURE: Evidence of Migration?



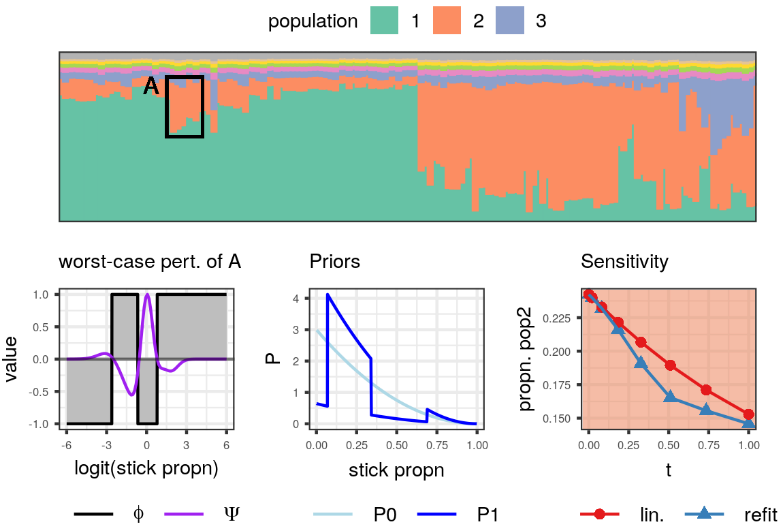
fastSTRUCTURE: Functional Sensitivity



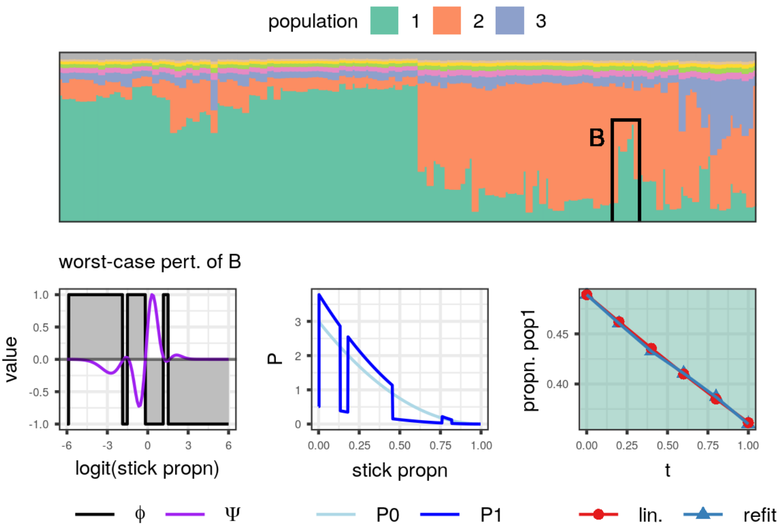
fastSTRUCTURE: Functional Sensitivity



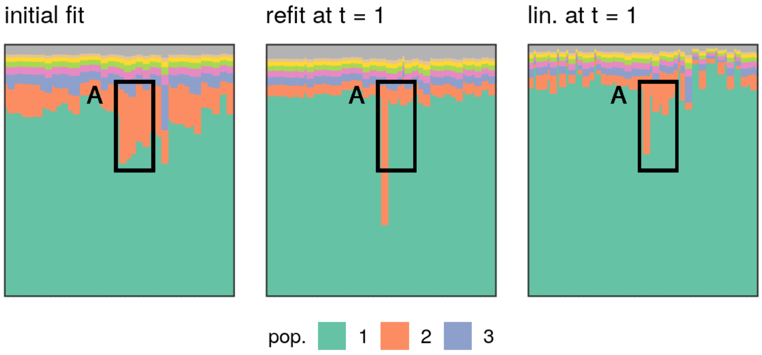
fastSTRUCTURE: Functional Sensitivity



fastSTRUCTURE: Functional Sensitivity

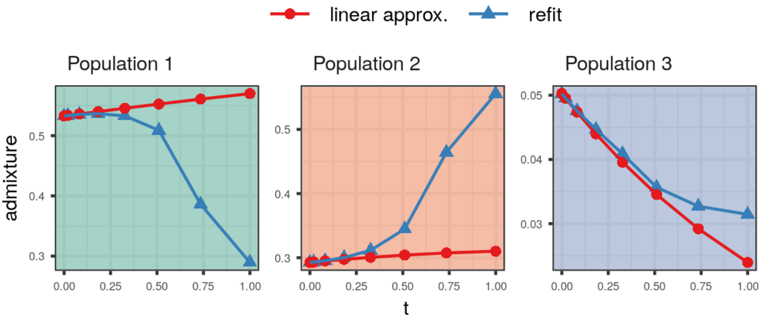


Limitations of Local Sensitivity

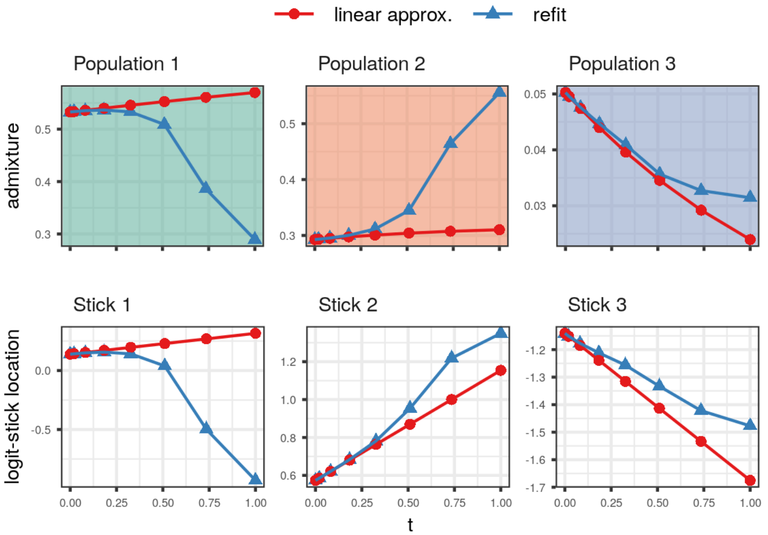


Inferred admixtures after the worst-case perturbation to individuals A. Individual $n = 26$ had a large increase in admixture proportion of population 2 after the refit.

Limitations of Local Sensitivity



Limitations of Local Sensitivity



Computational Complexity

Compute time of results on the Taita thrush dataset.

	time (seconds)
Initial fit	7
Hessian solve for α sensitivity	0.3
Linear approx. $\eta^{lin}(\alpha)$ for $\alpha = 1, ..., 7$	0.006
Refits $\eta(\alpha)$ for $\alpha = 1, ..., 7$	30
The influence function	0.6
Hessian solve for perturbation ϕ	0.4
Linear approx. $\eta^{lin}(\epsilon)\|_{\epsilon=1}$ for perturbation ϕ	0.001
Refit $\eta(\epsilon)\|_{\epsilon=1}$ for perturbation ϕ	10

Conclusions

- We provide a tool to efficiently evaluate the sensitivity of the variational posterior to prior choices.
- Linearizing the variational parameters provides a reasonable alternative to re-optimizing the variational approximation after model perturbations.
- For variational approximations based on KL divergence, one should express functional perturbations multiplicatively.
- The influence function can provide guidance for finding particularly sensitive model perturbations which can be investigated by re-fitting.

Links and references

Runjing Liu, Ryan Giordano, Michael I. Jordan, Tamara Broderick.
"Evaluating Sensitivity to the Stick Breaking Prior in Bayesian Nonparametrics."
https://arxiv.org/pdf/1810.06587.pdf

JAX: composable transformations of Python+NumPy programs
https://github.com/google/jax

D. Blei and M. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 2006.

P. Galbusera, L. Lens, T. Schenck, E. Waiyaki, and E. Matthysen. Genetic variability and gene flow in the globally, critically-endangered taita thrush. *Conservation Genetics*, 1:45–55, March 2000.

R. Giordano, T. Broderick, and M. Jordan. Covariances, robustness and variational Bayes. *Journal of machine learning research*, 19(51), 2018.

P. Gustafson. Local sensitivity of posterior expectations. *Annals of Statistics*, 24(1): 174–195, 1996.

M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

A. Raj, M. Stephens, and J. Pritchard. fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589, 2014.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica sinica*, pages 639–650, 1994.