Evaluating Sensitivity to the Stick Breaking Prior in Bayesian Nonparametrics

Runjing (Bryan) Liu

April 26, 2021

University of California, Berkeley

Collaborators

Runjing (Bryan) Liu
UC Berkeley

Ryan Giordano
MIT

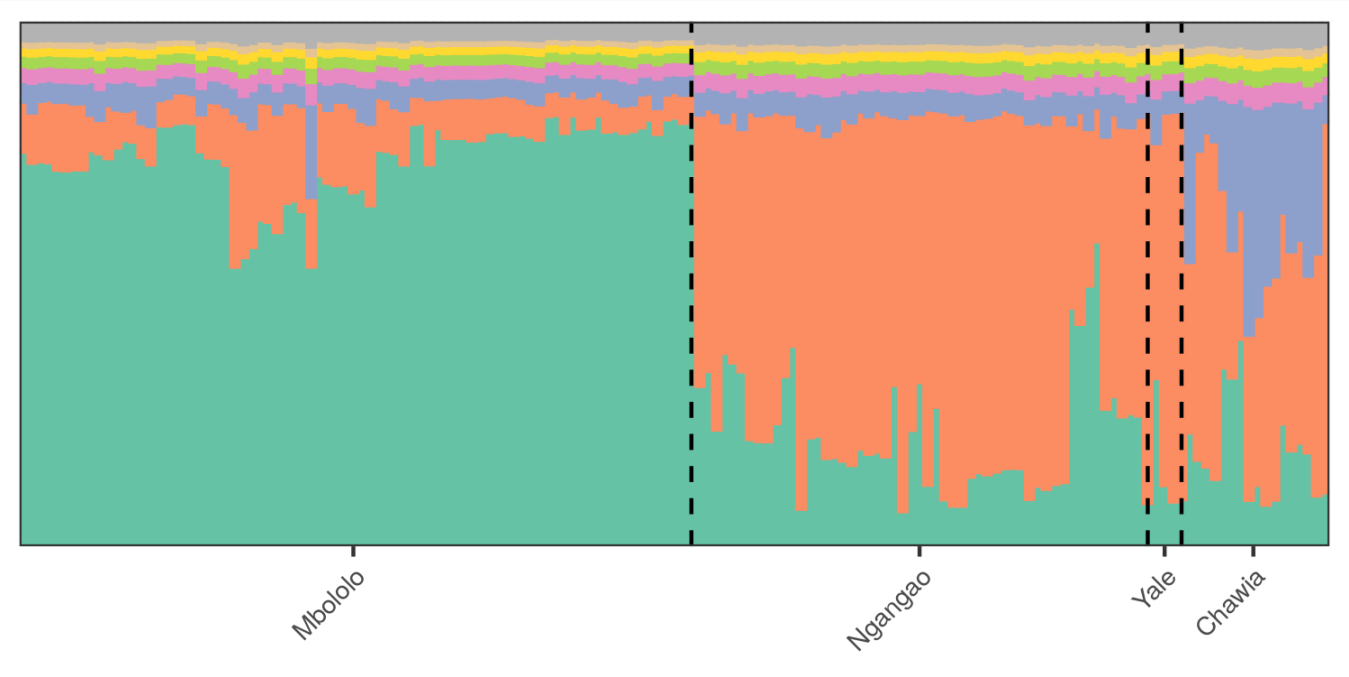
Michael I. Jordan
UC Berkeley

Tamara Broderick
MIT

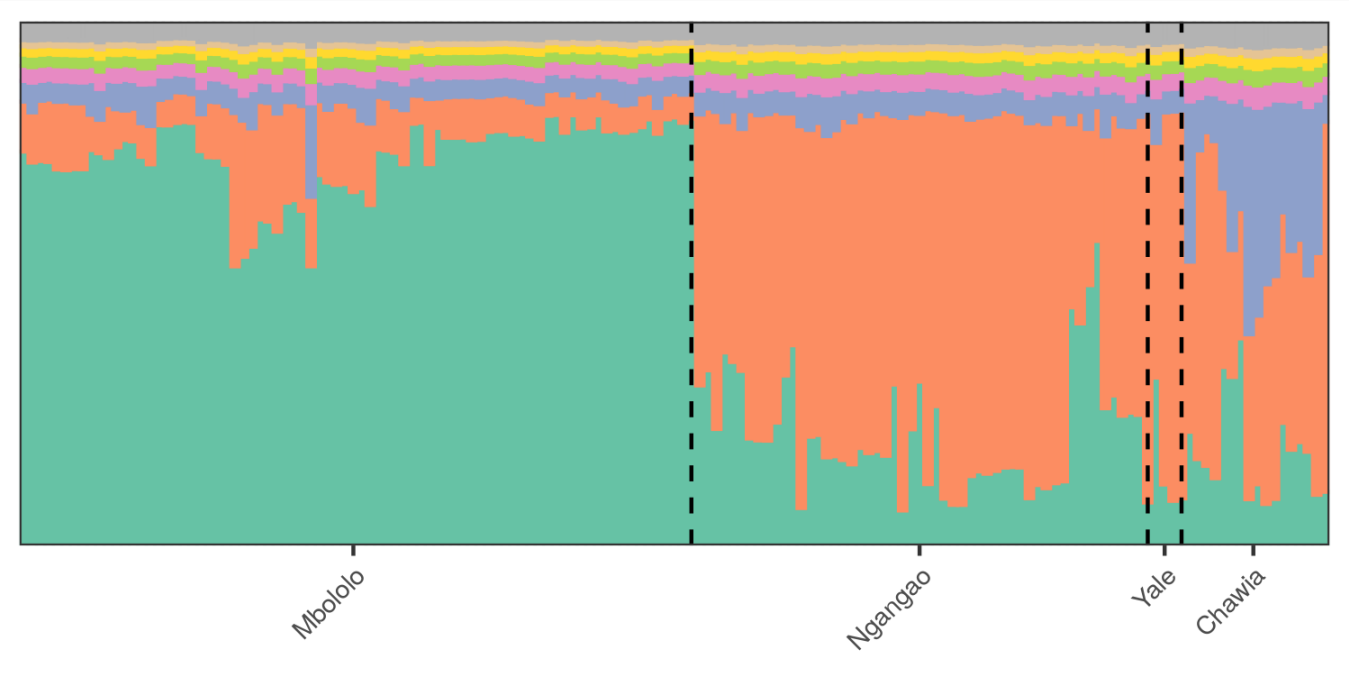
Motivation

Inferring population structure from genomic sequences.

- Genetic data from *Turdus helleri*, an endagered bird species (Pritchard et al. 2011).
- Microsatellites sequences of 155 individiuals at 7 loci.



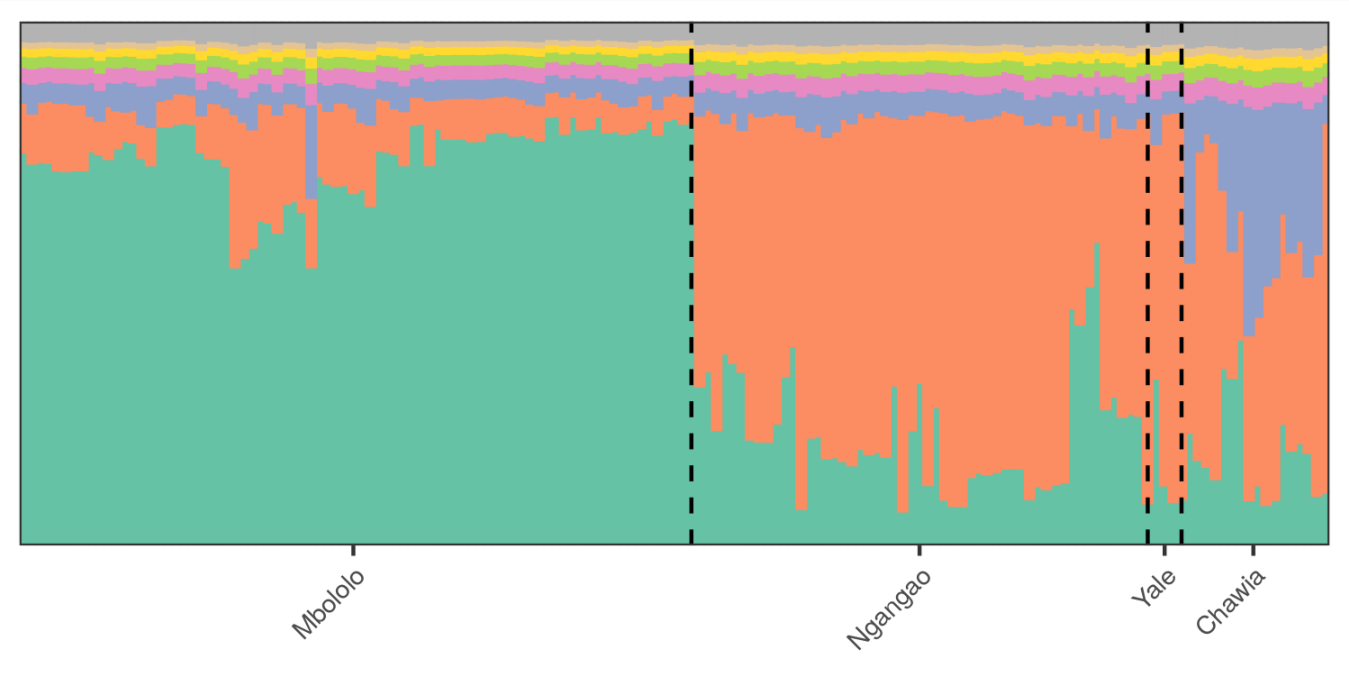
Motivation



Possible questions of interest:

1. How many latent populations—aka clusters—are present in the data set?

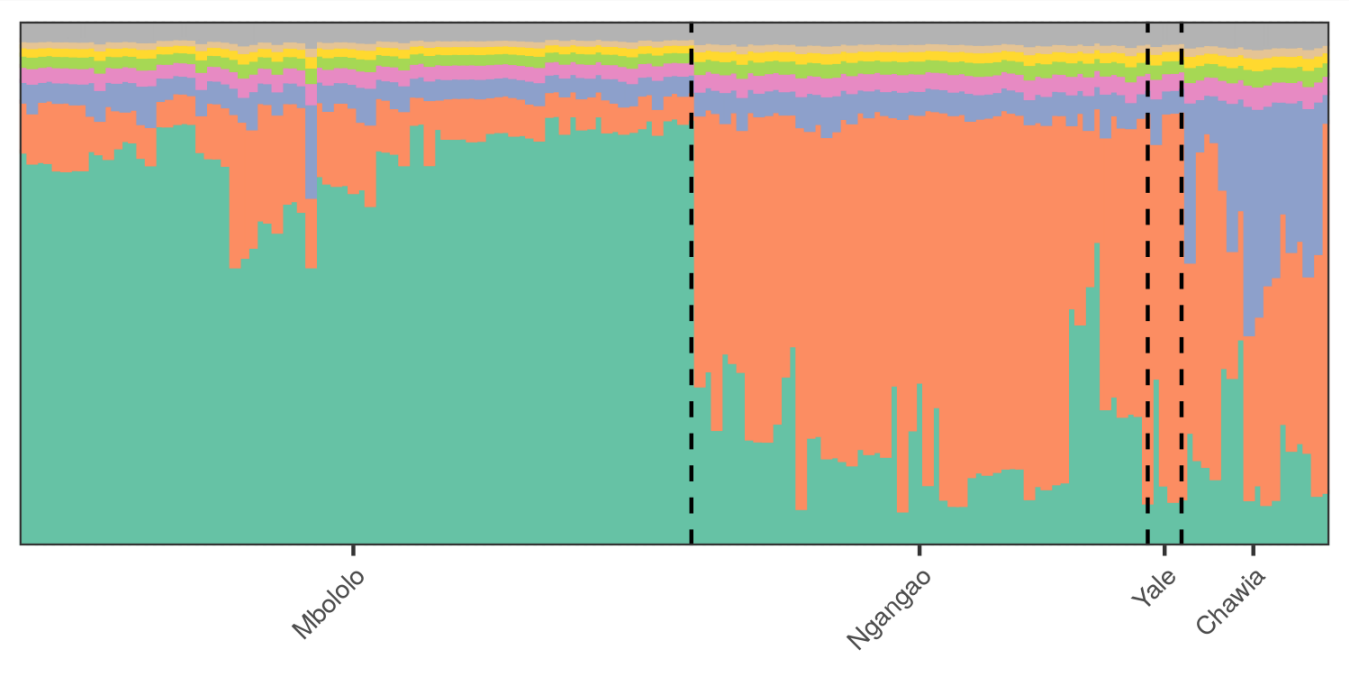
Motivation



Possible questions of interest:

1. How many latent populations—aka clusters—are present in the data set?
2. Which individuals cluster together?

Motivation



Possible questions of interest:

1. How many latent populations—aka clusters—are present in the data set?
2. Which individuals cluster together?
3. What are the unique characterstics of each cluster?

Research problem

A Bayesian nonparametric (BNP) model makes inferring the number of clusters amenable to Bayesian inference.

Research problem

A Bayesian nonparametric (BNP) model makes inferring the number of clusters amenable to Bayesian inference.

We approximate the exact posterior using variational Bayes.

Research problem

A Bayesian nonparametric (BNP) model makes inferring the number of clusters amenable to Bayesian inference.

We approximate the exact posterior using variational Bayes.

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

Research problem

A Bayesian nonparametric (BNP) model makes inferring the number of clusters amenable to Bayesian inference.

We approximate the exact posterior using variational Bayes.

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 Bayesian nonparametric (BNP) model makes inferring the number of clusters amenable to Bayesian inference.

We approximate the exact posterior using variational Bayes.

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 (to avoid expensive refitting).

Outline

- The BNP model
- The variational approximation
- Hyperparameter sensitivity
- Functional sensitivity and influence functions
- Results on data

The BNP Model

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

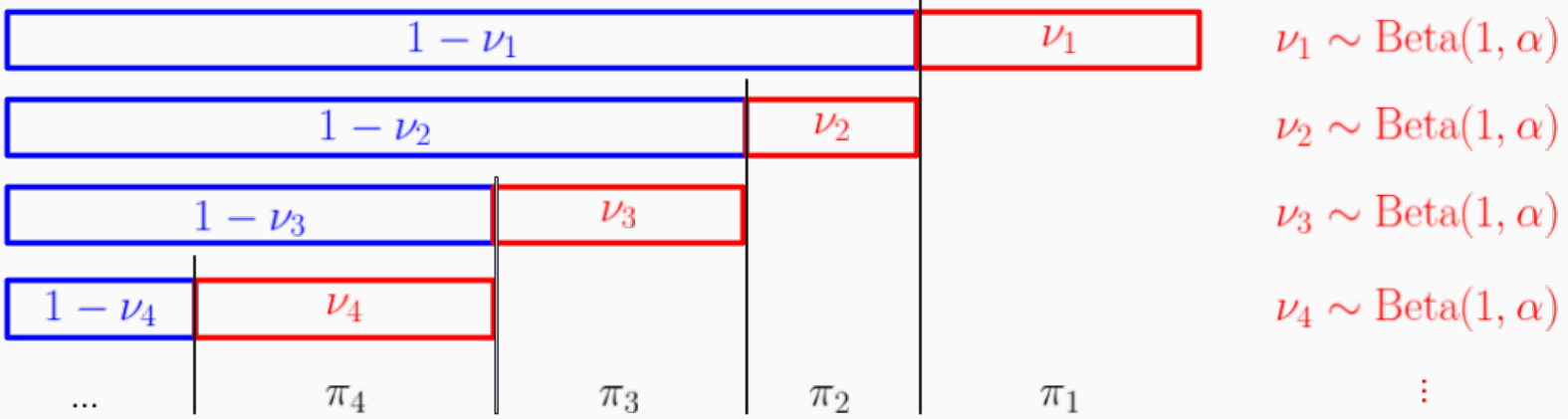


Figure 2: A schematic of the Dirichlet process prior

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

The BNP Model

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

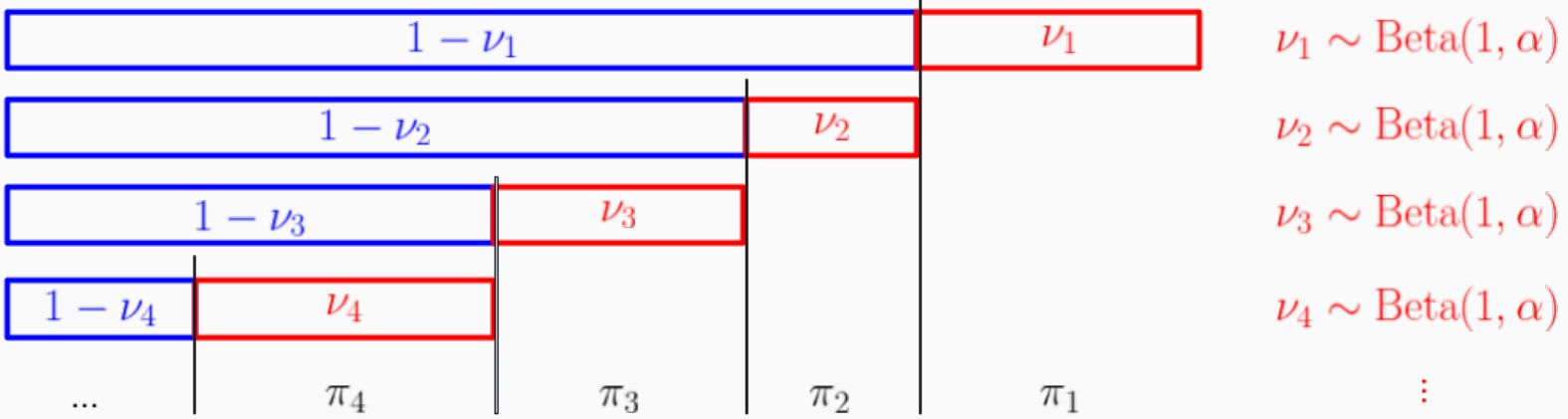


Figure 2: A schematic of the Dirichlet process prior

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

(1) How many clusters are in the *current* dataset?

The BNP Model

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

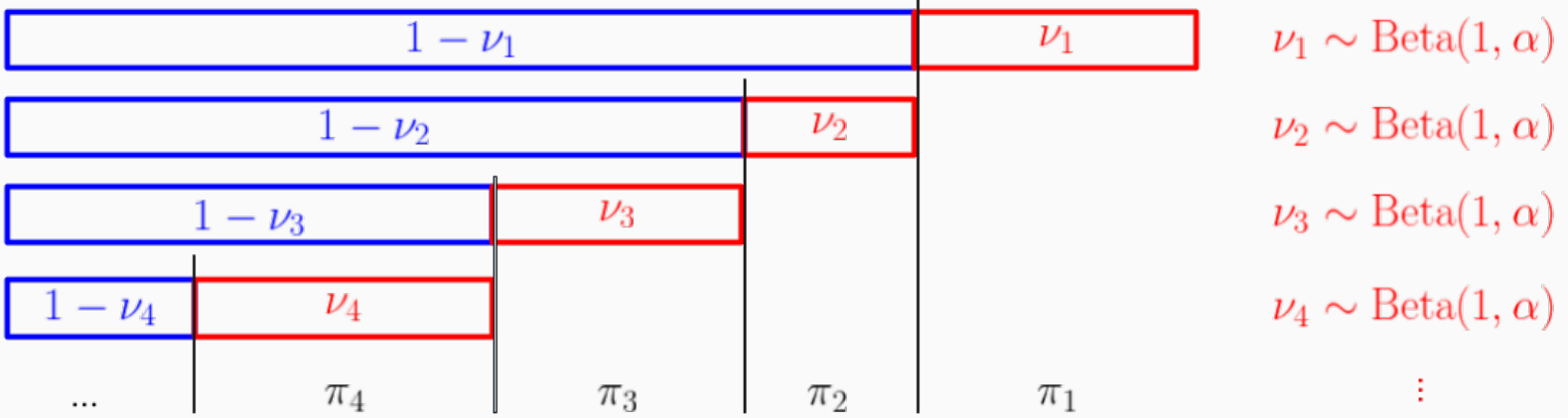


Figure 2: A schematic of the Dirichlet process prior

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

(1) How many clusters are in the *current* dataset?
(2) Given our current knowledge, how many clusters would we expect to see in a *new* dataset?

The BNP Model

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

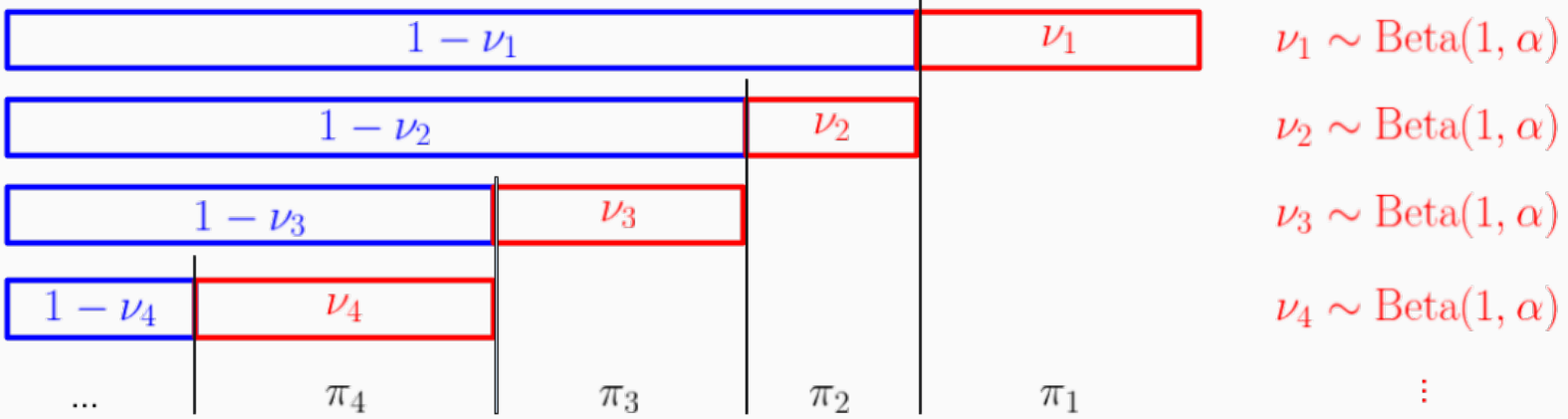


Figure 2: A schematic of the Dirichlet process prior

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

(1) How many clusters are in the *current* dataset?
(2) Given our current knowledge, how many clusters would we expect to see in a *new* dataset?

These quantities depend on the choice of stick-breaking prior.

The BNP Model

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

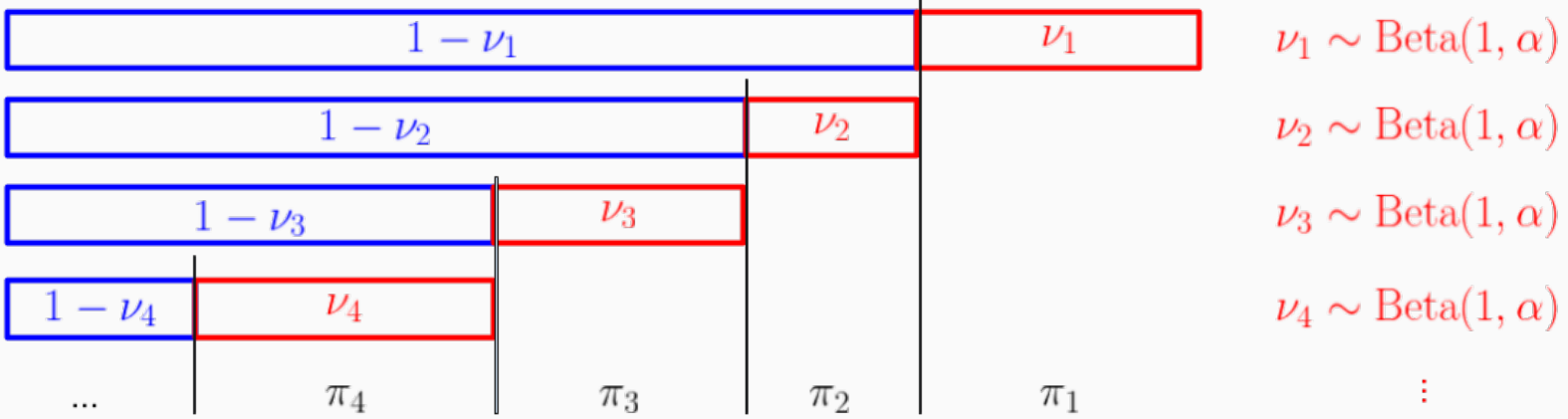


Figure 2: A schematic of the Dirichlet process prior

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

(1) How many clusters are in the *current* dataset?
(2) Given our current knowledge, how many clusters would we expect to see in a *new* dataset?

These quantities depend on the choice of stick-breaking prior.

What makes this stick-breaking prior a reasonable one?

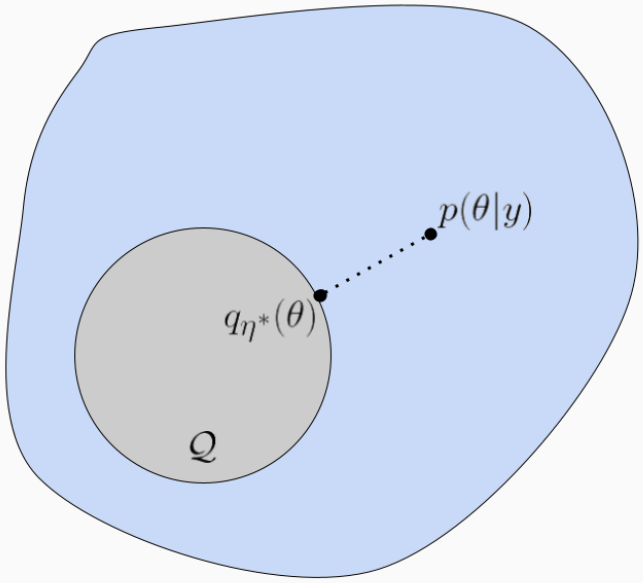
The Variational Approximation

Let θ be latent variables and y the observed data.

We posit a class of mean-field distributions parameterized by a real vector η.

We solve

$$\eta^* = \arg\min_{\eta} KL\left(q(\theta|\eta)\middle\|p(\theta|y)\right)$$



The Variational Approximation

Let θ be latent variables and y the observed data.

We posit a class of mean-field distributions parameterized by a real vector η.

We solve

$$\eta^* = \arg\min_{\eta} KL\left(q(\theta|\eta) \middle\| p(\theta|y)\right)$$

Note that

- The optimal variational parameters η^* depend on the prior through optimizing the KL objective.
- The approximate posterior quantities are then functions of η^*, e.g.

$$\eta^* \mapsto \mathbb{E}_{q_{\eta^*}}[\#\text{clusters}] \quad \text{or} \quad \eta^* \mapsto \mathbb{E}_{q_{\eta^*}}\left[\#\left\{\begin{smallmatrix}\text{clusters in}\\ \text{new dataset}\end{smallmatrix}\right\}\right].$$

The Variational Approximation

Let θ be latent variables and y the observed data.

We posit a class of mean-field distributions parameterized by a real vector η.

We solve

$$\eta^* = \arg\min_{\eta} KL\left(q(\theta|\eta) \middle\| p(\theta|y)\right)$$

Note that

- The optimal variational parameters η^* depend on the prior through optimizing the KL objective.
- The approximate posterior quantities are then functions of η^*, e.g.

$$\eta^* \mapsto \mathbb{E}_{q_{\eta^*}}[\#\text{clusters}] \quad \text{or} \quad \eta^* \mapsto \mathbb{E}_{q_{\eta^*}}\left[\#\left\{\begin{smallmatrix}\text{clusters in}\\ \text{new dataset}\end{smallmatrix}\right\}\right].$$

How do these approximate posterior quantities depend on the DP prior?

Hyperparameter Sensitivity

Let ϵ be a real-valued hyperparameter for the stick-breaking distribution (e. g., this could be the α concentration parameter, or it could parameterize a functional shape).

Hyperparameter Sensitivity

Let ϵ be a real-valued hyperparameter for the stick-breaking distribution (e. g., this could be the α concentration parameter, or it could parameterize a functional shape).

Main idea: We approximate the dependence of η^* on ϵ with a first-order Taylor expansion:

$$\eta^*(\epsilon) \approx \eta^*(0) + \frac{d\eta^*(\epsilon)}{d\epsilon^T}\Big|_{\epsilon=0} \epsilon$$

Hyperparameter Sensitivity

Let ϵ be a real-valued hyperparameter for the stick-breaking distribution (e. g., this could be the α concentration parameter, or it could parameterize a functional shape).

Main idea: We approximate the dependence of η^* on ϵ with a first-order Taylor expansion:

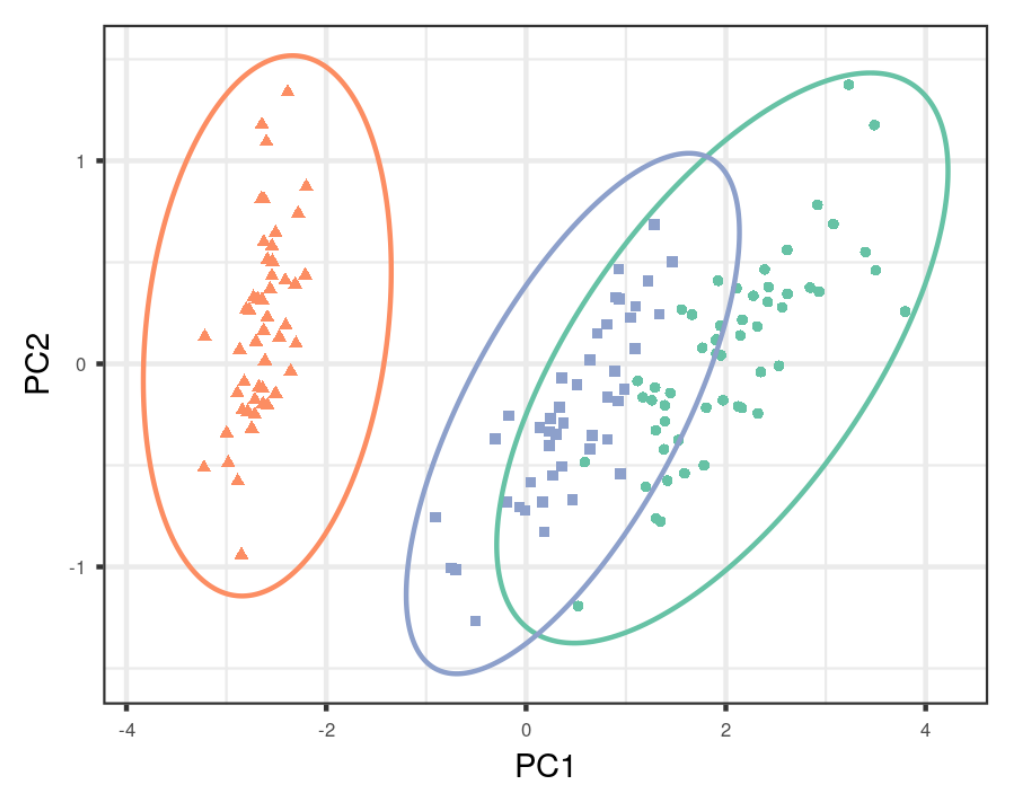
$$\eta^*(\epsilon) \approx \eta^*(0) + \frac{d\eta^*(\epsilon)}{d\epsilon^T}\Big|_{\epsilon=0} \epsilon$$

Notes:

- Evaluation of the derivative can be done efficiently using formulas from Giordano et al. 2018 and modern automatic differentiation tools.
- We only use a linear approximation for the map $\epsilon \mapsto \eta^*(\epsilon)$. We retain nonlinearities in the map $\eta^* \mapsto \mathbb{E}_{q_{\eta^*}}$ [#clusters].

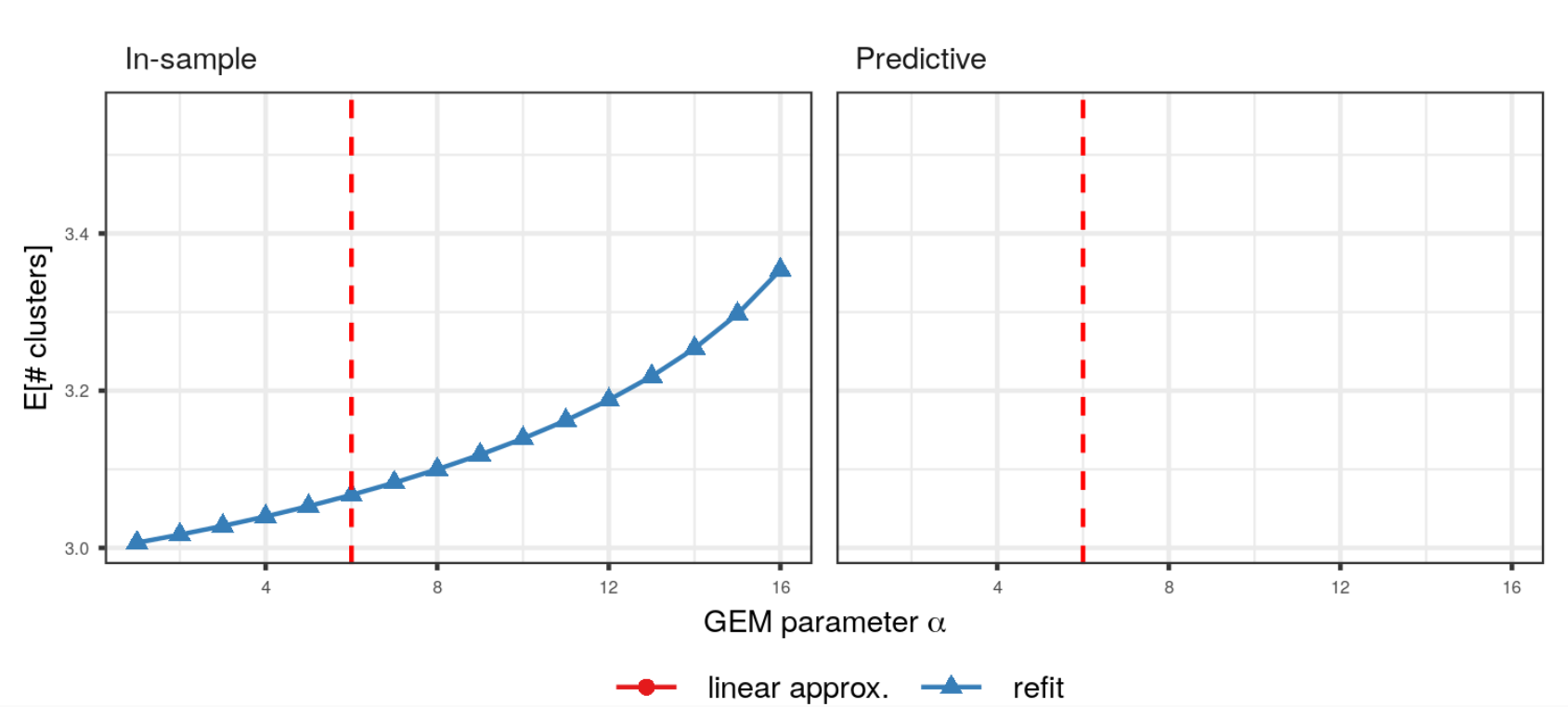
A simple example: iris data

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



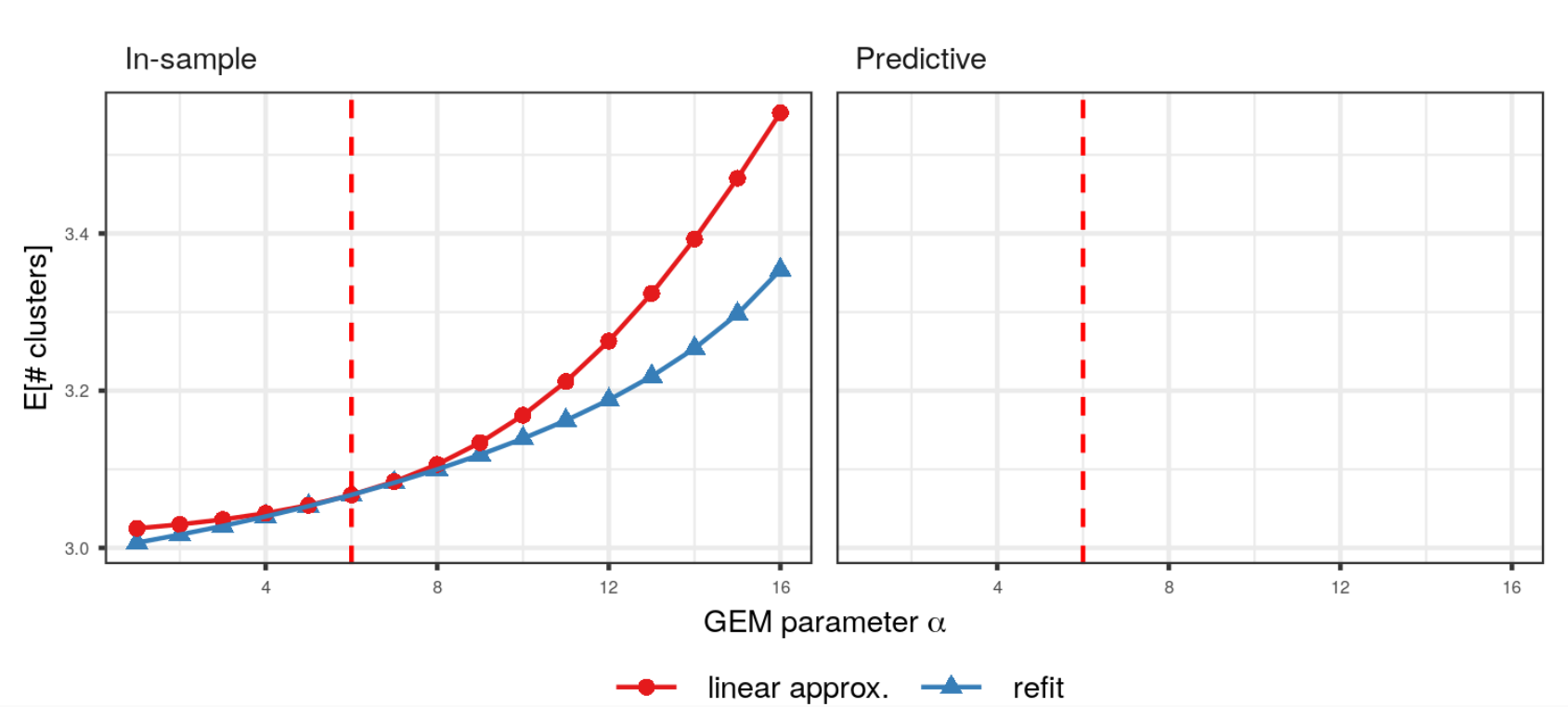
The iris data in principal component space and GMM fit at $\alpha = 6$.

iris data: parametric sensitivity



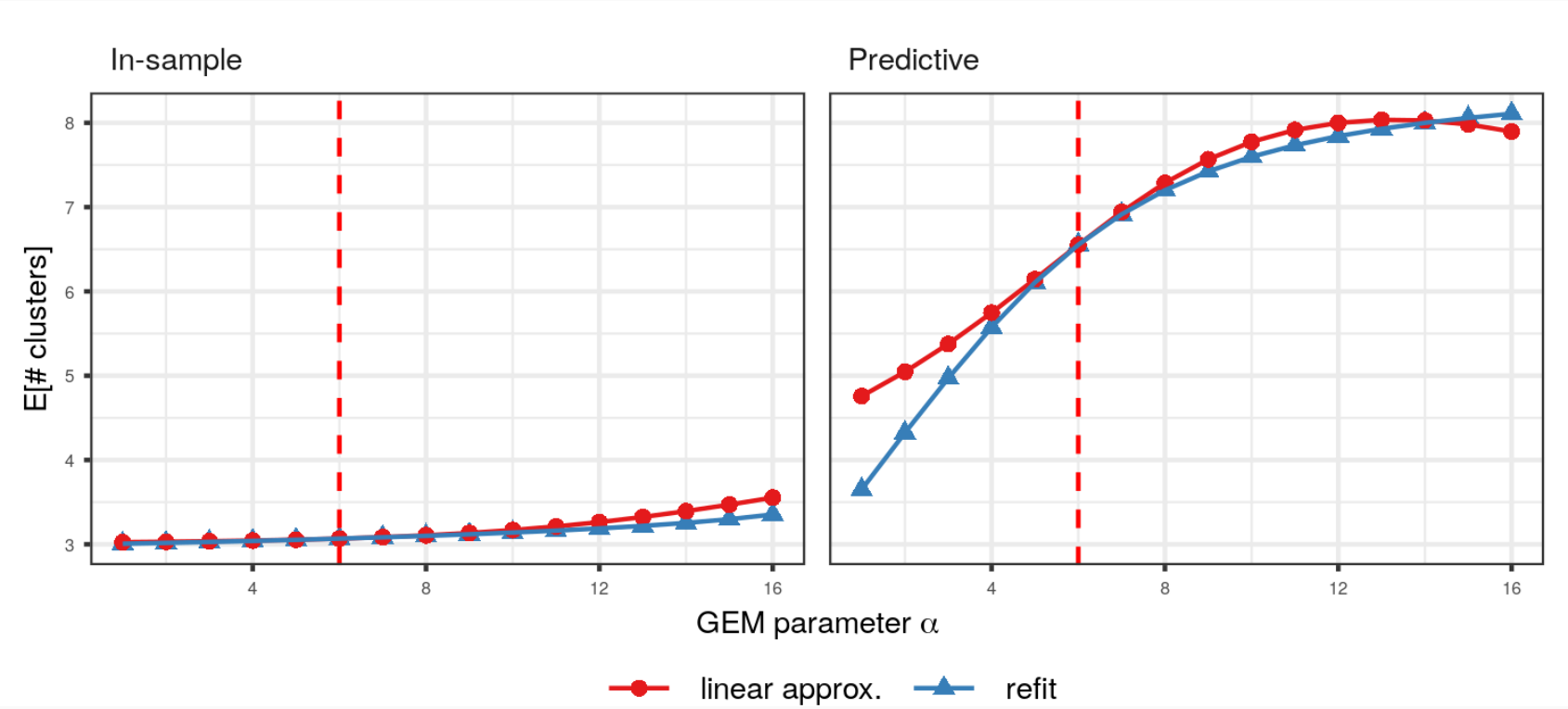
The expected number of posterior clusters in the iris data as α varies.

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.

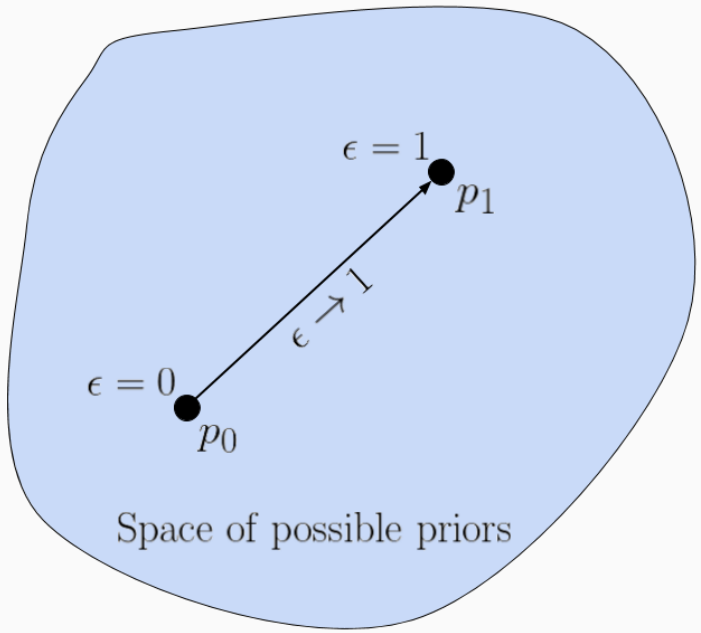
Functional sensitivity

Let $p_0(\nu_k)$ be the original Beta prior on sticks.

Suppose we wish to replace p_0 with another distribution p_1. Define the "contaminated" prior as:

$$p_c(\nu_k|\epsilon) \propto p_0(\nu_k) \exp(\epsilon\phi(\nu_k))$$

where $\phi(\nu_k) = \log p_1(\nu_k) - \log p_0(\nu_k)$.



Functional sensitivity: influence functions

Consider a posterior statistic of interest $g(\eta)$, e.g.

$$g_{\mathrm{cl}}(\eta) = \mathbb{E}_{q_\eta}\left[\#\text{clusters}\right]$$

Let S_g be the *local sensitivity* of g with respect to a hyper-parameter ϵ

$$S_g := \frac{d}{d\epsilon}\, g(\eta(\epsilon))$$

Functional sensitivity: influence functions

Consider a posterior statistic of interest $g(\eta)$, e.g.

$$g_{\mathrm{cl}}(\eta) = \mathbb{E}_{q_\eta}\left[\#\text{clusters}\right]$$

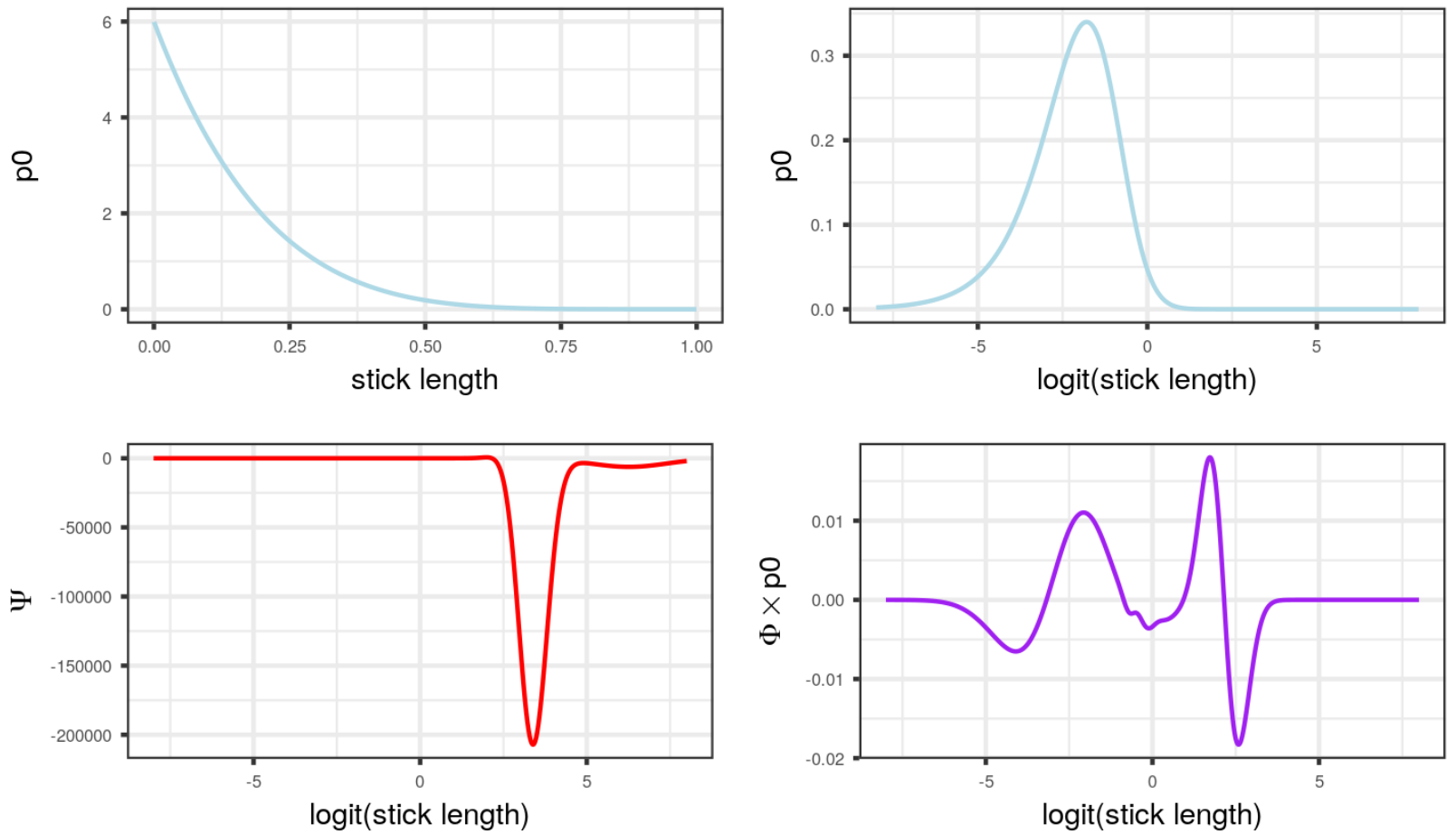
Let S_g be the *local sensitivity* of g with respect to a hyper-parameter ϵ

$$S_g := \frac{d}{d\epsilon}\, g(\eta(\epsilon))$$

The local sensitivity can be expressed as an inner-product between an *influence function* Ψ and the functional perturbation ϕ in an appropriate Hilbert space:

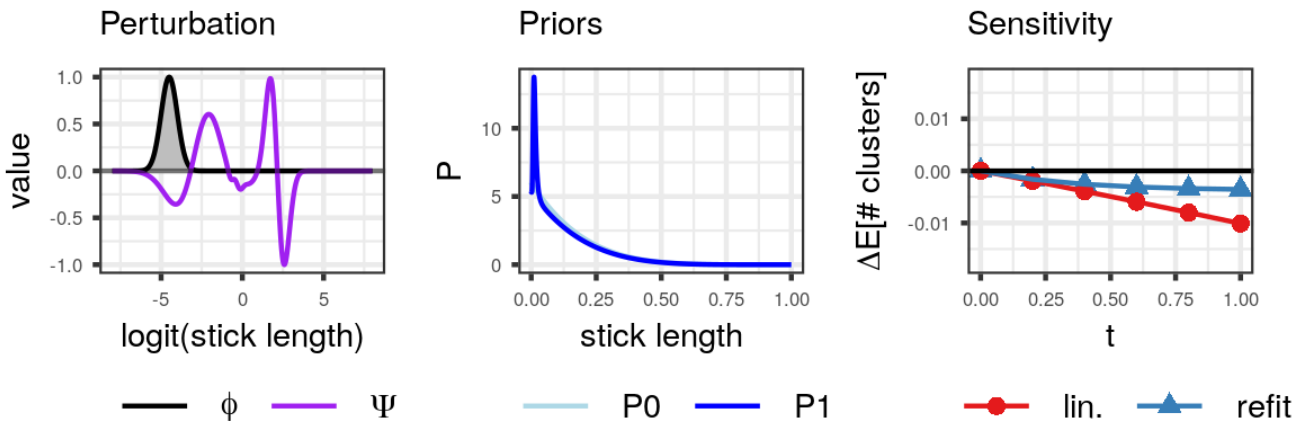
$$\begin{aligned} S_g &= \langle \Psi, \phi \rangle \\ &= \int \Psi(\nu)\phi(\nu)p_0(\nu)\, d\nu \end{aligned}$$

Iris data: influence functions

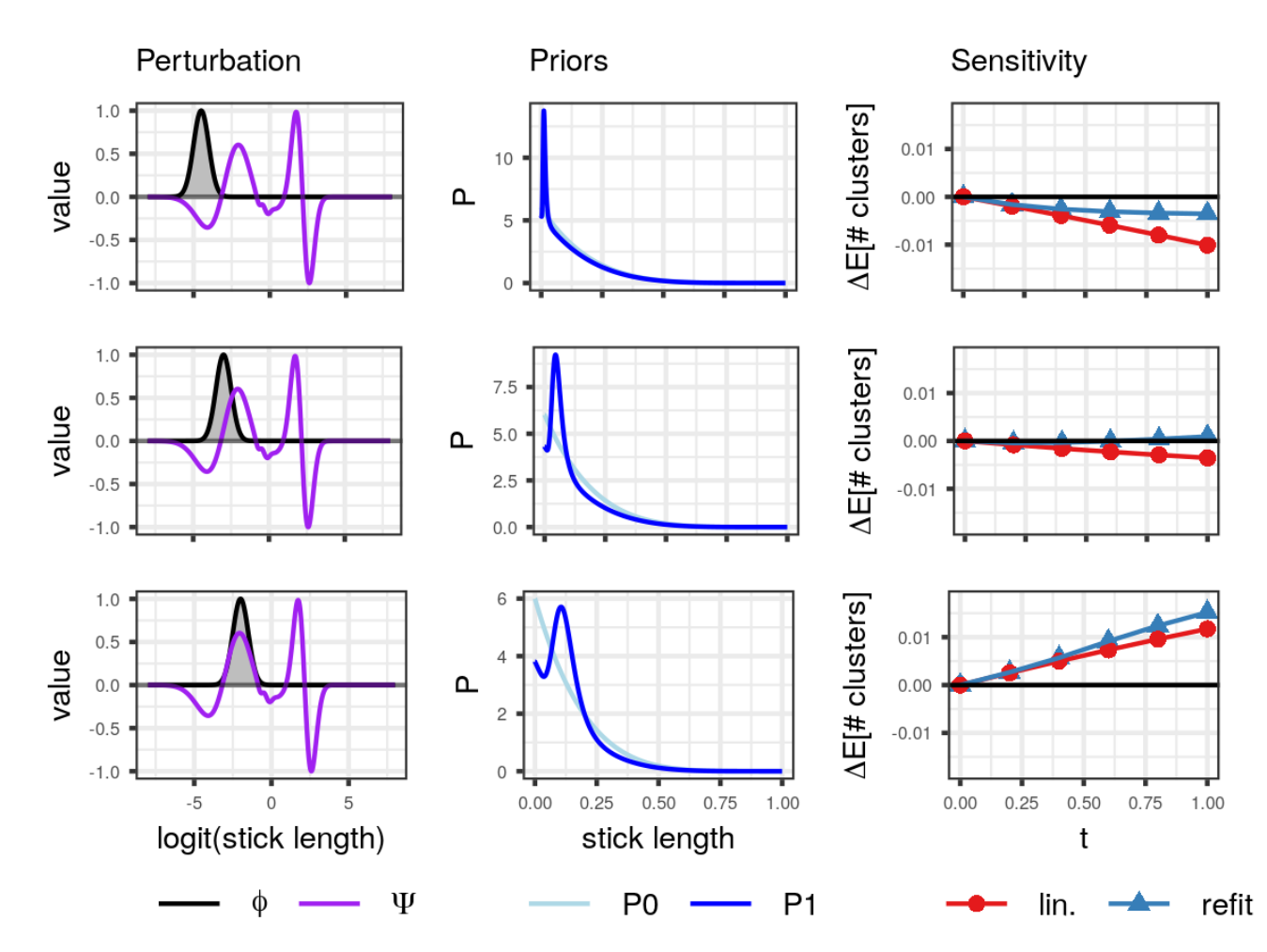


The influence function for the number of clusters, g_{cl}.

Iris data: functional perturbations



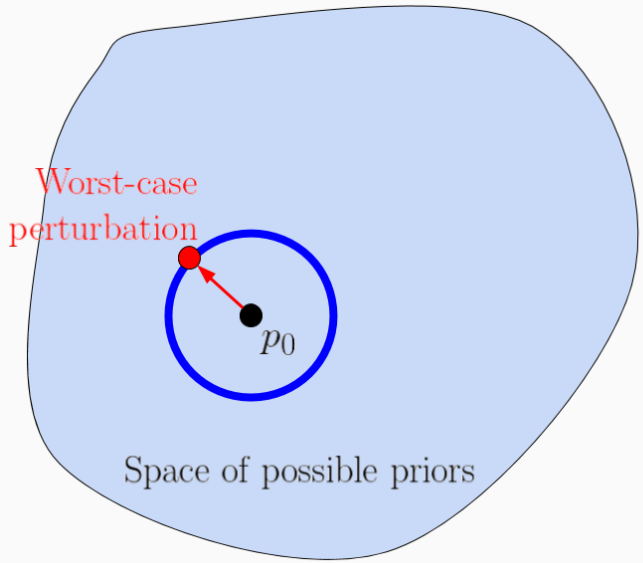
Iris data: functional perturbations



Functional perturbations: worst-case

There are many possible choices for p_1.

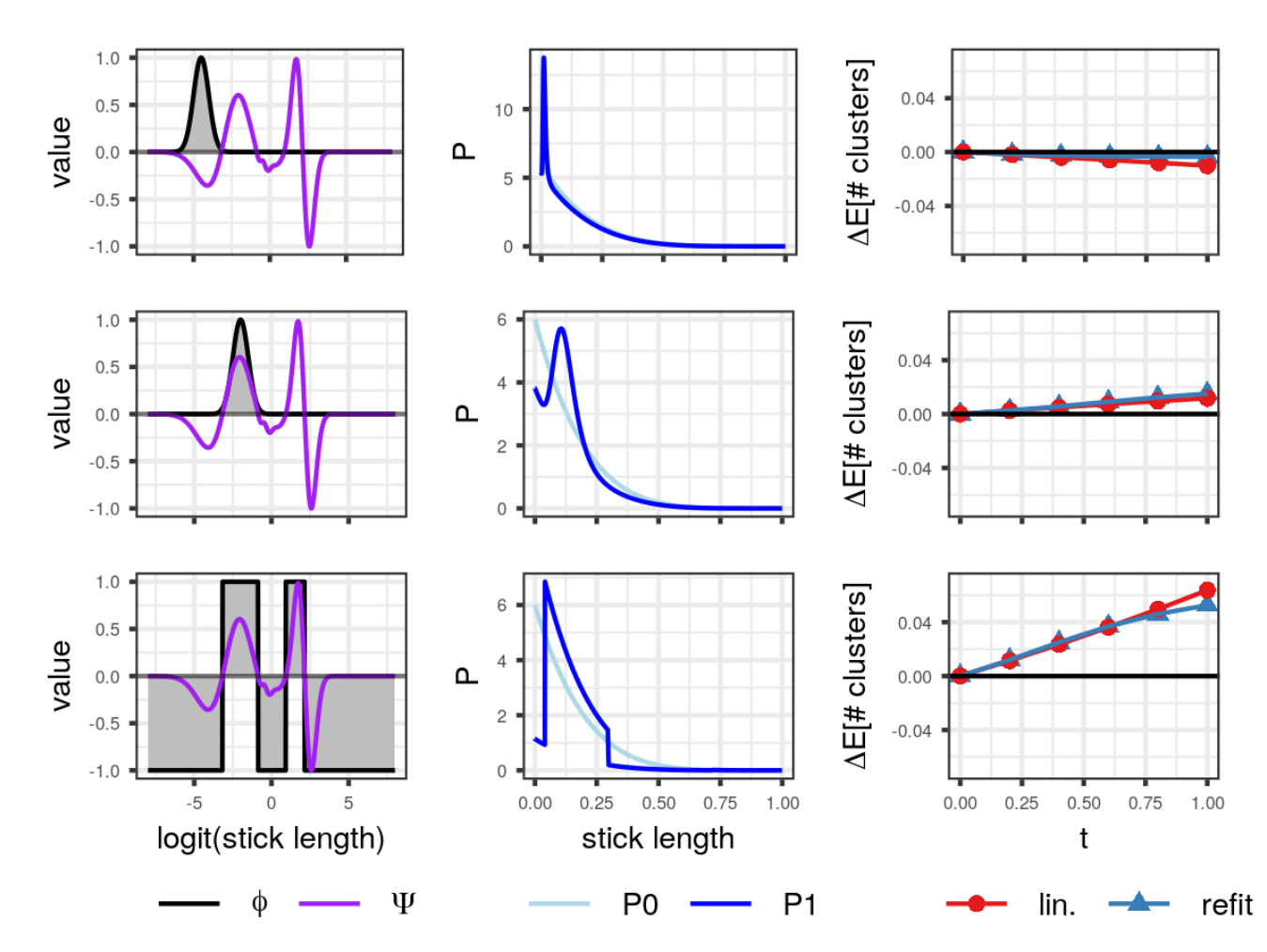
Given a posterior quantity g, we can find the *worst-case* perturbation in a ball of radius δ, that is, find the direction such that $|S_g|$ is maximized.



Specifically, we consider the L-infinity ball of radius δ:

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

Iris data: worst-case perturbation



Results on STRUCTURE

We adapt STRUCTURE (Pritchard et al. 2011; Raj et al. 2014) , a Bayesian model for population genetics, to include a BNP prior.

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

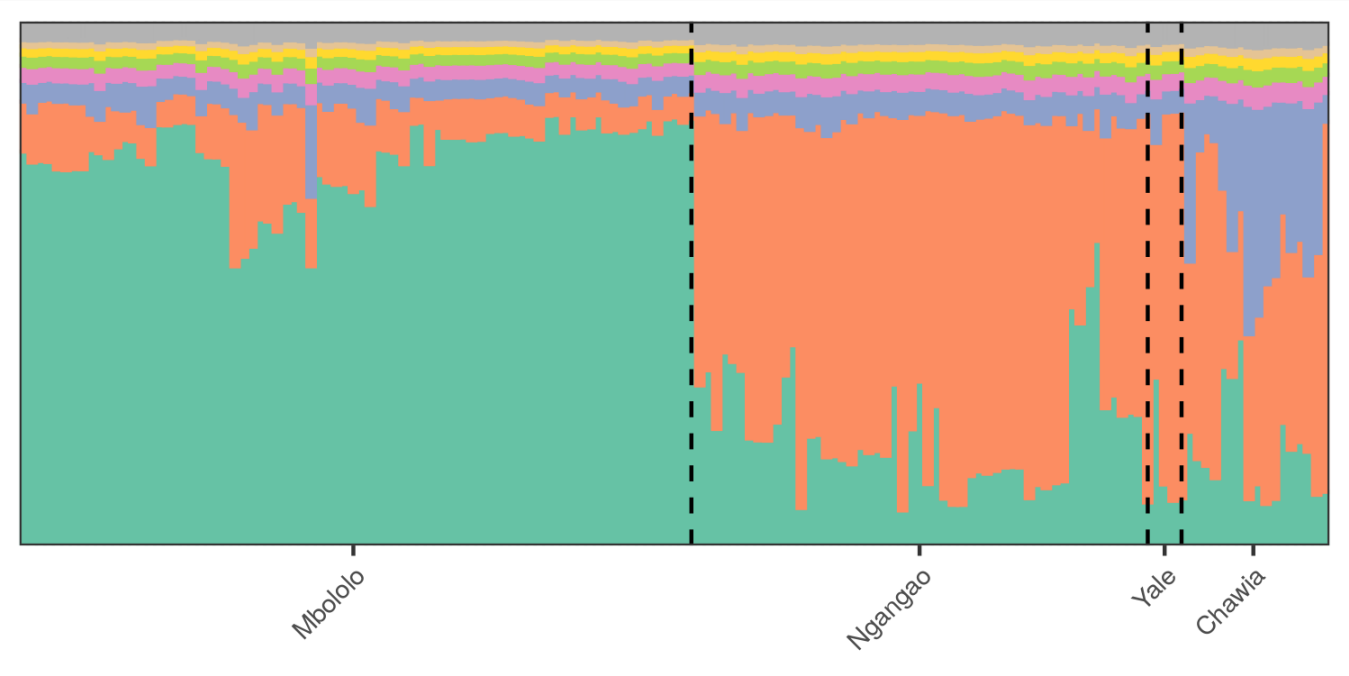
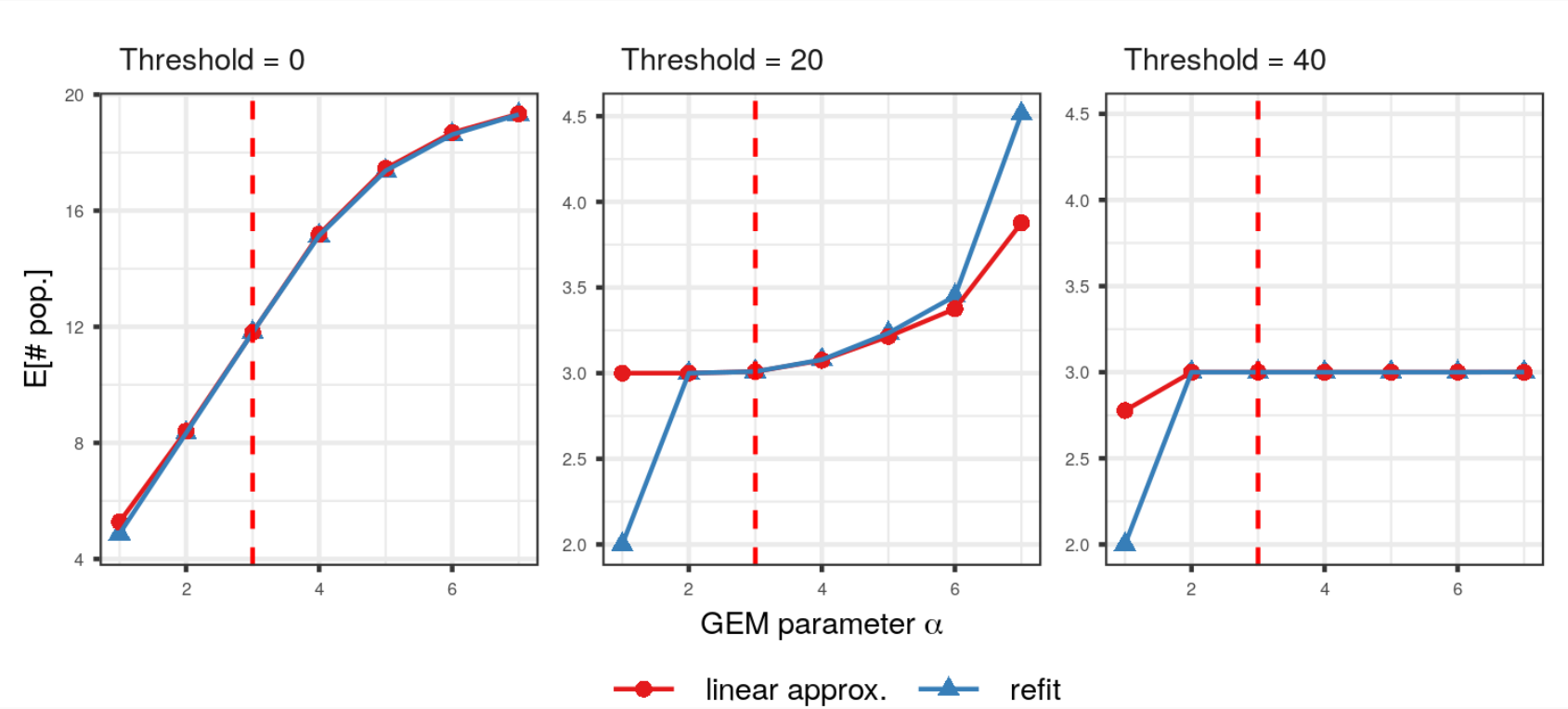


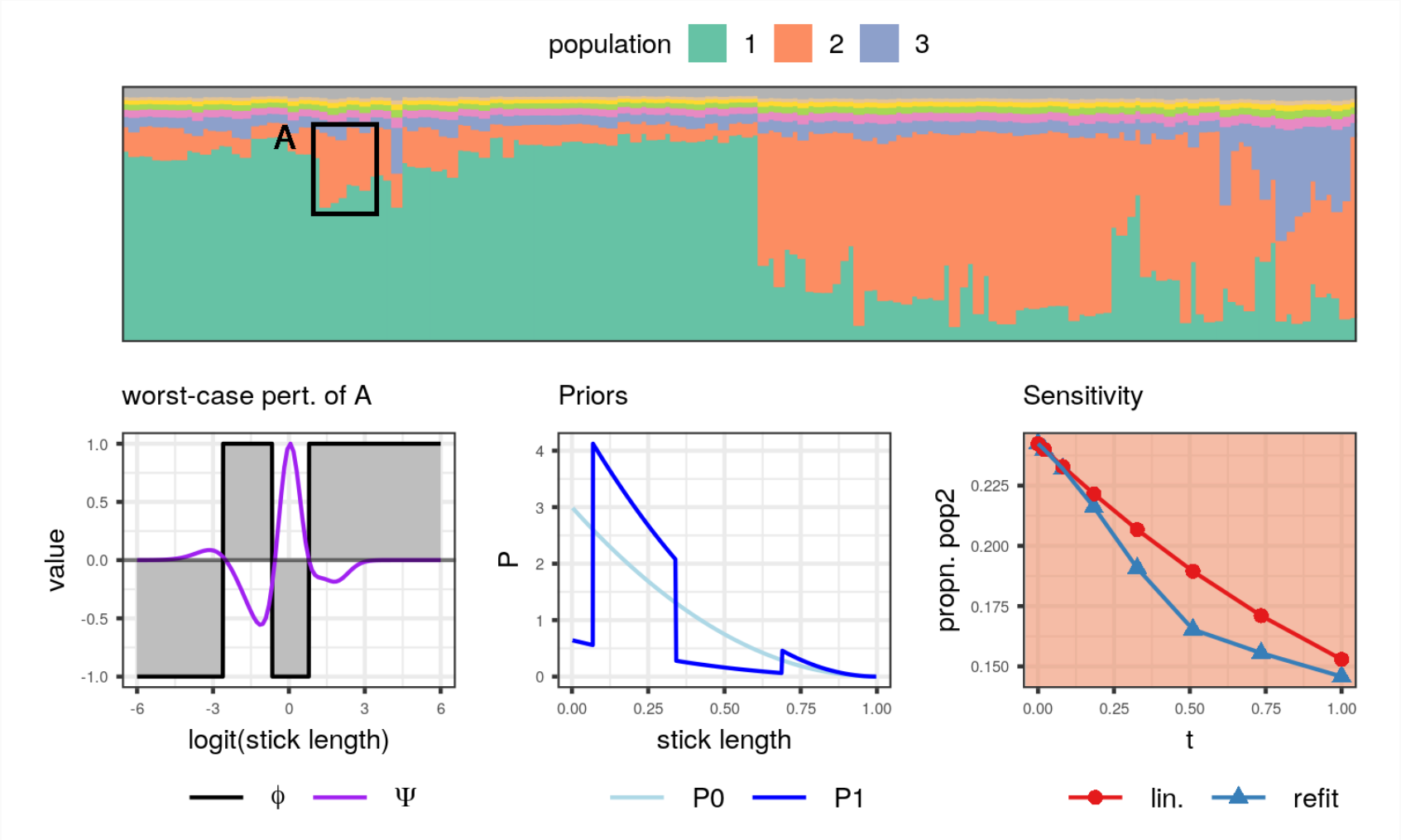
Figure 3: The intitial fit at $\alpha = 3$.

STRUCTURE: parametric sensitivity

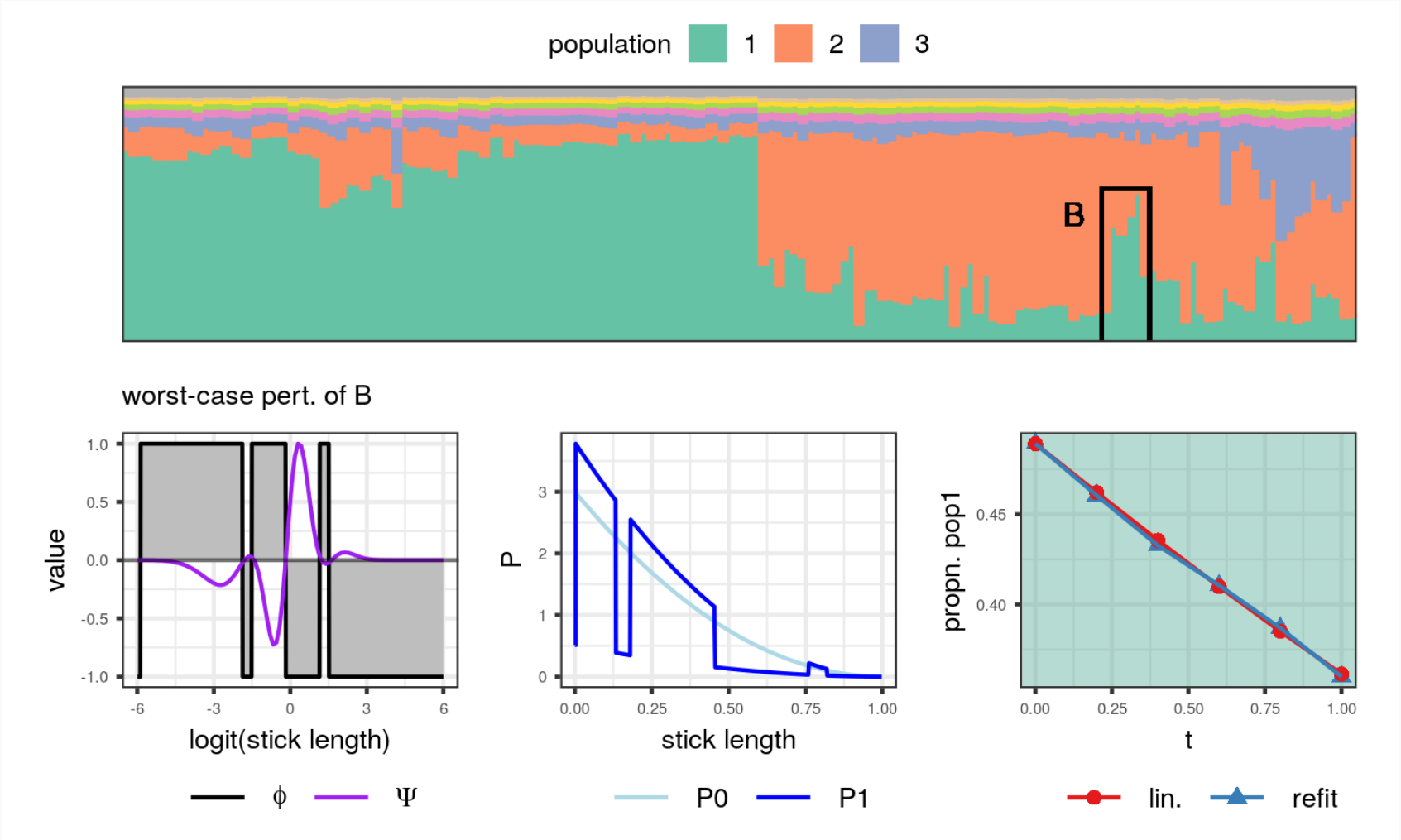


The expected number of posterior in-sample clusters in the thrush data as α varies.

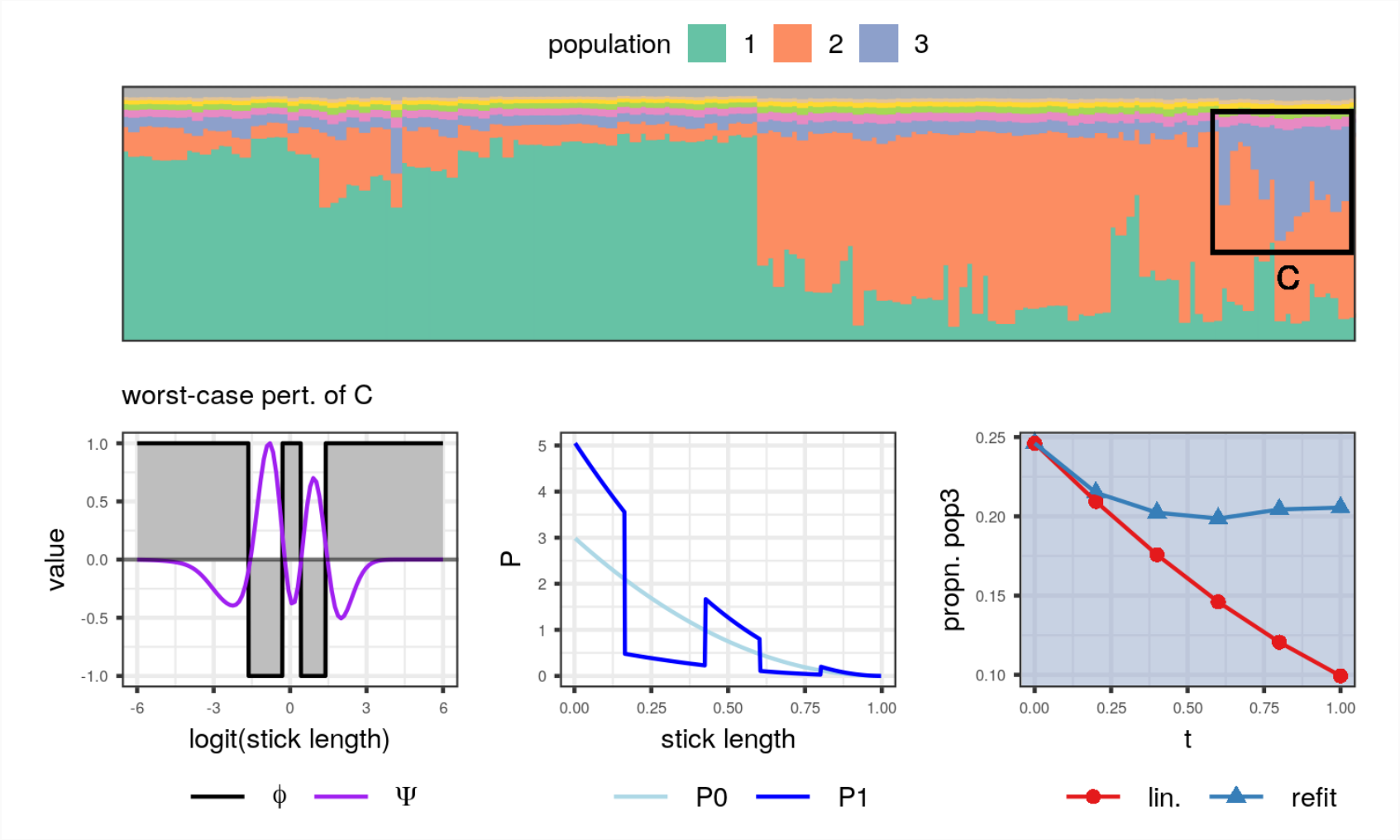
STRUCTURE: functional sensitivity



STRUCTURE: functional sensitivity



STRUCTURE: functional sensitivity

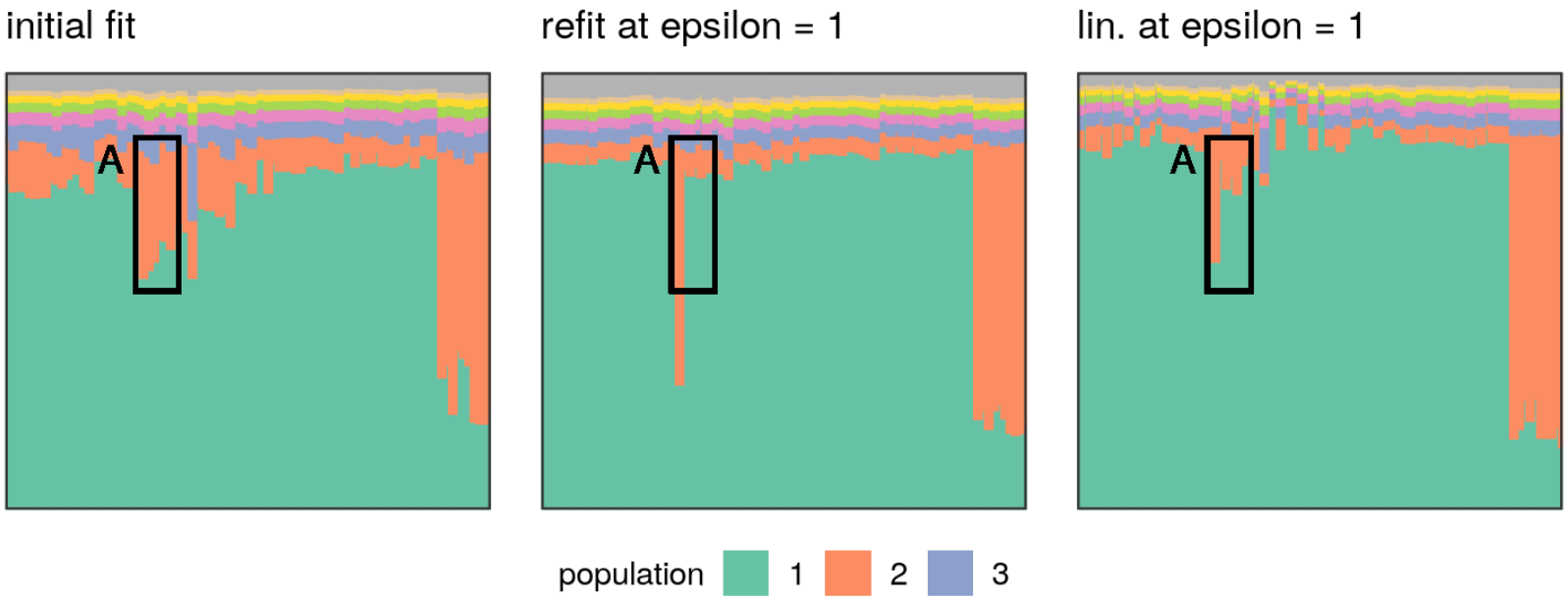


Computational times

Compute time of results on the Taita thrush dataset.

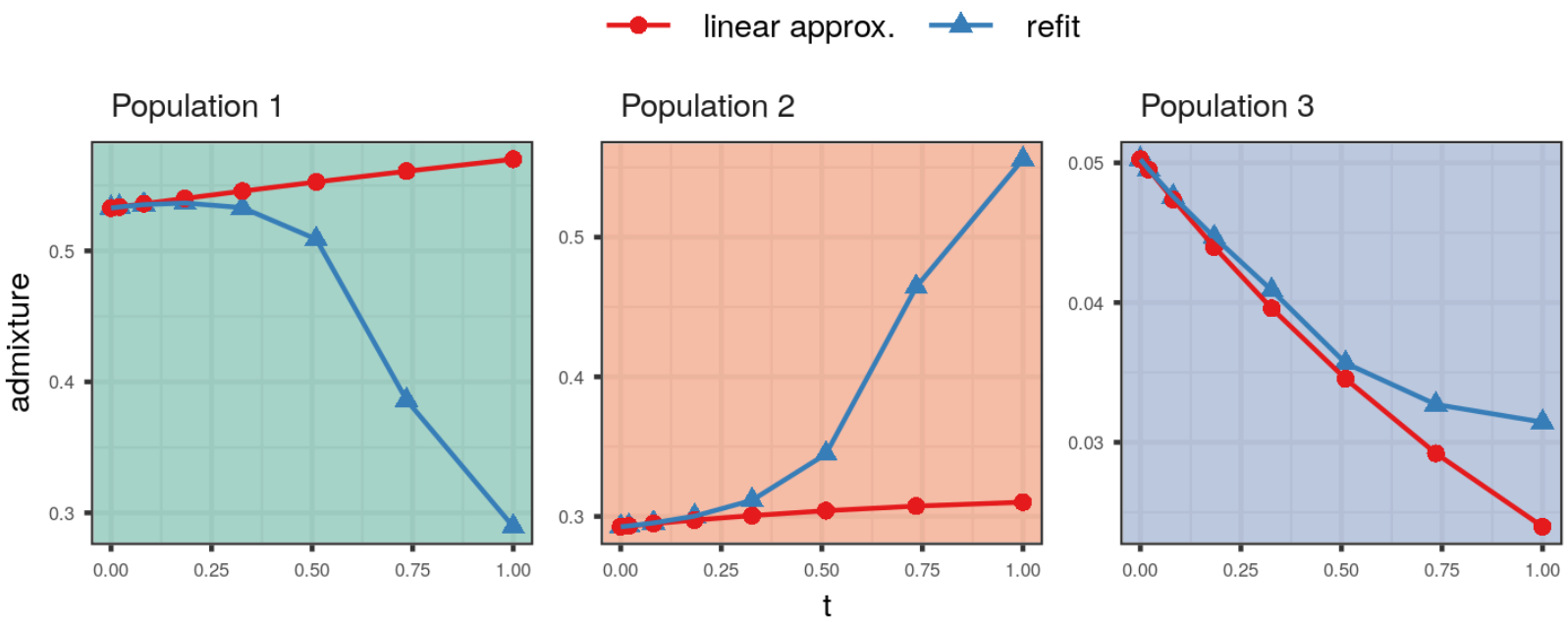
	time (seconds)
Initial fit	7
Hessian solve for α sensitivity	0.3
Linear approx. $\eta^{lin}(\alpha)$ for $\alpha = 1, ..., 7$	0.006
Refits $\eta(\alpha)$ for $\alpha = 1, ..., 7$	30
The influence function	0.6
Hessian solve for worst-case ϕ	0.4
Linear approx. $\eta^{lin}(\epsilon)\|_{\epsilon=1}$ for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

Limitations of local sensitivity

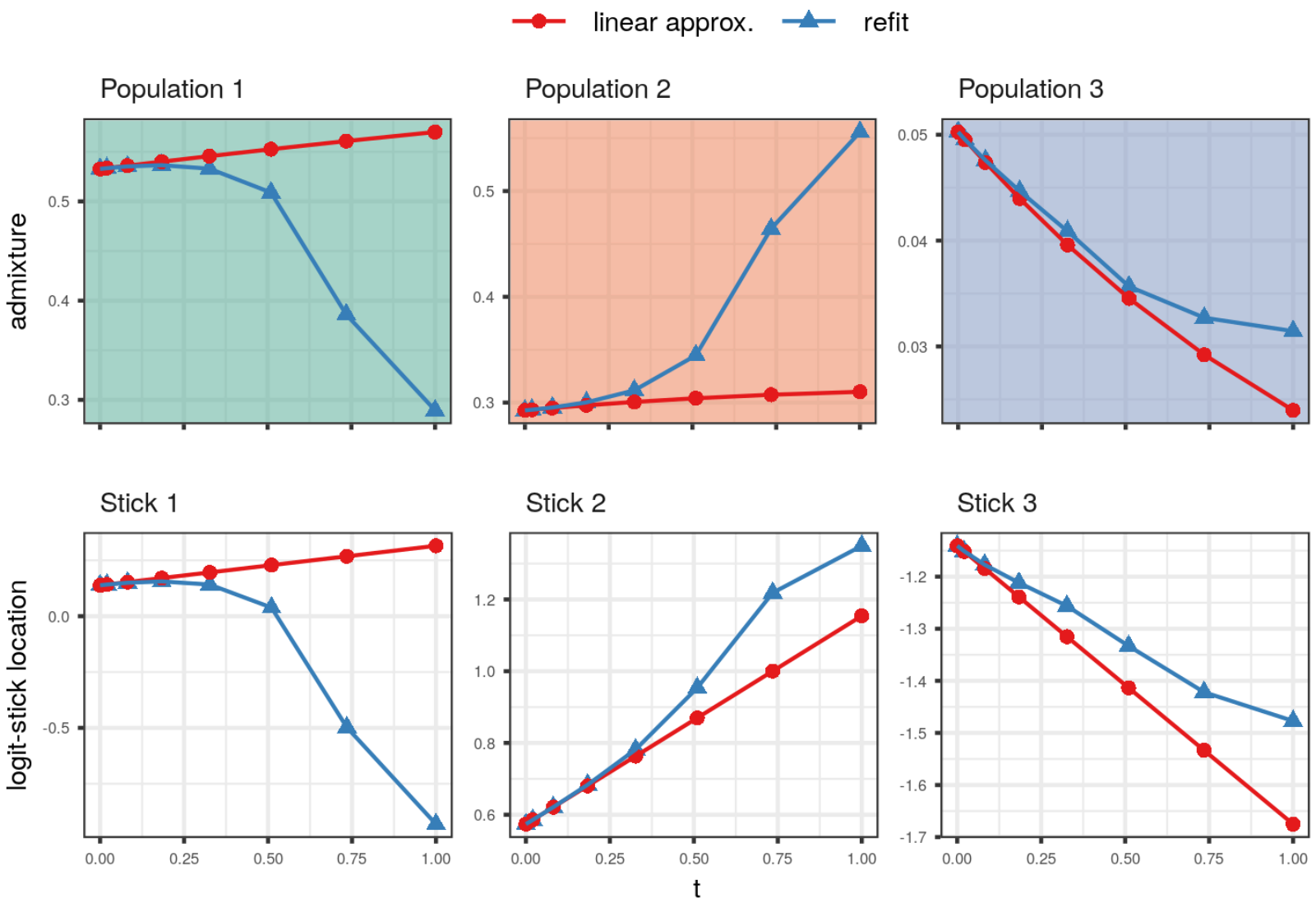


Inferred admixtures after the worst-case perturbation to individuals A. Individual $n = 26$ had a large increase in admixture proportion of population 2 after the refit.

Limitations of local sensitivity



Limitations of local sensitivity



Conclusions

- We provide a tool to efficiently evaluate the sensitivity of the variational posterior to prior chioces.
- Linearizing the variational parameters provides a reasonable alternative re-optimizing the variational approximation after model perturbations.
- The influence function can provide guidance to find particularly sensitive model perturbations.

References

A workshop paper:

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

Code:

Paragami: parameter folding and flattening for optimization problems

https://github.com/rgiordan/paragami

Vittles: library for sensitivity analysis in optimization problems

https://pypi.org/project/vittles/

JAX: composable transformations of Python+NumPy programs

https://github.com/google/jax