

Bayesian Nonparametrics

A tutorial with applications in Brain Imaging Analysis

Michele Guindani

Department of Biostatistics
University of California, Los Angeles

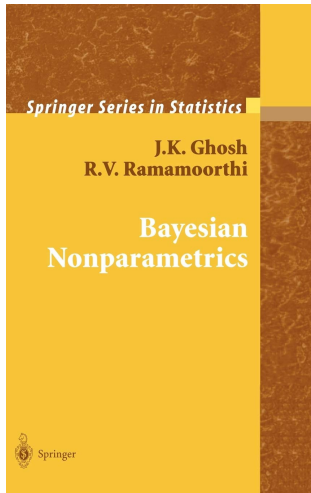
O'Bayes 2022

Slides available here: http://bit.ly/BNP_OBayes2022

- A tutorial!! Daunting task: more qualified people have already done much better!

- ❑ A tutorial!! Daunting task: more qualified people have already done much better! And we have less than an hour! 😓

- ❑ A tutorial!! Daunting task: more qualified people have already done much better! And we have less than an hour! 😓



Bayesians believe that all inference and more is Bayesian territory.

So, it is natural that a Bayesian should explore non-parametrics and other infinite-dimensional problems.

However, putting a prior, which is always a delicate and difficult exercise in Bayesian analysis, poses special conceptual, mathematical, and practical difficulties in infinite-dimensional problems.

Can one really have a subjective prior based on knowledge and belief, in an infinite-dimensional space?

*Even if one settles for a largely non-subjective prior, it is mathematically difficult to construct prior distributions on such sets as the space of all distribution functions or the space of all probability density functions and ensure that they have large support, which is a minimum requirement because **a largely non-subjective prior should not put too much mass on a small set.***

J.K. Ghosh & R.V. Ramamoorthi (2003),
Bayesian Nonparametrics, Springer

- At its essence, a Bayesian non-parametric approach can be broadly described by models that put a prior on an infinite-dimensional object.

In the literature, one can see the term applied to:

- At its essence, a Bayesian non-parametric approach can be broadly described by models that put a prior on an infinite-dimensional object.

In the literature, one can see the term applied to:

- 👉 **Functional data analysis/functional regression:**

$$Y_i = f(X_i) + \epsilon_i,$$

e.g. by using Gaussian processes

- At its essence, a Bayesian non-parametric approach can be broadly described by models that put a prior on an infinite-dimensional object.

In the literature, one can see the term applied to:

- 👉 **Functional data analysis/functional regression:**

$$Y_i = f(X_i) + \epsilon_i,$$

e.g. by using Gaussian processes

- 👉 **Bayesian Additive Regression Trees (BART):** flexible modeling of the relationships between covariates and outcomes

- At its essence, a Bayesian non-parametric approach can be broadly described by models that put a prior on an infinite-dimensional object.



In the literature, one can see the term applied to:




- 👉 **Functional data analysis/functional regression:**




$$Y_i = f(X_i) + \epsilon_i,$$

e.g. by using Gaussian processes

- 👉 **Bayesian Additive Regression Trees (BART):** flexible modeling of the relationships between covariates and outcomes
- 👉 **Density Estimation:** provide the flexibility necessary to analyze complex data beyond simple parametric assumptions:
 - Dirichlet Processes (DP)
 - Polya Tree priors
and their generalizations (Dependent DP, Normalized Random Measures...)




-  Parametric models make restrictive assumptions about the data-generating mechanism (e.g. the data are generated from a Normal distribution)
-  If the data do not follow the assumed DGM (they rarely do), the distributional assumption may cause serious biases in the inference

-  Parametric models make restrictive assumptions about the data-generating mechanism (e.g. the data are generated from a Normal distribution)
-  If the data do not follow the assumed DGM (they rarely do), the distributional assumption may cause serious biases in the inference
-  A parametric model $X|\theta \sim p_\theta$ for $\theta \in \Theta \subset \mathbb{R}^d$ with a prior specification $\theta \sim \pi$

-  Parametric models make restrictive assumptions about the data-generating mechanism (e.g. the data are generated from a Normal distribution)
-  If the data do not follow the assumed DGM (they rarely do), the distributional assumption may cause serious biases in the inference
-  A parametric model $X|\theta \sim p_\theta$ for $\theta \in \Theta \subset \mathbb{R}^d$ with a prior specification $\theta \sim \pi$ can be also described as follows:


$$X|p \sim p, \quad p \sim \Pi$$

where Π is a prior distribution on the set of all possible densities with the property that $\Pi(\{p_\theta : \theta \in \Theta\}) = 1$.

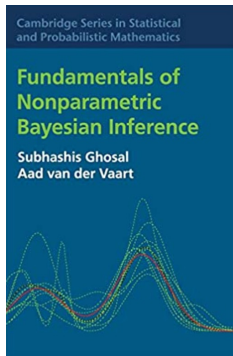
-  Parametric models make restrictive assumptions about the data-generating mechanism (e.g. the data are generated from a Normal distribution)
-  If the data do not follow the assumed DGM (they rarely do), the distributional assumption may cause serious biases in the inference
-  A parametric model $X|\theta \sim p_\theta$ for $\theta \in \Theta \subset \mathbb{R}^d$ with a prior specification $\theta \sim \pi$ can be also described as follows:

$$X|p \sim p, \quad p \sim \Pi$$

where Π is a prior distribution on the set of all possible densities with the property that $\Pi(\{p_\theta : \theta \in \Theta\}) = 1$.

-  Thus parametric modeling insists on a prior that assigns probability one to a very small subset of all densities.

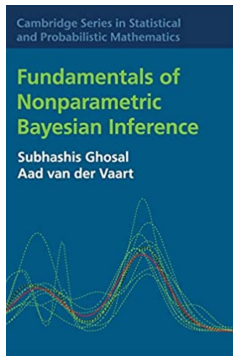
There are many different resources to learn BNP:



Foundational and
Theoretical Aspects

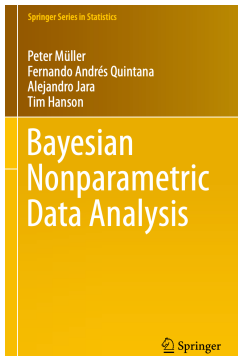
Large sample behavior of the posterior distribution: understanding the behavior of posteriors is critical to selecting priors that work

There are many different resources to learn BNP:



Foundational and
Theoretical Aspects

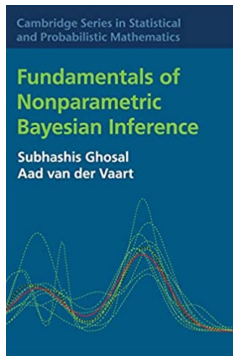
Large sample behavior of the posterior distribution: understanding the behavior of posteriors is critical to selecting priors that work



Methodological Aspects

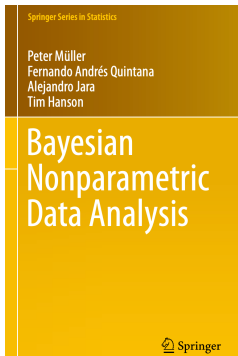
Organized by traditional data analysis problems. Shows how nonparametric Bayesian models are used to implement inference in a given data analysis problem

There are many different resources to learn BNP :



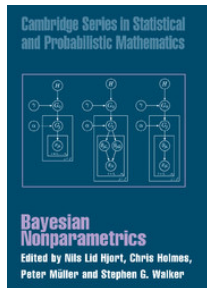
Foundational and
Theoretical Aspects

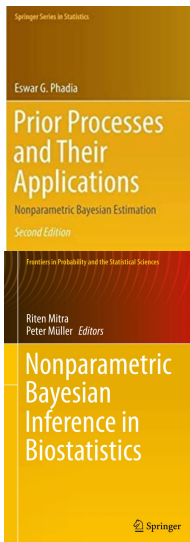
Large sample behavior of the posterior distribution: understanding the behavior of posteriors is critical to selecting priors that work



Methodological Aspects

Organized by traditional data analysis problems. Shows how nonparametric Bayesian models are used to implement inference in a given data analysis problem





Bayesian Analysis (2013)

8, Number 2, pp. 269–302

Bayesian Nonparametric Inference – Why and How

Peter Müller^{*} and Riten Mitra[†]

Abstract. We review inference under models with nonparametric Bayesian (BNP) priors. The discussion follows a set of examples for some common inference problems. The examples are chosen to highlight problems that are challenging for standard parametric inference. We discuss inference for density estimation, clustering, regression and for mixed effects models with random effects distributions. While we focus on arguing for the need for the flexibility of BNP models, we also review some of the more commonly used BNP models, thus hopefully answering a bit of both questions, why and how to use BNP.

Keywords: Nonparametric models, Dirichlet process, Polya tree, dependent Dirichlet process

Wiley StatsRef:
Statistics Reference Online



Bayesian Nonparametrics

By Antonio Canale^{1,2}, Antonio Lijoi^{2,3}, and Igor Prünster⁴

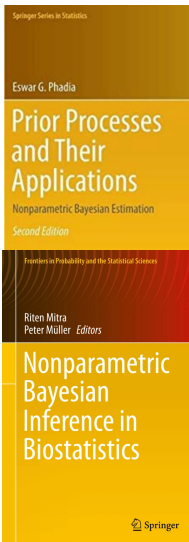
Stat. Methods Appl. (2018) 27:175–206
<https://doi.org/10.1007/s10260-017-0405-z>



ORIGINAL PAPER

Nonparametric Bayesian inference in applications

Peter Müller¹ · Fernando A. Quintana² ·
Garritt Page³



Bayesian Analysis (2013)

8, Number 2, pp. 269–302

Bayesian Nonparametric Inference – Why and How

Peter Müller* and Riten Mitra†

Abstract. We review inference under models with nonparametric Bayesian (BNP) priors. The discussion follows a set of examples for some common inference problems. The examples are chosen to highlight problems that are challenging for standard parametric inference. We discuss inference for density estimation, clustering, regression and for mixed effects models with random effects distributions. While we focus on arguing for the need for the flexibility of BNP models, we also review some of the more commonly used BNP models, thus hopefully answering a bit of both questions, why and how to use BNP.

Keywords: Nonparametric models, Dirichlet process, Polya tree, dependent Dirichlet process

Wiley StatsRef:
Statistics Reference Online



Bayesian Nonparametrics

By Antonio Canale^{1,2}, Antonio Lijoi^{2,3}, and Igor Prünster⁴

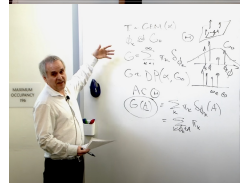
Stat. Methods Appl. (2018) 27:175–206
<https://doi.org/10.1007/s10260-017-0405-z>



ORIGINAL PAPER

Nonparametric Bayesian inference in applications

Peter Müller¹ · Fernando A. Quintana² · Garritt Page³



There is an increasing recognition that brain functioning is heterogeneous and varies greatly both **within and between individuals**:

- differences in **activation** to different stimuli
- differences in **connectivity** to different stimuli
- the different brain activity patterns may be **associated to a clinical outcome or different behaviors** (e.g., large brain responses to food-related cues predict cue-induced eating, Versace et al, 2019)

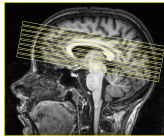


A Bayesian Nonparametric Approach can be used to account for such heterogeneity

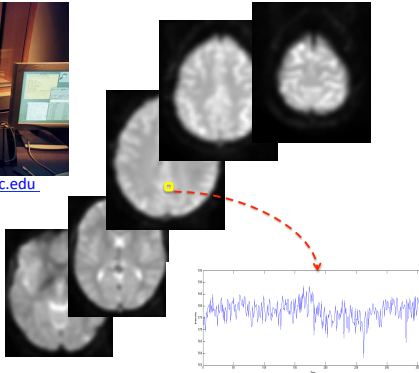
Capturing within-subject heterogeneity



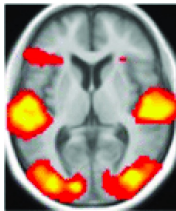
Source: photos.uc.wisc.edu



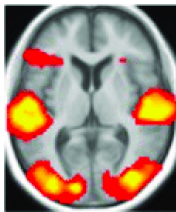
Source: www.anc.ed.ac.uk



- Indirect measure of brain activity as changes in blood flow, typically collected during a sensorimotor task.
- ☐ Observed data ➤ time series of the blood oxygenation level dependent (BOLD) response, at each voxel in the brain.



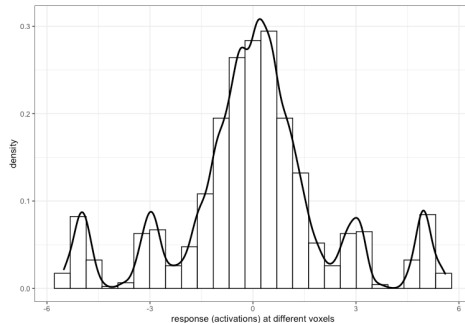
$$Y_V = \mu_V + \varepsilon_V, \quad \varepsilon_V \sim N_T(0, \sigma)$$



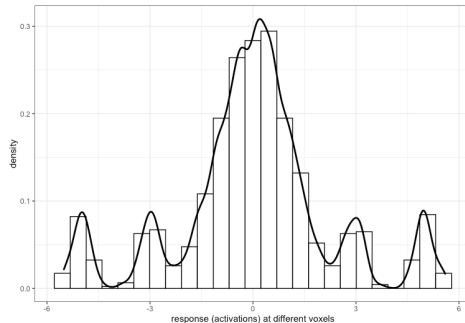
$$Y_V = \mu_V + \varepsilon_V, \varepsilon_V \sim N_T(0, \sigma)$$

- Y_V , BOLD response summarized at the v th voxel in a subject
- μ_V , random effect to capture activations at different voxels
- ε_V , is an error term.

- To fix ideas, we can think that the response is characterized by different levels of activations at different voxels:



- To fix ideas, we can think that the response is characterized by different levels of activations at different voxels:



- 👉 We can think at a mixture model to describe the responses:

$$f(y_\nu | \sigma) = \int \mathcal{N}(y_\nu; \mu_\nu, \sigma) d\tilde{p}(\mu_\nu)$$

We can rewrite the previous model in hierarchical form as:

$$Y_\nu \mid \mu_\nu \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{Y}_\nu; \mu_\nu), \quad \nu = 1, \dots, V$$
$$\mu_\nu \mid \tilde{\rho} \stackrel{\text{iid}}{\sim} \tilde{\rho}$$

for some choice of the mixing distribution $\tilde{\rho}$.

We can rewrite the previous model in hierarchical form as:

$$Y_\nu \mid \mu_\nu \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{Y}_\nu; \mu_\nu), \quad \nu = 1, \dots, V$$
$$\mu_\nu \mid \tilde{\rho} \stackrel{\text{iid}}{\sim} \tilde{\rho}$$

for some choice of the mixing distribution $\tilde{\rho}$.

□ **Desiderata:**

- ① The model should be **tractable**, i.e., it should be **easily computed**, either analytically or through simulations.

We can rewrite the previous model in hierarchical form as:

$$Y_\nu \mid \mu_\nu \stackrel{\text{ind}}{\sim} \mathcal{N}(Y_\nu; \mu_\nu), \quad \nu = 1, \dots, V$$
$$\mu_\nu \mid \tilde{\rho} \stackrel{\text{iid}}{\sim} \tilde{\rho}$$

for some choice of the mixing distribution $\tilde{\rho}$.

□ **Desiderata:**

- ① The model should be **tractable**, i.e., it should be **easily computed**, either analytically or through simulations.
- ② The model should be **rich**, in the sense of having a **large enough support**.

We can rewrite the previous model in hierarchical form as:

$$Y_\nu \mid \mu_\nu \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{Y}_\nu; \mu_\nu), \quad \nu = 1, \dots, V$$
$$\mu_\nu \mid \tilde{\rho} \stackrel{\text{iid}}{\sim} \tilde{\rho}$$

for some choice of the mixing distribution $\tilde{\rho}$.

□ **Desiderata:**

- ① The model should be **tractable**, i.e., it should be **easily computed**, either analytically or through simulations.
- ② The model should be **rich**, in the sense of having a **large enough support**.
- ③ The hyperparameters in the model should be easily **interpretable**.

(Ferguson, 1973)

- One natural choice is to assume:

$$\tilde{\rho} = \sum_{k=1}^K \pi_k \delta_{\mu_k^*}$$

where $1 \leq K \leq \infty$, π_k are weights ($\sum_{k=1}^K \pi_k = 1$) and the μ_k^* 's can be thought of as "*centroids*" of the set of responses

- One natural choice is to assume:

$$\tilde{p} = \sum_{k=1}^K \pi_k \delta_{\mu_k^*}$$

where $1 \leq K \leq \infty$, π_k are weights ($\sum_{k=1}^K \pi_k = 1$) and the μ_k^* 's can be thought of as “centroids” of the set of responses

- Ishwaran and James (2001) propose a **stick-breaking prior**:

$$\mu_k^* \stackrel{iid}{\sim} G_0 \quad \text{👉} \quad E(\tilde{p}(A)) = G_0(A) \quad (\text{centering distribution}).$$

- One natural choice is to assume:

$$\tilde{p} = \sum_{k=1}^K \pi_k \delta_{\mu_k^*}$$

where $1 \leq K \leq \infty$, π_k are weights ($\sum_{k=1}^K \pi_k = 1$) and the μ_k^* 's can be thought of as “centroids” of the set of responses

- Ishwaran and James (2001) propose a **stick-breaking prior**:

$$\mu_k^* \stackrel{iid}{\sim} G_0 \quad \leftarrow E(\tilde{p}(A)) = G_0(A) \quad (\text{centering distribution}).$$

$$\pi_1 = V_1 \quad \text{and} \quad \pi_k = (1 - V_1)(1 - V_2) \cdots (1 - V_{k-1}) V_k, \quad k \geq 2$$

with $V_k \stackrel{ind}{\sim} \text{Beta}(a_k, b_k)$.

If $K < \infty$, $V_K = 1 \Leftrightarrow \sum_{k=1}^K \pi_k = 1$.

- One natural choice is to assume:

$$\tilde{p} = \sum_{k=1}^K \pi_k \delta_{\mu_k^*}$$

where $1 \leq K \leq \infty$, π_k are weights ($\sum_{k=1}^K \pi_k = 1$) and the μ_k^* 's can be thought of as “centroids” of the set of responses

- Ishwaran and James (2001) propose a **stick-breaking prior**:

$$\mu_k^* \stackrel{iid}{\sim} G_0 \quad \leftarrow E(\tilde{p}(A)) = G_0(A) \quad (\text{centering distribution}).$$

$$\pi_1 = V_1 \quad \text{and} \quad \pi_k = (1 - V_1)(1 - V_2) \cdots (1 - V_{k-1}) V_k, \quad k \geq 2$$

with $V_k \stackrel{ind}{\sim} \text{Beta}(a_k, b_k)$.

If $K < \infty$, $V_K = 1 \Leftrightarrow \sum_{k=1}^K \pi_k = 1$.

🐎 “The Pony Process”

The stick-breaking formulation for the weights generalizes the Sethuraman's (1994) construction of the weights of the Dirichlet Process. Indeed, setting $K = \infty$,

👉 $a_k = 1$ and $b_k = \alpha$

⇒ $DP(\alpha, G_0)$ (Ferguson, 1973; Sethuraman, 1994)

👉 $a_k = 1 - a, b_k = b + ka$, with discount parameter $0 \leq a < 1$ and strength parameter $b > -a$

⇒ $\mathcal{PY}(a, b, G_0)$ (Pitman and Yor, 1997)

- The discount parameter plays a role on the induced distribution of the number of clusters in the data, the larger being a the flatter and less informative the prior


- 👉 Based on the priors above, the model for the data becomes

$$y_\nu \stackrel{iid}{\sim} \sum_{k=1}^K \pi_k f(y_\nu | \mu_k^*, \sigma)$$


that is, a **univariate location PY mixture model** (Ferguson, 1983)

- We can assume $G_0 \equiv N(m_0, \sigma_0)$, and $\sigma \sim \pi(\sigma)$, e.g. $\sigma^2 \sim \text{IGa}(a_0, b_0)$.
- We can further assume $m_0 | \sigma_0^2 \sim N(m_1, \sigma_0^2/k_1)$ and $\sigma_0^2 \sim \text{IGa}(a_1, b_1)$.

Two major types of MCMC algorithms have been proposed:

-  **Marginal Samplers** (Escobar and West, 1995 and Müller et al., 1996) Based on the Polya-Urn scheme of Blackwell and MacQueen (1973)
 - ⇒ Computationally slow for high-dimensional data

Two major types of MCMC algorithms have been proposed:

 **Marginal Samplers** (Escobar and West, 1995 and Müller et al., 1996) Based on the Polya-Urn scheme of Blackwell and MacQueen (1973)

⇒ Computationally slow for high-dimensional data


 **Conditional Samplers:**

① **Blocked Gibbs Sampler** (Ishwaran and James, 2001) Based on finite-dimensional truncations

⇒ the error in approximating the infinite-dimensional posterior can be hard to control for many models (Griffin, 2016)


② **Slice Sampler** (Walker 2007; Kalli, Griffin, and Walker 2011), uses a sequence of auxiliary random variables to describe the non-empty mixture components

Two major types of MCMC algorithms have been proposed:


-  **Marginal Samplers** (Escobar and West, 1995 and Müller et al., 1996) Based on the Polya-Urn scheme of Blackwell and MacQueen (1973)
 - ⇒ Computationally slow for high-dimensional data

Conditional Samplers:

- ① **Blocked Gibbs Sampler** (Ishwaran and James, 2001) Based on finite-dimensional truncations
 - ⇒ the error in approximating the infinite-dimensional posterior can be hard to control for many models (Griffin, 2016)
- ② **Slice Sampler** (Walker 2007; Kalli, Griffin, and Walker 2011), uses a sequence of auxiliary random variables to describe the non-empty mixture components


 DPpackage (A. Jara, long gone in R) 


Two major types of MCMC algorithms have been proposed:

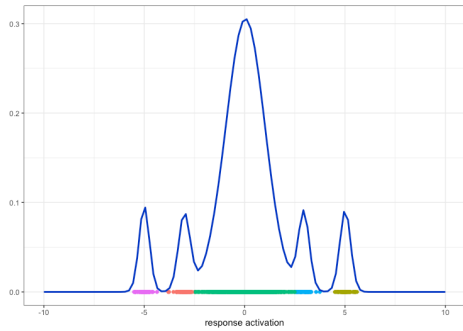
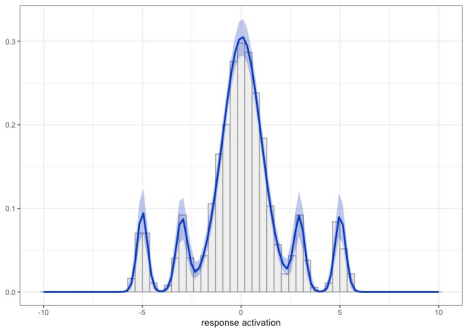
-  **Marginal Samplers** (Escobar and West, 1995 and Müller et al., 1996) Based on the Polya-Urn scheme of Blackwell and MacQueen (1973)
 - ⇒ Computationally slow for high-dimensional data

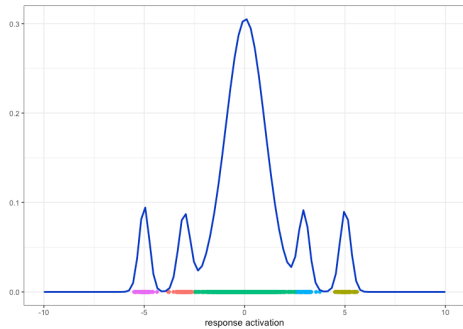
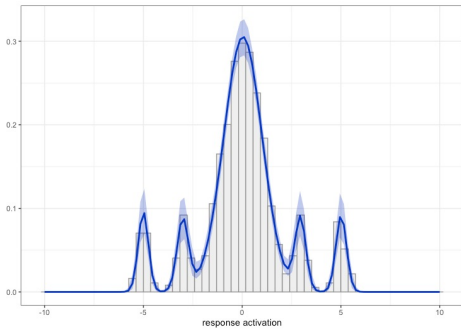
Conditional Samplers:

- ① **Blocked Gibbs Sampler** (Ishwaran and James, 2001) Based on finite-dimensional truncations
 - ⇒ the error in approximating the infinite-dimensional posterior can be hard to control for many models (Griffin, 2016)
- ② **Slice Sampler** (Walker 2007; Kalli, Griffin, and Walker 2011), uses a sequence of auxiliary random variables to describe the non-empty mixture components

 DPpackage (A. Jara, long gone in R) 🙏

 BNPmix (Corradin, R., Canale, A., and Nipoti, B. 2021) (C++, Rcpp)

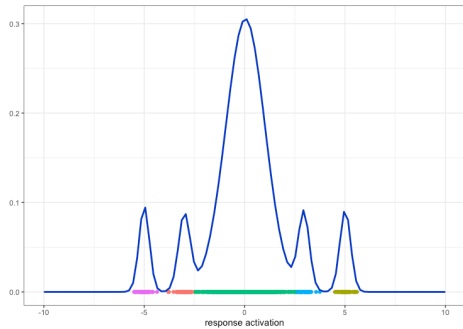
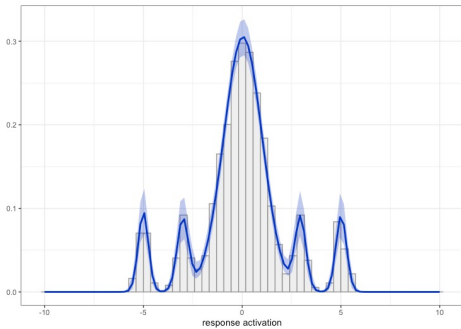




Mixture models are well-suited for density estimation (describing heterogeneity).

Caution on inferences about the number of components:

Induced Posterior on the number of components is inconsistent (Miller and Harrison, 2014)



Mixture models are well-suited for density estimation (describing heterogeneity).

Caution on inferences about the number of components:

Induced Posterior on the number of components is inconsistent (Miller and Harrison, 2014)

Ill-posed problem ? **➡** Finding the number of clusters is essentially a decision problem

- ? What is an appropriate point estimate of the clustering structure based on the posterior distribution?
- ? What is an appropriate loss function on the space of clusterings?
- Let $L(c, \hat{c})$ be a loss function which measures the loss of estimating the true clustering c with \hat{c} .
- Since the true clustering is unknown, and the posterior weights the possible clustering configurations, the optimal cluster configurations can be obtained as

$$c^* = \underset{\hat{c}}{\operatorname{argmin}} \mathbb{E} [L(c, \hat{c}) \mid \mathbf{y}] = \underset{\hat{c}}{\operatorname{argmin}} \sum_c L(c, \hat{c}) p(c \mid \mathbf{y})$$

- **Binder's loss (1978)** is invariant to permutations of the data points indices and cluster labels
- It penalizes the two errors of allocating two observations to different clusters when they should be in the same cluster or allocating them to the same cluster when they should be in different clusters:

$$B(c, \hat{c}) = \sum_{n < n'} l_1 1(c_n = c_{n'}) 1(\hat{c}_n \neq \hat{c}_{n'}) + l_2 1(c_n \neq c_{n'}) 1(\hat{c}_n = \hat{c}_{n'})$$

If the errors have the same penalty, then it results in a quadratic function of the counts in the two clusters penalized by disagreements between the true and estimated clusterings.

- The **variation of information loss (VI)** has been proposed by Wade and Ghahramani (2018) and Meilă (2007)
- It measures the amount of information lost and gained in changing from one clustering partition to another

$$VI(c, \hat{c}) = H(c) + H(\hat{c}) - 2I(c, \hat{c})$$

where $H(\cdot)$ measures the entropy of a partition (zero if there is only one cluster) and $I(\cdot)$ is a measure of mutual information between the two clustering (sort of distance between the two clustering structures)

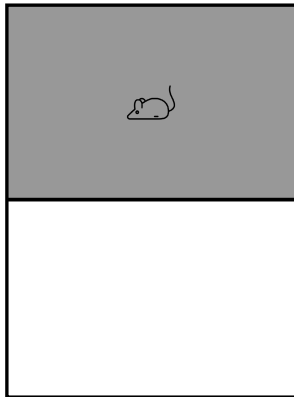
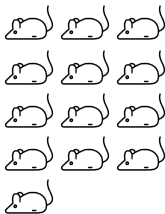
- WG (2018) show how the VI is able to better represent the idea of closest set of partitions to a true partition
- ⇒ They obtain point estimates and credible balls to reflect uncertainty on the partitions.

Implemented in the packages `mclust` and `BNPmix`.

Bayesian NP mixtures for screening in large-scale testing

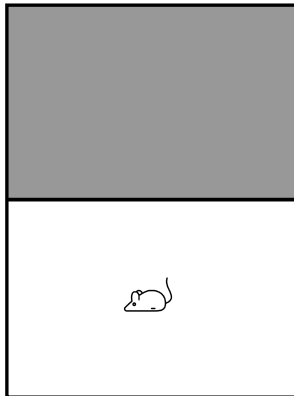
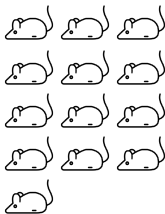
The light-sheet fluorescence microscopy dataset

- Fourteen mice were **individually housed in the dark** for 24 hours to establish baseline visual activity



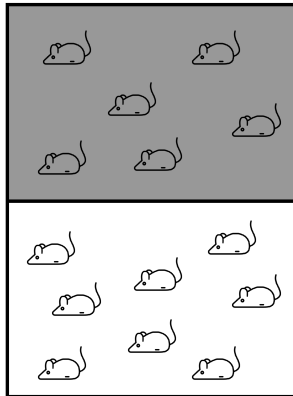
The light-sheet fluorescence microscopy dataset

- Fourteen mice were **individually housed in the dark** for 24 hours to establish baseline visual activity
- Mice were then transferred into a new cage **exposed to ambient light**

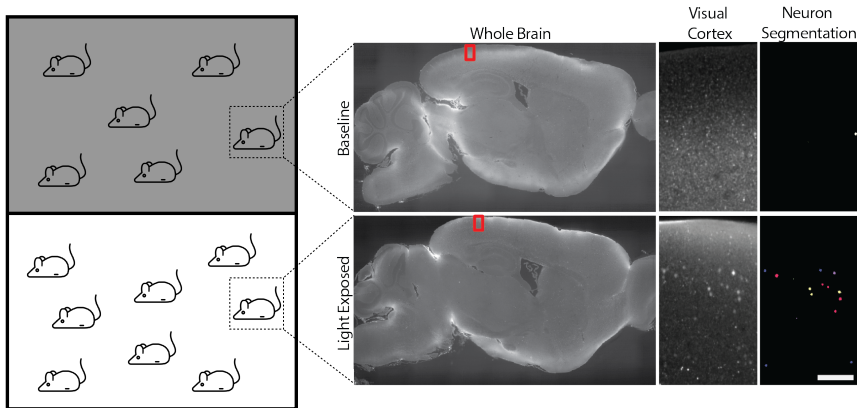


The light-sheet fluorescence microscopy dataset

- Fourteen mice were **individually housed in the dark** for 24 hours to establish baseline visual activity
- Mice were then transferred into a new cage **exposed to ambient light**
- The brains of six mice were examined **0-15 minutes** (i.e., no light) after light exposure to serve as the **baseline** group
- The brains of another eight mice were examined 30-120 minutes after light exposure, within the window of **Npas4** protein up-regulation (Ramamoorthi et al., 2011)



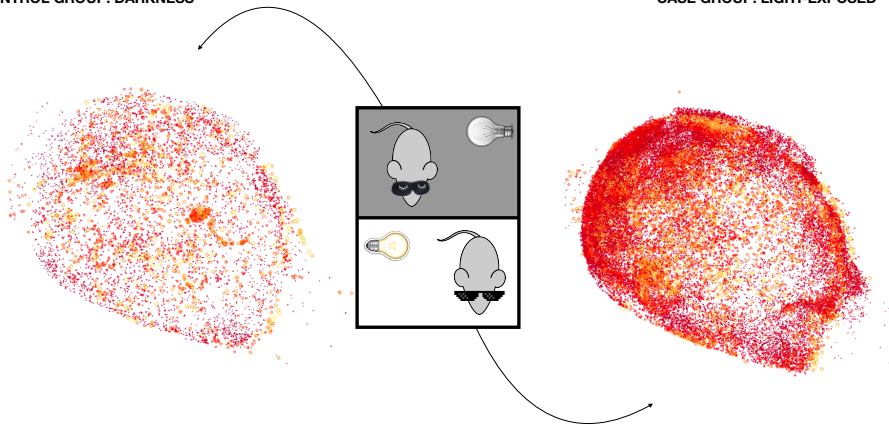
The light-sheet fluorescence microscopy dataset



- The light-sheet fluorescence microscopy imaging techniques allows the detection of activated cells at high resolution in vivo in the whole-brain fo the mouse.

CONTROL GROUP: DARKNESS

CASE GROUP: LIGHT-EXPOSED



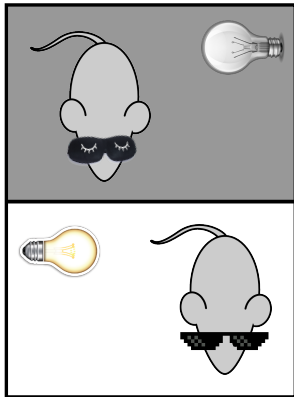
- “Activation” seems to be linked to both fluorescence intensity and frequency of neurons

The light-sheet fluorescence microscopy dataset

• GOAL of the study:

- Assess **differentially activated regions** by comparing the baseline and light-exposed groups
- The activation is measured in terms of Npas4 expression (we will refer to this as **fluorescence**)
- We expect that light exposure induces widespread, visually evoked activity in terms of **fluorescence intensity**
- Data are pre-processed eventually organized into 281 brain regions of interest and z-scores

$$Z_\nu = \beta_\nu + \varepsilon_\nu, \quad \varepsilon_\nu \sim N_T(0, \sigma)$$



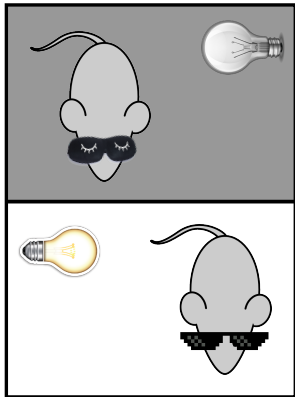
The light-sheet fluorescence microscopy dataset

• GOAL of the study:

- Assess **differentially activated regions** by comparing the baseline and light-exposed groups
- The activation is measured in terms of Npas4 expression (we will refer to this as **fluorescence**)

BH discovers 142 regions (50%) too liberal!

The local FDR method (Efron, 2004) flags only 38 brain regions as relevant, however missing many regions known to be associated with the visual task.



7.4. Adopting More Holistic Approaches

McShane, B., Gal, D., Gelman, A., Robert, C., and Tackett, J.,
Abandon Statistical Significance

1. Treat p -values (and other purely statistical measures like confidence intervals and Bayes factors) continuously rather than in a dichotomous or thresholded manner. In doing so, bear in mind that it seldom makes sense to calibrate evidence as a function of p -values or other purely statistical measures because they are, among other things, typically defined relative to the generally uninteresting and implausible null hypothesis of zero effect and zero systematic error.

5. Accept uncertainty and embrace variation in effects: we can learn much (indeed, more) about the world by forsaking the false promise of certainty offered by dichotomous declarations of truth or falsity—binary statements about there being “an effect” or “no effect”—based on some p -value or other statistical threshold being attained.

**THE
AMERICAN
STATISTICIAN**

A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION

VOLUME 73 • NUMBER S1 MARCH 2019



Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

Moving to a World Beyond “ $p < 0.05$ ”

THE AMERICAN STATISTICIAN

2019, VOL. 73, NO. S1, 1–19: Editorial

<https://doi.org/10.1080/00031305.2019.1583913>

- 👉 **Continuous scale mixtures of Gaussians** (Carvalho et al, 2010, Polson et al 2012) do not lead to an immediate “selection” of relevant parameters

$$\beta_\nu \mid \tau, \lambda_\nu \sim \mathcal{N}_1(0, \tau^2 \cdot \lambda_\nu^2)$$

with

$\tau \sim g$ a **global shrinkage parameter**

and

$\lambda_\nu \sim h_\nu$ a **local shrinkage parameter**

- However, the decisions on the “significance” of the β 's coefficients are typically **dichotomized** (e.g., based on 90% credible intervals or shrinkage factor)

or other decision theoretic-based procedures (Chandra, Mueller, Sarkar, 2022+; Lee et al, 2022+)

- We can consider a mixture:

$$\beta_\nu \mid \tau, \lambda_K, \pi, \sigma^2 \sim \sum_{k=1}^K \pi_k \phi(\beta_\nu; 0, \sigma^2 \cdot \tau^2 \cdot \lambda_k^2)$$

where λ_k^2 is a mixture shrinkage component.

The smallest variance component is typically such that $\tau \lambda_{(1)} \approx 0$ and represents the null distribution

The other components can be sorted according to the magnitudes of λ_k 's.

The alternative distribution gets segmented into different levels

- 👉 One can rank the β_ν 's into **tiers of relevance**

- We can consider a mixture:

$$\beta_\nu \mid \tau, \lambda_K, \pi, \sigma^2 \sim \sum_{k=1}^K \pi_k \phi(\beta_\nu; 0, \sigma^2 \cdot \tau^2 \cdot \lambda_k^2)$$

where λ_k^2 is a mixture shrinkage component.

The smallest variance component is typically such that $\tau \lambda_{(1)} \approx 0$ and represents the null distribution

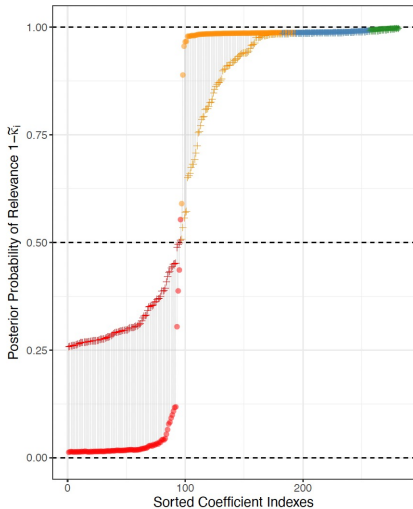
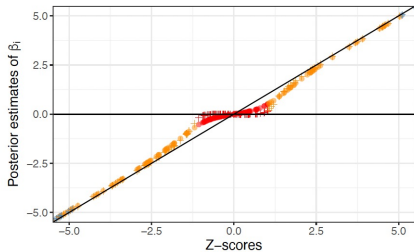
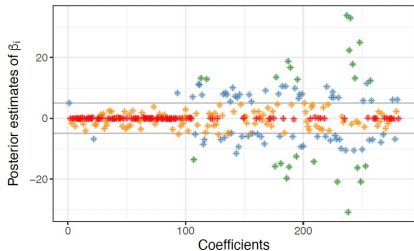
The other components can be sorted according to the magnitudes of λ_k 's.

The alternative distribution gets segmented into different levels

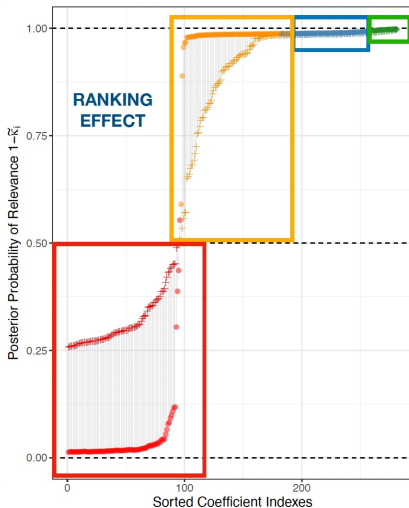
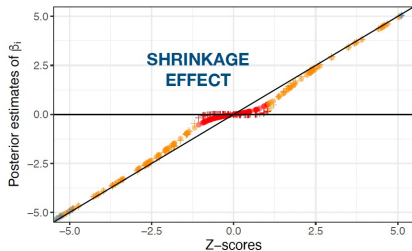
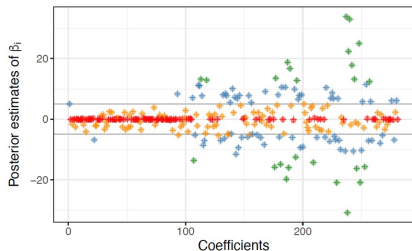
- ☞ One can rank the β_ν 's into **tiers of relevance**
- ☞ One can choose a **Half-Cauchy prior** for the mixture shrinkage component,

$$\lambda_l \sim \text{Cauchy}^+(0, 1), \forall l \quad (\text{Horseshoe pit})$$

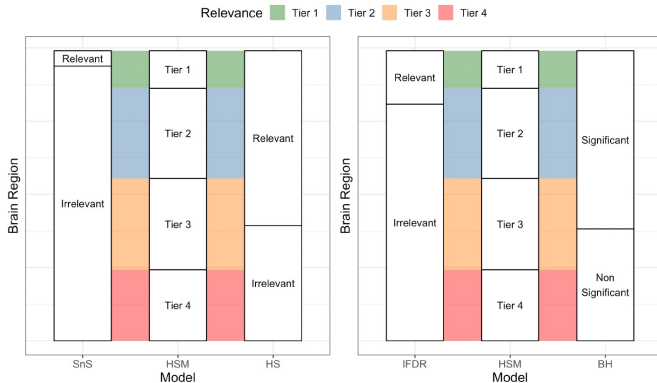
We can segment the results into 4 tiers of activation, from high-activity (**Tier 1**) to no activity (**Tier 4**)



We can segment the results into 4 tiers of activation, from high-activity (**Tier 1**) to no activity (**Tier 4**)



We compare the findings with other well-known methods: Local-FDR (IFDR), Horseshoe prior (HS), Spike-and-Slab (SnS), and Benjamini-Hochberg (BH)

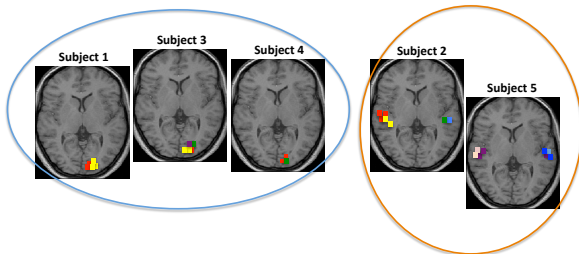


- The HSM model mediates between the more conservative IFDR and SnS methods and the numerous discoveries of the BH and HS models.
- Denti et al (2022+), *A Horseshoe mixture model for Bayesian screening with an application to light sheet fluorescence microscopy in brain imaging*, Submitted. <https://arxiv.org/abs/2106.08281> ✈

Capturing Between-subjects heterogeneity

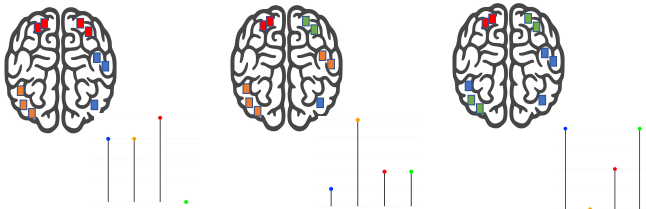
- Hierarchical mixtures are widely used in Bayesian Nonparametrics to cluster together observations from different groups (Camerlenghi et al, 2019; Bassetti et al, 2020; Argiento et al, 2019)

- Hierarchical mixtures are widely used in Bayesian Nonparametrics to cluster together observations from different groups (Camerlenghi et al, 2019; Bassetti et al, 2020; Argiento et al, 2019)
- **Basic Idea: Two-level mixtures:** a mixture is used to cluster subjects showing similar brain patterns; a lower-level mixture captures individual specific features

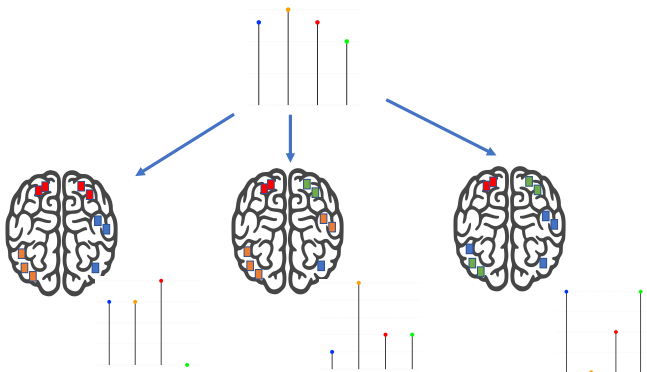


$$Y_{iv} = X_{iv}\beta_{iv} + \varepsilon_{iv}, \varepsilon_{iv} \sim N_T(0, \Sigma_{iv})$$

- Objective: capture activation patterns in response to a stimulus within and across subjects.



- Objective: capture activation patterns in response to a stimulus within and across subjects.



- More specifically, we use a Hierarchical Dirichlet Process (HDP, Teh et al, 2006) to define a multi-subject spike-and-slab nonparametric prior,

$$\beta_{iv} | \gamma_{iv}, \mathbf{G}_i \sim \gamma_{iv} \mathbf{G}_i + (1 - \gamma_{iv}) \delta_0$$

- More specifically, we use a Hierarchical Dirichlet Process (HDP, Teh et al, 2006) to define a multi-subject spike-and-slab nonparametric prior,

$$\beta_{iv} | \gamma_{iv}, \mathbf{G}_i \sim \gamma_{iv} \mathbf{G}_i + (1 - \gamma_{iv}) \delta_0$$

- More specifically, we use a Hierarchical Dirichlet Process (HDP, Teh et al, 2006) to define a multi-subject spike-and-slab nonparametric prior,

$$\beta_{iv} | \gamma_{iv}, G_i \sim \gamma_{iv} G_i + (1 - \gamma_{iv}) \delta_0$$

G_i is a **subject-specific** probability distribution that induces clustering of the β'_v s **within** subjects

- More specifically, we use a Hierarchical Dirichlet Process (HDP, Teh et al, 2006) to define a multi-subject spike-and-slab nonparametric prior,

$$\beta_{iv} | \gamma_{iv}, \mathbf{G}_i \sim \gamma_{iv} \mathbf{G}_i + (1 - \gamma_{iv}) \delta_0$$

\mathbf{G}_i is a **subject-specific** probability distribution that induces clustering of the β'_{iv} s **within** subjects

The \mathbf{G}_i are built by “picking” hierarchically the atoms in their support from a **common underlying (discrete) distribution**

$$\mathbf{G}_i | \eta_1, \mathbf{G}_0 \sim DP(\eta_1, \mathbf{G}_0)$$

$$\mathbf{G}_0 | \eta_2, P_0 \sim DP(\eta_2, P_0)$$

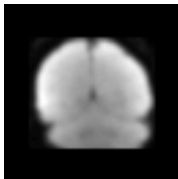
$$P_0 = N(0, \tau)$$

- ☆ η_1, η_2 : concentration parameters, controlling the variability
 P_0 : base measure, generating the global components which are shared within and across subjects

Real fMRI data collected by Versace's lab (MDACC):

- Data Dimension: 27 subjects, 286 time points, 2 slices of interest, 64×64 voxels per slice

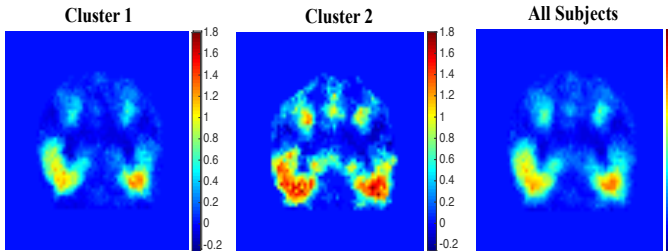
Occipital Slice
($y = -60$ mm)



Frontal Slice
($y = +38$ mm)



- Event-related design
- Goal:** detecting (differential) brain activity in response to visual scenes: emotional pictures (vs neutral pictures)



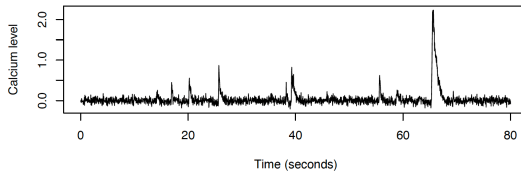
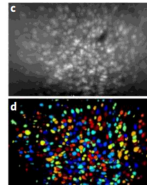
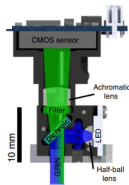
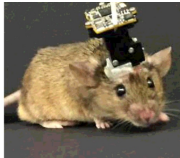
- Two groups of subjects characterized by different levels of activations
- Subjects who show decreased responses to certain emotional pictures may show lower reward sensitivity in surveys' responses (e.g., higher dissatisfaction, affecting mechanisms connected to reward processing)

Capturing Distributional heterogeneity

- ⇒ The Hierarchical Dirichlet Process assumes subject-specific distributions but does not allow **clustering distributions**

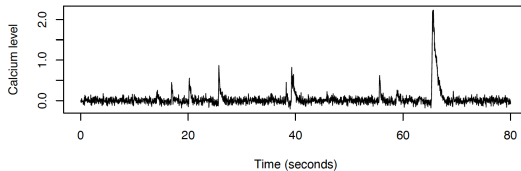
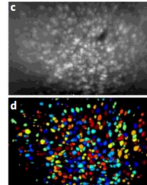
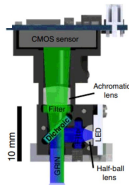
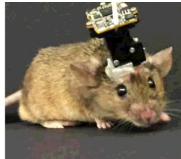
- ⇒ The Hierarchical Dirichlet Process assumes subject-specific distributions but does not allow **clustering distributions**

NATURE METHODS | VOL 16 | JANUARY 2019 | 9-15 | www.nature.com/naturemethods



- ⇒ The Hierarchical Dirichlet Process assumes subject-specific distributions but does not allow **clustering distributions**

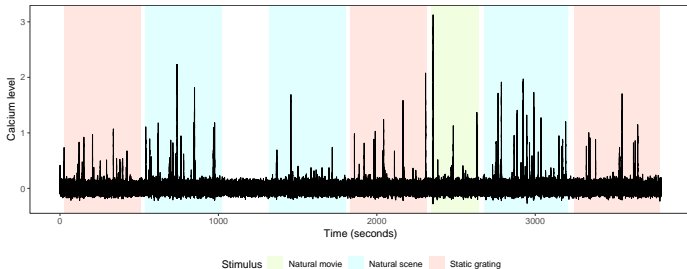
NATURE METHODS | VOL 16 | JANUARY 2019 | 9-15 | www.nature.com/naturemethods



- 👉 Study **how individual neurons react to stimulation** and how they encode information by **deconvolving the calcium traces** and identify the **precise spike times** of the observable neurons

Usually, the experiment involves multiple stimuli (e.g. visual stimuli, or odors):

- the interest is to understand how the different types of stimuli affect the neuronal activity \Rightarrow investigate similarities and differences in the **distribution of spikes** over time and conditions.



The distribution of the spikes can be very similar across the conditions of an experiment \Rightarrow Clustering

- Approaches for clustering distributional features directly are sparse.

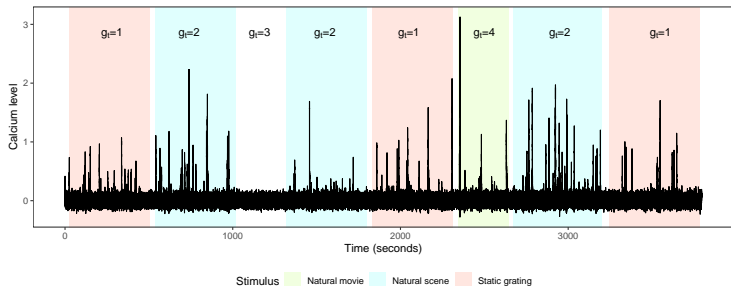
- Approaches for clustering distributional features directly are sparse.
- Clustering methods in symbolic statistics (Irpino and Verde, 2015; Batagelj et al., 2015) do not allow for a probabilistic assessment of cluster uncertainty.

- Approaches for clustering distributional features directly are sparse.
- Clustering methods in symbolic statistics (Irpino and Verde, 2015; Batagelj et al., 2015) do not allow for a probabilistic assessment of cluster uncertainty.
- The Nested Dirichlet process (nDP, Rodriguez et al, 2008) and its extensions have been widely employed to identify distributional groups in Bayesian Nonparametric model-based approaches.
- The nDP leads to a **two-layered clustering**: first, it allows grouping together similar distributions (**distributional clustering**), and then it clusters similar observations within each distributional cluster (**observational clustering**).

We are interested in characterizing the neural activity under different experimental conditions.

We are interested in characterizing the neural activity under different experimental conditions.

We introduce a categorical variable g_t taking values in $\{1, \dots, J\}$, with J the number of different experimental settings.



For each $t = 1, \dots, T$, $g_t = j$ indicates that the neural activity at time t is observed under condition j .

A popular model¹ to relate the observed trace y_t to the underlying true calcium concentration c_t , and to the neuronal activity A_t :

$$\begin{aligned}y_t &= b + c_t + \epsilon_t & \epsilon_t &\sim N(0, \sigma^2) \\c_t &= \gamma c_{t-1} + A_t + \omega_t & \omega_t &\sim N(0, \tau^2)\end{aligned}$$

for $t = 1, \dots, T$; with b baseline level, ϵ_t measurement error.

- In **absence of neuronal activity**: $A_t = 0$ and the calcium level follows a AR(1) process controlled by the parameter γ ;
- when a **spike** occurs: $A_t > 0$ and the concentration increases instantaneously with the spike amplitude A_t .

¹ Vogelstein et al. (2010). Fast nonnegative deconvolution for spike train inference from population calcium imaging. *Journal of Neurophysiology* **104**, 3691–3704

To allow the response to vary according to the condition, we assume that the spikes A_t come from **stimulus-specific distributions**: for

$j = 1, \dots, J$

$$A_t \mid g_t = j, G_j \sim G_j.$$

To allow the response to vary according to the condition, we assume that the spikes A_t come from **stimulus-specific distributions**: for $j = 1, \dots, J$

$$A_t \mid g_t = j, G_j \sim G_j.$$

To model the G_j 's we adopt a Bayesian **nested** finite mixture model:

nested structure → reconstruct the distribution within each experimental condition + borrow information between groups (distributional clustering)

mixture formulation → cluster the A_t across and within distributions
⇒ discover similarities in the activation response to different stimuli.

To allow the response to vary according to the condition, we assume that the spikes A_t come from **stimulus-specific distributions**: for $j = 1, \dots, J$

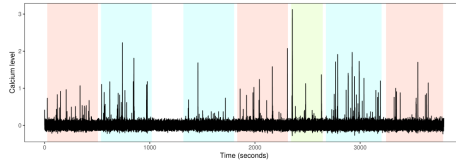
$$A_t \mid g_t = j, G_j \sim G_j.$$

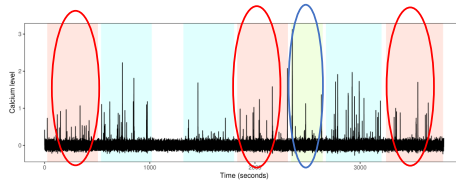
To model the G_j 's we adopt a Bayesian **nested** finite mixture model:

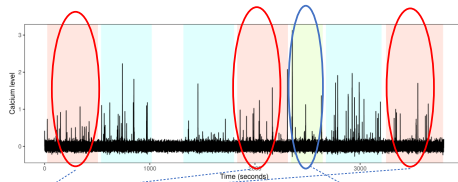
nested structure → reconstruct the distribution within each experimental condition + borrow information between groups (distributional clustering)

mixture formulation → cluster the A_t across and within distributions
⇒ discover similarities in the activation response to different stimuli.

The model allows to represent the data through two-layers: at the first level clusters of distributions across conditions, and at the second level a convenient representation of the distributions via models

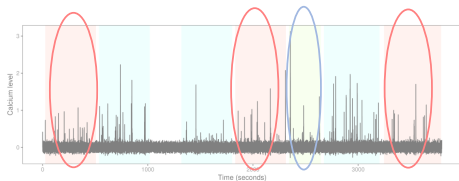






$$A_t \mid g_t = 1 \sim G_1.$$

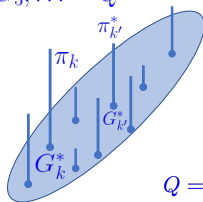
$$A_t \mid g_t = 3 \sim G_3$$



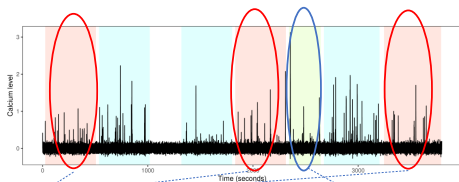
$$A_t \mid g_t = 1 \sim G_1.$$

$$A_t \mid g_t = 3 \sim G_3$$

$$G_1, G_2, G_3, \dots \sim Q$$



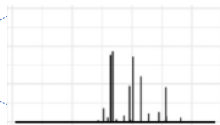
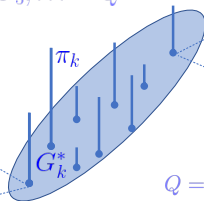
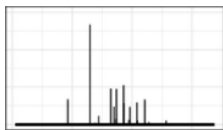
$$Q = \sum_{k=1}^K \pi_k \delta_{G_k^*}$$



$$A_t \mid g_t = 1 \sim G_1.$$

$$A_t \mid g_t = 3 \sim G_3$$

$$G_1, G_2, G_3, \dots \sim Q$$



$$Q = \sum_{k=1}^K \pi_k \delta_{G_k^*} \quad G_k^*(\cdot) = \sum_{l=1}^{\infty} \omega_{lk} \delta_{\theta_{lk}}(\cdot)$$

- Camerlenghi et al (2019) have recently proved that the inference obtained using the nDP may be affected by a *degeneracy* property:

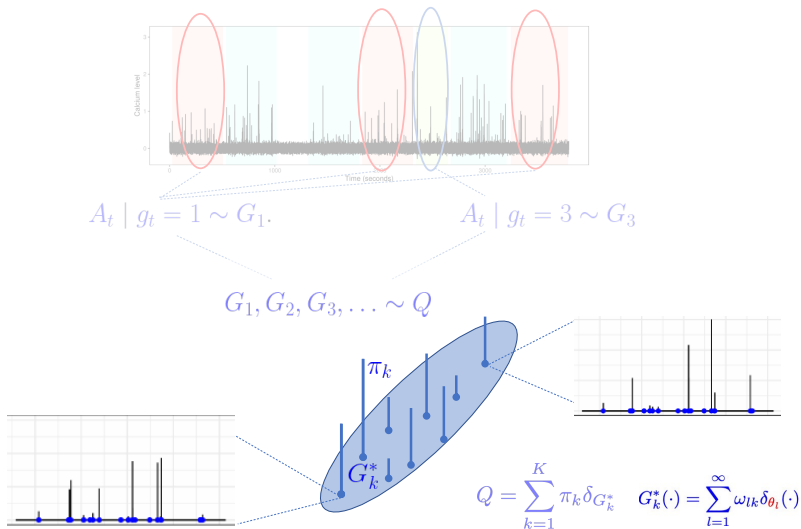
⇒ If two distributions share **even only one atom in their support**, the two distributions are automatically assigned to the same cluster.

More precisely, the partially exchangeable partition probability function (pEPPF), i.e. the function which describes the probability of each clustering allocation for partially exchangeable data modeled with a nDP, collapses to a fully exchangeable case when ties are present among the observational atoms.

The problem persists with nDP mixture model formulations

- Camerlenghi et al (2019) propose a class of latent nested processes, which relies on estimating a latent **mixture of shared and idiosyncratic processes** \Rightarrow very computationally complex, only small datasets with few groups.

- Camerlenghi et al (2019) propose a class of latent nested processes, which relies on estimating a latent **mixture of shared and idiosyncratic processes** \Rightarrow very computationally complex, only small datasets with few groups.
- Beraha et al (2021) propose a variation of the hierarchical DP, where the **baseline distribution** is itself a **mixture of a DP and a non-atomic measure** (semi-HDP). They further combine the semi-HDP prior with a random partition model that allows different populations to be grouped in clusters that are internally homogeneous, i.e. arising from the same distribution.
- Denti, Camerlenghi, Guindani & Mira (2022+) show that the degeneracy is avoided if the prior explicitly models **commonality of atoms** between groups.
- Lijoi, Pruenster, Rebaudo (2022+) move this idea further along by essentially combining the NDP and the HDP into a *hidden hierarchical Dirichlet Process*.



- For computational efficiency (long time series), one can employ the **generalized mixtures of finite mixtures (gMFM)** of Frühwirth-Schnatter et al. (BA, 2021) where the nested structure is based on the **common atom model**:

$$A_t \mid g_t = j, G_j \sim G_j.$$

$$G_1, \dots, G_J \mid Q \sim Q, \quad Q = \sum_{k=1}^K \pi_k \delta_{G_k^*}$$

where G_k^* are distributions (identifying clusters of distributions across conditions/experimental settings)

- For computational efficiency (long time series), one can employ the **generalized mixtures of finite mixtures (gMFM)** of Frühwirth-Schnatter et al. (*BA, 2021*) where the nested structure is based on the **common atom model**:

$$A_t \mid g_t = j, G_j \sim G_j.$$

$$G_1, \dots, G_J \mid Q \sim Q, \quad Q = \sum_{k=1}^K \pi_k \delta_{G_k^*}$$

where G_k^* are distributions (identifying clusters of distributions across conditions/experimental settings)

- More specifically, we assume:

$$\pi_1, \dots, \pi_K \mid K, \alpha \sim \text{Dir}_K\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$$

$$K - 1 \sim \text{beta-negative-binomial}$$

$$\alpha \sim F$$

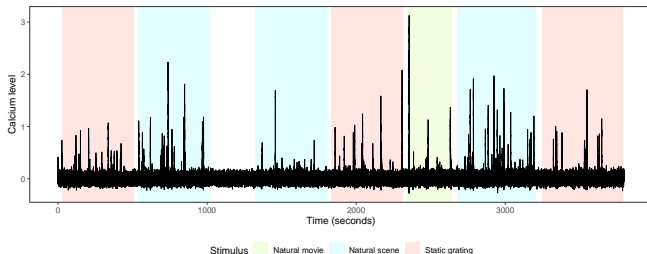
(Frühwirth-Schnatter et al., 2021)

- Also for the **distributional atoms** G_k^* , for $k = 1, \dots, K$ we assume a mixture

$$G_k^* = \sum_{l=1}^L \omega_{l,k} \delta_{A_l^*}$$

where the set of atoms A^* is **common** across the distributions G_1^*, \dots, G_K^* and they are obtained as i.i.d. draws from a base measure G_0 .

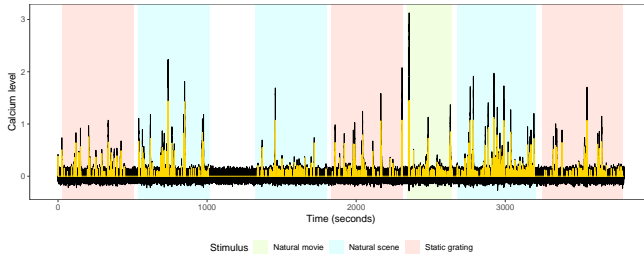
- The distributions G_k^* differ by the weight given to each atom (some weights $\omega_{l,k}$ can be zero for some k)

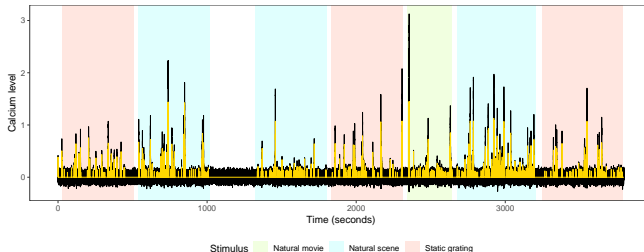


4 experimental conditions:

- 3 stimuli of increasing complexity (static grating, natural scene, natural movie)
- period of spontaneous activity (absence of stimuli)

⁶Allen Institute for Brain Science (2016). Allen brain observatory.
<http://observatory.brain-map.org/visualcoding>.

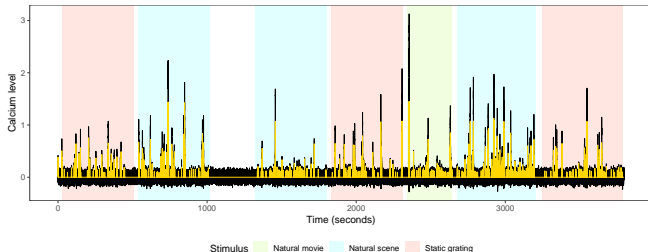




Cluster 1 = { Natural scene, Natural movie }

Cluster 2 = { Static grating }

Cluster 3 = { Absence of stimuli }



Cluster 1 = { Natural scene, Natural movie }

Cluster 2 = { Static grating }

Cluster 3 = { Absence of stimuli }

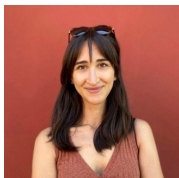
D'Angelo et al (2022+) Bayesian nonparametric analysis for the detection of spikes in noisy calcium imaging data, Biometrics, to appear

<https://arxiv.org/abs/2102.09403>

- We have discussed old and recent modeling frameworks in BNP with an emphasis on applications to neuroimaging.
- **What did I leave out?**

- We have discussed old and recent modeling frameworks in BNP with an emphasis on applications to neuroimaging.
- **What did I leave out?**
- What if we have information the partitions?
 - Smith and Allenby (2020), Paganin et al (2021), Dhal, Warr et al (2022+)

- We have discussed old and recent modeling frameworks in BNP with an emphasis on applications to neuroimaging.
- **What did I leave out?**
- What if we have information the partitions?
 - Smith and Allenby (2020), Paganin et al (2021), Dhal, Warr et al (2022+)
- **Dependent random measures** (MacEachern, 2000; Quintana et al, 2022)
 - e.g., adding covariates/clustering dependent on external stimuli/information about the environment.
- **Computational Challenges**
 - Dimension reduction
 - Approximate computational methods



Laura D'Angelo
Universita' Bicocca, U. Padova, Italy



Francesco Denti
U. Cattolica, Milan, Italy



Jaylen Lee
UCI – soon Facebook



Marina Vannucci
Rice University



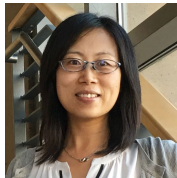
Senior Data Scientist
ExxonMobil



Antonio Canale
U. Padova, Italy



Babak, Shababa
UCI



Zhaoxia Yu
UCI



Xiangmin Xu
Anatomy & Neurobiology
UCI

Bibliography

- Blackwell D, MacQueen JB (1973). "Ferguson Distributions Via Polya Urn Schemes." *The Annals of Statistics*, 1(2), 353–355
- Camerlenghi F, Dunson D. B., Lijoi A., Prünster I., Rodríguez A. (2019a). Latent Nested Nonparametric Priors (with Discussion). *Bayesian Analysis*, 14, (4), 1303-1356.
- Corradin, R., Canale, A., & Nipoti, B. (2021). BNPmix: An R Package for Bayesian Nonparametric Modeling via Pitman-Yor Mixtures. *Journal of Statistical Software*, 100(15), 1–33.
- Escobar MD, West M (1995). "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association*, 90(430), 577–588.
- Ferguson T (1973). "A Bayesian Analysis of some Nonparametric Problems." *The Annals of Statistics*, 1(2), 209–230.
- Ferguson, T.S., 1983. Bayesian density estimation by mixtures of normal distributions. In: Rizvi H., Rustagi J., (Eds.), *Recent Advances in Statistics*. Academic Press, New York, 287-302.
- Griffin, J.E. An adaptive truncation method for inference in Bayesian nonparametric models. *Stat Comput* **26**, 423–441 (2016).
- Kalli M, Griffin J, Walker S (2011). "Slice sampling mixture models." *Statistics and Computing*, 21, 93–105.

- Ishwaran, Hemant, and Lancelot F. James. "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association*, vol. 96, no. 453, 2001, pp. 161–73.
- Lo AY (1984). "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates." *The Annals of Statistics*, 12(1), 351–357.
- MacEachern, Steven N., and Peter Müller. "Estimating Mixture of Dirichlet Process Models." *Journal of Computational and Graphical Statistics*, vol. 7, no. 2, 1998, pp. 223–38.
- MacEachern (2000), "Dependent Dirichlet Processes", Technical report.
- Müller P, Erkanli A, West M (1996). "Bayesian Curve Fitting Using Multivariate Normal Mixtures." *Biometrika*, 83(1), 67–79.
- Fernando A. Quintana, Peter Müller, Alejandro Jara, Steven N. MacEachern "The Dependent Dirichlet Process and Related Models," *Statistical Science, Statist. Sci.* 37(1), 24-41
- Pitman J, Yor M (1997). "The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator." *The Annals of Probability*, 25(2), 855–900.
- Sethuraman J (1994). "A Constructive Definition of Dirichlet Priors." *Statistica Sinica*, 4(2) 639–650
- Sara Wade, Zoubin Ghahramani "Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)," *Bayesian Analysis, Bayesian Anal.* 13(2), 559-626, (June 2018)

- *Beraha M., Guglielmi A., Quintana F. A. (2021). The Semi-Hierarchical Dirichlet Process and Its Application to Clustering Homogeneous Distributions. Bayesian Analysis, 16, (4), 1187-1219.*
- *Denti F., Camerlenghi F., Guindani M., Mira A. (2021). A Common Atoms Model for the Bayesian Nonparametric Analysis of Nested Data. Journal of the American Statistical Association.*
- *Lijoi A, Pruenster I, Ribaudo G (2022) Flexible clustering via hidden hierarchical Dirichlet priors, Scandinavian Journal of Statistics*
- *Frühwirth-Schnatter S., Malsiner-Walli G., Grün B. Generalized mixtures of finite mixtures and telescoping sampling, in: Bayesian Analysis, 2021*
- *Adam N. Smith & Greg M. Allenby (2020) Demand Models With Random Partitions, Journal of the American Statistical Association, 115:529, 47-65*
- *Sally Paganin, Amy H. Herring, Andrew F. Olshan, David B. Dunson "Centered Partition Processes: Informative Priors for Clustering (with Discussion)," Bayesian Analysis, Bayesian Anal. 16(1), 301-370, (March 2021)*

Bayesian
BOREDOM

Non
NOT

Parametric
PERMITTED