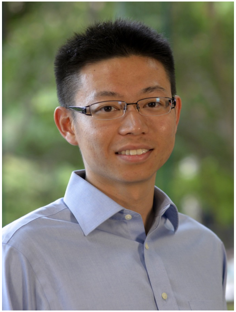
Permuted and Augmented Stick-Breaking Multinomial Regression

Mingyuan Zhou

IROM Department, McCombs School of Business
The University of Texas at Austin

39th Annual ISMS Marketing Science Conference
University of Southern California, June 8, 2017

Joint work with



Quan Zhang

McCombs PhD student
in Risk Analysis and Decision Making

Overview

Stick-breaking construction

- Provide a size-biased random permutation of the random draw from a Dirichlet process (Sethuraman 1994).
- Stick success probabilities can depend on covariates (Dunson and Park 2008, Chung and Dunson 2009, Ren et al 2011).
- Can be used to model the dependencies between multinomial probabilities by logit-Gaussian distribution/process (Linderman et al 2015).

Stick-breaking for multinomial

- Drawing $y_i \sim \text{Multinomial}(p_{i1}, \ldots, p_{iS})$ is equivalent to drawing a sequence of binary random variables as

$$b_{is} \,\big|\, \{b_{ij}\}_{j<s} \sim \text{Bernoulli}\left[\left(1 - \textstyle\sum_{j<s} b_{ij}\right) \pi_{is}\right],$$
$$\pi_{is} = \frac{p_{is}}{1 - \sum_{j<s} p_{ij}}, \quad s = 1, 2, \ldots, S.$$

- If we define $y_i = s$ if and only if $b_{is} = 1$ and $b_{ij} = 0$ for all $j \neq s$, then we have

$$\begin{aligned}
p_{is} &= P(y_i = s \,|\, \{\pi_{is}\}_{1,S}) \\
&= P(b_{is} = 1) \prod_{j \neq s} P(b_{ij} = 0) \\
&= (\pi_{is})^{\mathbf{1}(s \neq S)} \prod_{j<s} (1 - \pi_{ij}).
\end{aligned}$$

Augmented stick-breaking

Theorem (1)

Suppose $y_i \sim \sum_{s=1}^{S} p_{is}\delta_s$, *where* $[p_{i1}, \ldots, p_{iS}]$ *is a multinomial probability vector whose elements are constructed as*

$$p_{is} = (\pi_{is})^{\mathbf{1}(s \neq S)} \prod_{j<s}(1 - \pi_{ij}),$$

then y_i *can be equivalently generated from* ***augmented stick-breaking (aSB)*** *as*

$$y_i \sim \sum_{s=1}^{S} \left\{ \mathbf{1}(b_{is} = 1)^{\mathbf{1}(s \neq S)} \prod_{j<s} \mathbf{1}(b_{ij} = 0) \right\} \delta_s,$$

$$b_{is} \sim \text{Bernoulli}(\pi_{is}), \quad s \in \{1, \ldots, S\}.$$

- Augmented stick-breaking transforms the problem of multinomial regression with S categories into the problem of S conditionally independent binary regressions.
- Gibbs sampling:
  - Sample b_{is} for $s \in \{1, \dots, S\}$:
    - $b_{is} = 0$ if $s < y_i$
    - $b_{is} = 1$ if $s = y_i$
    - $b_{is} \sim \text{Bernoulli}(\pi_{is})$ if $s > y_i$
  - Solve $b_{is} \sim \text{Bernoulli}(\pi_{is})$ for $s \in \{1, \dots, S\}$, where the covariate-dependent stick probability π_{is} for the sth stick/category is a deterministic function of $\boldsymbol{x}_i' \boldsymbol{\beta}_s$.
- Any binary regression model (with cross entropy loss) can be generalized to a multinomial one under augmented stick-breaking, but a naive combination may not work well.

Example: augmented stick-breaking logistic regrssion

- If we let $\pi_{is} = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}_s}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_s}}$, which means $logit(\pi_{is}) = \boldsymbol{x}_i'\boldsymbol{\beta}_s$, then we have

$$p_{is} = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}_s}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_s}} \prod_{j<s} \frac{1}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j}}.$$

- Augmented logistic stick-breaking:

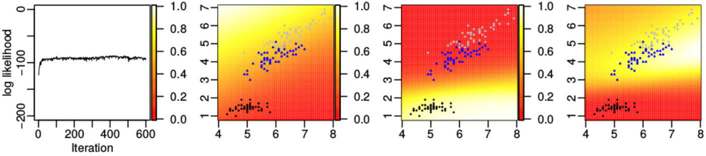
$$y_i \sim \sum_{s=1}^{S} \left\{ \mathbf{1}(b_{is}=1)^{\mathbf{1}(s\neq S)} \prod_{j<s} \mathbf{1}(b_{ij}=0) \right\} \delta_s,$$

$$b_{is} \sim \text{Bernoulli}\left(\pi_{is} = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}_s}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_s}} \right), \quad s \in \{1,\ldots,S\}.$$

- Gibbs sampling:
  - Sample b_{is} for $s \in \{1, \dots, S\}$:
    - $b_{is} = 0$ if $s < y_i$
    - $b_{is} = 1$ if $s = y_i$
    - $b_{is} \sim \text{Bernoulli}(\pi_{is})$ if $s > y_i$
  - Solve $b_{is} \sim \text{Bernoulli}\left(\pi_{is} = \frac{e^{\boldsymbol{x}_i' \boldsymbol{\beta}_s}}{1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_s}}\right)$:
    - $\boldsymbol{\beta}_s$ can be inferred with the Polya-Gamma data augmentation.
    - Closed-form Gibbs sampling update equations.
- Problem solved? End of the talk?? Not really...

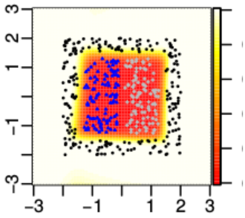
- The number of geometric constraints increases in s.
  - $p_{i1} = \left(1 + e^{-\boldsymbol{x}_i'\boldsymbol{\beta}_1}\right)^{-1}$ is larger than 0.5 if
$$\boldsymbol{x}_i'\boldsymbol{\beta}_1 > 0.$$
  - $p_{i2} = \left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}_1}\right)^{-1}\left(1 + e^{-\boldsymbol{x}_i'\boldsymbol{\beta}_2}\right)^{-1}$ is possible to be larger than 0.5 only if both
$$\boldsymbol{x}_i'\boldsymbol{\beta}_1 < 0 \text{ and } \boldsymbol{x}_i'\boldsymbol{\beta}_2 > 0$$
  - $p_{i3} = \left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}_1}\right)^{-1}\left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}_2}\right)^{-1}\left(1 + e^{-\boldsymbol{x}_i'\boldsymbol{\beta}_3}\right)^{-1}$ is possible to be larger than 0.5 only if
$$\boldsymbol{x}_i'\boldsymbol{\beta}_1 < 0,\ \boldsymbol{x}_i'\boldsymbol{\beta}_2 < 0,\ \text{and } \boldsymbol{x}_i'\boldsymbol{\beta}_3 > 0.$$
  - $\cdots$

- Augmented logistic stick-breaking is not invariant to label permutation, and may or may not work depending on how the categories are labeled (ordered).
- For the Iris data (sepal and petal lengths as covaraites), it does not work well if
  - 1st category/stick: blue points (middle)
  - 2nd category/stick: black points (bottom)
  - 3rd category/stick: gray points (top)



- It works well as long as the blue points are not labeled as the first category (stick).

- If some categories are not linearly separable, augmented logistic stick-breaking may not work well no matter how the categories are labeled.



- How to address the sensitivity to label permutation?
- How to separate the categories that are not linearly separable?

Permuted and augmented stick-breaking (paSB)

Denote $\boldsymbol{z} = (z_1, \ldots, z_S)$ as a permutation of $(1, \ldots, S)$

- $z_s \in \{1, \ldots, S\}$ is the index of the latent stick that category s is uniquely mapped to.
- $S!$ possible permutations
- $6! = 720$; $7! = 5,040$; $\ldots$; $10! = 3,628,800$; $\ldots$
- Fortunately, the effective search space could be significantly smaller than $S!$. We will discuss why and provide examples.

Theorem (2)

Suppose $y_i \sim \sum_{s=1}^S p_{is}(\boldsymbol{z})\delta_s$, where $[p_{i1}(\boldsymbol{z}), \ldots, p_{iS}(\boldsymbol{z})]$ is a multinomial probability vector whose elements are constructed as

$$p_{is}(\boldsymbol{z}) = (\pi_{iz_s})^{\mathbf{1}(z_s \neq S)} \prod_{j<z_s}(1-\pi_{ij}),$$

then y_i can be equivalently generated under the permuted and augmented stick-breaking (paSB) construction as

$$y_i \sim \sum_{s=1}^S \left\{[\mathbf{1}(b_{iz_s}=1)]^{\mathbf{1}(z_s \neq S)} \prod_{j<z_s} \mathbf{1}(b_{ij}=0)\right\}\delta_s\,,$$
$$b_{ij} \sim \mathrm{Bernoulli}(\pi_{ij}), \quad j \in \{1, \ldots, S\}.$$

- S categories are randomly one-to-one mapped to S sticks.
- $\{b_{is}\}_s$ given $\{\pi_{is}\}_s$ are mutually independent in the prior.

- Gibbs sampling for paSB:
  - Sample b_{is} for $s \in \{1, \dots, S\}$:
    - Category y_i is mapped to stick z_{y_i}.
    - $b_{is} = 0$ if $s < z_{y_i}$
    - $b_{is} = 1$ if $s = z_{y_i}$
    - $b_{is} \sim \text{Bernoulli}(\pi_{is})$ if $s > z_{y_i}$
  - Solve $b_{is} \sim \text{Bernoulli}(\pi_{is})$ for $s \in \{1, \dots, S\}$, where the covariate-dependent stick probability π_{is} for the s*th* category is a deterministic function of $\boldsymbol{x}_i' \boldsymbol{\beta}_s$.
  - Sample $(z_1, \dots, z_S)$, the one-to-one mapping between the category and stick indices, using Metropolis-Hastings.

- Sample the label-stick one-to-one mapping:
  - Let $(z_1, \dots, z_S)$ be uniformly at random selected from the $S!$ possible permutations in the prior.
  - Propose to change $\boldsymbol{z} = (z_1, \dots, z_j, \dots, z_{j'}, \dots, z_S)$ to $\boldsymbol{z}' = (z'_1, \dots, z'_S) := (z_1, \dots, z_{j'}, \dots, z_j, \dots, z_S)$.
  - Accept the proposal with probability

$$\min\left\{\prod_i \frac{\prod_{s=1}^S [p_{iz'_s}]^{\mathbf{1}(y_i=s)}}{\prod_{s=1}^S [p_{iz_s}]^{\mathbf{1}(y_i=s)}},\ 1\right\}$$

$$= \min\left\{\prod_i \frac{\prod_{s=1}^S \left[(\pi_{iz'_s})^{\mathbf{1}(z'_s\neq S)} \prod_{j<z'_s}(1-\pi_{ij})\right]^{\mathbf{1}(y_i=s)}}{\prod_{s=1}^S \left[(\pi_{iz_s})^{\mathbf{1}(z_s\neq S)} \prod_{j<z_s}(1-\pi_{ij})\right]^{\mathbf{1}(y_i=s)}},\ 1\right\}.$$

- Proposing two indices z_j and $z_{j'}$ to switch in each iteration is effective for escaping from the set of poor mappings.
- The probability of a z_j not proposed to switch is $[(S-2)/S]^t$ after t MCMC iterations. Even if $S = 100$, this probability is less than 10^{-8} at $t = 1000$.
- $S/2$ is the expected number of iterations for a z_j to be proposed to switch.

paSB logistic regression

- Model:

$$y_i \sim \sum_{s=1}^{S} \left\{ \mathbf{1}(b_{iz_s} = 1)^{\mathbf{1}(z_s \neq S)} \prod_{j<z_s} \mathbf{1}(b_{ij} = 0) \right\} \delta_s,$$

$$b_{is} \sim \text{Bernoulli}\left(\pi_{is} = \frac{e^{\boldsymbol{x}_i' \boldsymbol{\beta}_s}}{1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_s}} \right), \quad s \in \{1, \dots, S\}.$$

- The number of geometric constraints increases in z_s.
  - If $z_s = 1$, then $p_{is} = \left(1 + e^{-\boldsymbol{x}_i'\boldsymbol{\beta}_1}\right)^{-1}$ is larger than 0.5 if
  $$\boldsymbol{x}_i'\boldsymbol{\beta}_1 > 0.$$
  - If $z_s = 2$, then $p_{is} = \left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}_1}\right)^{-1}\left(1 + e^{-\boldsymbol{x}_i'\boldsymbol{\beta}_2}\right)^{-1}$ is possible to be larger than 0.5 only if both
  $$\boldsymbol{x}_i'\boldsymbol{\beta}_1 < 0 \text{ and } \boldsymbol{x}_i'\boldsymbol{\beta}_2 > 0$$
  - If $z_s = 3$, then $p_{i3} = \left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}_1}\right)^{-1}\left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}_2}\right)^{-1}\left(1 + e^{-\boldsymbol{x}_i'\boldsymbol{\beta}_3}\right)^{-1}$ is possible to be larger than 0.5 only if
  $$\boldsymbol{x}_i'\boldsymbol{\beta}_1 < 0, \boldsymbol{x}_i'\boldsymbol{\beta}_2 < 0, \text{ and } \boldsymbol{x}_i'\boldsymbol{\beta}_3 > 0.$$
  - ...

Sequential decision making and relaxing "independence of irrelevant alternative" (IIA) assumption

A *one-vs-remaining* decision at each of the stick breaking steps.

Lemma

Under the paSB construction, the probability ratio of two choices are influenced by the success probabilities of the sticks that lie between these two choices' corresponding sticks. In other words, the probability ratio of two choices will be influenced by some other choices if they are not mapped to adjacent sticks.

paSB multinomial logistic regression as a discrete choice model

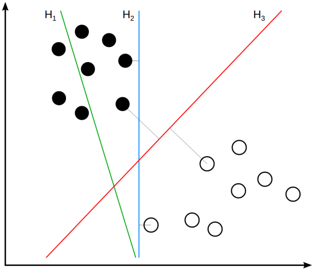
The paSB multinomial logistic regression that assigns choice $s \in \{1, \dots, S\}$ for individual i with probability $p_{is} = (\pi_{is})^{\mathbf{1}(s \neq S)} \prod_{j<s}(1 - \pi_{ij})$, $\pi_{is} = 1/(1 + e^{-W_{is}})$, is equivalent to a sequential random utility maximization model which selects choice s once $U_{is} > \sum_{j \geq s} U_{ij}$ is observed, where

$$
\begin{aligned}
U_{i1} &= U_{i2} + \cdots + U_{iS} + W_{i1} + \varepsilon_{i1}, \\
&\cdots \\
U_{is} &= \sum_{j>s} U_{ij} + W_{is} + \varepsilon_{is}, \\
&\cdots \\
U_{i(S-1)} &= W_{i(S-1)} + \varepsilon_{i(S-1)}, \\
U_{iS} &= 0,
\end{aligned}
$$

and $\varepsilon_{is} \stackrel{i.i.d.}{\sim} Logistic(0, 1)$.

Binary SVM

$$l(\beta, \nu) = \sum_{i=1}^{n} max(1 - y_i \mathbf{x}_i' \beta, 0) + \nu R(\beta), \text{ where } y_i \in \{-1, 1\}$$



https://en.wikipedia.org/wiki/Support_vector_machine

Bayesian Binary SVM [Polson & Scott (2011)]

- Mixture representation:

$$L(y_i \,|\, \boldsymbol{x}_i', \boldsymbol{\beta}) = \exp\left\{-2max(1 - y_i\boldsymbol{x}_i'\boldsymbol{\beta}, 0)\right\}$$
$$= \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left(-\frac{1}{2}\frac{(1+\lambda_i - y_i\boldsymbol{x}_i'\boldsymbol{\beta})^2}{\lambda_i}\right) d\lambda_i.$$

- Gibbs sampling for $\boldsymbol{\beta}$ is available under data augmentation.
- Decision rule [sollich (2002) and Mallick, et al. (2005)]:

$$P(y_i = 1 \,|\, \boldsymbol{x}_i, \boldsymbol{\beta}) = \begin{cases} \dfrac{1}{1 + e^{-2y_i\boldsymbol{x}_i\boldsymbol{\beta}}}, & \text{for } |\boldsymbol{x}_i'\boldsymbol{\beta}| \le 1; \\ \dfrac{1}{1 + e^{-y_i[\boldsymbol{x}_i\boldsymbol{\beta} + \mathrm{sign}(\boldsymbol{x}_i'\boldsymbol{\beta})]}}, & \text{for } |\boldsymbol{x}_i'\boldsymbol{\beta}| > 1; \end{cases}$$

paSB multinomial SVM

Under the paSB construction, given the covariate vector $\boldsymbol{x}_i$ and category-stick mapping $\boldsymbol{z}$, multinomial support vector machine (MSVM) parameterizes p_{is}, the multinomial probability of category s, as

$$p_{is}(\boldsymbol{z}) = [\pi_{iz_s,\mathrm{svm}}(\boldsymbol{x}_i, \boldsymbol{\beta}_s)]^{\mathbf{1}(z_s \neq S)} \prod\nolimits_{j: z_j < z_s} \pi_{iz_j,\mathrm{svm}}(\boldsymbol{x}_i, \boldsymbol{\beta}_j),$$

where

$$\pi_{iz_j,\mathrm{svm}}(\boldsymbol{x}_i, \boldsymbol{\beta}_j) = \begin{cases} \dfrac{1}{1+e^{-2\boldsymbol{x}_i\boldsymbol{\beta}_j}}, & \text{for } |\boldsymbol{x}_i'\boldsymbol{\beta}_j| \leq 1; \\ \dfrac{1}{1+e^{-\boldsymbol{x}_i\boldsymbol{\beta}_j - \mathrm{sign}(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}}, & \text{for } |\boldsymbol{x}_i'\boldsymbol{\beta}_j| > 1. \end{cases}$$

Softplus regression [Zhou (2016)]

$b_{is} \sim$ Bernoulli

$$\left[1 - \prod_{k=1}^{K} \left(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_{sk}^{(T+1)}} \ln\left\{1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_{sk}^{(T)}} \ln\left[1 + \ldots \ln\left(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_{sk}^{(2)}}\right)\right]\right\}\right)^{-r_{sk}}\right].$$

Equivalently,

$$\theta_{isk}^{(T)} \sim \text{Gamma}\left(r_{sk}, e^{\boldsymbol{x}_i' \boldsymbol{\beta}_{sk}^{(T+1)}}\right),$$

$$\cdots$$

$$\theta_{isk}^{(t)} \sim \text{Gamma}\left(\theta_{isk}^{(t+1)}, e^{\boldsymbol{x}_i' \boldsymbol{\beta}_{sk}^{(t+1)}}\right),$$

$$\cdots$$

$$\theta_{isk}^{(1)} \sim \text{Gamma}\left(\theta_{isk}^{(2)}, e^{\boldsymbol{x}_i' \boldsymbol{\beta}_{sk}^{(2)}}\right),$$

$$b_{is} = \mathbf{1}(m_{is} \geq 1), \;\; m_{is} = \sum_{k=1}^{K} m_{isk}^{(1)}, \;\; m_{isk}^{(1)} \sim \text{Pois}(\theta_{isk}^{(1)}),$$

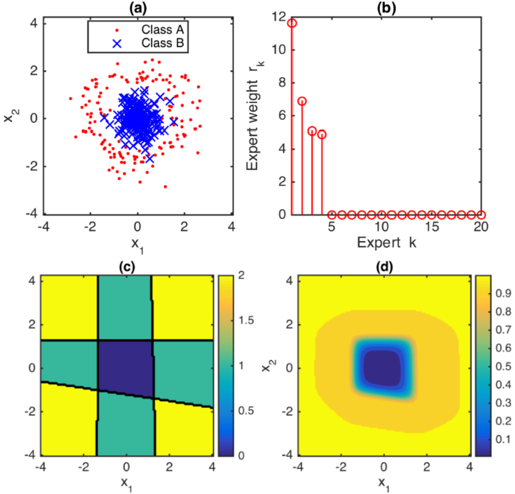
$K \to \infty$ is supported by the gamma process.

Properties of binary softplus regression

- $K > 1$ and $T = 1$: using the interaction of up to K hyperplanes to enclose negative examples
- $K = 1$ and $T > 1$: using the interaction of up to T hyperplanes to enclose positive examples
- $K >$ and $T > 1$: using the union of convex-polytope-like confined space to enclose positive examples
- K and T together control the nonlinear capacity of the model
- $K \to \infty$ is supported by the gamma process.

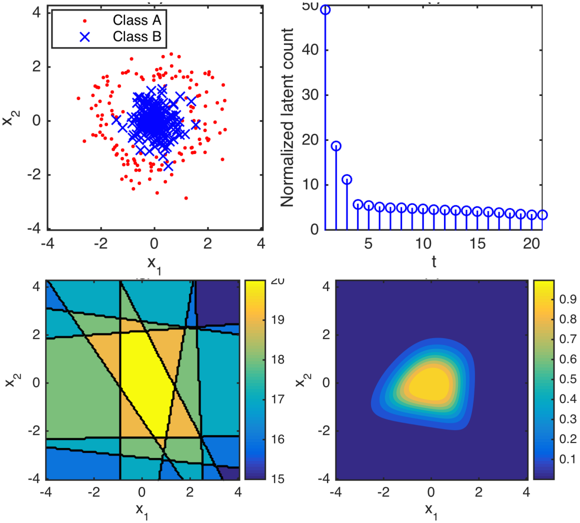
Binary softplus regression $K = 20,\ T = 1$

Label Class A as 1 and Class B as 0:



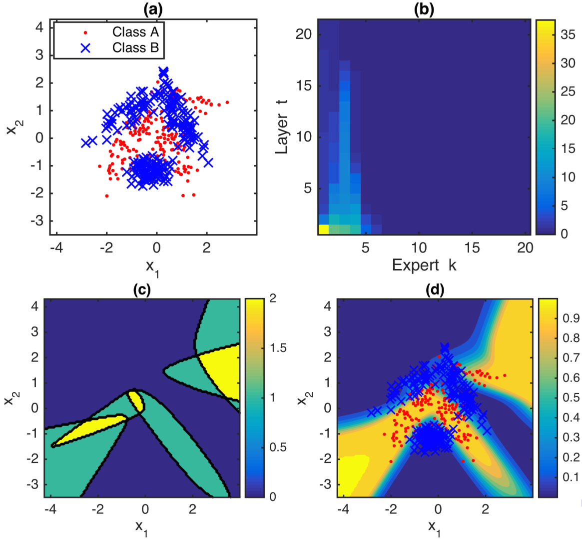
Binary softplus regression $K = 1,\ T = 20$

Label Class A as 0 and Class B as 1:



Binary softplus regression $K = 20,\ T = 20$

Label Class A as 1 and Class B as 0:



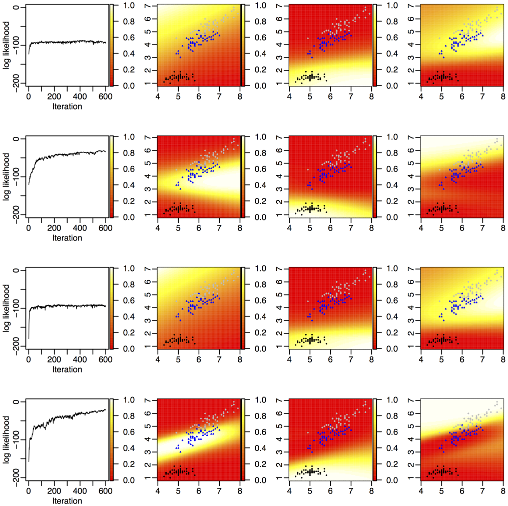
Multinomial softplus regression (MSR)

With a draw from a gamma process for each category that consists of countably infinite atoms $\beta_{sk}^{(2:T+1)}$ with weights $r_{sk} > 0$, where $\beta_{sk}^{(t)} \in \mathbb{R}^{P+1}$, given the covariate vector $\boldsymbol{x}_i$ and category-stick mapping $\boldsymbol{z}$, MSR parameterizes p_{is}, the multinomial probability of category s, under the paSB construction as

$$p_{iz_s} =$$

$$\left[1 - \prod_{k=1}^{\infty}\left(1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{sk}^{(T+1)}}\ln\left\{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{sk}^{(T)}}\ln\left[1+\ldots\ln\left(1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{sk}^{(2)}}\right)\right]\right\}\right)^{-r_{sk}}\right]^{\mathbf{1}(z_s \neq S)}$$

$$\times \prod_{j:z_j<z_s}\left[\prod_{k=1}^{\infty}\left(1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{jk}^{(T+1)}}\ln\left\{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{jk}^{(T)}}\ln\left[1+\ldots\ln\left(1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{jk}^{(2)}}\right)\right]\right\}\right)^{-r_{jk}}\right].$$

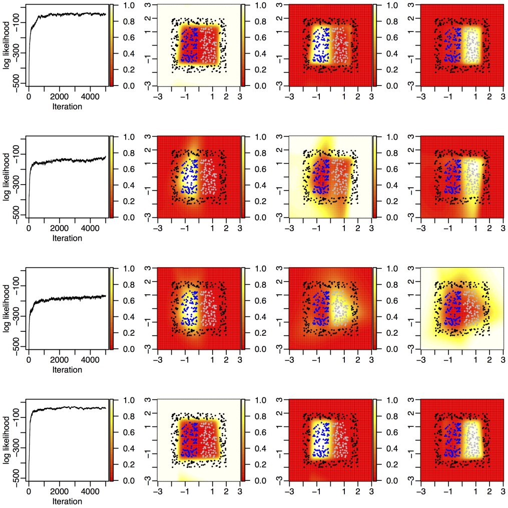


Label blue, black, and gray (middle, bottom, and top) points as classes 1, 2, and 3, respectively.

Fix $\boldsymbol{z} = [1, 2, 3]$ for paSB multinomial softplus regression.

Row 1: $K = 1$, $T = 1$. Row 2: $K = 1$, $T = 3$.

Row 3: $K = 5$, $T = 1$. Row 4: $K = 5$, $T = 3$.



Label blue, black, and gray (inside left, outside, and inside right) points as classes 1, 2, and 3, respectively.

paSB multinomial softplus regression with $K = T = 10$.

Row 1: $\boldsymbol{z} = [1, 2, 3]$. Row 2: $\boldsymbol{z} = [2, 1, 3]$

Row 3: $\boldsymbol{z} = [3, 1, 2]$. Row 4: sample $\boldsymbol{z}$ during MCMC iterations

Model comparison

Table: Comparison of classification error rate of paSB-MSVM, MSR with different K and T, L_2-MLR, SVM and AMM.

	paSB-MSVM	$K=1$ $T=1$	$K=1$ $T=3$	$K=5$ $T=1$	$K=5$ $T=3$	L_2-MLR	SVM	AMM
square	0	13.49	0.79	0	0	53.17	4.76	16.67
iris	3.33	4.00	4.00	4.00	3.33	3.33	4.00	4.67
wine	2.78	2.78	1.11	3.33	1.11	3.89	2.78	3.89
glass	29.30	29.30	28.37	31.16	30.70	35.81	28.84	37.67
vehicle	21.65	20.08	19.29	18.11	19.29	22.44	14.17	21.89
waveform	15.76	16.93	16.91	15.38	15.80	15.80	15.02	18.54
segment	7.98	6.16	6.83	6.25	5.87	9.04	5.77	12.47
vowel	36.36	49.78	47.84	48.48	48.05	58.87	37.23	52.47
dna	3.96	4.64	5.40	4.81	4.55	5.23	4.55	5.43
satimage	8.90	13.45	12.95	12.10	11.50	17.95	8.50	15.31
ANER	0.97	1.07	1.04	1.06	0.97	2.27	1	1.71

Number of active experts in paSB multinomial softplus regression

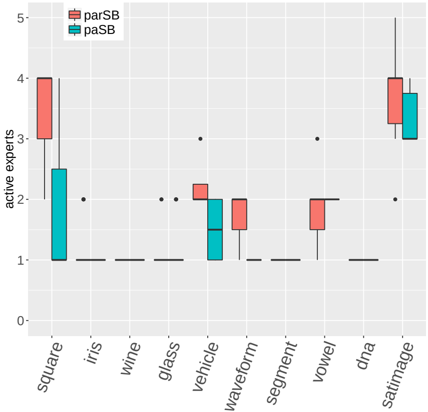


Figure: Boxplots of the number of active experts.

Likelihood for $S!$ different permulations

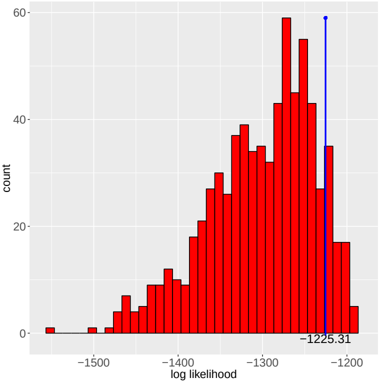


Figure: Histogram of $S! = 6! = 720$ log-likelihoods for Satimage, using augmented stick-breaking multinomial softplus regression (MSR) with $K = 5$ and $T = 3$. The blue line indicates the average log-likelihood of the collected MCMC samples of paSB MSR, with the permutation $\boldsymbol{z}$ sampled via the proposed Metropololis-Hasting step.

Softplus regression with support hyperplanes

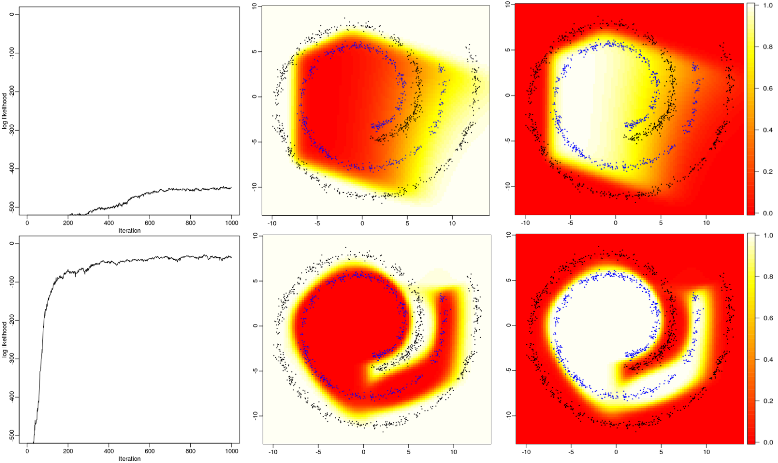


Figure: Row 1: Softplus regression with $K = 5$, $T = 3$. Row 2: Softplus regression with $K = 5$, $T = 3$ and data transformation of support hyperplanes.

Discussions

- A general framework to transform a binary classifier to a multi-class one.
- Fully Bayesian inference via data augmentation.
- The regression coefficient vectors of different categories can be sampled in parallel in each MCMC iteration.
- Not invariant to label permutation if the label-stick mapping is fixed.
- Asymmetric geometric constraints (more constraints for a category mapped to a larger-indexed stick).

Thank You!