Bayesian Nonparametric Modeling of Suicide Attempts

Francisco J. R. Ruiz, Isabel Valera,

Carlos Blanco, Fernando Perez-Cruz

- **Indian Buffet process:** $\mathbf{Z} \sim \mathrm{IBP}(\alpha)$.

- **Discrete observations:** $x_{nd} \in \{1, \ldots, R\}$.

$$p(x_{nd} = \text{`red'}|\mathbf{z}_{n\cdot}, \mathbf{B}^d) \propto \exp\left(\mathbf{z}_{n\cdot}\mathbf{b}^d_{\cdot\text{red}}\right),$$
$$p(x_{nd} = \text{`blue'}|\mathbf{z}_{n\cdot}, \mathbf{B}^d) \propto \exp\left(\mathbf{z}_{n\cdot}\mathbf{b}^d_{\cdot\text{blue}}\right),$$
$$\vdots$$
$$p(x_{nd} = r|\mathbf{z}_{n\cdot}, \mathbf{B}^d) = \frac{\exp\left(\mathbf{z}_{n\cdot}\mathbf{b}^d_{\cdot r}\right)}{\displaystyle\sum_{r'=1}^{R} \exp\left(\mathbf{z}_{n\cdot}\mathbf{b}^d_{\cdot r'}\right)}, \quad r = 1, \ldots, R,$$

where $\mathbf{b}^d_{\cdot r} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_b = \sigma_B^2 \mathbf{I})$.

Inference

- **Gibbs sampling** algorithm (requires $p(\mathbf{X}|\mathbf{Z})$).
- The posterior of $\mathbf{B}^1, \ldots, \mathbf{B}^D$ factorizes as

$$\overbrace{p(\mathbf{B}^1, \ldots, \mathbf{B}^D | \mathbf{X}, \mathbf{Z})}^{\text{Non Gauss}} = \prod_{d=1}^{D} p(\mathbf{B}^d | \mathbf{x}_{\cdot d}, \mathbf{Z}) = \prod_{d=1}^{D} \frac{\overbrace{p(\mathbf{x}_{\cdot d} | \mathbf{B}^d, \mathbf{Z})}^{\text{Non Gauss}} \overbrace{p(\mathbf{B}^d)}^{\text{Gauss}}}{p(\mathbf{x}_{\cdot d} | \mathbf{Z})}.$$

- We define $\psi(\mathbf{B}^d) = \log p(\mathbf{x}_{\cdot d} | \mathbf{B}^d, \mathbf{Z}) + \log p(\mathbf{B}^d)$.
- We want to compute

$$p(\mathbf{x}_{\cdot d} | \mathbf{Z}) = \int \exp\left(\psi(\mathbf{B}^d)\right) d\mathbf{B}^d.$$

Laplace approximation:

Approximate $\psi(\mathbf{B}^d)$ by its second-order Taylor series expansion.
Assume $p(\mathbf{B}^d | \mathbf{X}, \mathbf{Z})$ is a Gaussian distribution.

Results

- ▶ NESARC database (*National Epidemiologic Survey on Alcohol and Related Conditions*):
  - ▶ Samples the U.S. population.
  - ▶ Multiple-choice questions ('Blank', 'Unknown', 'Yes' and 'No').
  - ▶ We select the **20 questions** with the highest mutual information with the question about having **attempted suicide**.
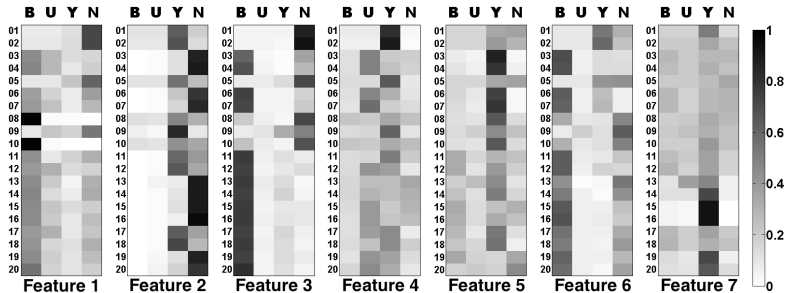

Figure: Probability of each response to the 20 selected questions.

NESARC database

- Suicide attempt probability in the whole database: $\sim 8\%$.

Latent features							Suicide attempt probability		Number of cases	
							Train	Hold-out	Train	Hold-out
1	-	-	-	-	-	-	6.74%	5.55%	430	8072
-	1	-	-	-	-	-	10.56%	11.16%	322	6083
-	-	1	-	-	-	-	**3.72%**	**4.60%**	457	8632
-	-	-	1	-	-	-	**25.23%**	**22.25%**	111	2355
-	-	-	-	1	-	-	8.64%	9.69%	301	5782
-	-	-	-	-	1	-	6.90%	7.18%	464	8928
-	-	-	-	-	-	1	14.29%	14.18%	91	1664
-	-	0	0	-	-	-	30.77%	28.55%	26	571
-	-	0	1	-	-	-	**82.35%**	**61.95%**	17	297
-	-	1	0	-	-	-	**0.83%**	**0.87%**	363	6574
-	-	1	1	-	-	-	14.89%	16.52%	94	2058

Visit us on Thursday 2:00 - 6:00pm, Poster Th41