Bayesian Nonparametrics for Marathon Modeling

Melanie F. Pradier[1], Francisco J. R. Ruiz[1,2], Fernando Perez-Cruz[1,3]

[1]University Carlos III in Madrid
[2]Columbia University
[3]Bell Labs, Alcatel-Lucent

Outline

Sports Statistical Modelling

What for?

1. Building a team
   - Who to hire and what for.
2. Betting.
   - Paramutual.
   - Book.
3. Ranking Teams or Players.
4. Training.
4. Analyzing oponents.

Sports

- Horse racing.
- Baseball.
- football.
- Football (aka futbol).
- Basketball.
- Chess.
- Tennis.
- ...

Why Marathon?

What were we looking for?

- Fairly compare runners of different ages and sex, because:
  - Oversubscribed popular marathons award entry by time (Boston the only way).
  - World Masters Athletics rankings.
- Understand the best strategy to run a marathon.

Why Generative Modeling?

- It has less discriminative performance.
- It needs more data.
- Its harder to do inference.
- It is not straightforward.
- It does not give you a number.

Why Generative Modeling?

- It has less discriminative performance.
- It needs more data.
- Its harder to do inference.
- It is not straightforward.
- It does not give you a number.

- We can interpret the solution.
- We can answer any question.
- We can find errors in our data.
- We can find what we were not looking for.

Why Marathon?

What were we looking for?

- Fairly compare runners of different ages and sex, because:
  - Oversubscribed popular marathons award entry by time (Boston the only way).
  - World Masters Association rankings.
- Understand the best strategy to run a marathon.

What else did we find?

- Women age differently than man.
- Men are riskier than women.
- Novice runners do not know their limits.

Our data

- NYC Marathon finishing time 2006-2011.
- Boston Marathon finishing time 2010-2011.
- London Marathon finishing time 2010-2011.
- NYC and Boston Marathon age of participants.
- London Marathon age group of the participants.
- Male and Female information of all participants.
- NYC Marathon intermediate times (every 5Km and half).
- Over 366,000 records.

Dependent Dirichlet Process

- Dirichlet Process for clustering:

$$p(\mathbf{x}) = \sum_i \pi_i f_i(\mathbf{x}|\boldsymbol{\theta}_i)$$

  Open-ended number of degrees of freedom.

- Dependent Dirichlet Process for clustering:

$$p(\mathbf{x}|d) = \sum_i \pi_i(d) f_i(\mathbf{x}|\boldsymbol{\theta}_i(d))$$

  Use side information d to make parameters depend on it.

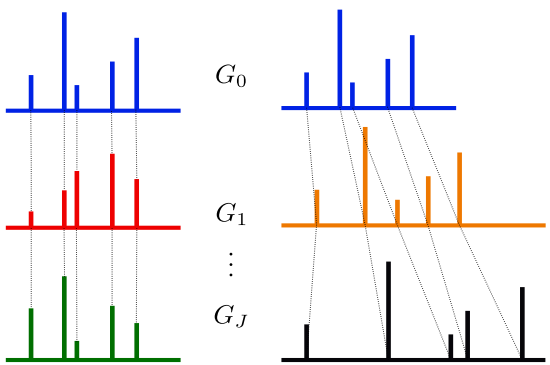
Particular Dependent Dirichlet Processes

- Hierarchical Dirichlet Process for clustering:

$$p(\mathbf{x}|d) = \sum_i \pi_i(d) f_i(\mathbf{x}|\boldsymbol{\theta}_i)$$

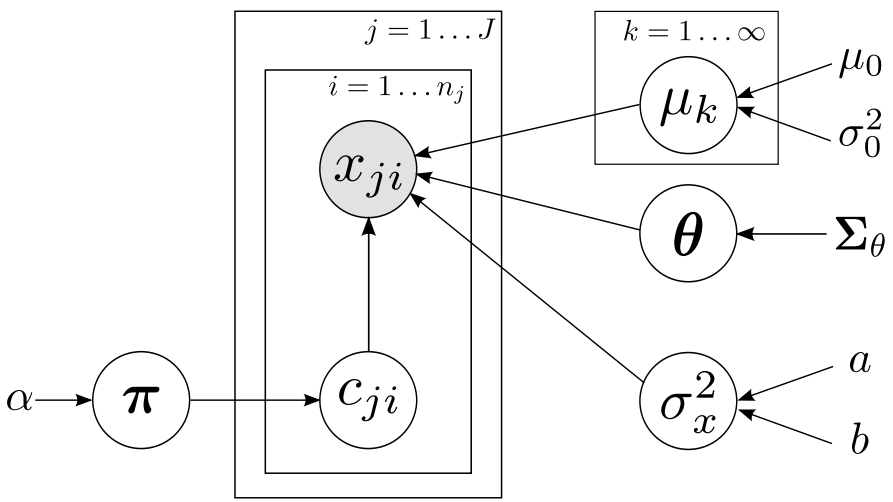
- Single-p Dependent Dirichlet Process for clustering:

$$p(\mathbf{x}|d) = \sum_i \pi_i f_i(\mathbf{x}|\boldsymbol{\theta}_i(d))$$



Our Basic Model

$$x_{ji}|c_{ji} = k, \mu_k, \theta_j, \sigma_x^2 \sim \mathcal{N}\left(x_{ji}|\mu_k + \theta_j, \sigma_x^2\right)$$
$$\boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right) \quad \boldsymbol{\theta} = [\theta_1, \ldots, \theta_J]$$
$$(\boldsymbol{\Sigma}_\theta)_{\ell j} = \sigma_\theta^2 \cdot \exp\left(-(\ell - j)^2/2\nu^2\right) + \kappa\delta(\ell - j)$$
$$\mu_k \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right),$$
$$c_{ji} = k|\pi_k \sim \pi_k,$$
$$\boldsymbol{\pi}|\alpha \sim \mathrm{GEM}(\alpha),$$
$$\sigma_x^2 \sim \mathcal{IG}\left(a, b\right)$$



Hyperparameter Setting

$\mu_0 = 5$ hours

$\sigma_0 = 1$ hour

$\sigma_\theta^2 = 15$ minutes

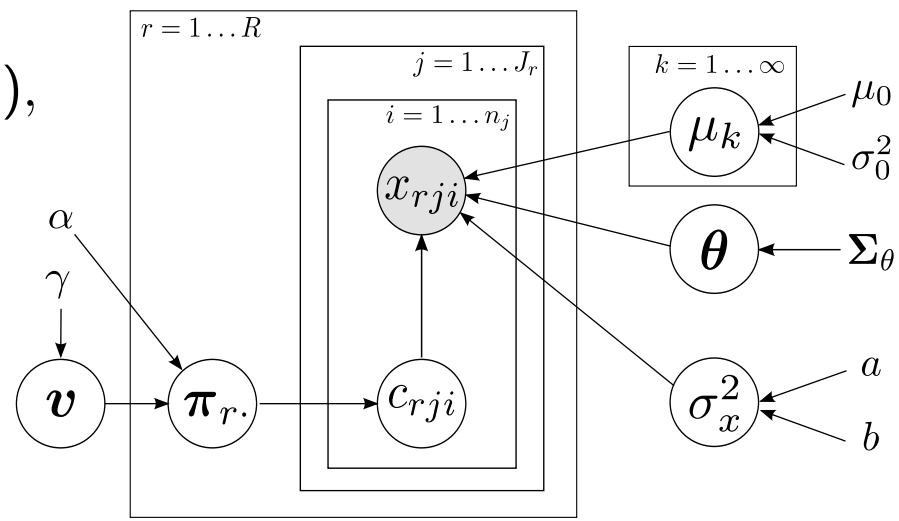
$a = 1$ and $b = 1$

$\nu = 10$ and $\kappa = 10^{-6}$

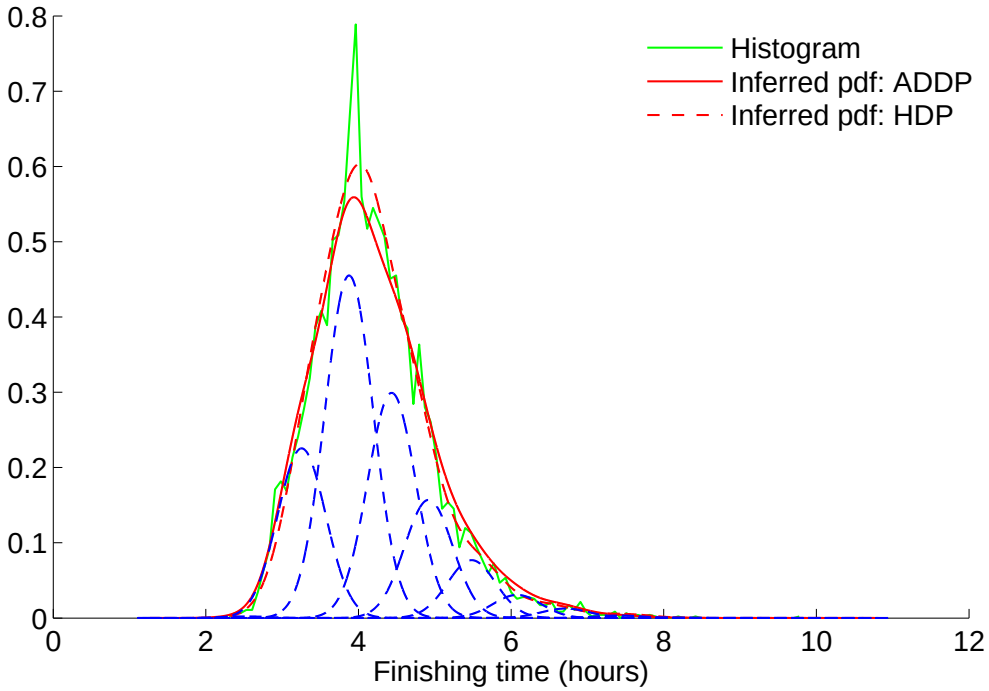
$\alpha \sim \Gamma(1, 10)$

Multiple Races

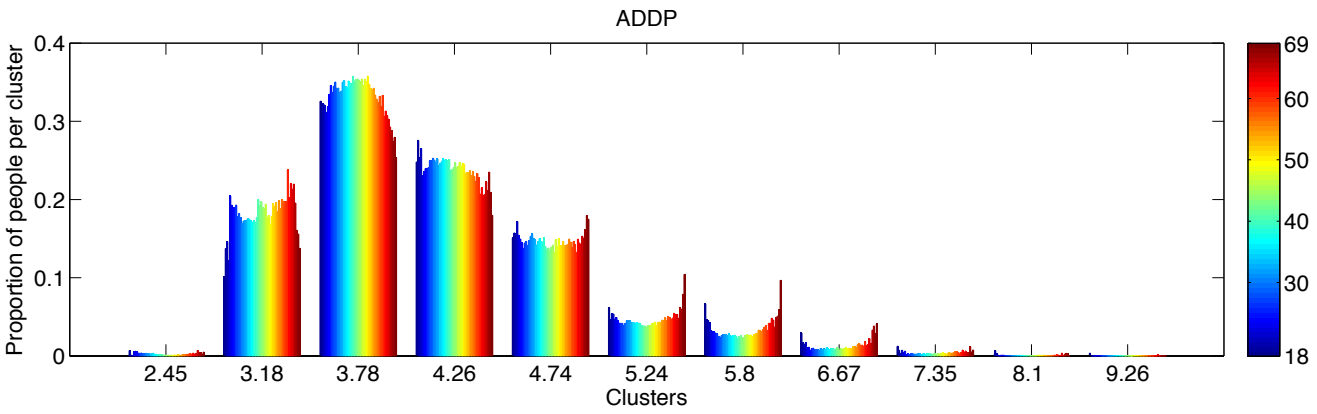
$$x_{rji}|c_{rji} = k, \mu_k, \theta_j, \sigma_x^2 \sim \mathcal{N}\left(x_{rji}|\mu_k + \theta_j, \sigma_x^2\right),$$
$$\mu_k \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right),$$
$$\boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right),$$
$$\sigma_x^2 \sim \mathcal{IG}\left(a, b\right),$$
$$c_{rji} = k|\pi_{rk} \sim \pi_{rk},$$
$$\boldsymbol{\pi}_{r.}|\boldsymbol{v}, \alpha \sim \mathrm{DP}(\alpha, \boldsymbol{v}),$$
$$\boldsymbol{v}|\gamma \sim \mathrm{GEM}(\gamma),$$



Overall fit for 28 year-old male runners

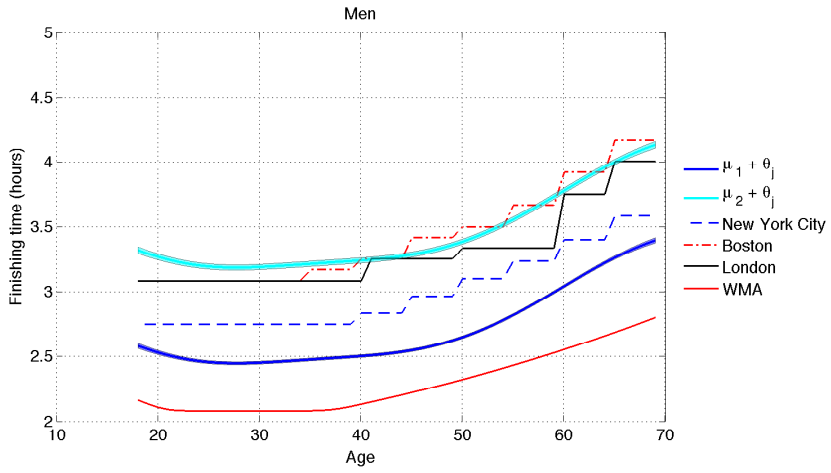


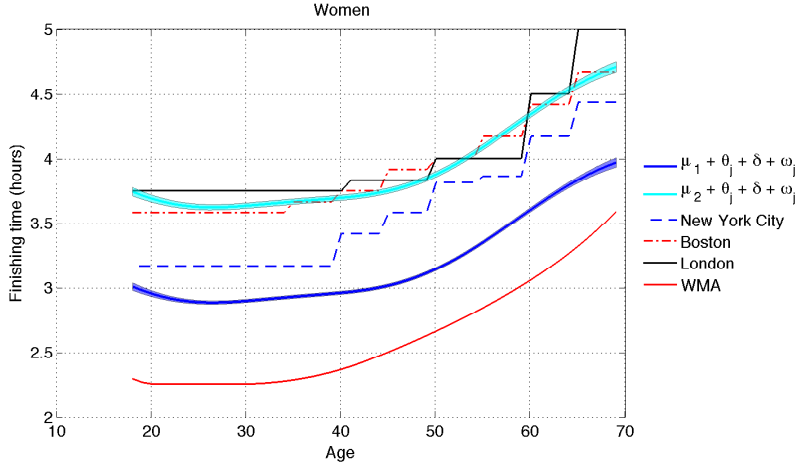
Proportions (28-year-old male runners)



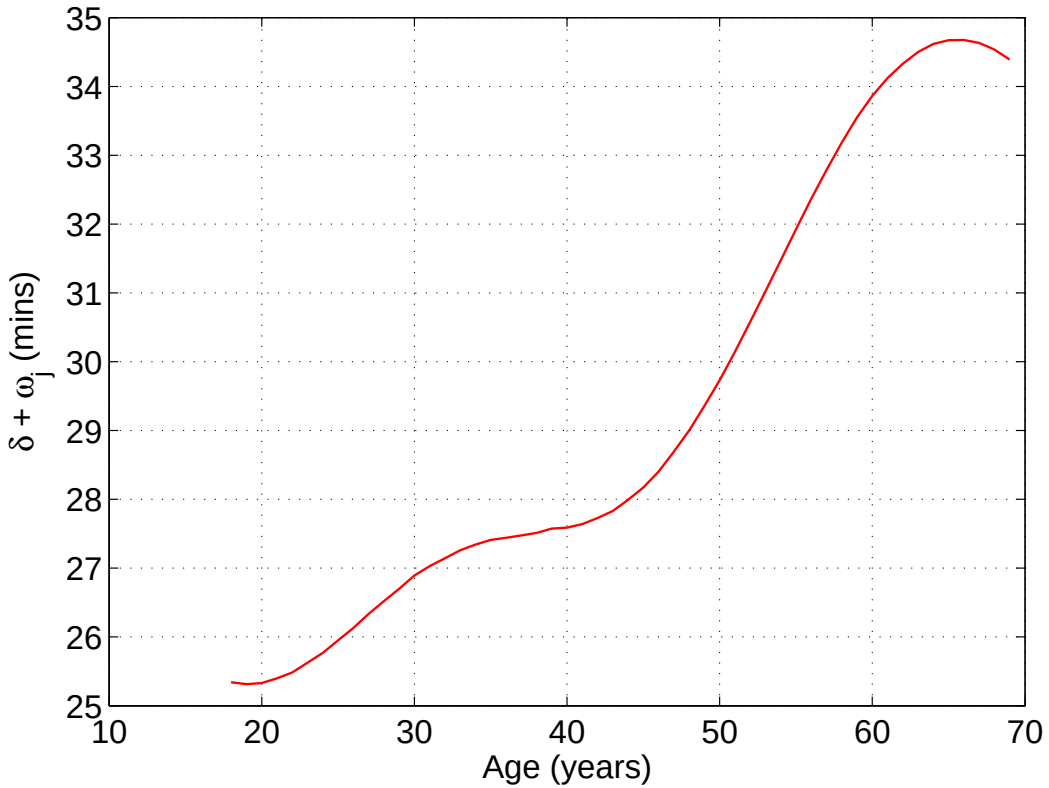
age	18	23	28	33	38	43	48	53	58	63	68
	8.00	1.66	0.00	1.09	2.65	4.61	9.09	17.74	30.21	43.80	55.09

Comparison with Marathon entry requirements

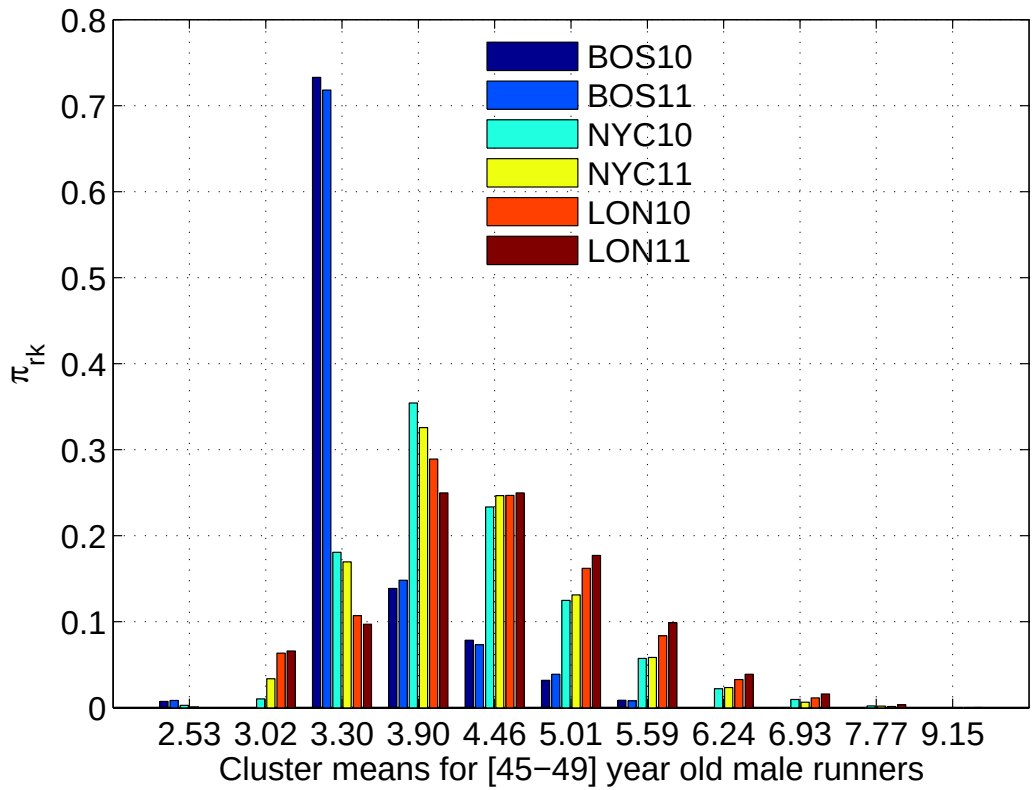




Delay for female runners



Different races

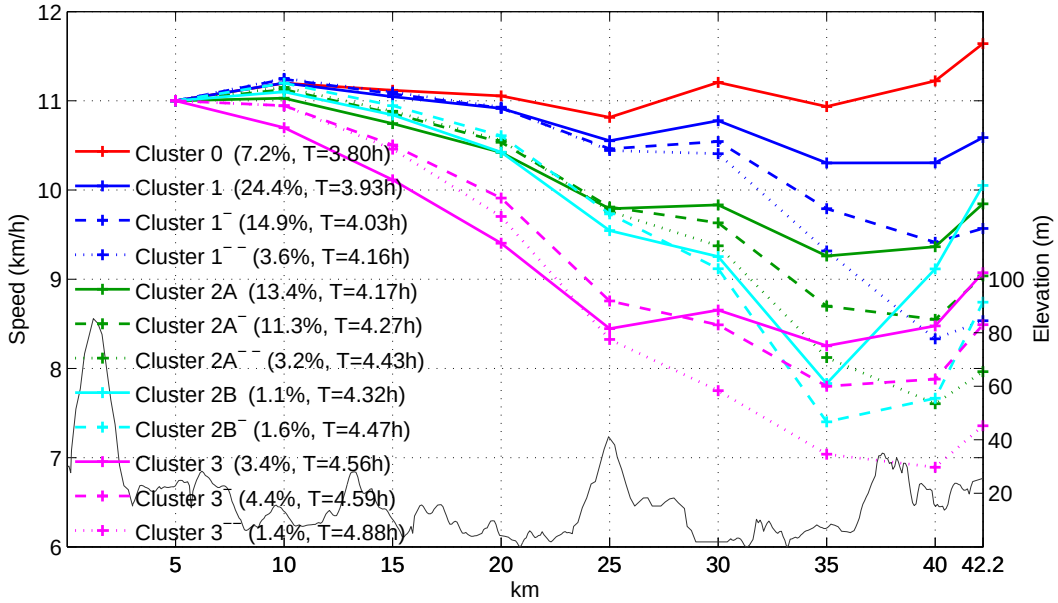


Intermediate time model

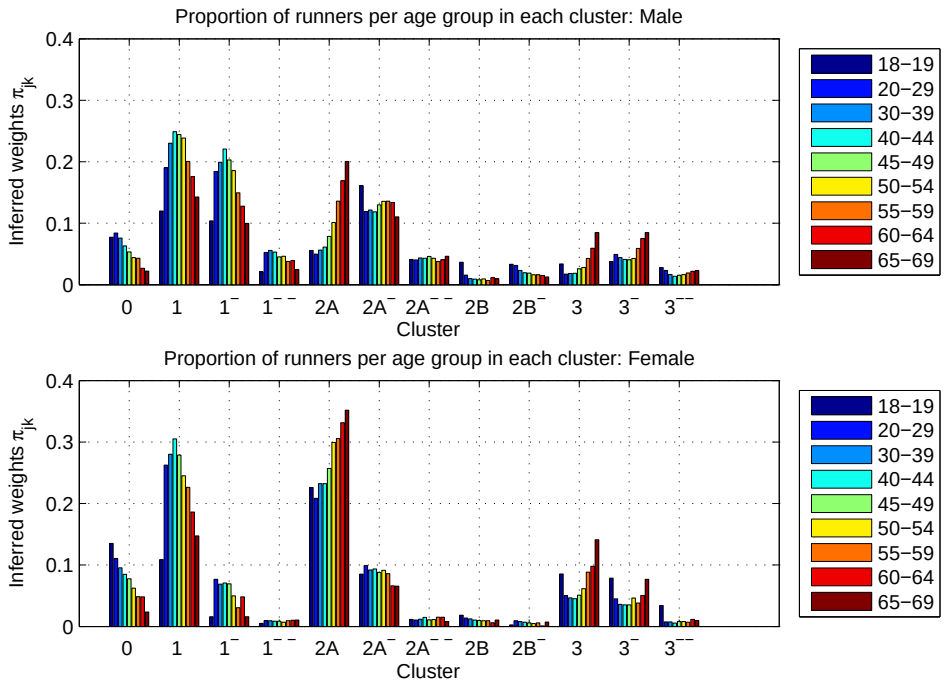
- Each participant has a time for each 5Km and half and full Marathons.
- We convert each record to a proportion of time spent at each interval.
- We use an HDP to cluster this proportions.

$$\mathbf{x}_{ji}|c_{ji} = k, \mathbf{p}_k \sim \text{Dirichlet}\left(\tau p_{k1}, \ldots, \tau p_{kD}\right)$$
$$\mathbf{p}_k \sim \text{Dirichlet}\left(\epsilon \ell_1, \ldots, \epsilon \ell_D\right)$$

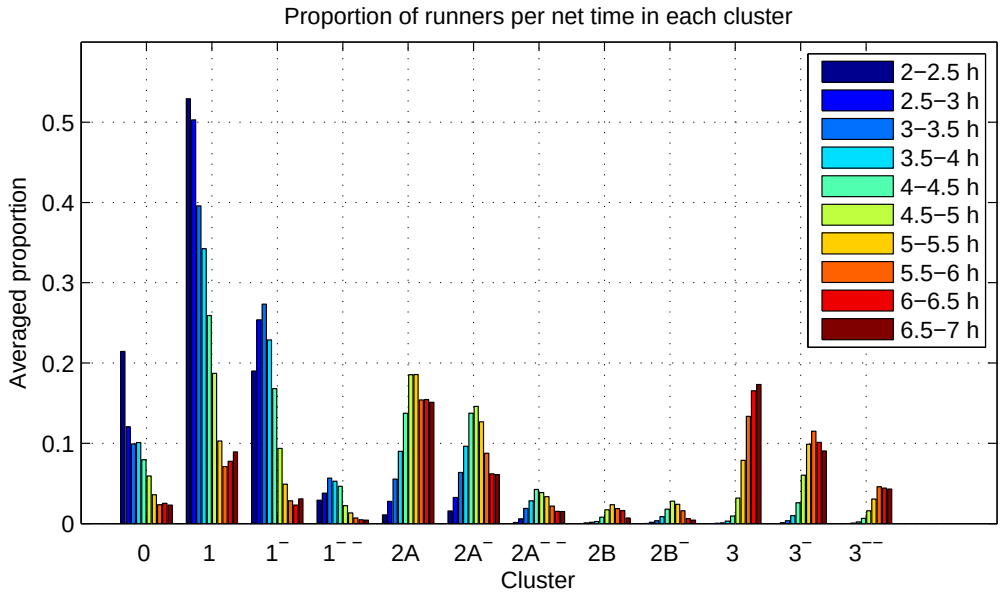
Hierarchical Dirichlet Process



Proportions age and sex



Proportions finishing times



Thanks!

Questions?