Вариационный вывод для непараметрической скрытой марковской модели.

Сокурский Юрий

Факультет ВМК МГУ им. Ломоносова

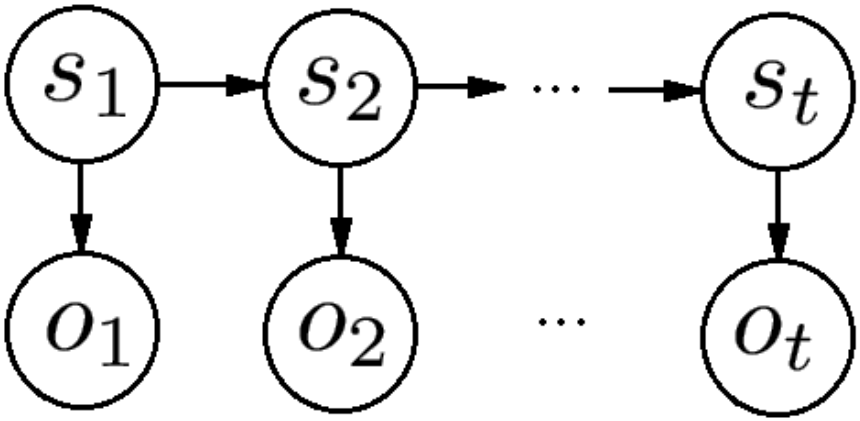
29.04.2014

Outline

1. Hidden Markov Model
2. Stick breaking
3. Infinite Hidden Markov Model
4. Variational Inference
5. Эксперименты
6. Выводы
7. Подробные формулы.

Hidden Markov Model

Joint Distribution



$$p(O, S|\beta) =$$

$$\prod_{k=1}^{K} p(s_{1,k} = 1, \pi)^{s_{1,k}} \prod_{t=2}^{Tm} \prod_{i=1}^{K} \prod_{j=1}^{K} p(s_{t,j} = 1|s_{t-1,i} = 1, T)^{s_{t-1,i}s_{t,j}}$$

$$\prod_{t=1}^{Tm} \prod_{k=1}^{K} p(o_t|\theta_k)^{s_{t,k}} = \prod_{k=1}^{K} \pi_k{}^{s_{1,k}} \prod_{t=2}^{Tm} \prod_{i=1}^{K} \prod_{j=1}^{K} T_{i,j}{}^{s_{t-1,i}s_{t,j}} \prod_{t=1}^{Tm} \prod_{k=1}^{K} p(o_t|\theta_k)^{s_{t,k}}$$

$$\beta = \{\pi, T, \theta\}$$

$logp(O|\beta) \to \max\limits_{\beta}$

Общая схема:

- E-step: $p(S|O,\hat{\beta})$
- M-step: $\hat{\beta}_{new} = \operatorname{argmax}\limits_{\beta} \mathbb{E}_{p(s|o,\hat{\beta})} log\ p(O,S|\beta)$

Hidden Markov Model

Observation distribution

Распределение на наблюдаемые переменные:
$p(o_t|\mu_k, \sigma_k) = \prod_{n=1}^{N} p(o_t^n|\mu_k^n, \sigma_k^n)$
$p(o_t^n|\mu_k^n, \sigma_k^n) = \mathcal{N}(\mu_k^n, (\sigma_k^n)^2)$

Параметры распределения: $\theta_k = \{\mu_k^n, \sigma_k^n\}_{n=1}^{N}$

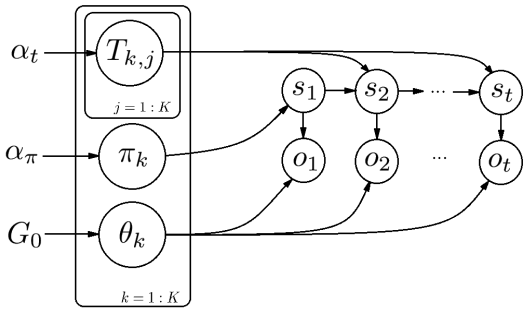
Пусть наблюдения O_{test} следуют прямо за O_{train}
Тогда качество работы алгоритма можно оценить с помощью:

$$p_K(O_{train}|\beta) = \int\limits_s p(O_{train}, S_{train}|\beta)dS$$

$$p_K(O_{test}|O_{train}, \beta) = \frac{\int\limits_s p(O_{train}, O_{test}, S_{train}, S_{test}|\beta)dS}{\int\limits_s p(O_{train}, S_{train}|\beta)dS}$$

HMM with prior

HMM с априорым распределением на T, π, θ:



$T_k \sim Dir(\alpha_t)\ k = \{1, ...K\}$

$\pi \sim Dir(\alpha_\pi)$

$\theta_k \sim N - \Gamma^{-1}(\alpha, \beta, \lambda, \mu)$: $\mu_k^n \sim \mathcal{N}(\mu, (\sigma_k^n)^2/\lambda)$; $(\sigma_k^n)^2 \sim \Gamma^{-1}(\alpha, \beta)$

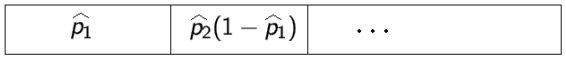
Dirichlet Process

Процесс Дирихле над пространством S: случайный процесс, с помощью которого можно породить вероятностные меры над S.
Пусть G_0 - вероятностная мера в пространстве S, $\alpha > 0$
$X \sim DP(\alpha, G_0)$, если
Для любого разбиения пространства S: $\{B_i\}_{i=1}^N$
$(X(B_1), ..., X(B_N)) \sim Dirichlet(\alpha G_0(B_1), ..., \alpha G_0(B_N))$

Stick breaking

TSBP

Stick breaking process:

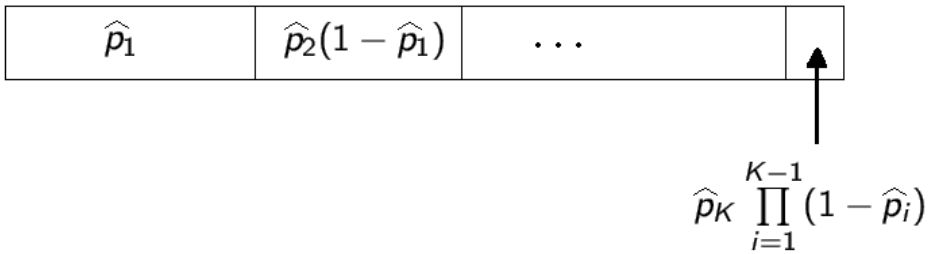


$$\hat{p}_j \sim Beta(1, \alpha)\ j = \{1, ...\infty\},\ \theta_j \sim G_0$$

$$p_j = \hat{p}_j \prod_{i=1}^{j-1}(1 - \hat{p}_i)\ j = \{1, ...\infty\}$$

$$p(\theta) = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}(\theta)$$

Truncated stick breaking process (TSBP):



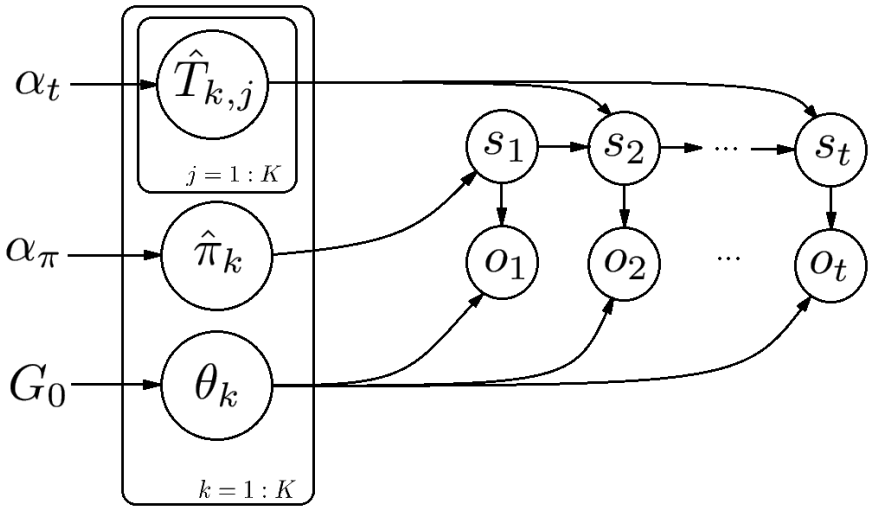
$$\hat{p}_j \sim Beta(1, \alpha)\ j = \{1, ...K-1\},\ \hat{p}_K = 1$$

$$p_j = \hat{p}_j \prod_{i=1}^{j-1}(1 - \hat{p}_i)\ j = \{1, ...K\}$$

iHMM

based on stick breaking

Схема iHMM:



$$\hat{T}_{k,j} \sim Beta(1, \alpha_t)\ k = \{1, ...K\}, j = \{1, ...K-1\}$$

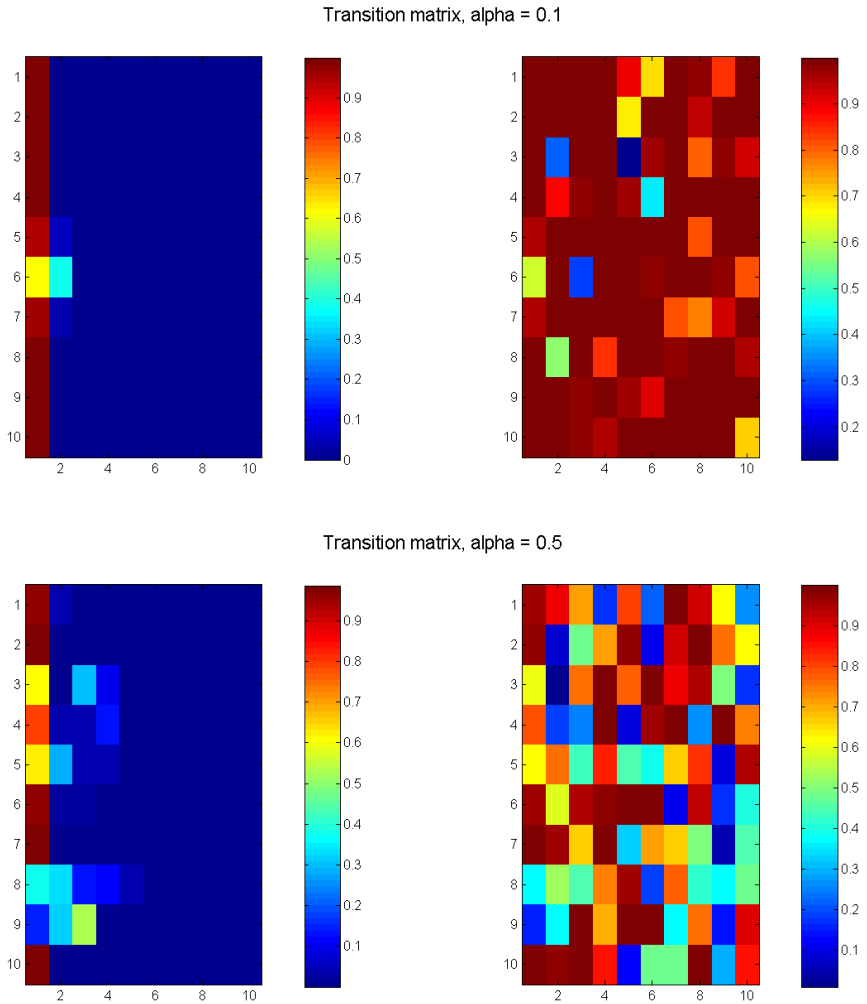
$$\hat{\pi}_k \sim Bets(1, \alpha_\pi) k = \{1, ...K-1\}$$

$$T_{k,j} = \hat{T}_{k,j} \prod_{i=1}^{j-1} (1 - \hat{T}_{k,i})\ k, j = \{1, ...K\}$$

$$\pi_k = \hat{\pi}_k \prod_{i=1}^{k-1} (1 - \hat{\pi}_i)\ k = \{1, ...K\}$$

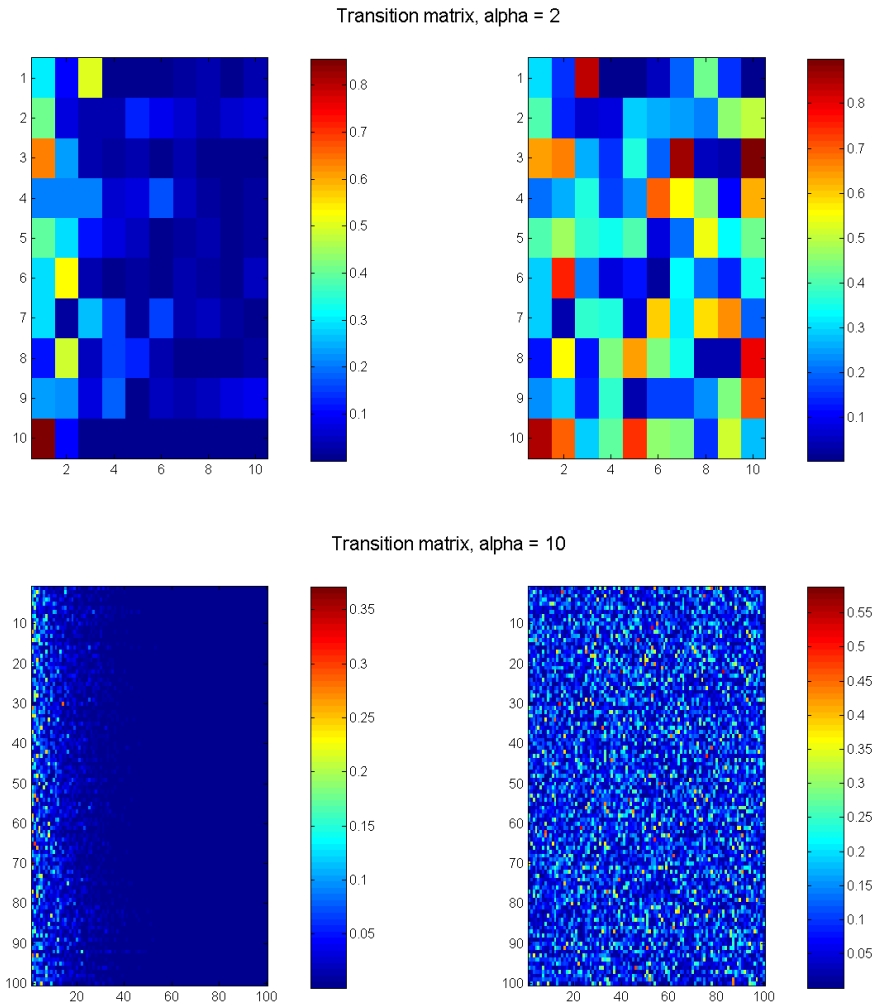
iHMM

sample transition matrix



iHMM

sample transition matrix



Variational Inference

Мотивация

Наблюдаемые переменные: o, скрытые: h

$p(h|o) = \frac{p(h,o)}{p(o)}$

$p(o) = \int\limits_h p(o,h)dh$ - экспоненциальная сложность.

Преимущество вариационного вывода: быстрая монотонная сходимость

Variational Inference

ELBO

$p(o) = \frac{p(h,o)}{p(h|o)}$. Приблизим $p(h|o)$ с помощью $q(h) = \prod_{m=1}^{M} q(h_m)$.

$KL(q(h)||p(h|o)) \to \min_q$

$log\, p(o) = log \int_h p(o,h)dh \geq$

$\mathbb{E}_q log \frac{p(o,h)}{q(h)} = \mathbb{E}_q log\, p(o,h) - \mathbb{E}_q log\, q(h) = ELBO(o,q)$

{evidence lower bound - нижняя оценка на $log\, p(o)$}

Изветно, что: $KL(q(h)||p(h|o)) \to \min_q \Leftrightarrow ELBO(o, q(h)) \to \max_q$,

где $KL(q(h)||p(h|o)) = \mathbb{E}_q log \frac{q(h)}{p(h|o)}$

Максимизация ELBO:

$$\mathcal{L} = ELBO(o, q(h)) = \int_h \prod_{i=1}^{M} q(h_i) log\, p(h, o) dh - \sum_{i=1}^{M} \mathbb{E}_i log\, q(h_i)$$

$(\mathbb{E}_i = \mathbb{E}_{q(h_i)})$.

Используя $p(h, o) = p(o) \prod_{i=1}^{M} p(h_i | h_{1:(i-1)}, o)$.

Получим $\mathcal{L} = \mathcal{L}_j + const(h_j)$, где $\mathcal{L}_j = \mathbb{E}_q log\, p(h_j | h_{-j}, o) - \mathbb{E}_{q_j} log\, q(h_j)$

По очереди будем максимизировать $\mathcal{L}_j$. $\frac{d\mathcal{L}_j}{dq(h_j)} = 0 \Rightarrow$

$q^*(h_j) \propto exp\{\mathbb{E}_{-j} log\, p(h_j | h_{-j}, o)\}$.

Variational Inference

iHMM

Скрытые переменные: $h = \{s, \hat{T}, \hat{\pi}, \theta\}$

Вариационное приближение: $q(h) = \prod_{t=1}^{Tm} \prod_{k=1}^{K} q(s_t^k = 1|\gamma_t^k)^{s_t^k} \prod_{k=1}^{K-1} q(\hat{\pi}_k|\beta_\pi^k)$

$\prod_{k=1}^{K} \prod_{j=1}^{K-1} q(\hat{T}_{k,j}|\beta_T^{k,j}) \prod_{k=1}^{K} q(\theta_k|G_k)$

$q(\hat{\pi_k}) \sim Beta$

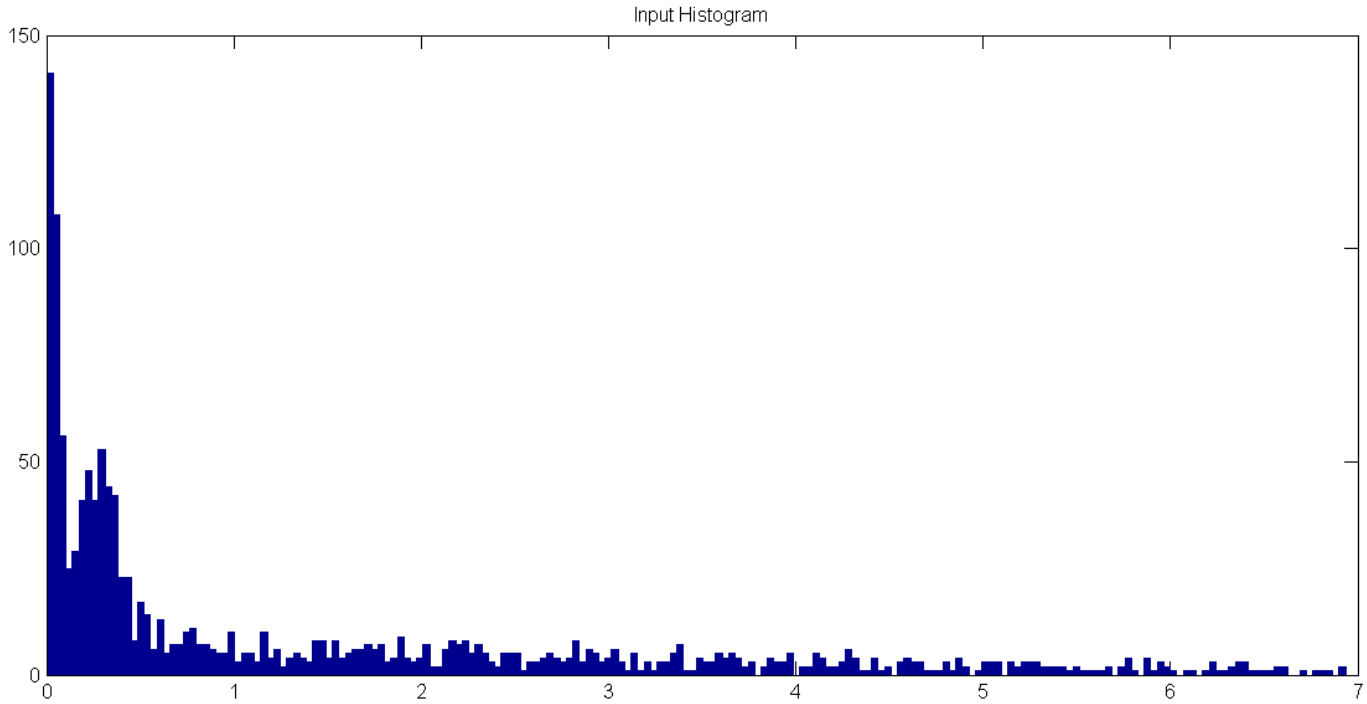
$q(\hat{T_{k,j}}) \sim Beta$

$q(s_t^k = 1|\gamma_t^k) = \gamma_t^k$

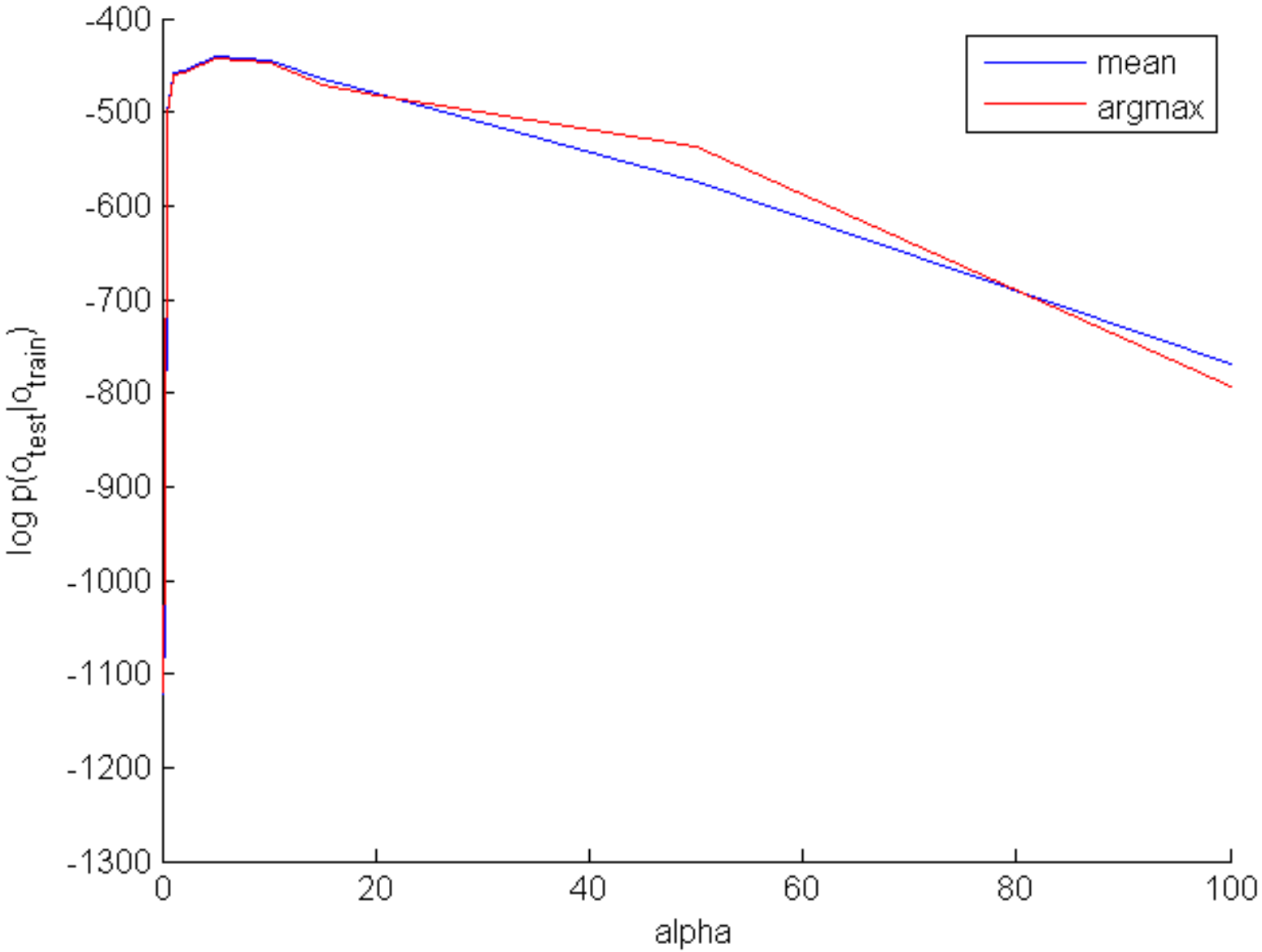
$q(\theta_k|G_k) = G_k(\theta_k)$

Данные

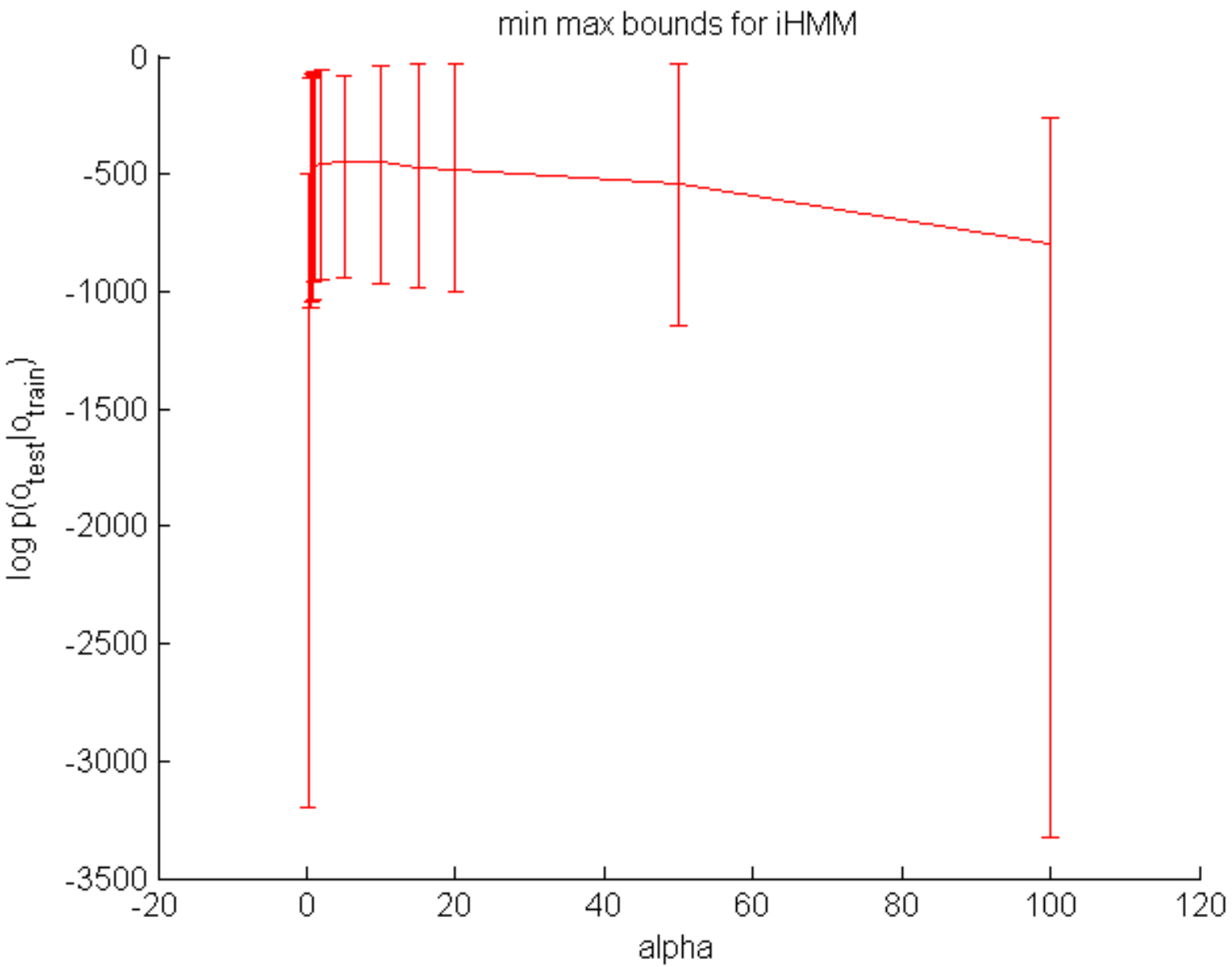
Распределение на наблюдаемую переменную:



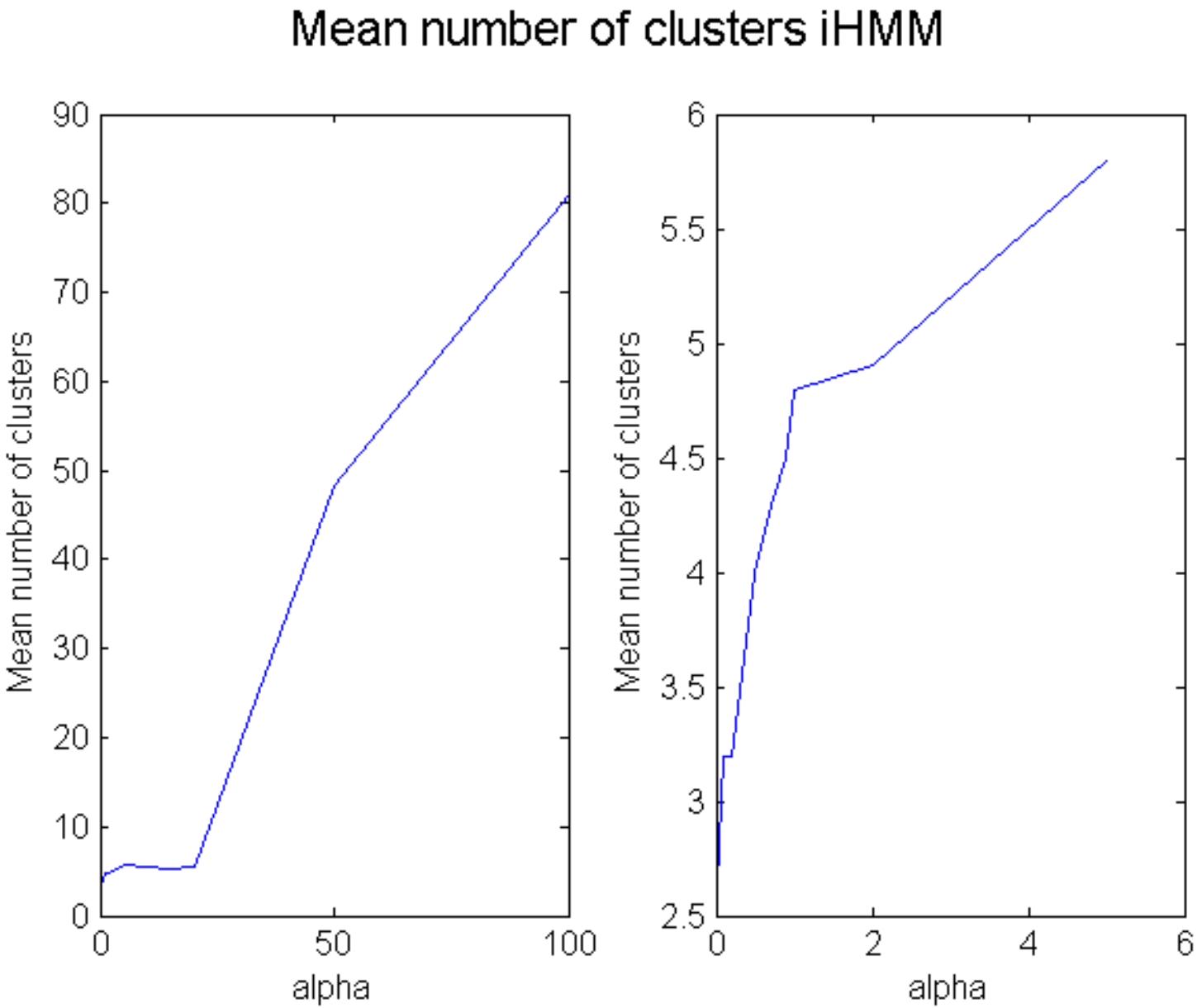
iHMM подбор α_t



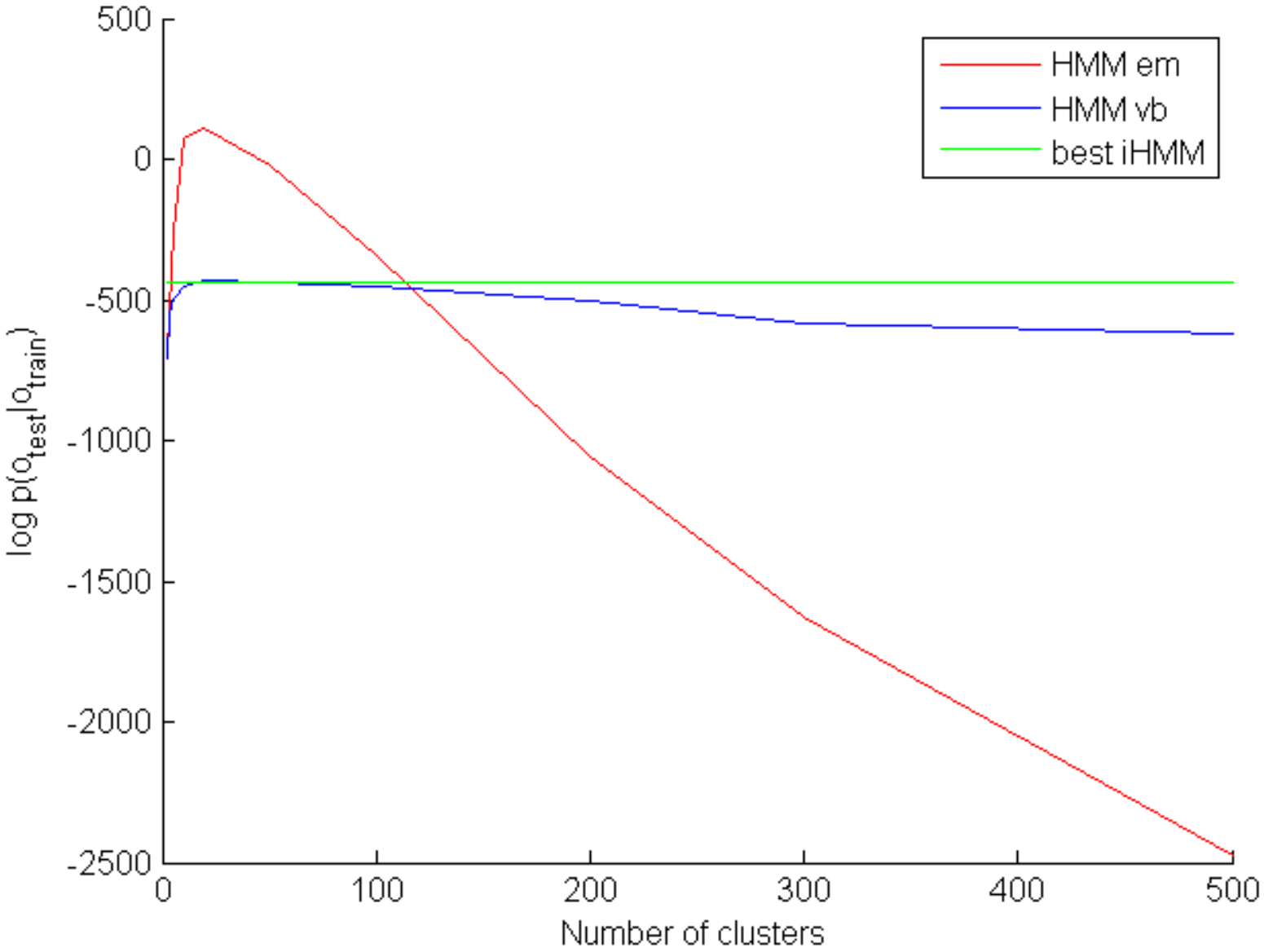
iHMM крайние значения на 10 запусках



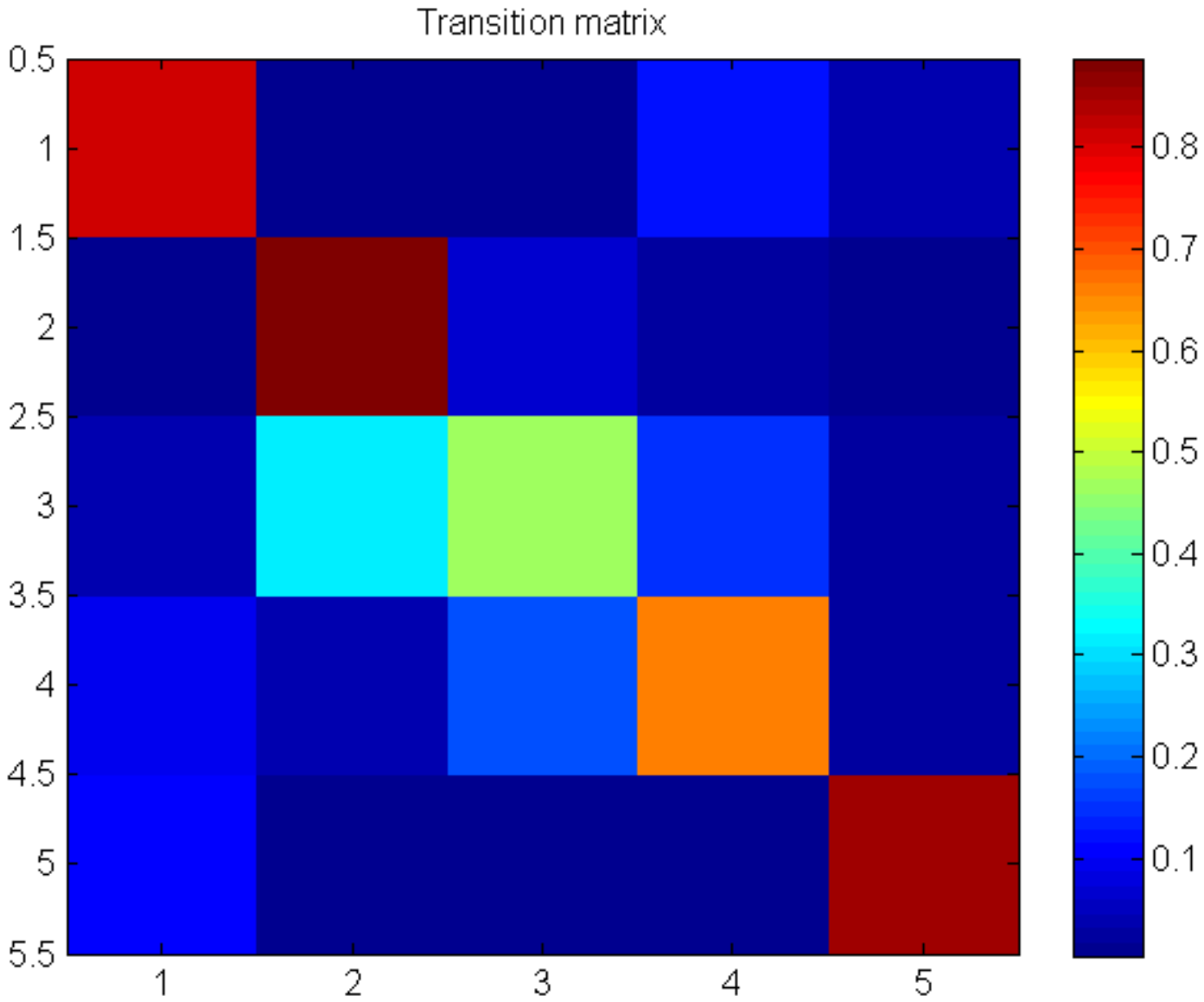
iHMM среднее кол-во кластеров



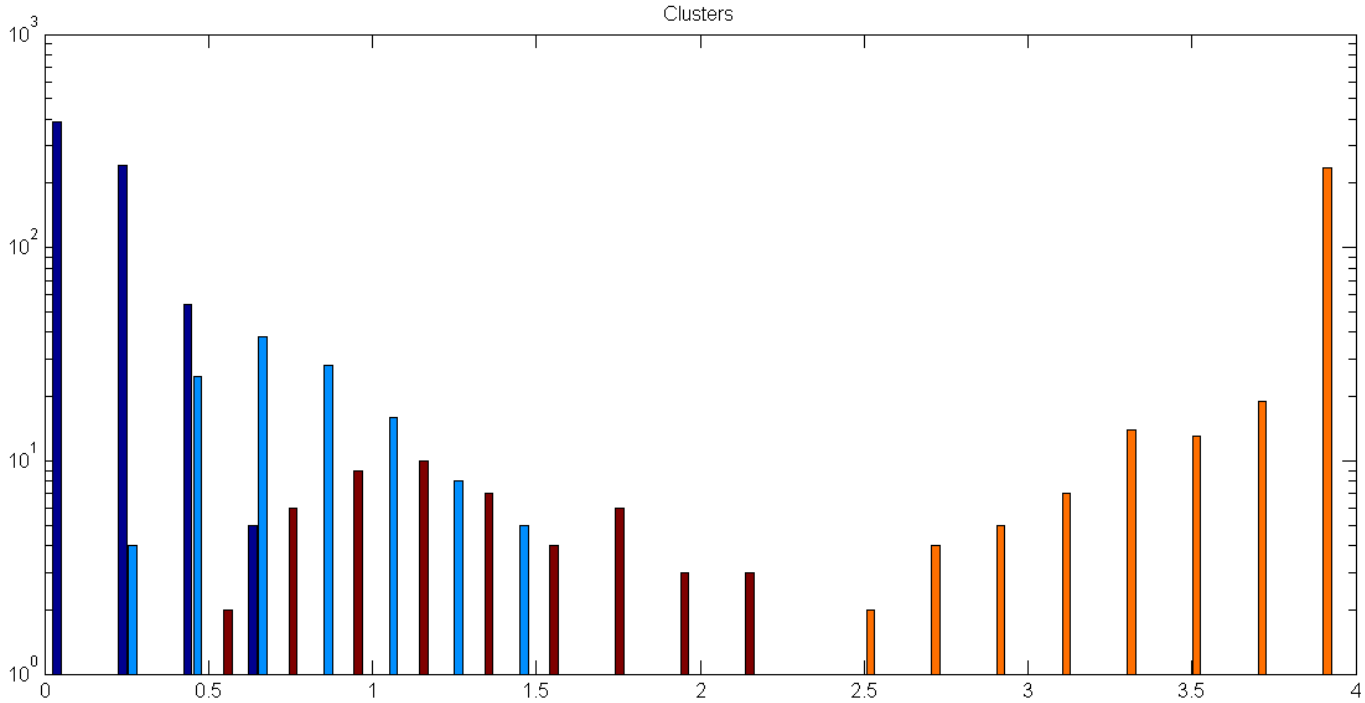
Сравнение с HMM



Матрица переходов для лучшего α_t



Кластеризация для лучшего α_t



Выводы

- iHMM позволяет оценить количество кластеров
- Вариационное приближение не сильно понизило качество
- Вариационное приближение повысило устойчивость к некорректным параметрам (HMM).

EM algorithm for HMM

Формулы для параметров:

- E-step: $\gamma_t(s_t) = p(s_t)$, $\xi_{t-1,t}(s_{t-1}, s_t) = p(s_{t-1}, s_t)$ (маргинальные распределения считаются с помощью message passing)
- M-step: $T_{i,j} = \frac{\sum_{t=2}^{Tm} \xi_{t-1,t}(i,j)}{\sum_{t=2}^{Tm} \gamma_{t-1}(i)}$ $\pi_k = \gamma_1(k)$,

$$\hat{\mu}_k^n = \frac{\sum_{t=1}^{Tm} \gamma_t^k o_t^n}{\sum_{t=1}^{Tm} \gamma_t^k}, \; (\hat{\sigma}_k^n)^2 = \frac{\sum_{t=1}^{Tm} \gamma_t^k (o_t^n - \hat{\mu}_k^n)^2}{\sum_{t=1}^{Tm} \gamma_t^k}.$$

Variational inference for iHMM

Совместное распределение: $log\ p(o, s, \hat{T}, \hat{\pi}, \theta | \alpha_t, \alpha_\pi, G_0) =$

$$\sum_{t=1}^{T} \sum_{k=1}^{K} s_t^k log\ p(o_t|\theta_k) + \sum_{k=1}^{K} s_1^k log\ p(s_1^k = 1|\hat{pi}) +$$

$$\sum_{t=1}^{T} \sum_{i=1}^{K} \sum_{j=1}^{K} s_{t-1}^i s_t^j log\ p(s_t^j = 1|s_{t-1}^i = 1, \hat{T}) + \sum_{k=1}^{K} log\ p(\theta_k|G_0) +$$

$$\sum_{k=1}^{K-1} log\ p(\pi_k|\alpha_\pi) + \sum_{k=1}^{K-1} \sum_{j=1}^{K} log\ p(\hat{T}_{j,k}|\alpha_t)$$