

1 General Notations

- N : Number of users
- n : Number of items, $n = \begin{cases} 2m - 1, & \text{if } n \text{ is odd} \\ 2m, & \text{otherwise} \end{cases}$
- \mathcal{P}_n : the space of permutation of n items
- $\mathbf{R}^1, \dots, \mathbf{R}^N$: full rankings given by the users
- $\mathbf{R}^j \in \mathcal{P}_n = \{R_1^j, \dots, R_n^j\} \sim \text{Mallows}(\boldsymbol{\rho}^0, \alpha^0)$, defined as $P(\mathbf{R}^j | \alpha^0, \boldsymbol{\rho}^0) = \frac{\exp\{-\frac{\alpha^0}{n} d(\mathbf{R}^j, \boldsymbol{\rho}^0)\}}{\sum_{\mathbf{r} \in \mathcal{P}_n} \exp\{-\frac{\alpha^0}{n} d(\mathbf{r}, \boldsymbol{\rho}^0)\}}$
- $P(\boldsymbol{\rho} | \mathbf{R}^1, \dots, \mathbf{R}^N, \alpha^o)$: Mallows posterior
- $\{i_1, \dots, i_n\}$: a ranking of n items that determines the sequence following which the items are to be sampled. i.e. $i_j = k$ indicates that item j is the k -th item is to be sampled
- $\{o_1, \dots, o_n\}$: an ordering of n items that corresponds to $\{i_1, \dots, i_n\}$ s.t. $i_{o_k} = k$. $\{o_1, \dots, o_n\}$ and $\{i_1, \dots, i_n\}$ have a one-to-one relationship
- $Q(\tilde{\boldsymbol{\rho}} | \cdot) = \sum_{\{i_1, \dots, i_n\} \in \mathcal{P}_n} q(\tilde{\boldsymbol{\rho}} | i_1, \dots, i_n, \alpha_0, \mathbf{R}^1, \dots, \mathbf{R}^N) \cdot g(i_1, \dots, i_n | \dots)$: pseudolikelihood that approximates the Mallows posterior
- $q(\tilde{\boldsymbol{\rho}} | i_1, \dots, i_n, \alpha^0, \mathbf{R}^1, \dots, \mathbf{R}^N) = q(\tilde{\boldsymbol{\rho}} | o_1, \dots, o_n, \alpha^0, \mathbf{R}^1, \dots, \mathbf{R}^N)$
 $= q(\tilde{\rho}_{o_1} | \alpha^0, o_1, R_{o_1}^1, \dots, R_{o_1}^N) \cdot q(\tilde{\rho}_{o_2} | \alpha^0, o_2, \tilde{\rho}_{o_1}, R_{o_2}^1, \dots, R_{o_2}^N) \cdot \dots \cdot$
 $q(\tilde{\rho}_{o_{n-1}} | \alpha^0, o_{n-1}, \tilde{\rho}_{o_1}, \dots, \tilde{\rho}_{o_{n-2}}, R_{o_{n-1}}^1, \dots, R_{o_{n-1}}^N) \cdot q(\tilde{\rho}_{o_n} | \alpha^0, o_n, \tilde{\rho}_{o_1}, \dots, \tilde{\rho}_{o_{n-1}}, R_{o_n}^1, \dots, R_{o_n}^N)$
 $= \frac{\exp\{-\frac{\alpha_0}{n} \sum_{j=1}^N d(R_{o_1}^j, \tilde{\rho}_{o_1})\} \mathbb{1}_{\tilde{\rho}_{o_1} \in \{1, \dots, n\}}}{\sum_{\tilde{r}_{o_1} \in \{1, \dots, n\}} \exp\{-\frac{\alpha_0}{n} \sum_{j=1}^N d(R_{o_1}^j, \tilde{r}_{o_1})\}}$

$$- q(\tilde{\rho}_{o_k} | \alpha^0, o_k, \tilde{\rho}_{o_1}, \dots, \tilde{\rho}_{o_{k-1}}, R_{o_k}^1, \dots, R_{o_k}^N) = \frac{\exp\{-\frac{\alpha_0}{n} \sum_{j=1}^N d(R_{o_k}^j, \tilde{\rho}_{o_k})\} \mathbb{1}_{\tilde{\rho}_{o_k} \in \{1, \dots, n\} \setminus \{\tilde{\rho}_{o_1}, \dots, \tilde{\rho}_{o_{k-1}}\}}}{\sum_{\tilde{r}_{o_k} \in \{1, \dots, n\} \setminus \{\tilde{\rho}_{o_1}, \dots, \tilde{\rho}_{o_{k-1}}\}} \exp\{-\frac{\alpha_0}{n} \sum_{j=1}^N d(R_{o_k}^j, \tilde{r}_{o_k})\}}$$

for $k = 2, \dots, n$

- \mathbf{o}^0 : a set of ordering that corresponds to $\boldsymbol{\rho}^0$ s.t. $\rho^{0^{-1}}(m) = o_m^0$
- Define the “v-function” $f_v(\cdot)$ such that $f_v(\boldsymbol{\rho}^0) = \mathcal{V}_{\boldsymbol{\rho}^0}$, where

$$- \mathcal{V}_{\boldsymbol{\rho}^0} = \begin{cases} \{\mathbf{r} \in \mathcal{P}_n : r_{o_m^0} = 1, r_{o_{m \pm k}^0} \in \{2k, 2k+1\}, k = 1, \dots, m-1\}, & \text{if } n \text{ is odd} \\ \{\mathbf{r} \in \mathcal{P}_n : \{r_{o_{m-k}^0}, r_{o_{m+k+1}^0}\} \in \{2k+1, 2k+2\}, k = 0, \dots, m\}, & \text{if } n \text{ is even} \end{cases}$$

2 Theorems and Lemmas

2.1

Lemma 2.1.1 *Given there are odd number of items, i.e. $n = 2m - 1$. $\forall \alpha^0 \in (0, \infty)$,*

1. $\mathbb{E}(R_{o_m^0} | \boldsymbol{\rho}^0, \alpha^0) = \rho_{o_m^0}^0 = m$
2. $\forall j \in [1, m-2], j < \mathbb{E}[R_{o_j^0} | \boldsymbol{\rho}^0, \alpha^0] < \mathbb{E}[R_{o_{j+1}^0} | \boldsymbol{\rho}^0, \alpha^0] < m$
3. $\forall j \in [m+2, 2m-1], m < \mathbb{E}[R_{o_{j-1}^0} | \boldsymbol{\rho}^0, \alpha^0] < \mathbb{E}[R_{o_j^0} | \boldsymbol{\rho}^0, \alpha^0] < j$

Similarly, if n is even, i.e. $n = 2m$, $\forall \alpha^0 \in (0, \infty)$,

1. $\forall j \in [1, m-1], j < \mathbb{E}[R_{o_j^0} | \boldsymbol{\rho}^0, \alpha^0] < \mathbb{E}[R_{o_{j+1}^0} | \boldsymbol{\rho}^0, \alpha^0]$
2. $\forall j \in [m+2, 2m], \mathbb{E}[R_{o_{j-1}^0} | \boldsymbol{\rho}^0, \alpha^0] < \mathbb{E}[R_{o_j^0} | \boldsymbol{\rho}^0, \alpha^0] < j$

Note that for both cases, it satisfies that $\forall 1 \leq j < k \leq n$ and $\forall \alpha > 0$, $\mathbb{E}[R_{o_j^0} | \boldsymbol{\rho}^0, \alpha^0] < \mathbb{E}[R_{o_k^0} | \boldsymbol{\rho}^0, \alpha^0]$

Lemma 2.1.2 *As $N \rightarrow \infty$, $\frac{1}{N} \sum_{j=1}^N R_i^j \rightarrow \mathbb{E}[R_i | \boldsymbol{\rho}^0, \alpha^0]$, $\forall i = 1, \dots, n$*

Definition 1 *Given a vector of length n , i.e. $\{x_1, \dots, x_n\}$, the function $\text{rank}(x_1, \dots, x_n)$ is defined as $\text{rank}(x_1, \dots, x_n) = \{r_1, \dots, r_n\}$ such that $x_{(r_k)} = x_k$, $\forall k = 1, \dots, n$*

Theorem 2.1.3 *As $N \rightarrow \infty$, and $\forall \alpha > 0$,*

$$\text{rank}(\frac{1}{N} \sum_{j=0}^N R_1^j, \dots, \frac{1}{N} \sum_{j=0}^N R_n^j) \rightarrow \text{rank}(\mathbb{E}[R_1 | \boldsymbol{\rho}^0, \alpha_0], \dots, \mathbb{E}[R_n | \boldsymbol{\rho}^0, \alpha_0]) = \boldsymbol{\rho}^0$$

To rephrase, as N approaches infinity, the Mallows consensus parameter $\boldsymbol{\rho}^0$ can be inferred from the data by taking the marginal mean for each item and then apply the rank function

to these marginal means.

2.2

Theorem 2.2.1 For a function g defined on \mathcal{P}_n which can depend on $\boldsymbol{\rho}^0$, for any n ,

$$\begin{aligned} & \arg \min_{g \in \mathcal{D}_{\boldsymbol{\rho}^0}} \lim_{N \rightarrow \infty} KL(P(\boldsymbol{\rho}|\alpha^0, \mathbf{R}^1, \dots, \mathbf{R}^N) || \sum_{\{i_1, \dots, i_n\} \in \mathcal{P}_n} q(\tilde{\boldsymbol{\rho}}|\alpha^0, \mathbf{R}^1, \dots, \mathbf{R}^N, i_1, \dots, i_n) g(i_1, \dots, i_n | \boldsymbol{\rho}^0)) \\ & = g^*(i_1, \dots, i_n | \mathcal{V}_{\boldsymbol{\rho}^0}), \\ & \text{where} \end{aligned}$$

- $\mathcal{D}_{\boldsymbol{\rho}^0}$ is the set of all distributions on \mathcal{P}_n , which can depend on $\boldsymbol{\rho}^0$
- $g^*(i_1, \dots, i_n | \mathcal{V}_{\boldsymbol{\rho}^0})$ is a distribution whose density is concentrated on $\boldsymbol{\rho}^0$, defined as
$$\begin{cases} g^*(i_1, \dots, i_n | \mathcal{V}_{\boldsymbol{\rho}^0}) = |\mathcal{V}_{\boldsymbol{\rho}^0}|^{-1} > 0, & \text{if } \{i_1, \dots, i_n\} \in \mathcal{V}_{\boldsymbol{\rho}^0} \\ g^*(i_1, \dots, i_n | \mathcal{V}_{\boldsymbol{\rho}^0}) = 0, & \text{if } \{i_1, \dots, i_n\} \notin \mathcal{V}_{\boldsymbol{\rho}^0} \end{cases}, \text{ where } |\mathcal{V}_{\boldsymbol{\rho}^0}| = \begin{cases} 2^{m-1}, & \text{if } n \text{ is odd} \\ 2^m, & \text{otherwise} \end{cases}$$

That is to say, for a set of distributions g , which are defined on the space of permutation of n items \mathcal{P}_n , as the number of users $N \rightarrow \infty$, the distribution g^* that minimizes the KL-divergence between the Mallows posterior and the pseudolikelihood defined above, is a uniform distribution with its density concentrated on $\mathcal{V}_{\boldsymbol{\rho}^0}$

2.3

For a given $N < \infty$, define $\hat{\boldsymbol{\rho}}^0$ as $\text{rank}(\frac{1}{N} \sum_{j=0}^N R_1^j, \dots, \frac{1}{N} \sum_{j=0}^N R_n^j)$ and $\mathcal{V}_{\hat{\boldsymbol{\rho}}^0} = f_v(\hat{\boldsymbol{\rho}}^0)$.

Theorem 2.3.1 $\exists \sigma \geq 0$ and $g'(i_1, \dots, i_n | \mathcal{V}_{\hat{\boldsymbol{\rho}}^0}, \sigma)$ such that

$$\begin{aligned} & KL(P(\boldsymbol{\rho}|\alpha^0, \mathbf{R}^1, \dots, \mathbf{R}^N) || \sum_{\{i_1, \dots, i_n\} \in \mathcal{P}_n} q(\tilde{\boldsymbol{\rho}}|\alpha^0, \mathbf{R}^1, \dots, \mathbf{R}^N, i_1, \dots, i_n) g^*(i_1, \dots, i_n | \mathcal{V}_{\hat{\boldsymbol{\rho}}^0})) \geq \\ & KL(P(\boldsymbol{\rho}|\alpha^0, \mathbf{R}^1, \dots, \mathbf{R}^N) || \sum_{\{i_1, \dots, i_n\} \in \mathcal{P}_n} q(\tilde{\boldsymbol{\rho}}|\alpha^0, \mathbf{R}^1, \dots, \mathbf{R}^N, i_1, \dots, i_n) g'(i_1, \dots, i_n | \mathcal{V}_{\hat{\boldsymbol{\rho}}^0}, \sigma)) \end{aligned}$$

where $g'(i_1, \dots, i_n | \mathcal{V}_{\hat{\boldsymbol{\rho}}^0}, \sigma) = \sum_{\hat{\mathbf{v}} \in \mathcal{V}_{\hat{\boldsymbol{\rho}}^0}} \{g^*(\hat{\mathbf{v}} | \mathcal{V}_{\hat{\boldsymbol{\rho}}^0}) \int_{\mathbf{x}} \mathcal{F}_r(i_1, \dots, i_n | x_1, \dots, x_n) \prod_{i=1}^n \mathcal{N}(x_i | \hat{v}_i, \sigma) d\mathbf{x}\}$, and

- $\hat{\mathbf{v}} \sim g^*(\hat{\mathbf{v}} | \mathcal{V}_{\hat{\boldsymbol{\rho}}^0})$
- $x_i \sim \mathcal{N}(x_i | \hat{v}_i, \sigma)$ for $i = 1, \dots, n$
- $i_1, \dots, i_n \sim \mathcal{F}_r(i_1, \dots, i_n | x_1, \dots, x_n)$, where $\mathcal{F}_r = \begin{cases} 1, & \text{if } \{i_1, \dots, i_n\} = \text{rank}(x_1, \dots, x_n) \\ 0, & \text{otherwise} \end{cases}$

As N is limited, $\boldsymbol{\rho}^0$ and therefore, $\mathcal{V}_{\boldsymbol{\rho}^0}$ usually cannot be accurately inferred from the data. We can however, sample for i_1, \dots, i_n by sampling for each item i from a univariate Gaussian

distribution centered on \hat{v}_i with a fixed variance σ for all items, and then obtain a ranking using the rank function. By introducing the variance, a smaller KL divergence from the Mallows posterior can be achieved.

2.4

Theorem 2.4.1 *With the usage of $g'(i_1, \dots, i_n | \mathcal{V}_{\hat{\rho}^0}, \sigma)$, the value of σ that minimizes the KL-divergence between the Mallows posterior and the resulted pseudolikelihood is*

$$\sigma = \begin{cases} 0, & \text{if } \delta(\alpha^0, n, N) \leq \delta^* \\ f(\alpha^0, n, N), & \text{otherwise} \end{cases}$$

In other words, σ should be 0 when $\delta(\alpha^0, n, N) \geq \delta^*$. Beyond this point, the optimal choice of σ should be greater than 0, and it follows a function $f(\alpha^0, n, N)$.

2.5

Theorem 2.5.1 *As $N \rightarrow \infty, \sigma = 0 \forall \alpha > 0$ and $n \geq 1$*

3 Evidence and proofs

3.1 Evidence for Theorem 2.3.1

In **Figure 3.2.1**, for each subfigure, we calculate and plot the KL-divergences between the Mallows posterior and the pseudo-likelihood, computed with different choices of σ . The left-most point on each sub-figure corresponds to the KL-divergence when no Gaussian variation is introduced, i.e. $\sigma = 0$. It can be observed that for most situations shown in the figure, the lowest KL-divergence is achieved when some level of Gaussian variation is introduced, especially as N and α^0 are relatively small. However, as N and α^0 increase, the optimal σ appears to decrease towards 0.

3.2 Evidence for Theorem 2.4.1

3.2.1 Optimal σ is determined by N, n, α^0

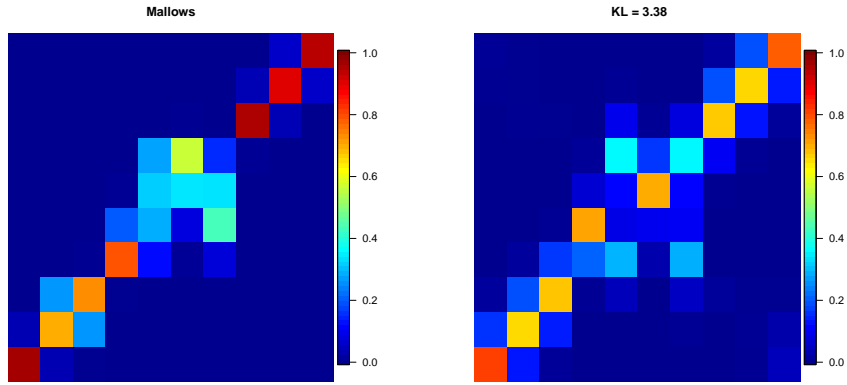
As shown in **Figure 3.2.1**, in each subfigure, the σ value that corresponds to the lowest KL-divergence is the optimal σ for its specific (N, n, α^0) set up. Each row of 3 figures shows a comparison of the optimal σ when one of the variables (N, n, α) changes. It can be observed that all three variables have an impact on the optimal value of σ . More specifically, the optimal choice of σ appear to decrease as N and α^0 increase, and as n decreases.

For each N, n and α^0 , we simulate 10 datasets, and for each dataset, a grid of σ values are tried out. In **Figure 3.2.1**, we plot α^0 on the x-axis, and its corresponded optimal

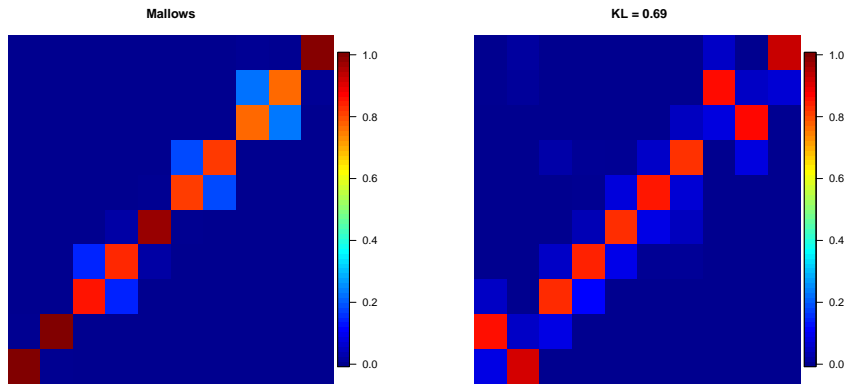
σ on the y-axis. It can be observed that when α^0 and N are large, and n is small, all σ values result in small KL-divergence. To simplify, in these situations, we can safely set $\sigma = 0$ to achieve small KL-divergence. As N and α^0 decreases and as n increases, optimal σ increases.

3.2.2 using data standard deviation / n as a proxy for α^0

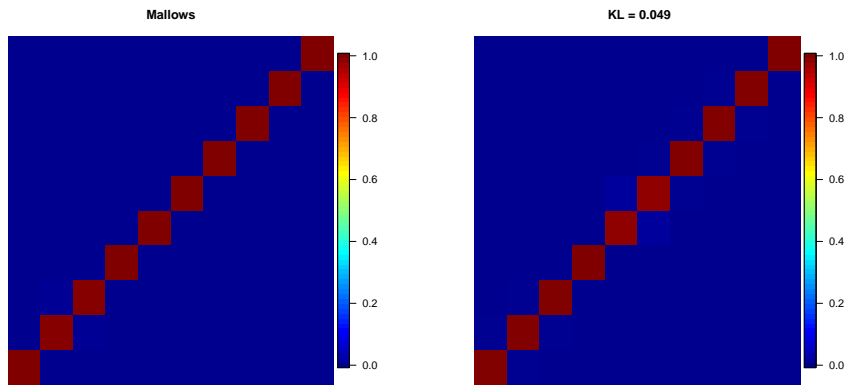
Under most situations, the value of α^0 is unknown, however, we can compute the marginal standard deviation of each item from the data, and normalize it by deviding it by n. The trend demonstrated in **Figure 3.2.1** is well-preserved, as shown in **Figure 3.2.2**.



(a)



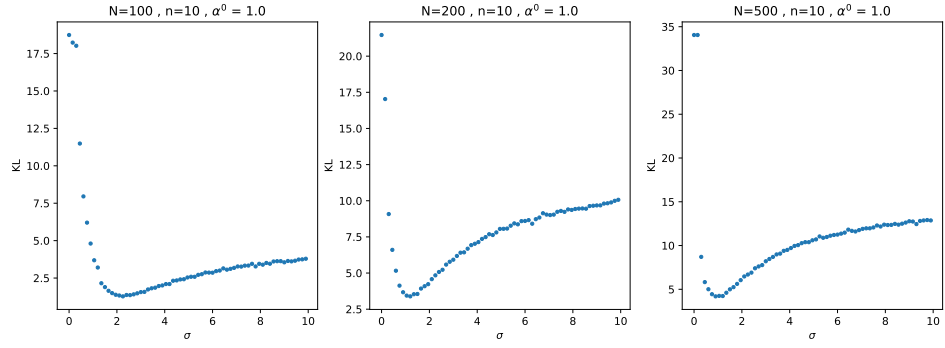
(b)



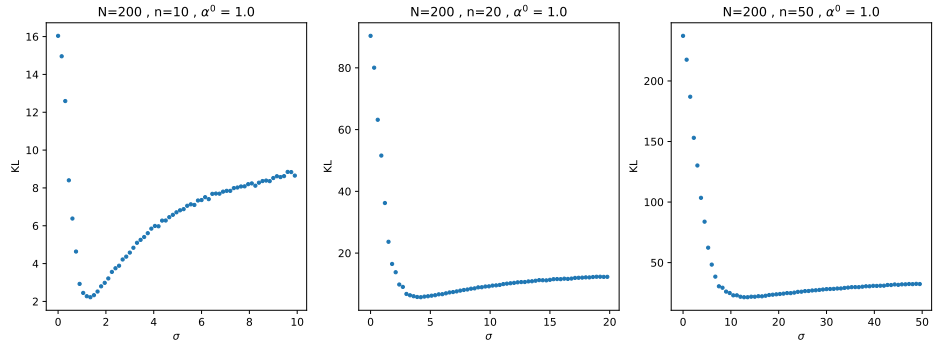
(c)

6

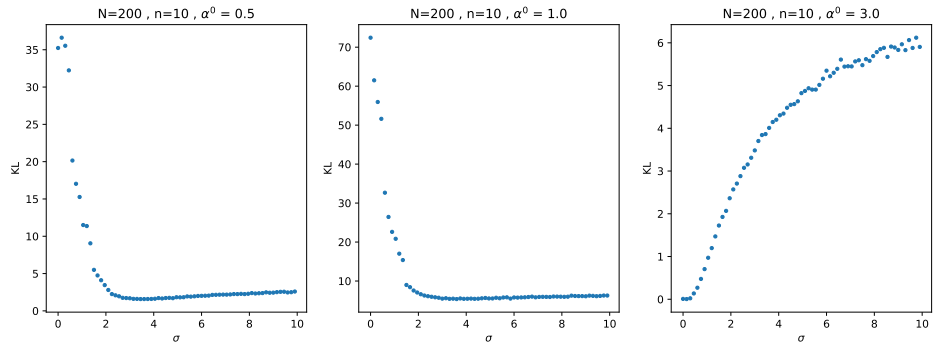
Figure 1: left column: Mallows, right column: pseudolikelihoods. $N = 200, n = 10$, $\alpha^0 = 1, 1.5, 2$ respectively, from top to bottom



(a)

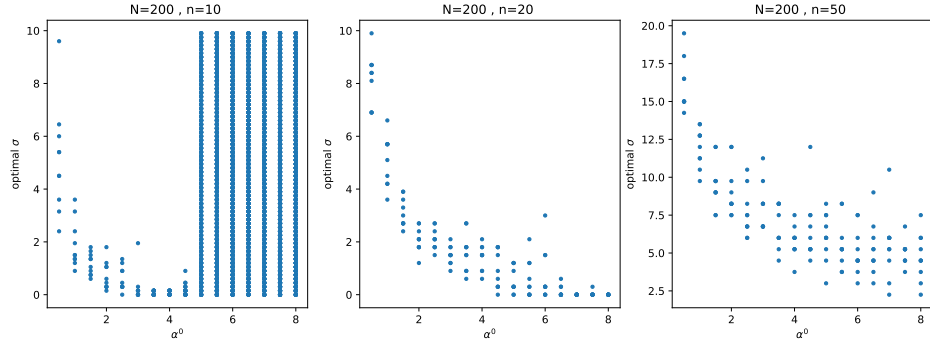


(b)

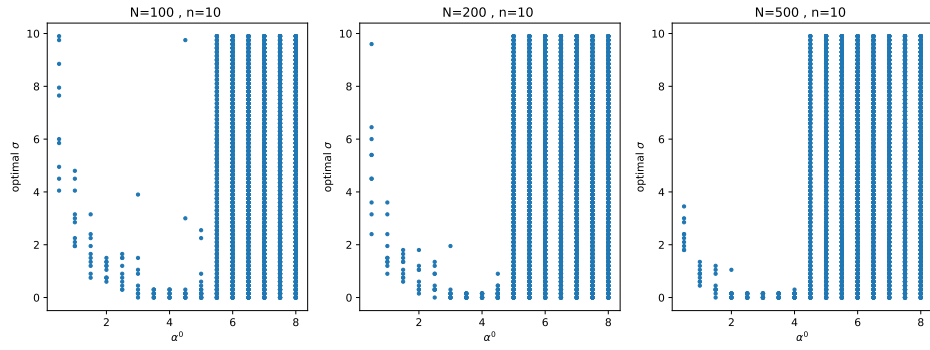


(c)

Figure 2: x-axis: σ , y-axis: KL-divergence



(a)



(b)

Figure 3: x-axis: α^0 , y-axis: optimal σ

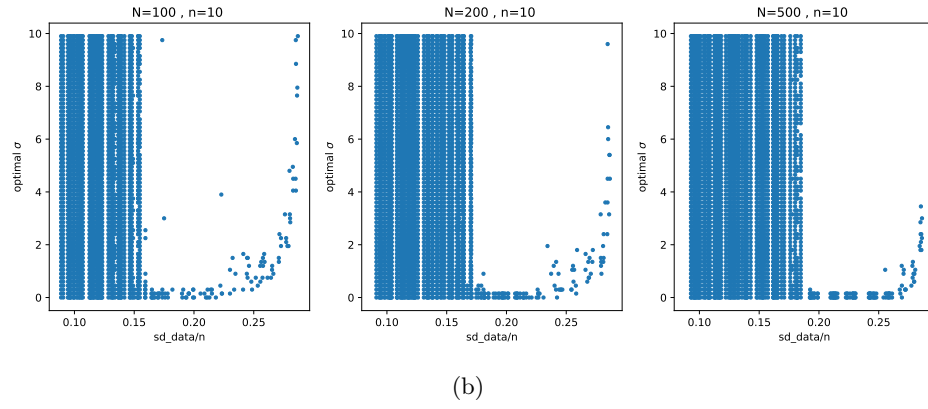
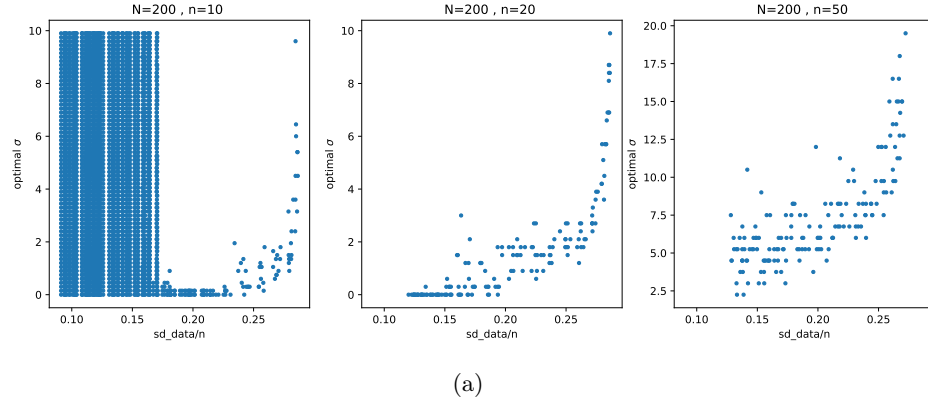


Figure 4: x-axis: α^0 , y-axis: optimal σ