Оцінювання характеристик частково спостережуваного ланцюга Маркова на двійкових послідовностях

А. В. Цибульник І. І. Ніщенко

Всеукраїнська науково-практична конференція студентів, аспірантів та молодих вчених

2023

План доповіді

План доповіді

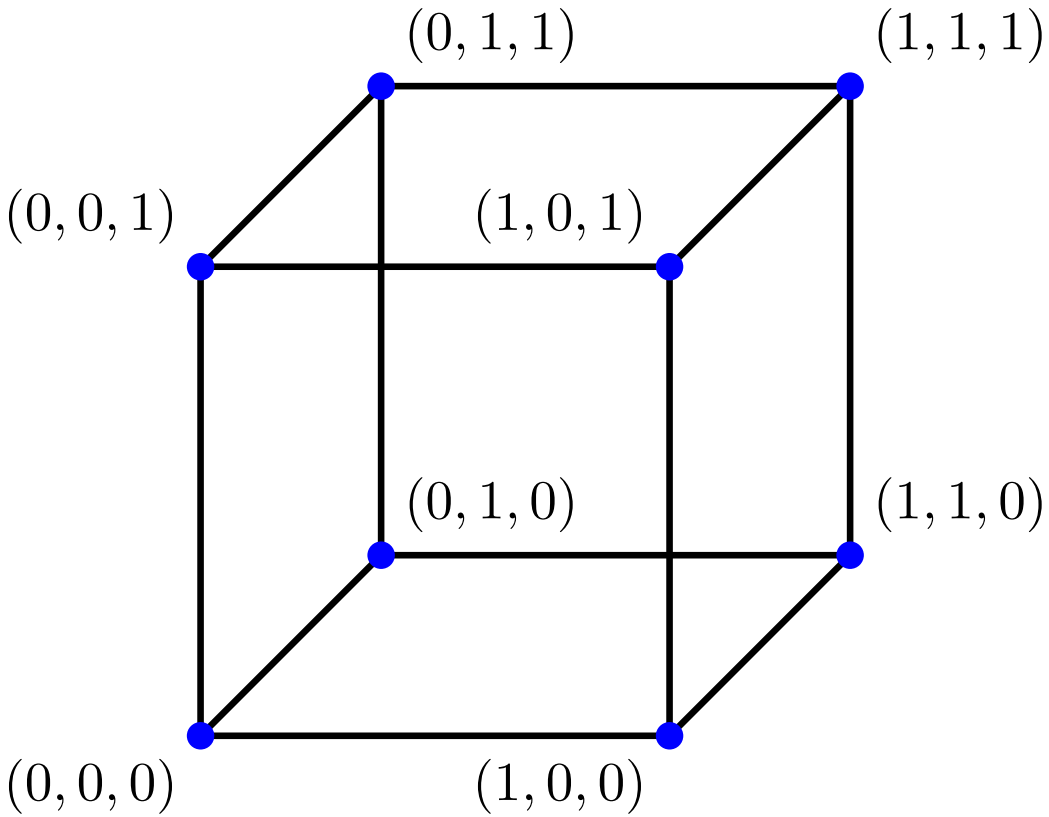
Сфери застосування бінарних послідовностей

- Еволюція ДНК в біології
- Бінарні послідовності в теорії інформації
- Спінові системи у фізиці

План доповіді

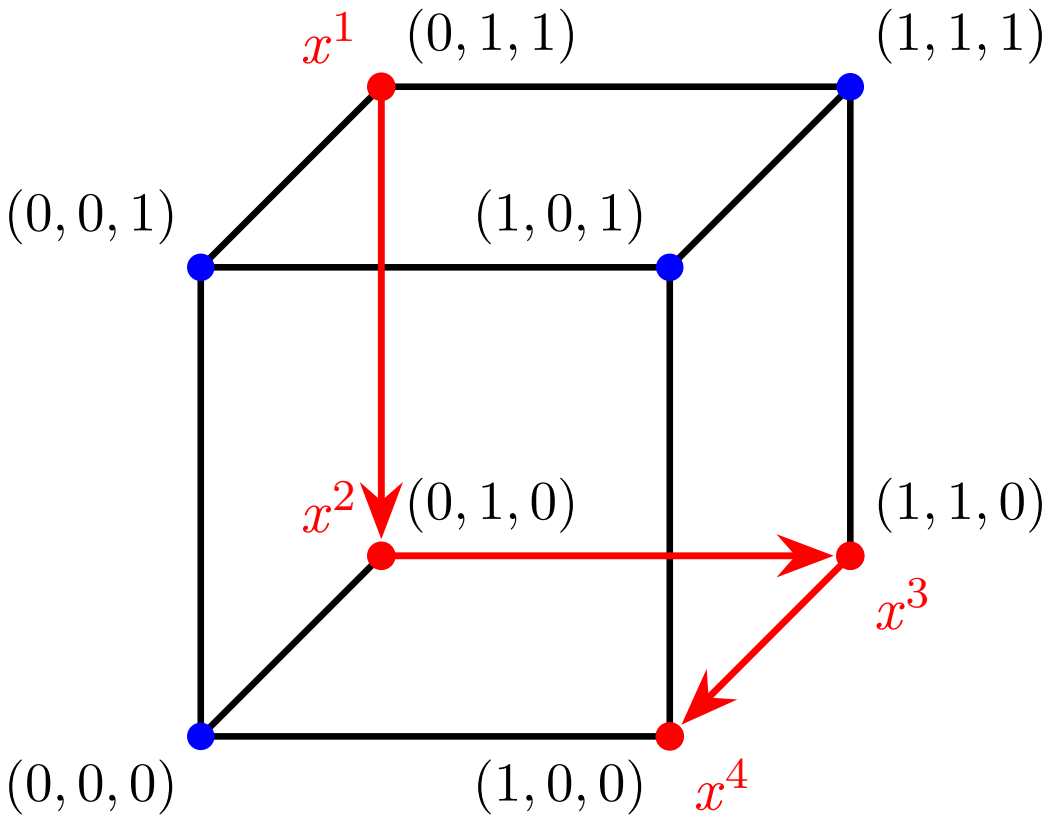
Моделювання об’єкту дослідження

Еволюцію бінарної послідовності довжини N уявимо як випадкове блукання вершинами N- вимірного куба.



Моделювання об’єкту дослідження

Ліниве блукання: з імовірністю p залишаємося на місці, з імовірністю $\frac{1-p}{N}$ переходимо в сусідню вершину.



	x^t
$t=1$	$(0,1,1)$
$t=2$	$(0,1,0)$
$t=3$	$(1,1,0)$
$t=4$	$(1,0,0)$

Моделювання об’єкту дослідження

$\{X^t\}_{t=\overline{1,T}}$ є ланцюгом Маркова зі станами в $E = \{0,1\}^N$ з початковим рівномірним розподілом π :

$$\pi_x = P\left(X^1 = x\right) = \frac{1}{2^N}$$

та матрицею перехідних імовірностей A :

$$A_{xx'} = P\left(X^{t+1} = x' \mid X^t = x\right) = \begin{cases} p, & d_H(x,x') = 0 \\ \frac{1-p}{N}, & d_H(x,x') = 1 \\ 0, & \text{інакше} \end{cases}$$

Моделювання об’єкту дослідження

Спостерігаємо набір функціоналів

$$Y^t = (Y_1^t, \ldots, Y_L^t) = (\sum_{i \in I_1} X_i^t, \ldots, \sum_{i \in I_L} X_i^t),$$

де $I_1, \ldots, I_L$ є заданими підмножинами $\{1, 2, \ldots, N\}$.

Моделювання об'єкту дослідження

Спостерігаємо набір функціоналів

$$Y^t = (Y_1^t, \ldots, Y_L^t) = (\sum_{i \in I_1} X_i^t, \ldots, \sum_{i \in I_L} X_i^t),$$

де $I_1, \ldots, I_L$ є заданими підмножинами $\{1, 2, \ldots, N\}$.

Наприклад, $N = 12$, $I_1 = (1, 2, 3)$, $I_2 = (6, 7, 10, 11, 12)$

	x^t	y^t
$t = 1$	010011101101	$(1, 4)$
$t = 2$	011011101101	$(2, 4)$
$t = 3$	011011111101	$(2, 4)$
$t = 4$	011011111111	$(2, 5)$

Моделювання об’єкту дослідження

Твердження

Послідовність $\{(X^t, Y^t)\}_{t=\overline{1,T}}$ *утворює приховану марковську модель* (π, A, B)*, де*

$$B_{xy} = P\left(Y^t = y \,|\, X^t = x\right) = \prod_{k=1}^{L} \mathbb{1}\left(y_k = \sum_{i \in I_k} x_i\right)$$

План доповіді

Постановка задачі

1. Оцінити параметр p за набором спостережень та декодувати послідовність станів прихованого ланцюга;

Постановка задачі

1. Оцінити параметр p за набором спостережень та декодувати послідовність станів прихованого ланцюга;

Метод максимальної правдоподібності

$$P\left(Y = y \,|\, p\right) = \sum_{x \in E^T} P\left(X = x,\, Y = y \,|\, p\right) \longrightarrow \max$$

Функція повної правдоподібності

$$L_{p,x,y} = P\left(X = x,\, Y = y \,|\, p\right)$$

Відтак

$$\widehat{p} = \operatorname*{argmax}_{p} \sum_{x \in E^T} L_{p,x,y}$$

Побудова оцінок невідомих параметрів

Ітераційний алгоритм Баума-Велша:

$$Q\left(p^{(n)}, p\right) = \sum_{x \in E^T} L_{p^{(n)},x,y} \cdot \ln L_{p,x,y} \longrightarrow \max$$

Тож починаючи з деякого $p^{(0)}$

$$p^{(n+1)} = \operatorname*{argmax}_p Q\left(p^{(n)}, p\right)$$

Побудова оцінок невідомих параметрів

Формула переоцінки параметра p :

$$p^{(n+1)} = p^{(n)} \cdot \frac{\sum\limits_{t=1}^{T-1} \sum\limits_{x \in E} \alpha_t(x)\, B_{xy^{t+1}}\, \beta_{t+1}(x)}{\sum\limits_{t=1}^{T-1} \sum\limits_{x \in E} \alpha_t(x)\, \beta_t(x)},$$

де

$$\alpha_t(x) = P\left(Y^1 = y^1, \ldots, Y^t = y^t,\, X^t = x \,|\, p^{(n)}\right)$$
$$\beta_t(x) = P\left(Y^{t+1} = y^{t+1}, \ldots, Y^T = y^T \,|\, X^t = x,\, p^{(n)}\right)$$

Побудова оцінок невідомих параметрів

Алгоритм Вітербі: пошук такої послідовності прихованих станів $\widehat{X}^1, \widehat{X}^2, \ldots, \widehat{X}^T$, яка найкращим чином описує наявні спостереження:

$$\widehat{X} = \underset{x \in E^T}{\operatorname{argmax}} P\left(X = x \,\middle|\, Y = y, \widehat{p}\right)$$

Постановка задачі

2. Спостерігаємо значення $Y^t_{I_*} = \sum\limits_{i \in I_*} X^t_i$, де I_* — деяка невідома підмножина множини індексів.

Постановка задачі

2. Спостерігаємо значення $Y^t_{I_*} = \sum\limits_{i \in I_*} X^t_i$, де I_* — деяка невідома підмножина множини індексів.

Наприклад, $N = 12$, $I_1 = (1, 2, 3)$, $I_2 = (6, 7, 10, 11, 12)$, $I_* = ?$

	x^t	y^t	$y^t_{I_*}$
$t = 1$	010011101101	$(1, 4)$	3
$t = 2$	011011101101	$(2, 4)$	3
$t = 3$	011011111101	$(2, 4)$	4
$t = 4$	011011111111	$(2, 5)$	4

Постановка задачі

2. Спостерігаємо значення $Y^t_{I_*} = \sum\limits_{i \in I_*} X^t_i$, де I_* — деяка невідома підмножина множини індексів.

Наприклад, $N = 12$, $I_1 = (1, 2, 3)$, $I_2 = (6, 7, 10, 11, 12)$, $I_* = ?$

	x^t	y^t	$y^t_{I_*}$
$t = 1$	010011101101	$(1, 4)$	3
$t = 2$	011011101101	$(2, 4)$	3
$t = 3$	011011111101	$(2, 4)$	4
$t = 4$	011011111111	$(2, 5)$	4

Яким чином можна відтворити елементи множини I_* за спостереженнями $Y^1_{I_*}, \ldots, Y^T_{I_*}$?;

Побудова оцінок невідомих параметрів

Твердження

Змістовною і незміщеною оцінкою потужності множини I_* є статистика

$$\widehat{|I_*|} = \frac{N}{1-p}\left(1 - \frac{1}{T-1}\sum_{t=1}^{T-1} \mathbb{1}\left(Y_{I_*}^{t} = Y_{I_*}^{t+1}\right)\right)$$

Побудова оцінок невідомих параметрів

Визначення компонент множини I_* :

$$\widehat{I} = \underset{1 \leqslant k \leqslant C_N^{\widehat{|I_*|}}}{\operatorname{argmin}} \; d\left(\widehat{Y}_{\mathtt{I_k}}, Y_{I_*}\right),$$

тут

$$\widehat{Y}_{\mathtt{I_k}} = \sum_{i \in \mathtt{I_k}} \widehat{X}_i^t$$

є сумою від декодованих елементів прихованого ланцюга.

Побудова оцінок невідомих параметрів

Визначення компонент множини I_* :

$$\widehat{I} = \underset{1 \leqslant k \leqslant C_N^{\widehat{|I_*|}}}{\operatorname{argmin}} \; d\left(\widehat{Y}_{\mathtt{I_k}}, Y_{I_*}\right),$$

тут

$$\widehat{Y}_{\mathtt{I_k}} = \sum_{i \in \mathtt{I_k}} \widehat{X}_i^t$$

є сумою від декодованих елементів прихованого ланцюга.

Що обрати в ролі міри близькості d?

Побудова оцінок невідомих параметрів

Середньоквадратична відстань

$$d_S\left(\widehat{Y}_{\mathtt{I_k}}, Y_{I_*}\right) = \sum_{t=1}^{T}\left(\widehat{Y}^{t}_{\mathtt{I_k}} - Y^{t}_{I_*}\right)^2$$

Зважена відстань Жаккара

$$d_J\left(\widehat{Y}_{\mathtt{I_k}}, Y_{I_*}\right) = 1 - \frac{\sum_{t=1}^{T}\min\left(\widehat{Y}^{t}_{\mathtt{I_k}}, Y^{t}_{I_*}\right)}{\sum_{t=1}^{T}\max\left(\widehat{Y}^{t}_{\mathtt{I_k}}, Y^{t}_{I_*}\right)}$$

Постановка задачі

3. Спостереження на множинах $I_1, \ldots, I_L$ спотворюються ймовірностями $q_1, \ldots, q_L$:

$$Y^t = \left(Y_k^t\right)_{k=\overline{1,L}} = \left(\sum_{i \in I_k} \widetilde{X}_i^t\right)_{k=\overline{1,L}}$$

де для $i \in I_k$

$$\widetilde{X}_i^t = \begin{cases} 1 - X_i^t, & \text{з імовірністю } q_k \\ X_i^t, & \text{з імовірністю } 1 - q_k \end{cases}$$

Оцінити невідомий параметр моделі p та ймовірності спотворень $q_1, q_2, \ldots, q_L$.

Побудова оцінок невідомих параметрів

Наприклад, $N = 12$, $I_1 = (1, 2, 3)$, $I_2 = (6, 7, 10, 11, 12)$

	x^t	$\widetilde{x}^t$	y^t	q
$t = 1$	010011101101	$0\mathbf{0}0011101101$	$(0, 4)$	(q_1, q_2)
$t = 2$	011011101101	$01\mathbf{0}011101101$	$(1, 4)$	(q_1, q_2)
$t = 3$	011011111101	$\mathbf{1}11011111\mathbf{1}1$	$(3, 5)$	(q_1, q_2)
$t = 4$	011011111111	$011011111\mathbf{0}\mathbf{0}$	$(2, 3)$	(q_1, q_2)

Побудова оцінок невідомих параметрів

Твердження

Якщо множини $I_1, \ldots, I_L$ є попарно неперетинними, то утворена послідовність $\{(X^t, Y^t)\}_{t=\overline{1,T}}$ є прихованою марковською моделлю (π, A, B^q), де

$$B^q_{xy} = P\left(Y^t = y \,|\, X^t = x\right) = \prod_{k=1}^{L} P\left(\xi^k_{01}(x) + \xi^k_{11}(x) = y_k\right),$$

$$\xi^k_{01}(x) \sim Bin\left(|I_k| - \sum_{i \in I_k} x_i,\ q_k\right),\ \xi^k_{11}(x) \sim Bin\left(\sum_{i \in I_k} x_i,\ 1 - q_k\right),\ k = \overline{1, L}$$

Побудова оцінок невідомих параметрів

Починаючи з деякого наближення моделі $\left(\pi, A^{(0)}, B^{q^{(0)}}\right)$, формула переоцінки параметра p :

$$p^{(n+1)} = p^{(n)} \cdot \frac{\sum\limits_{t=1}^{T-1} \sum\limits_{x \in E} \alpha_t(x)\, B^{q^{(n)}}_{xy^{t+1}}\, \beta_{t+1}(x)}{\sum\limits_{t=1}^{T-1} \sum\limits_{x \in E} \alpha_t(x)\, \beta_t(x)}$$

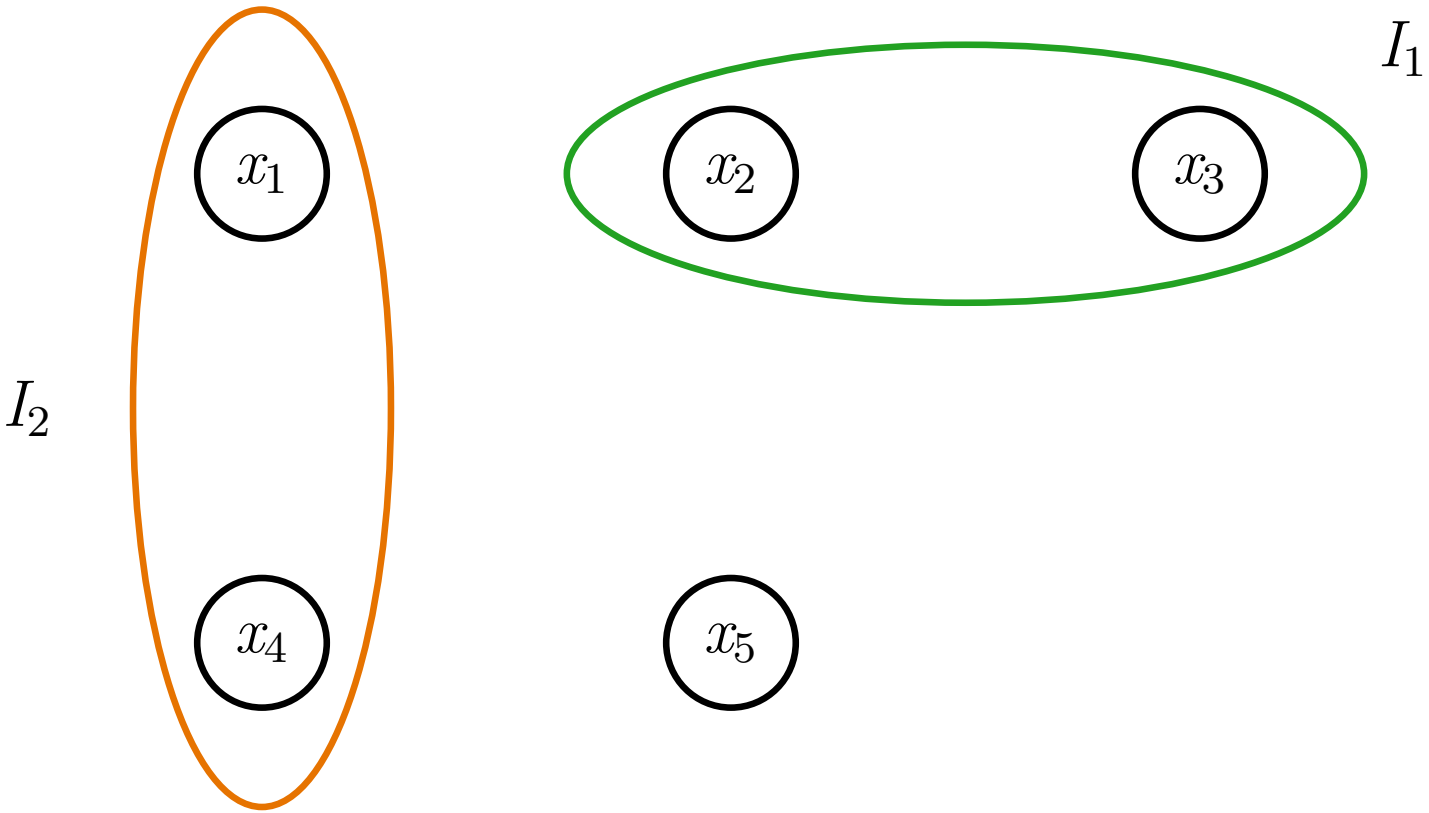
Формула переоцінки компонент вектора q :

$$q_k^{(n+1)} = q_k^{(n)} \cdot \frac{\sum\limits_{t=1}^{T} \sum\limits_{x \in E} \beta_t(x) \sum\limits_{x' \in E} \alpha_{t-1}(x')\, A^{(n)}_{x'x} \sum\limits_{i \in I_k} P^{q^{(n)}}_{x,i}}{|I_k| \sum\limits_{t=1}^{T} \sum\limits_{x \in E} \alpha_t(x)\, \beta_t(x)}$$

План доповіді

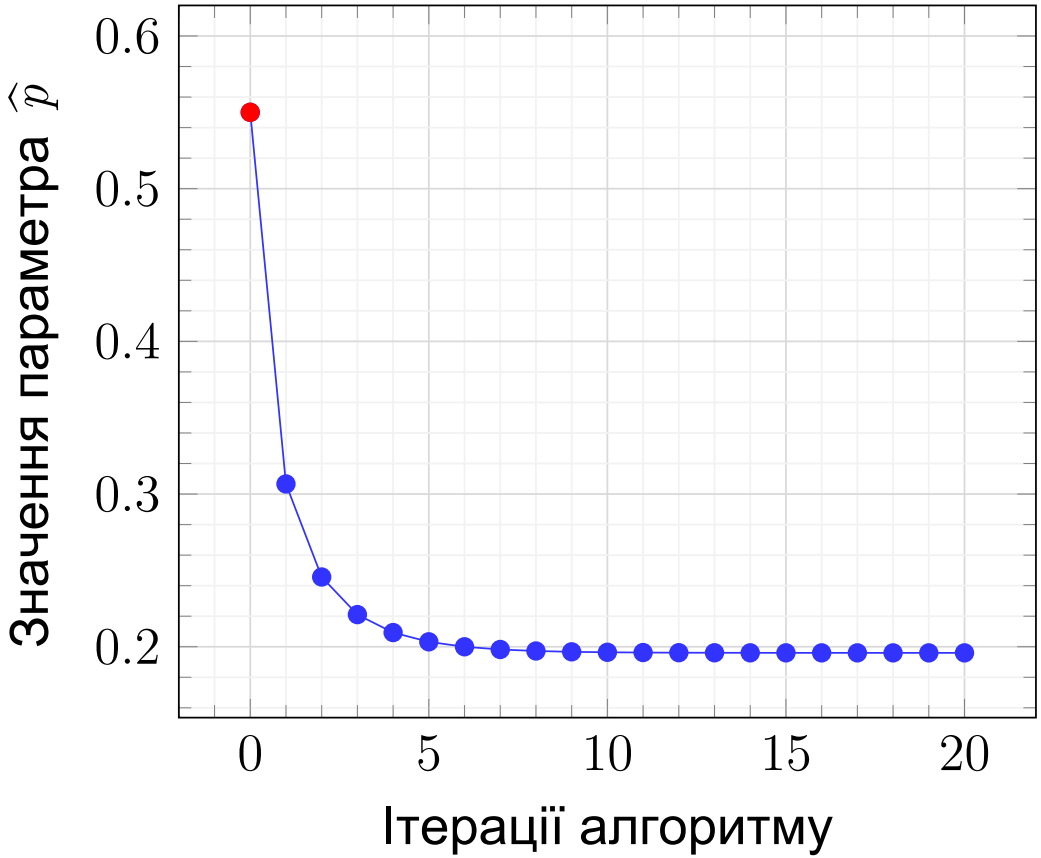
Результати чисельного експерименту

Було згенеровано прихований ланцюг Маркова протягом $T = 200$ моментів часу, $N = 5$ та $p = 0.2$. Множина спостережуваних індексів: $I = \{I_1, I_2\} = \{(2, 3), (1, 4)\}$.



Результати чисельного експерименту

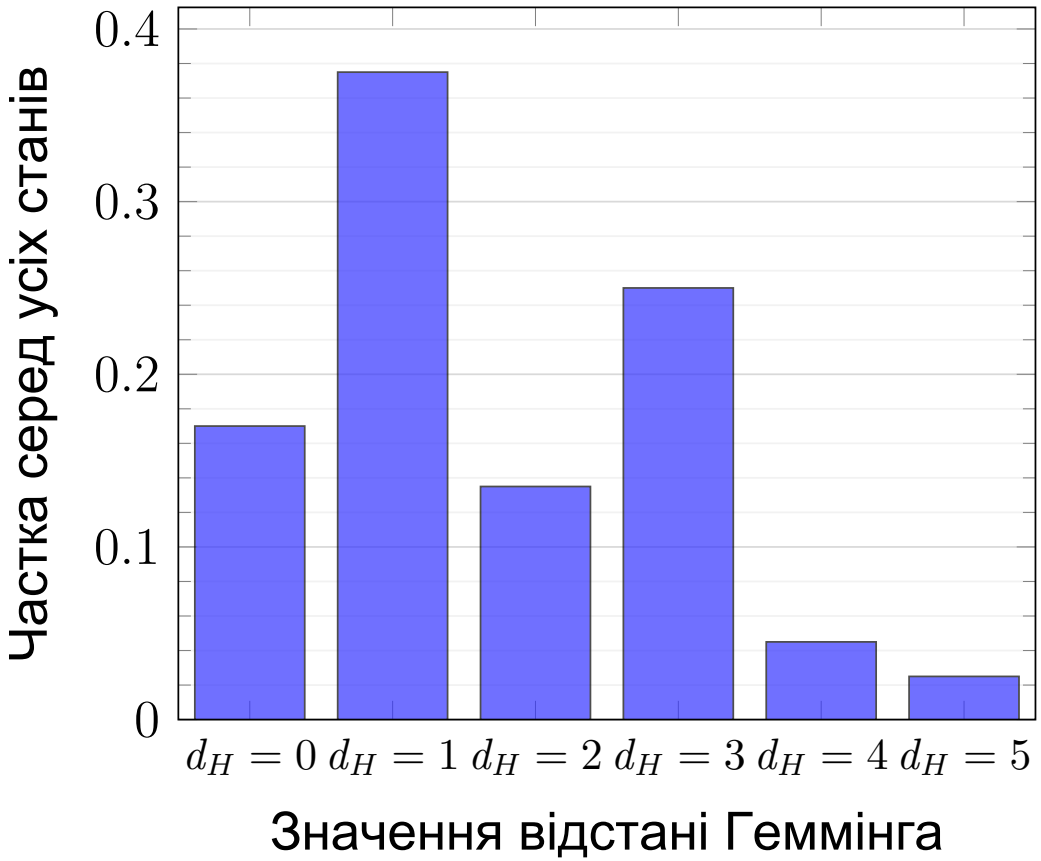
1. Оцінити параметр p за набором спостережень та декодувати послідовність станів прихованого ланцюга;



p	$\widehat{p}$	$\lvert p - \widehat{p} \rvert$
0.2	0.1959	0.0041

Результати чисельного експерименту

1. Оцінити параметр p за набором спостережень та декодувати послідовність станів прихованого ланцюга;

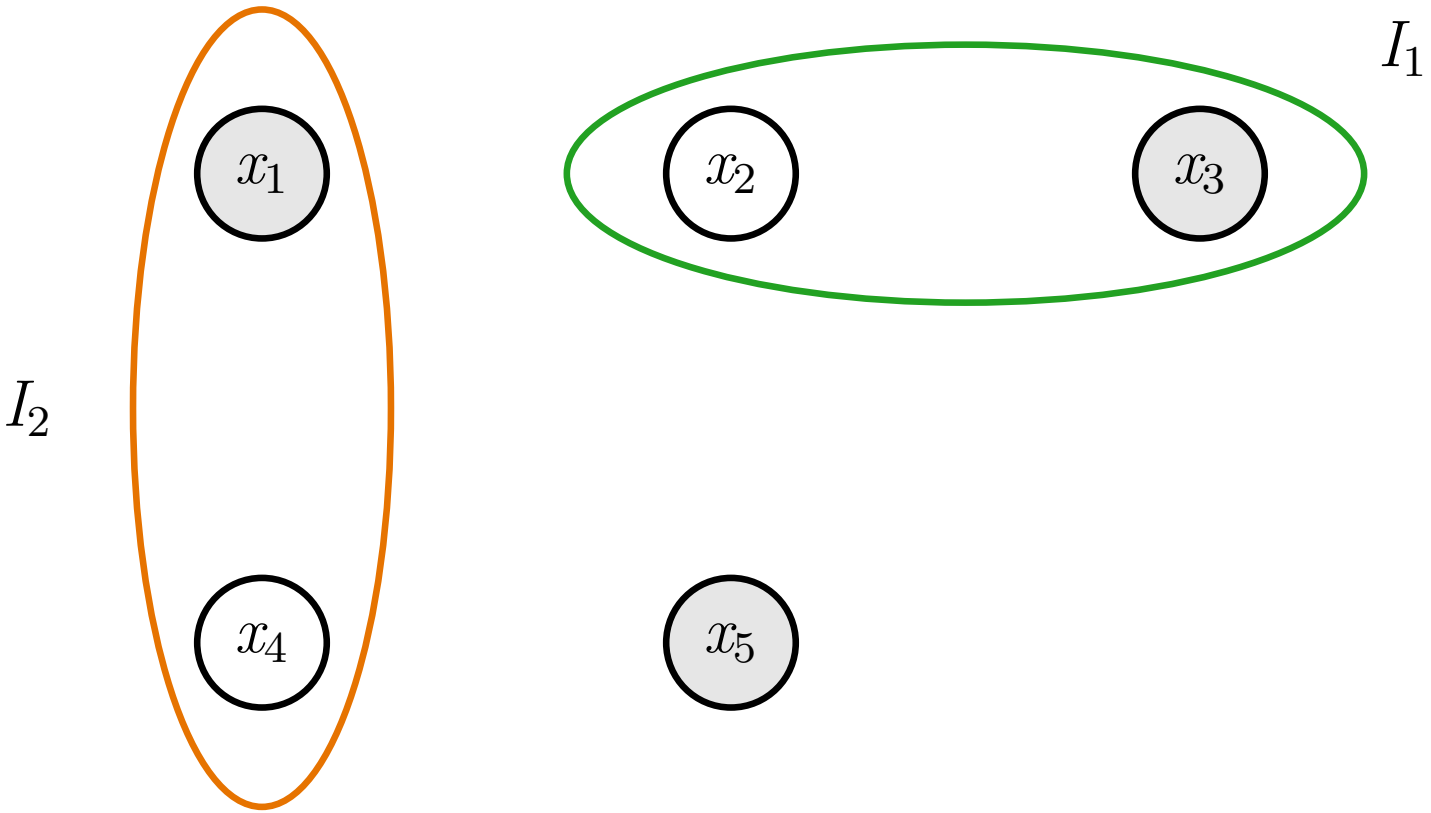


$$d_H\left(X^t, \widehat{X^t}\right) = \sum_{i=1}^{N} \mathbb{1}\left(X_i^t \neq \widehat{X_i^t}\right)$$

Результати чисельного експерименту

2. Відтворити елементи «множини неявних індексів» I_*;

Спостереження $I = \{I_1, I_2\} = \{(2,3),(1,4)\}$, неявні індекси покладемо $I_* = (1,3,5)$:



Результати чисельного експерименту

2. Відтворити елементи «множини неявних індексів» I_*;

Залежність значення оцінки від довжини ланцюга

T	200	400	600	800	1000
$\widehat{p}$	0.1959	0.1823	0.1882	0.2099	0.2092
$\widehat{\lvert I_* \rvert}$	2	2	2	3	3

Результати чисельного експерименту

2. Відтворити елементи «множини неявних індексів» I_*;

Залежність значення оцінки від довжини ланцюга

T	200	400	600	800	1000
$\widehat{p}$	0.1959	0.1823	0.1882	0.2099	0.2092
$\widehat{\lvert I_* \rvert}$	2	2	2	3	3

При малих N можна використати оцінку

$$\widehat{|I_*|}_{\max} = \max_{1 \leqslant t \leqslant T} Y_{I_*}^t$$

Результати чисельного експерименту

2. Відтворити елементи «множини неявних індексів» I_*;

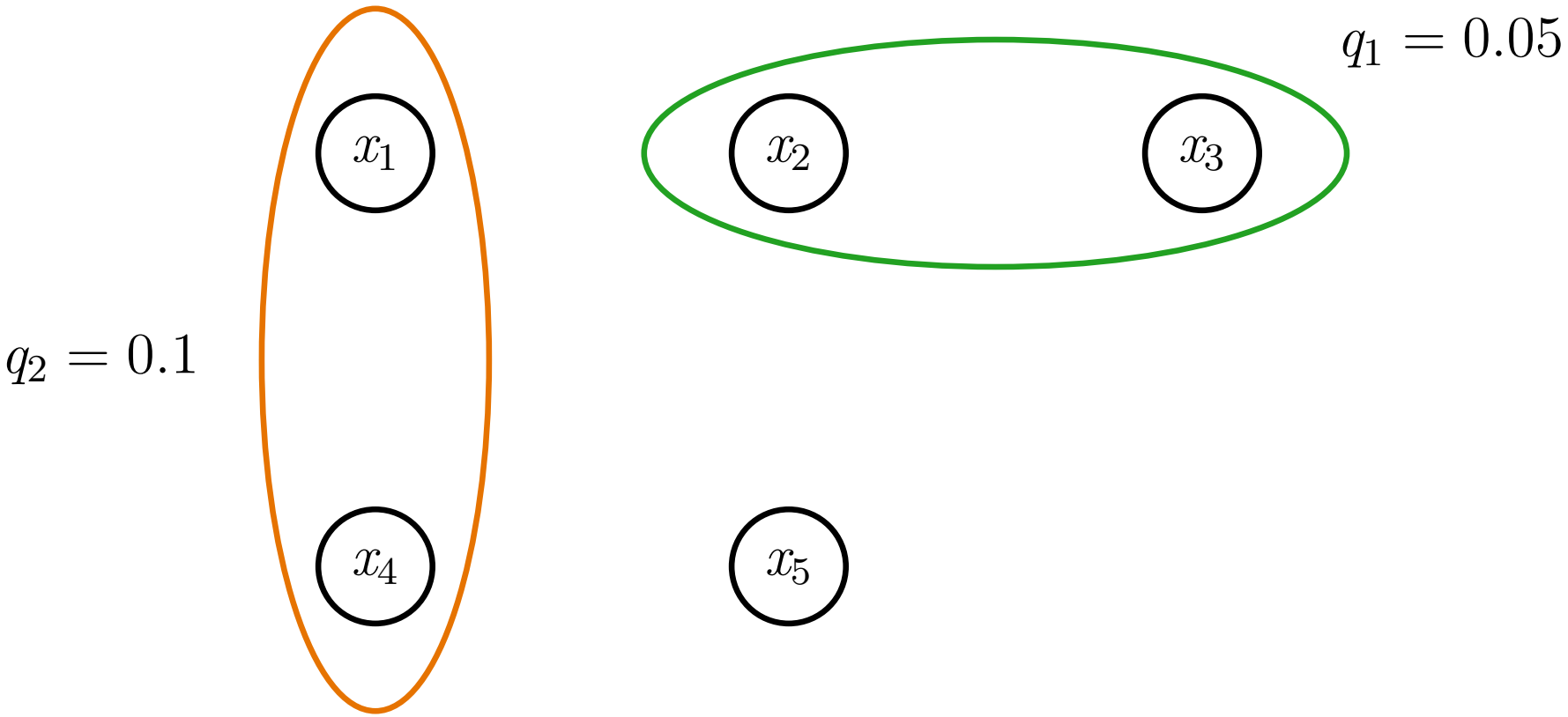
Отримані результати:

Істинна множина I_*	$(1, 3, 5)$
Оцінка $\widehat{I}_S$ за середньоквадратичною відстанню	$(1, 2, 5)$
Оцінка $\widehat{I}_J$ за зваженою відстанню Жаккара	$(1, 2, 3)$

Результати чисельного експерименту

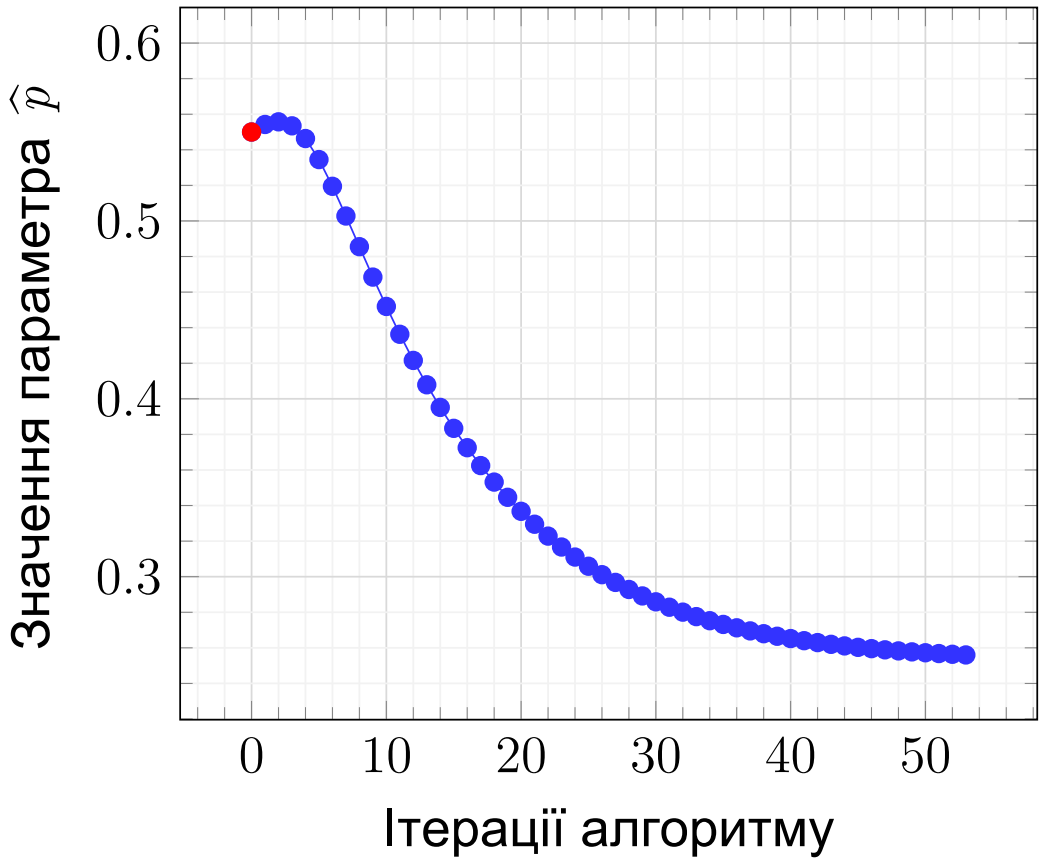
3. Оцінити невідомий параметр моделі p при ймовірностях спотворення $q_1, q_2, \ldots, q_L$.

Для I_1, I_2 було задано такі коефіцієнти спотворення:



Результати чисельного експерименту

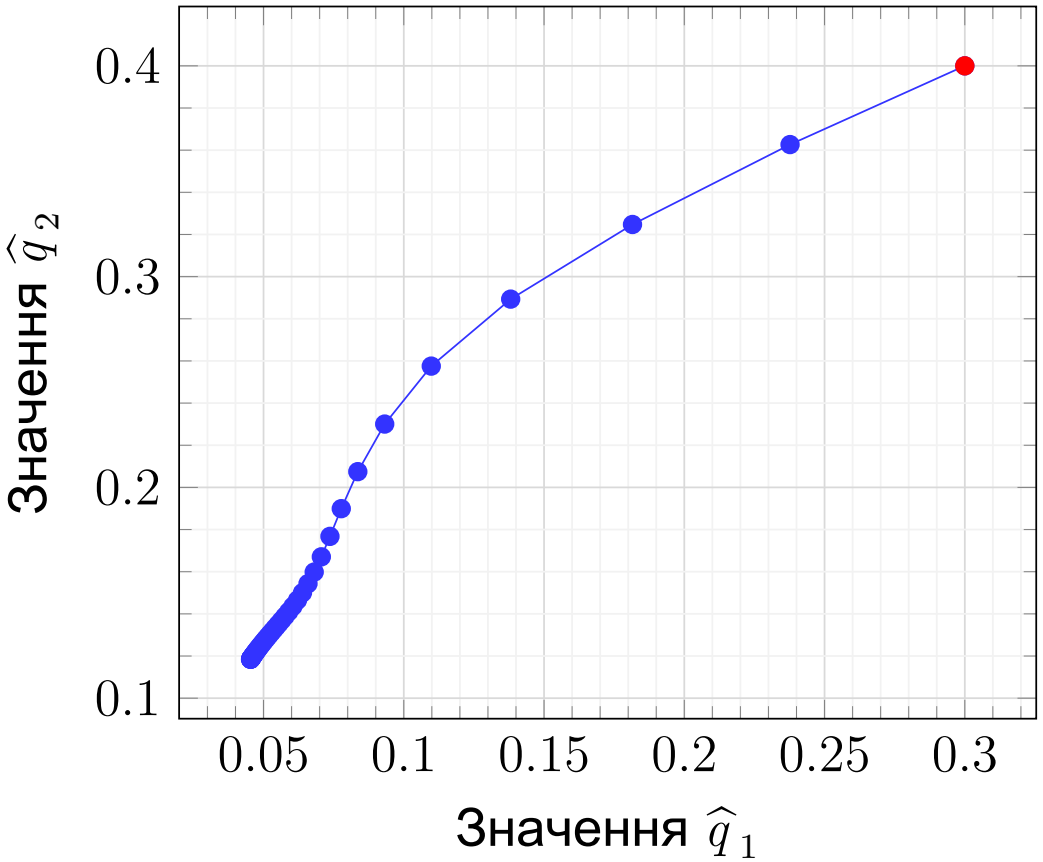
3. Оцінити невідомий параметр моделі p при ймовірностях спотворення $q_1, q_2, \ldots, q_L$.



p	$\widehat{p}$	$\lvert p - \widehat{p} \rvert$
0.2	0.2559	0.0559

Результати чисельного експерименту

3. Оцінити невідомий параметр моделі p при ймовірностях спотворення $q_1, q_2, \ldots, q_L$.



q_1	$\widehat{q}_1$	$\lvert q_1 - \widehat{q}_1 \rvert$
0.05	0.0454	0.0046

q_2	$\widehat{q}_2$	$\lvert q_2 - \widehat{q}_2 \rvert$
0.1	0.1184	0.0184

Висновки

Невідомі параметри заданої моделі були оцінені

- або шляхом побудови змістовних та незміщених статистичних оцінок;
- або за допомогою ітераційного алгоритму Баума-Велша.

Результати чисельного експерименту продемонстрували ефективність використаних методів, зокрема збіжність побудованих оцінок до істинних значень параметрів при збільшенні кількості спостережень.