Димитров Д.В. (Москва, Россия) — **Статистические методы определения неоднородностей волокнистых материалов с помощью статистик ближайших соседей.**

В работе рассматривается применение оценок k-ближайших соседей дивергенции Кульбака–Лейблера [5] к поиску областей неоднородности в материалах. Асимптотические свойства таких оценок (а именно: асимптотическая несмещенность и L^2 - состоятельность) были доказаны в недавней работе [2]. В статьях [1, 3] оценки дифференциальной энтропии Шеннона применяются для обнаружения неоднородностей в волокнистом материале, заполняющем параллелепипед $\Pi \subset \mathbb{R}^3$. Используется естественный подход, при котором выбирается скользящее окно достаточно малого размера, в пределах которого оценивается некоторая характеристика материала (фактически дифференциальная энтропия Шеннона распределения направлений волокон). Затем на основе собранных характеристик делается статистический вывод о свойствах этого материала.

В статье [4] оценки дивергенции Кульбака-Лейблера используются для обнаружения многомерных пространственных кластеров в модели Бернулли. Нами развиваются и обобщаются идеи этой работы для обнаружения произвольных неоднородностей в волокнистых материалах. Вводится понятие *волокна* — пары (X, T), где X — случайный вектор со значениями в $\mathbb{R}^{d_1} \cap \Pi$, который по смыслу является центром волокна, а T — случайный вектор со значениями в $\mathbb{R}^{d_2}$, который по смыслу является меткой или некоторым свойством волокна (например, направлением в пространстве). Именно среди таких меток и будет искаться неоднородность. Здесь предполагается, что Π — некоторое объемлющее множество или материал, $\Pi \in \mathcal{B}(\mathbb{R}^{d_1})$, $0 < \mu_{\mathbb{R}^{d_1}}(\Pi) < \infty$, $\mu_{\mathbb{R}^{d_1}}$ — мера Лебега в пространстве $\left(\mathbb{R}^{d_1}, \mathcal{B}(\mathbb{R}^{d_1})\right)$; $\mathsf{P}(X \in B) = \frac{\mu_{\mathbb{R}^{d_1}}(B)}{\mu_{\mathbb{R}^{d_1}}(\Pi)}$ для любого борелевского $B \in \mathcal{B}(\mathbb{R}^{d_1}) \cap \Pi$, а T имеет следующую природу: $T = \xi \cdot \mathbb{1}\{X \in R_0\} + \eta \cdot \mathbb{1}\{X \in \Pi \setminus R_0\}$, где ξ, η – некоторые случайные векторы со значениями в $\mathbb{R}^{d_2}$, $\mathsf{law}(\xi) \neq \mathsf{law}(\eta)$, $p_\xi := \frac{d\mathsf{P}_\xi}{d\mu_{\mathbb{R}^{d_2}}}$, $p_\eta := \frac{d\mathsf{P}_\eta}{d\mu_{\mathbb{R}^{d_2}}}$, $R_0 \in \mathcal{B}(\mathbb{R}^{d_1}) \cap \Pi$, $0 < \mu_{\mathbb{R}^{d_1}}(R_0) < \mu_{\mathbb{R}^{d_1}}(\Pi)$. В этом случае назовём R_0 областью неоднородности. Все случайные элементы определены на пополненном вероятностном пространстве $(\Omega, \mathcal{F}, \mathsf{P})$.

Рассматривается модель (см. Рис. 1), в которой наблюдателю доступны $\{X_i, T_i, i \in \{1, \ldots, \zeta_n\}\}$, $\zeta_n \sim \mathrm{Pois}(\lambda_n \cdot \mu_{\mathbb{R}^{d_1}}(\Pi))$, где для любого $n \in \mathbb{N}$, $\lambda_n > 0$, $\lambda_n \to \infty$, $n \to \infty$. Отметим также, что $\{X_i, T_i, \zeta_i, i \in \mathbb{N}\}$ предполагаются независимыми, $\mathsf{law}(X_i) = \mathsf{law}(X)$, $\mathsf{law}(T_i) = \mathsf{law}(T)$ для любого $i \in \mathbb{N}$. Для каждого множества R из некоторого набора множеств $\mathcal{R}$ строится оценка $\widehat{T}_n(R) := \frac{1}{2}\left(\widehat{D}_n(R) + \widehat{D}_n(\overline{R})\right)$, где $\widehat{D}_n(A)$ – оценка k-ближайших соседей дивергенции Кульбака-Лейблера (см. статью [2]) между распределением меток T_j внутри окна A (то есть при $X_j \in A$) и вне его (то есть при $X_j \in \Pi \setminus A$). Оценка области неоднородности $\widehat{R}_n$ полагается равной аргументу максимума такой специальной статистики $\widehat{T}_n(R)$, взятого по всем окнам $R \in \mathcal{R}$, то есть $\widehat{R}_n := \operatorname*{argmax}_{R \in \mathcal{R}} \widehat{T}_n(R)$.

При широких условиях на плотности p_ξ и p_η удается доказать состоятельность предложенной процедуры.

Теорема 1 *Пусть для некоторых* $\varepsilon, R > 0$, $N \in \mathbb{N}$, $f \in \{p_\xi, p_\eta\}$, $g \in \{p_\xi, p_\eta\}$ *функционалы* $K_{f,g}(2, N)$, $Q_{f,g}(\varepsilon, R)$, $T_{f,g}(\varepsilon, R)$ *конечны. Тогда*

$$\mathsf{P}\left(\widehat{R}_n \in \operatorname*{argmax}_{R \in \mathcal{R}} d_\Pi(R, R_0)\right) \to 1, n \to \infty.$$

Здесь $d_\Pi(R, R_0) := \left|\frac{\mu(RR_0)}{\mu(R)} - \frac{\mu(\overline{R}R_0)}{\mu(\overline{R})}\right|$, $\overline{R} := \Pi \setminus R$, *а определения функционалов* K_{f_1,f_2}, Q_{f_1,f_2}, T_{f_1,f_2} *для плотностей* f_1, f_2 *можно найти в статье [2].*

Работа выполнена при поддержке гранта МГУ им.М.В.Ломоносова “Современные проблемы фундаментальной математики и механики”.

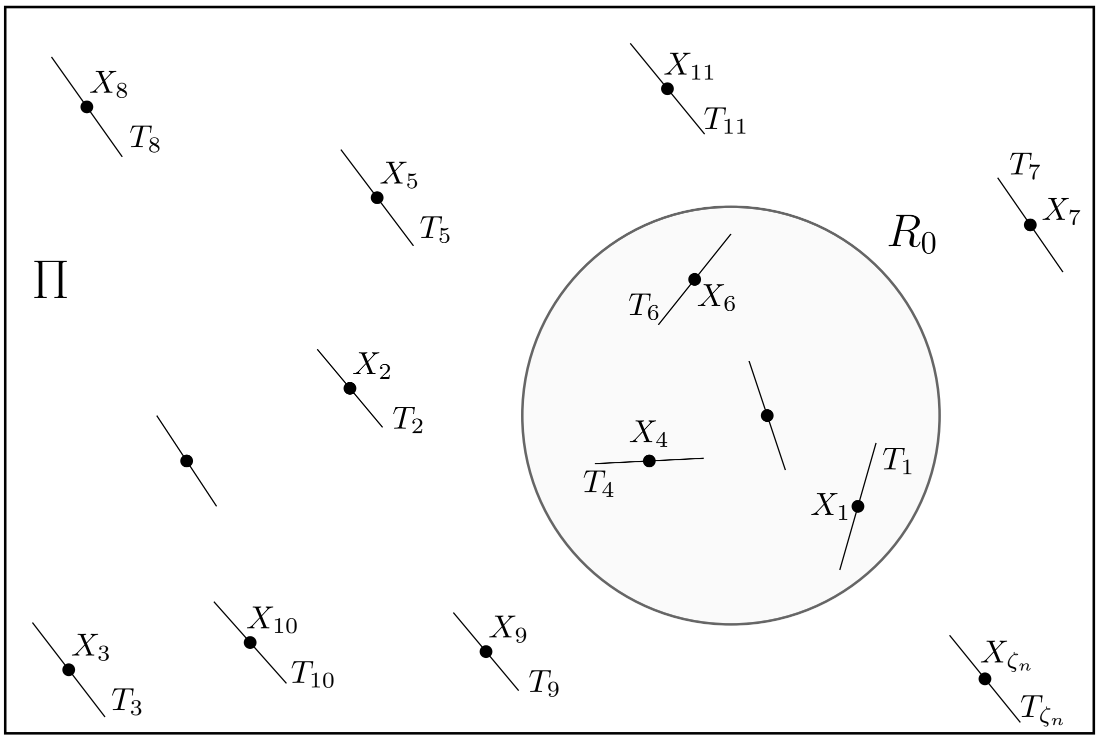

Рис. 1: Исследуемый волокнистый материал Π с областью неоднородности R_0

В процессе доказательства Теоремы 1 также устанавливаются новые результаты, касающиеся асимптотической несмещенности и L^2-состоятельности оценок $\widehat{D}_n(A)$ (в оценку входит сумма пуассоновского числа случайных величин; ранее в работе [2] асимптотические свойста были установлены только для оценок, вовлекающих суммы постоянного числа схожих случайных величин).

СПИСОК ЛИТЕРАТУРЫ

1. *Alonso-Ruiz, P., Spodarev, E.* (2018). Entropy-based inhomogeneity detection in fiber materials. *Methodology and Computing in Applied Probability.* **20**, 1223—1239.

2. *Bulinski, A., Dimitrov, D.* (2021). Statistical Estimation of the Kullback–Leibler Divergence. *Mathematics.* **9**, 544, 1–36.

3. *Dresvyanskiy, D., Karaseva, T., Makogin, V., Mitrofanov, S., Redenbach, C., Spodarev, E.* (2020). Detecting anomalies in fibre systems using 3-dimensional image data. *Statistics and Computing.* **30**, 4, 817–837.

4. *Walther, G.* (2010). Optimal and fast detection of spatial clusters with scan statistics. *The Annals of Statis-tics.* **38**, 1010–1033.

5. *Wang, Q., Kulkarni, S.R., Verdú, S.* (2009). Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory.* **55**, 2392–2405.