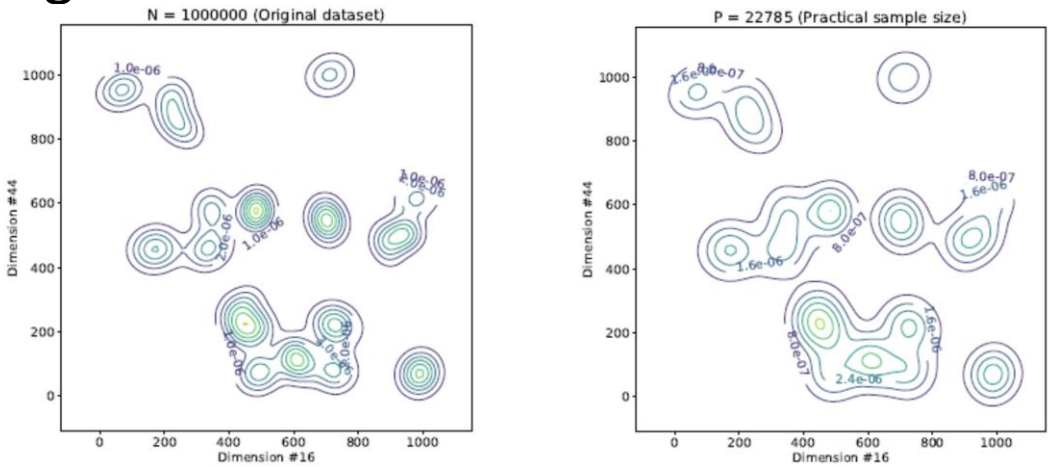
Density estimation-based method to determine sample size for random sample partition of big data

Yulin HE, Jiaqi CHEN, Jiaxing SHEN,
Philippe FOURNIER-VIGER, Joshua Zhexue HUANG

Frontiers of Computer Science, DOI: 10.1007/s11704-023-2356-x

Problems & Ideas

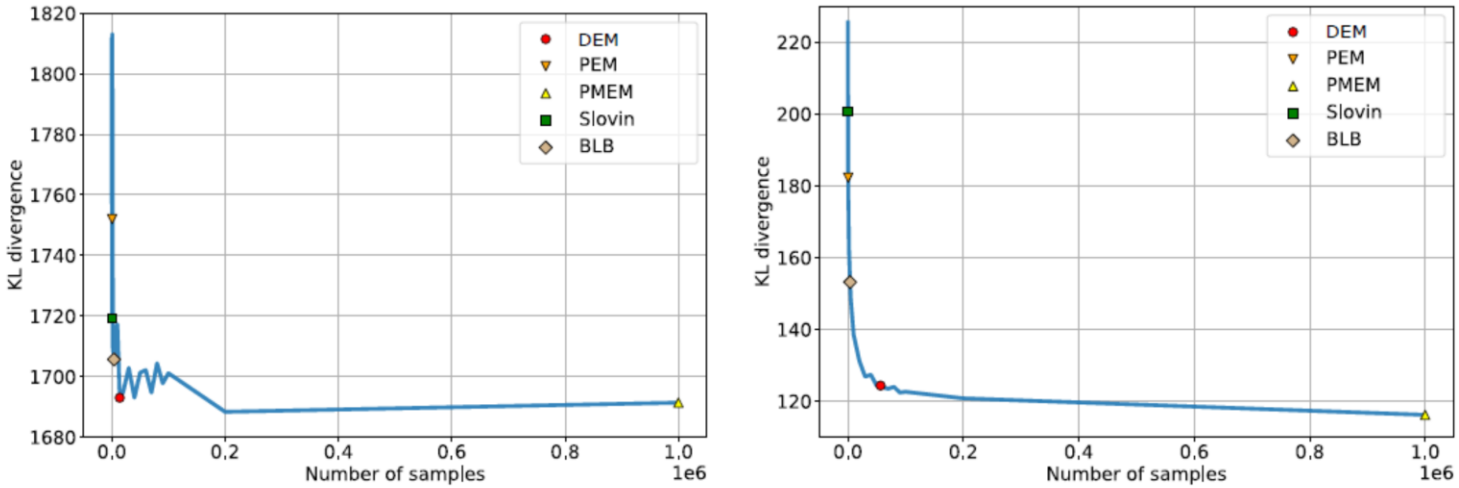
- Problems of sample size estimation for random sample partition of big data:
  - Existing sample size determination methods were designed for sampling problems based on the strong distribution assumption.
  - Moreover, these methods were proposed to deal with small and medium sized data sets and are not suitable for big data sets.
- Ideas: A novel density estimation-based method is proposed to determine the optimal sample size for RSP data blocks by minimizing the validation error of a kernel density estimator.



Two-dimensional probability density functions (*p.d.f.*s) estimated with kernel density estimation technology based on different sample sizes. Left: *p.d.f.* estimated with the whole sample points; Right: *p.d.f.* estimated with the partial sample points.

Main Contributions

- Contributions:
  - Calculation of theoretical sample size based on the multivariate Dvoretzky-Kiefer-Wolfowitz inequality by utilizing the fixed-point iteration method;
  - Determination of practical sample size by minimizing the validation error of the kernel density estimator constructed for an RSP data block;
  - Extensive experiments to validate the convergence of theoretical sample size, *p.d.f.* estimation quality with practical sample size , and applicability of big data analysis.



Change of KL divergence with increase of sample sizes.
Left: 10-mode-and-50-dimension big data set; Right: 20-mode-and-10-dimension big data set.