Hierarchical clustering of the pre-computed signals database to solve the parametric inverse light-scattering problem

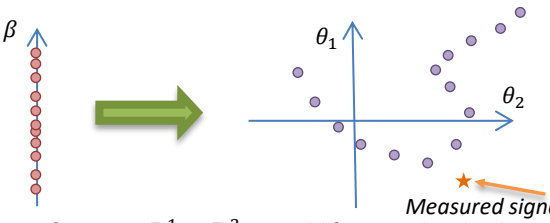
Artem R. Muliukov [a, b] *and Maxim A. Yurkin* [a, b,*]

[a] Physics Department, Novosibirsk State University, Novosibirsk, Russia, [b] Voevodsky Institute of Chemical Kinetics and Combustion SB RAS, Novosibirsk, Russia, *e-mail: yurkin@gmail.com

Parametric inverse problem

- Non-invasive characterization of single particles through the solution of the inverse light-scattering problem(LSP)
- Particle models with several free parameters [1]
- Preliminary computed database of simulated signals [2]

Problem statement



- Illustration is for a map $\mathbb{R}^1 \to \mathbb{R}^2$. In real-life examples the direct map is, e.g., g: $\mathbb{R}^4(\beta) \to \mathbb{R}^{61}(y)$ if the signal is an angle-resolved light-scattering pattern.
- Inverse problem is reduced to (slow) global optimization [1]:

$$\beta_*(x) \stackrel{\text{def}}{=} \underset{\beta}{\operatorname{argmin}} \|x - g(\beta)\|.$$

- Further approximated by nearest-neighbor interpolation using pre-computed database $Y = \{\beta_i, y_i \stackrel{\text{def}}{=} g(\beta_i)\}$ [2]:

$$\beta_*(x) \approx \beta_{i_*}, \qquad i_* \stackrel{\text{def}}{=} \underset{i}{\operatorname{argmin}}\, d_i, \qquad d_i \stackrel{\text{def}}{=} \|x - y_i\|.$$

- Statistical quantities (math. expectations, SDs, marginal probabilities, etc.) are obtained in the Bayesian framework

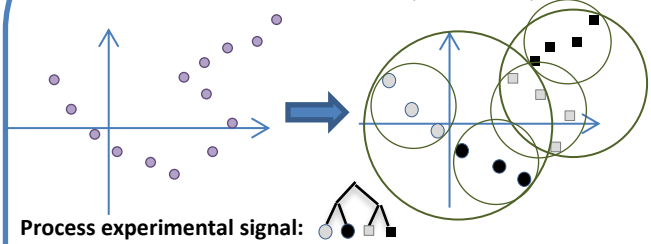
$$< f(\beta) > \stackrel{\text{def}}{=} \int_\beta \mathrm{d}\beta f(\beta) P(\|x - g(\beta)\|) \approx \sum_i f(\beta_i) P(d_i)$$

where f – any function, P – conditional probability.

- Evaluating all distances (each ~ 100 FLOPs) is still slow for real-time applications (100 signals/second). **Need to be accelerated!**

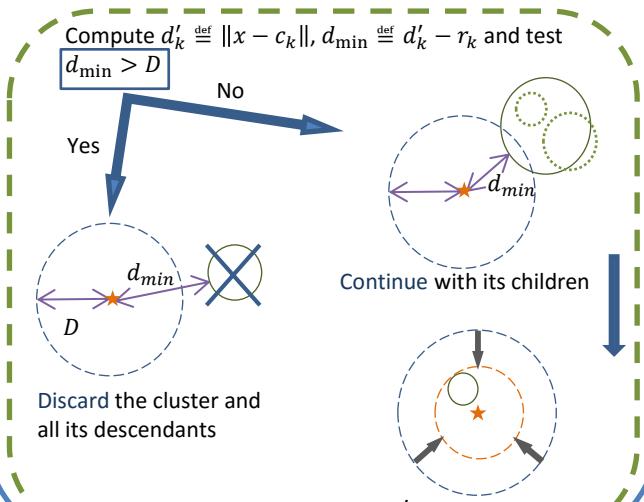
Main algorithm

Build a binary tree (with centers c_k and radii r_k):



Process experimental signal:

Traverse the tree starting from the top (root) node. For each node (cluster) k do the following (D – current upper bound for $\min d_i$):



Calculation of statistical quantities

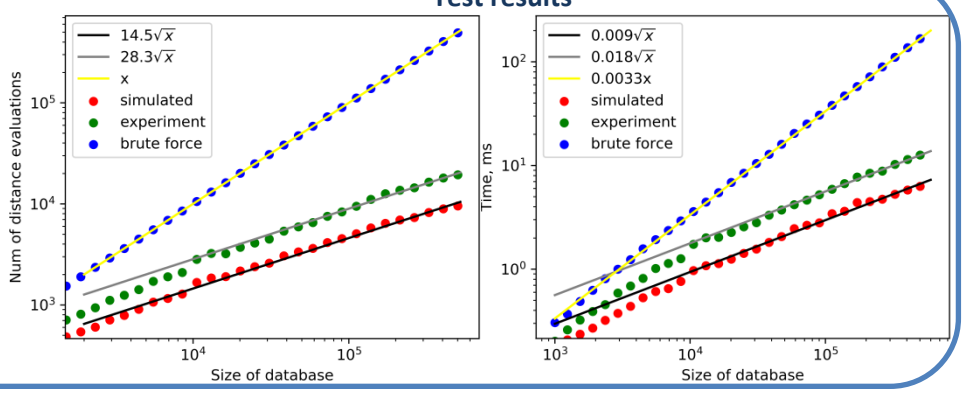
- Each discarded cluster decreases the number of distance evaluations ⇒ decreases the computational time
- Statistical quantities can still be estimated. For each discarded cluster A_k with N_k last-level descendants (elements of the original database)

$$\sum_{i \in A_k} f(\beta_i) P(d_i) \approx M_k f(\beta'_k) P(d'_k), \qquad \beta'_k \stackrel{\text{def}}{=} \frac{1}{M_k} \sum_{i \in A_k} \beta_i$$

- Error should not be large, since larger M_k are for larger d'_k, hence smaller $P(d'_k)$.

Test results

- Database of 5×10^5 LSP of blood platelets, characterized by 4 parameters (randomly distributed). LSP has 61 values in the range [10,70],corresponding to the measurements using the scanning flow cytometer.
- Test data are 1000 of either experimental or separately simulated LSPs. Results are averaged.
- Smaller databases are obtained as a subset of the original one.
- Both distance evaluations and whole time were measured in comparison with previous brute-force search.



Conclusion and plans

Presented algorithm significantly accelerates finding the nearest neighbor:

- Number of distance evaluations scales with square root of database size
- Up to 49/25 times decrease in this number for simulated/experimental data
- Total wall times decrease up to 25/13 times (cache issues?)

Further research:

- Optimal clustering of points (metric k-center problem)
- Other tree structures (e.g., cover tree)
- Open-source library, including computation of statistical quantities

References

1. Strokotov DI et al. J. Biomed. Opt. 2009;14:064036
2. Moskalensky AE et al. J. Biomed. Opt. 2013;18:017001