Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Electronic Supplementary Information for

Bayesian analysis of data from segmented super-resolution images for quantifying protein clustering

Tina Košuta,[a,b] Marta Cullell-Dalmau,[a] Francesca Cella Zanacchi,[c,d] and Carlo Manzo*[a]

This file includes:

[a] *the Quantitative BioImaging lab, Facultat de Ciències i Tecnologia, Universitat de Vic – Universitat Central de Catalunya (UVic-UCC), Vic, Spain.*
[b] *University of Ljubljana, Ljubljana, Slovenia.*
[c] *Nanoscopy and NIC@IIT, Istituto Italiano di Tecnologia, Genoa, Italy.*
[d] *Biophysics Institute (IBF), National Research Council (CNR), Genoa, Italy.*
∗ Corresponding author, e-mail: carlo.manzo@uvic.cat

Electronic Supplementary Information

Materials and Methods

Software implementation and simulations. The code for implementing the NS algorithm was written in both Matlab (The MathWorks, Inc., Natick, Massachusetts, United States) and R (R Foundation for Statistical Computing, Vienna, Austria). The program files, documentation, and example data are available on the GitHub repository hosting service at https://github.com/cmanzo/NS_multicomp.

The NS algorithm used for this study was implemented on 30 particles with a depth of 40 iterations. For the Metropolis-Hastings, the random motion of a particle is performed by adding a variable step in a random direction. To ensure an acceptance rate of ∼50%, the step length is drawn from a normal distribution with standard deviation σ_{step} initially set to 0.1 and updated at each iteration as (1, 2)

$$\sigma_{step} = \begin{cases} \sigma_{step} \cdot e^{1/A} & \text{if } A > R \\ \sigma_{step} \cdot e^{-1/R} & \text{if } A \leq R \end{cases}, \qquad [1]$$

where A and R are the numbers of accepted and rejected samples, respectively. We explored the behavior of the algorithm for different prior probabilities corresponding to the Dirichlet distribution with $\delta = 1$ (uniform distribution), $\delta = 0.5$ and $\delta = 1.5$. Data corresponding to simulations of localization counts obtained from monomeric clusters were generated considering a discretized lognormal *pdf*

$$f_1(n|\mu, \sigma) = \frac{1}{2}\left[\text{erf}\left(\frac{\mu - \log(n-1)}{\sqrt{2}\sigma}\right) - \text{erf}\left(\frac{\mu - \log(n)}{\sqrt{2}\sigma}\right)\right], \qquad [2]$$

with parameters (μ, σ) equal to $(3.349, 0.846)$ and $(3.227, 0.569)$, since they have been recently shown to accurately approximate the output of STORM imaging in different experimental conditions (3, 4). Localization counts in oligomeric clusters of m proteins were obtained by the sum of m random numbers obtained as described above. We generate data from mixtures of different number of species ($K_{gt} = 5, 7$) with decreasing ($\boldsymbol{\alpha}_{gt} = (0.33, 0.27, 0.20, 0.13, 0.07)$, $\boldsymbol{\alpha}_{gt} = (0.25, 0.21, 0.18, 0.14, 0.11, 0.07, 0.04)$) and bell-shaped weights ($\boldsymbol{\alpha}_{gt} = (0.11, 0.22, 0.33, 0.22, 0.11)$, $\boldsymbol{\alpha}_{gt} = (0.06,\ 0.12,\ 0.19,\ 0.25,\ 0.19,\ 0.12,\ 0.06)$). A comparison of the algorithm performance in all the explored conditions can be found in Figs. S1-S4.

Sample preparation for STORM microscopy. HeLa IC74-mfGFP stably transfected cell line (from Takashi Murayama lab, Department of Pharmacology, Juntendo University School of Medicine, Tokyo, Japan) were plated on 8-well Lab-tek 1 coverglass chamber (Nunc) and grown under standard conditions (DMEM, high glucose, pyruvate (Invitrogen 41966052) supplemented with 10% FBS, 2 mM glutamine and selected with 400 μg/mL Hygromycin). Cells were fixed with PFA (3% in PBS) at RT for 7 minutes. Cells were then incubated at RT with blocking buffer (3% (wt/vol) BSA (Sigma) in PBS and 0.2% Tryton. In HeLa IC74-mfGFP stably transfected cells, dynein intermediate chain-green fluorescent protein (GFP) was immuno-stained with primary antibody (chicken polyclonal anti GFP, Abcam 13970) diluted 1:2000 in blocking buffer for 45 minutes at room temperature. Cells were rinsed 3 times in blocking buffer for 5 minutes and incubated for 45 minutes with secondary antibodies donkey-anti chicken labeled with photoactivatable dye pairs for STORM (Alexa Fluor 405-Alexa Fluor 647).

STORM microscopy. Imaging was performed with an oil immersion objective (Nikon, CFI Apo TIRF 100x, NA 1.49, Oil), repeated cycles of activation (405 nm laser), and readout (647 nm laser) using TIRF illumination. During experiments the focus was locked through the Perfect Focus System (Nikon) and imaging was performed on an EMCCD camera (Andor iXon X3 DU-897, Andor Technologies). A commercial N-STORM microscope (Nikon Instruments) was used to acquire 40,000 frames at a 33 Hz frame rate. An excitation intensity of ∼0.9 kW/cm^2 for the 647 nm readout laser (300 mW MPB Communications, Canada) and an activation intensity of ∼35 W/cm^2 for the 405 nm activation laser (100 mW, Cube Coherent, CA) were used. STORM imaging buffer was used containing GLOX solution as oxygen scavenging system (40 mg/mL Catalase [Sigma], 0.5 mg/mL glucose oxidase, 10% glucose in PBS) and MEA 10 mM (Cysteamine MEA [SigmaAldrich, #30070-50G] in 360 mM Tris-HCl).

STORM data analysis. Localization and reconstruction of STORM images were performed using custom software (Insight3, kindly provided by Bo Huang, University of California) by Gaussian fitting of the single molecules images to obtain the localization coordinates. The final image is obtained plotting each identified molecule as a Gaussian spot with a width corresponding to the localization precision (10 nm) and corrected for drift. A custom code implementing a distance-based clustering algorithm, was used to identify spatial clusters of localizations. The localizations list was first binned to 20 nm pixel size images that were filtered with a square kernel (7×7 pixel^2) and thresholded to obtain a binary image. Only the localizations lying on adjacent (6-connected neighbours) non-zero pixels of the binary image were considered for further analysis. To select the sparse dynein contribution large clusters were filtered out setting a threshold on the maximum number of localizations (1000 localizations/cluster).

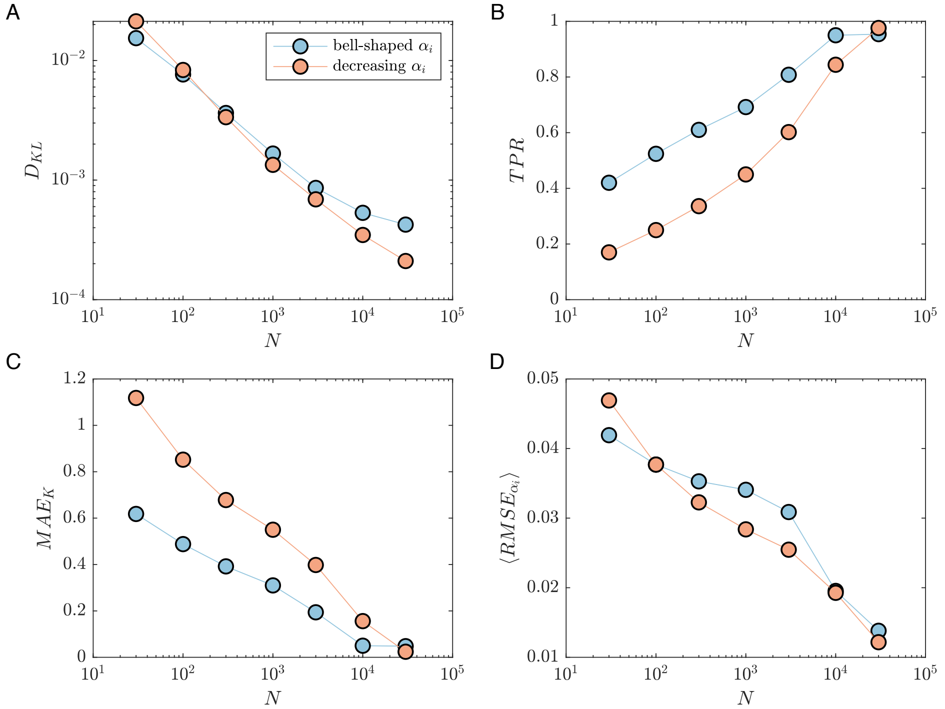

Fig. S1. Performance of the NS algorithm at varying the number of data points for different distribution of weights. (A) The Kullback-Leibler (D_{KL}) divergence. (B) The true positive rate TPR. (C) The mean absolute error of the number of components MAE_K. (D) The average root-mean square error of the weights $\langle RMSE_{\alpha_i} \rangle$. Each point correspond to 500 simulations with parameters: $\boldsymbol{\alpha} = (0.11, 0.22, 0.33, 0.22, 0.11)$ (bell-shaped) or $\boldsymbol{\alpha} = (0.33, 0.27, 0.20, 0.13, 0.07)$ (decreasing), $(\mu, \sigma) = (3.349, 0.846)$ and analyzed with a Dirichlet prior with $\delta = 1.5$.

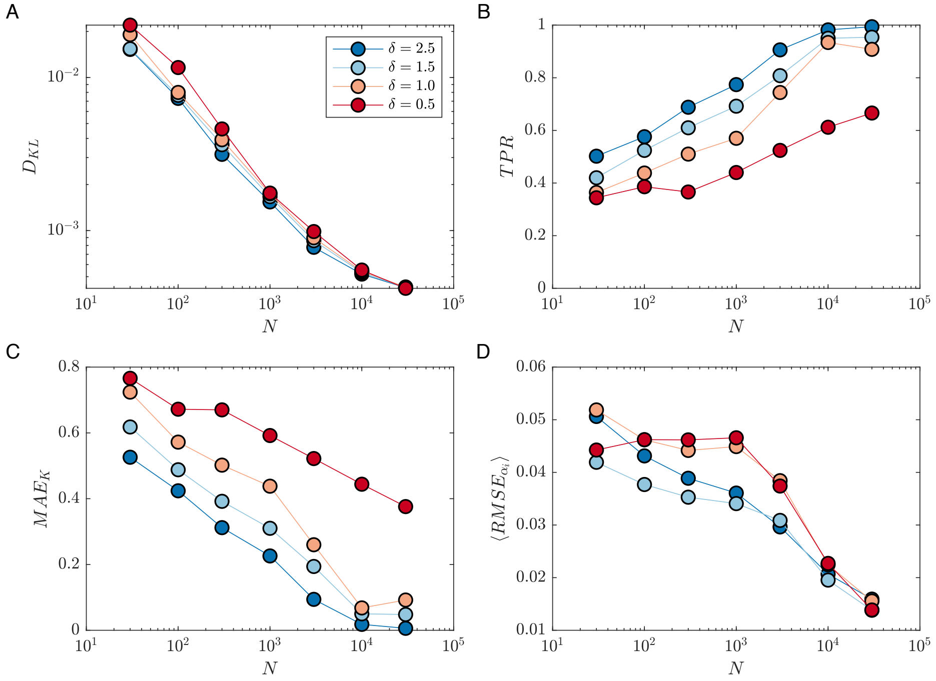

Fig. S2. Performance of the NS algorithm at varying the number of data points for different prior distributions. (A) The Kullback-Leibler (D_{KL}) divergence. (B) The true positive rate TPR. (C) The mean absolute error of the number of components MAE_K. (D) The average root-mean square error of the weights $\langle RMSE_{\alpha_i} \rangle$. Each point correspond to 500 simulations with parameters: $\boldsymbol{\alpha} = (0.11, 0.22, 0.33, 0.22, 0.11)$ (bell-shaped), $(\mu, \sigma) = (3.349, 0.846)$ and analyzed with a Dirichlet prior with $\delta = 0.5, 1, 1.5, 2.5$.

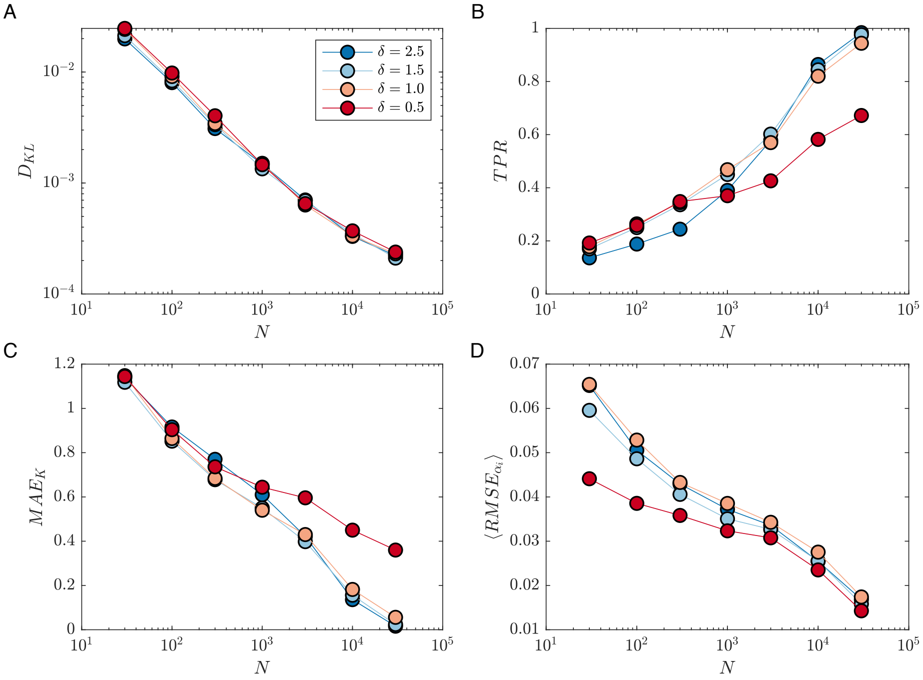

Fig. S3. Performance of the NS algorithm at varying the number of data points for different prior distributions. (A) The Kullback-Leibler (D_{KL}) divergence. (B) The true positive rate TPR. (C) The mean absolute error of the number of components MAE_K. (D) The average root-mean square error of the weights $\langle RMSE_{\alpha_i} \rangle$. Each point correspond to 500 simulations with parameters: $\boldsymbol{\alpha} = (0.33, 0.27, 0.20, 0.13, 0.07)$ (decreasing), $(\mu, \sigma) = (3.349, 0.846)$ and analyzed with a Dirichlet prior with $\delta = 0.5, 1, 1.5, 2.5$.

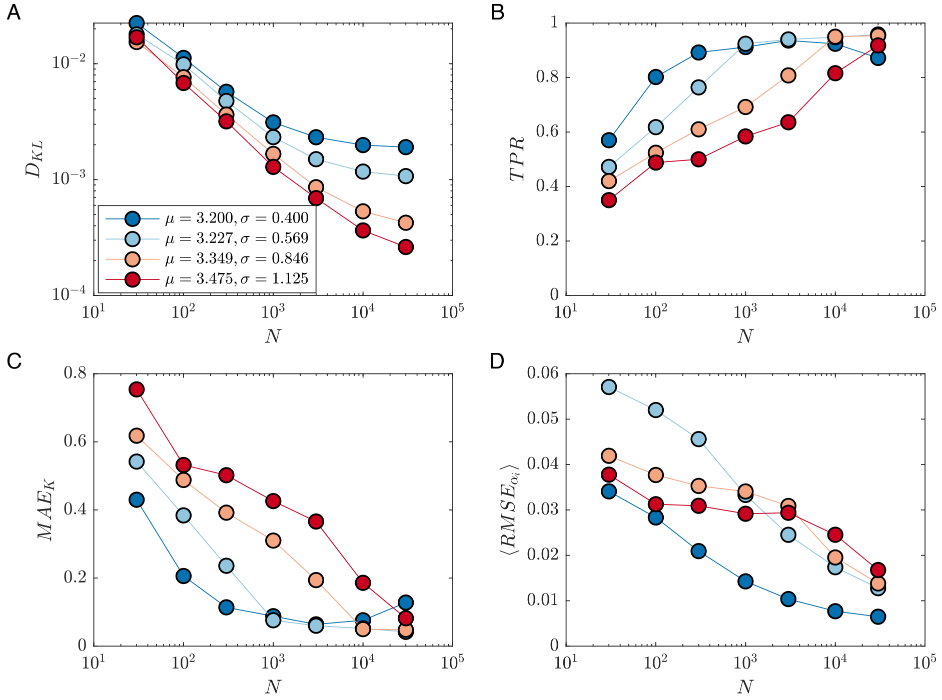

Fig. S4. Performance of the NS algorithm at varying the number of data points for different choices of the the parameters of the pdf. (A) The Kullback-Leibler (D_{KL}) divergence. (B) The true positive rate TPR. (C) The mean absolute error of the number of components MAE_K. (D) The average root-mean square error of the weights $\langle RMSE_{\alpha_i} \rangle$. Each point correspond to 500 simulations with parameters: $\boldsymbol{\alpha} = (0.11, 0.22, 0.33, 0.22, 0.11)$ (bell-shaped), $(\mu, \sigma) = (3.200, 0.400), (3.2270.569), (3.349, 0.846), (3.4751.125)$, and analyzed with a Dirichlet prior with $\delta = 1.5$.

References

1. D. Sivia and J. Skilling, *Data analysis: a Bayesian tutorial*, OUP Oxford, 2006.
2. F. Feroz and M. P. Hobson, *Monthly Notices of the Royal Astronomical Society*, 2008, **384**, 449–463.
3. F. Cella Zanacchi, C. Manzo, A. S. Alvarez, N. D. Derr, M. F. Garcia-Parajo and M. Lakadamyali, *Nature Methods*, 2017, **14**, 789.
4. F. Cella Zanacchi, C. Manzo, R. Magrassi, N. D. Derr and M. Lakadamyali, *Biophysical Journal*, 2019, **116**, 2195–2203.