Supporting Information 1: Complements about segclust2d

Initialising the Expectation-Maximization algorithm

The Expectation-Maximization (EM) algorithm used to estimate the distribution parameters in the segmentationclustering model is known to be sensitive to initialisation, so that it may converge to local maxima of the likelihood. This behaviour has some consequences on the parameters estimates but also makes the choice of the number of segments or states complicated. The classical initialisation solution consists in running the algorithm numerous times and select the point with the highest value of the log-likelihood, but this strategy is too computationally demanding. To minimize the risk of reaching local maxima within an unacceptable computation time, we use the following initialisation strategy: (1) perform a pure segmentation of the signal and (2) use a hierarchical cluster algorithm, based on the log-likelihood ratio distance to assign segments to states.

Even with smart initialisation points, however, the EM algorithm may still converge to local maxima. This situation appears when looking, for a given number of states M, at the log-likelihood as a function of the number of segments K: whereas it is expected to be somewhat regular, this function can be quite noisy. To solve this issue, we use new initial points for all 'non-reliable' solutions, i.e. for which an initialisation problem can be suspected. These new initial points are based on the parameter estimates of the distribution $(\pi_m, \boldsymbol{\mu}_m, \boldsymbol{\sigma}^2_m)$ obtained for all the 'reliable' solutions, which correspond to the points that lie on the convex hull of log-likelihood curve. New initial points are provided by cutting in half a segment or merging two segments. The improvement in terms of regularity of the log-likelihood curve obtained thanks to this procedure is illustrated in figure S1-1. Although the procedure does not formally guarantee the convergence, it would be unlikely that none of the first initialisation points had converged to a 'reliable' solution that could be propagated.

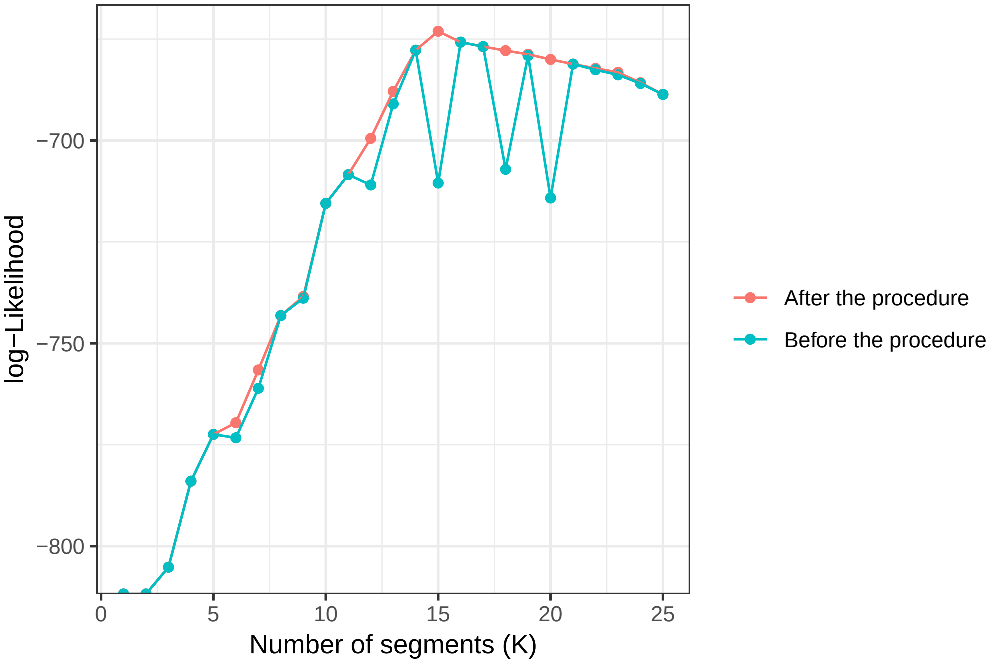

Figure S1-1: Maximum Likelihood estimates as a function of the number of segments before (in blue) and after (in red) the procedure. For instance the 'reliable' solution obtained with $K = 24$ segments was used to provide starting points for the EM algorithm for $K = 23$ and for $K = 25$, and this procedure gradually spreads over adjacent points.

Model selection

Choice of the number of segments K in the segmentation-only model

We used the adaptive model selection strategy proposed in Lavielle (2005) consisting in choosing the value of K that maximizes the following penalized log-likelihood : $\mathcal{L}_K - C\,K$ where $\mathcal{L}_K$ is the log-likelihood of the optimal segmentation in K segments and C is a unknown positive constant. The heuristic proposed by Lavielle (2005) makes it possible to bypass the estimation of C. It consists in detecting the value of K for which the log-likelihood ceases to increase significantly. More specifically, consider the normalised log-likelihood defined as:

$$\tilde{\mathcal{L}}_K = K_{max} - (K_{max} - 1)\frac{\mathcal{L}_{K_{max}} - \mathcal{L}_K}{\mathcal{L}_{K_{max}} - \mathcal{L}_1}$$

which ranges between 1 (for $K = 1$) and K_{max} (for $K = K_{max}$), K_{max} being the maximum number of segments considered. The optimal number of segments is then determined as the largest value of K for which one gets $\left(\tilde{\mathcal{L}}_K - \tilde{\mathcal{L}}_{K-1}\right) - \left(\tilde{\mathcal{L}}_{K+1} - \tilde{\mathcal{L}}_K\right)$ larger than a threshold, set to 0.75 in agreement with Lavielle (2005). As the selection relies on a predefined threshold, it is worth checking where the point corresponding to the selected number of segment lies on a plot of the log-likelihood curve (fig. S1-2). The optimal K value obtained in this way should correspond to a noticeable slope change.

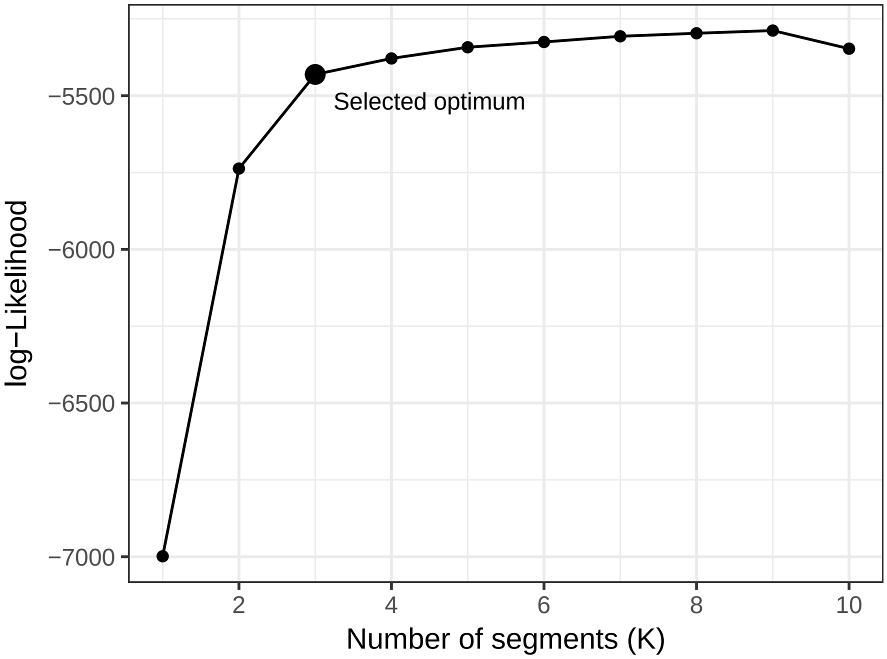

Figure S1-2: log-likelihood of a segmentation as a function of the number of segment. The optimum selected by the criterion from Lavielle (2005) should be located at a break in the increase of the curve.

Choice of the number of segments K and states M in the segmentationclustering model

Selection of the best segmentation-clustering model (i.e. of the best couple of K and M values) is a hard task as no method has been yet proposed for this purpose. A log-likelihood is expected to increase with the number of parameters. However, as explained in Picard et al. (2007), if the log-likelihood increases with the number of clusters M, it does not always increases with the number of segments K. Indeed a phenomenon of self-penalization occurs at the true number of segments when the detection of breakpoints is easy, stressing to choose K simply as the value that maximizes the log-likelihood. However when the detection of breakpoints is more difficult, choosing the maximum value would tend to overestimate K. Picard et al. (2007) suggested to add a penalty. A Bayesian Information Criterion (BIC)-based penalty appeared to be sufficient in this case, although it does not work in pure segmentation (Picard et al., 2005). As BIC is the most popular criterion to choose the optimal number of clusters in a mixture model (Frühwirth-Schnatter, 2006), we used the maximum value of the following BIC-based penalised likelihood $\mathcal{B}_{K,M}$ for the selection of both K and M parameters:

$$\mathcal{B}_{K,M} = \mathcal{L}_{K,M} - \frac{5 \times M - 1}{2} log(2n) - \frac{K-1}{2} log(2n),$$

where $\mathcal{L}_{K,M}$ stands for the log-likelihood for the optimal segmentation-clustering with K segments classified into M states. The penalization terms in the BIC criterion is half the number of parameters times the logarithm of the size of the dataset. For our model the number of parameters to be estimated is $2M$ means + $2M$ variances + $M-1$ proportions for the states, and $K-1$ breakpoints for the segments, and the size of the dataset for n bivariate values is $2n$.

Although this procedure appears to work well for choosing the optimal number of segments, it has been observed to be less reliable for choosing the optimal number of states, which tends to be overestimated. We therefore advise users to set an a priori number of states M, based on biological knowledge. We also advise to look at the plot of the BIC-penalized log-likelihood, as in figure 4b of main text, to check that the solution obtained makes sense.

References

Frühwirth-Schnatter, S., 2006. Finite mixture and markov switching models. Springer Science & Business Media.

Lavielle, M., 2005. Using penalized contrasts for the change-point problem. Signal Processing 85:1501–1510 . doi:10.1016/j.sigpro.2005.01.012.

Picard, F., S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin, 2005. A statistical approach for array CGH data analysis. BMC Bioinformatics 6:27. doi:10.1186/1471-2105-6-27.

Picard, F., S. Robin, E. Lebarbier, and J.-J. Daudin, 2007. A Segmentation/Clustering Model for the Analysis of Array CGH Data. Biometrics 63:758–766. doi:10.1111/j.1541-0420.2006.00729.x.