Supplementary Material

One case example

In Figure 1 we present a particular solution obtained with 10 different features and high SNR. In the first row of the graph, all biomarkers exhibit a single trajectory, whereas in the second row, the actual configuration involves two sub-trajectories. The estimated trajectories are represented by solid colored lines, and the true trajectories by shaded gray lines, while subjects are color-coded based on the estimated sub-group.

Our method successfully separates the different configurations and identifies distinct sub-populations for all biomarkers. We also observe that some biomarkers have a very high probability of following a sub-trajectory, such as the first one, while others have probabilities close to 50%; this occurs because the estimated noise level is slightly higher than the actual noise level (0.5), making the inclusion of a sub-trajectory less impactful on the posterior distribution.

The example also explains a curious behavior of the parameter describing the sub-populations. It was observed that the greatest uncertainty is in the first and last stages of the progression, while in the central part, uncertainties almost appreoches to zero. This happens because the shape of the Sigmoids causes the beginning and the end to be closer together than the middle part.

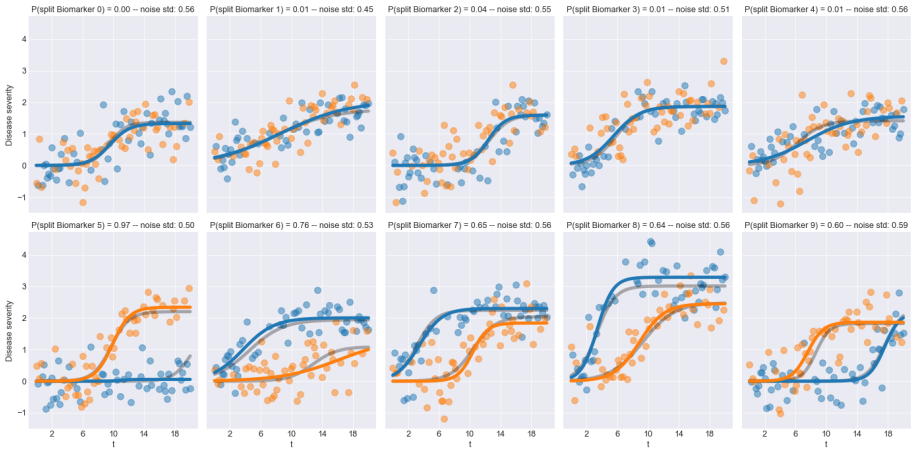

Fig. 1. The figure illustrates synthetic data: the true trajectory is in solid grey, and the estimated trajectories are in solid colors, while subjects are color-coded according to their estimated subgroup.

1 EM for the optimisation

In this first Section of the appendix we give the mathematical details for performing the EM-steps needed for the evaluation of the MAP of the posterior distribution.

We decided to exploit the two level mixture model structure to implement a multiple step approach. Indeed we perform iteratively the optimisation of the parameters θ via Gradient Descent (GD) and a classical EM step for the parameters ξ and π.

For what concern the EM-step for the parameter ξ, we follow the same reasoning used for deriving the classical EM-step, i.e. we first derive the posterior distribution for the split.

Let us introduce a new variable $z_b \in \{0, 1\}$ such that $p(z_b = 1) = \xi_b$. Therefore we can describe the posterior distribution for the auxiliary variable in terms of other quantities:

$$\gamma_b^n = p(z_b = 1 \mid x_b^j) = \frac{p(x_b^j \mid z_b = 1)p(z_b = 1)}{p(x_b^j)} = \frac{p(x_b^j \mid \theta_b^0)\xi_b}{p(x_b^j)} \tag{1}$$

$$1 - \gamma_b^n = p(z_b = 0 \mid x_b^j) = \frac{p(x_b^j \mid z_b = 0)p(z_b = 0)}{p(x_b^j)} = \frac{p(x_b^j \mid \theta_b^{1:2})(1 - \xi_b)}{p(x_b^j)} \tag{2}$$

where the conditioning on noise standard deviation is omitted for simplicity of notation.

For performing the EM-step, we can consider to evaluate the gradient of the posterior distribution w.r.t. the parameter of interest and setting it equal to zero:

$$\begin{aligned}
\partial_{\xi_b}\left(\ln(p(\theta, \sigma, \xi, \pi \mid \boldsymbol{x}))\right) &= \partial_{\xi_b}\left(\ln(p(\boldsymbol{x} \mid \theta, \sigma, \xi, \pi)) + \beta\xi_b + \beta_N\left(\ln(\sigma_b) - \frac{1}{\sigma_b}\right)\right) \\
&= \sum_n \left(\frac{\partial_{\xi_b} p(x_b^j \mid \theta_b, \sigma_b, \xi_b, \pi^n)}{p(x_b^j \mid \theta_b, \sigma_b, \xi_b, \pi^n)}\right) + \beta \\
&= \sum_n \left(\frac{p(x_b^j \mid \theta_b^0, \sigma_b, \xi_b, \pi^n) - p(x_b^j \mid \theta_b^{1:2}, \sigma_b, \xi_b, \pi^n)}{p(x_b^j)}\right) + \beta \\
&= \sum_n (\gamma_b^n) - N\xi_b + \beta(1 - \xi_b)\xi_b \\
&= \sum_n (\gamma_b^n) + (-N + \beta(1 - \xi_b))\xi_b
\end{aligned}$$

Thefereore, if we want to maximise the value for ξ_b we can equalise to zero the loss function derivative, obtaining an iterative way to update the parameter:

$$\xi_b^{(k)} = \frac{\sum_n \gamma_b^n}{N + (\xi_b^{(k-1)} - 1)\beta} \tag{3}$$

We observe that due to the fact that ξ_b is a mixture coefficient, it has to be a value in range between zero and one; therefore not all values for the prior

parameter β can be considered. In the next Session we are giving a sufficient condition to ensure ξ_b to be in an appropriate range.

For what concern the parameters π^n, the reasoning is similar, with the simplicty given by the fact that it is a common EM algorithm.

Let us introduce a new variable $\nu^n \in \{0, 1\}$ such that $p(\nu^n = 1) = \pi^n$. Therefore we can describe the posterior distribution for the auxiliary variable in terms of other quantities:

Following the same reasoning

$$\chi_b^n = p(\nu_n = 1 \mid x_b^j) = \frac{p(x_b^j \mid \nu^n = 1)p(\nu^n = 1)}{p(x_b^j)} = \frac{p(x_b^j \mid \theta_b^1)\pi^n}{p(x_b^j)} \tag{4}$$

$$1 - \chi_b^n = p(\nu^n = 0 \mid x_b^j) = \frac{p(x_b^j \mid v_n = 0)p(\nu^n = 0)}{p(x_b^j)} = \frac{p(x_b^j \mid \theta_b^2)(1 - \pi^n)}{p(x_b^j)} \tag{5}$$

For performing the EM-step, we can consider to evaluate the gradient of the loss function and setting it equal to zero:

$$\partial_{\pi^n} \left(\ln(p(\theta, \sigma, \xi, \pi \mid \boldsymbol{x}))\right) = \sum_b \frac{\partial_{\pi^n} p(x_b^j \mid \theta_b, \sigma_b, \xi_b, \pi^n)}{p(x_b^j \mid \theta_b, \sigma_b, \xi_b, \pi^n)} = \sum_b \chi_b^n - N\pi^n$$

Thefereore, if we want to maximise the value for π^n we can equalise to zero the loss function derivative, obtaining an iterative way to update the parameter:

$$\pi^{n(k)} = \frac{\sum_b \chi_b^n}{N} \tag{6}$$

1.1 Bounds for prior parameter β

In this Section we derive a sufficient condition on the prior parameter β to allow ξ to be a proper mixture coefficient for our model, i.e. between zero and one.

This Lemma is useful for the demonstration of the following result.

Lemma 1. *Let f_n and g_n positive functions for all $n = 1, \cdots, N$; then the following inequality holds:*

$$\frac{\sum_n f_n}{\sum_n g_n} < \sum_n \frac{f_n}{g_n}$$

Proof. The proof is straightforward, indeed, being g_n positive functions, it is true that

$$\frac{f_n}{\sum_n g_n} \leq \frac{f_n}{g_n} \Longrightarrow \frac{\sum_n f_n}{\sum_n g_n} = \sum_n \frac{f}{\sum_n g_n} \leq \sum_n \frac{f_n}{g_n}$$

Lemma 2. *Given the model defined by:*

$$p(\theta, \sigma, \xi, \pi \mid \boldsymbol{x}) \propto p(\theta, \sigma, \xi, \pi) \prod_{j,b} p(\boldsymbol{x}_b^j \mid \theta_b, \sigma_b, \xi_b, \pi_j), \tag{7}$$

if the likelihood with two sub-trajectories is in average better than the one with one trajectory, then

$$\frac{\sum_n p(x_b^j \mid \theta_b^{1:2}) - \Big(p(x_b^j \mid \theta_b^0) - p(x_b^j \mid \theta_b^{1:2})\Big)\,\xi_b^{(k-1)}}{\sum_n p(x_b^j \mid \theta_b^{1:2}) + \Big(p(x_b^j \mid \theta_b^0) - p(x_b^j \mid \theta_b^{1:2})\Big)\,\xi_b^{(k-1)}} > 1$$

Proof.

$$\begin{aligned}
&\sum_n p(x_b^j \mid \theta_b^{1:2}) > \sum_n p(x_b^j \mid \theta_b^0) \\
&\Leftrightarrow \sum_n -2p(x_b^j \mid \theta_b^0) + 2p(x_b^j \mid \theta_b^{1:2}) > 0 \\
&\Leftrightarrow \sum_n -p(x_b^j \mid \theta_b^0) + p(x_b^j \mid \theta_b^{1:2}) - p(x_b^j \mid \theta_b^0) + p(x_b^j \mid \theta_b^{1:2}) > 0 \\
&\Leftrightarrow \sum_n p(x_b^j \mid \theta_b^{1:2}) - \Big(p(x_b^j \mid \theta_b^0) - p(x_b^j \mid \theta_b^{1:2})\Big)\,\xi_b^0 - p(x_b^j \mid \theta_b^{1:2}) - \Big(p(x_b^j \mid \theta_b^0) - p(x_b^j \mid \theta_b^{1:2})\Big)\,\xi_b^0 > 0
\end{aligned}$$

from which we obtain the thesis.

Theorem 1. *Given the model (7), if the likelihood with two sub-trajectories is in average better than the one with one trajectory, $\xi^{(k-1)} \in [0,1]$, $\xi^{(k)}$ is given by (3), and $\beta \in [0, N]$, then $\xi_b^{(k)} \in [0,1]$.*

Proof. We start proving that $0 \le \beta \le N$ ensures $\xi_b^{(k)} \ge 0$. We observe that by equation (3) we have that

$$\xi_b^{(k)} \ge 0 \Leftrightarrow \frac{\sum_n \gamma_b^n}{N + (\xi_b^{(k-1)} - 1)\beta} \ge 0 \Leftrightarrow N + (\xi_b^{(k-1)} - 1)\beta \ge 0 \Leftrightarrow \beta \le \frac{N}{1 - \xi_b^{(k-1)}};$$

therefore, being $N \le N/(1 - \xi_b^{(k-1)})$, we obtain the sufficient condition $\beta \le N$.

We now prove that $0 \le \beta \le N$ ensures $\xi_b^{(k)} \le 1$.

We observe again that by equation (3) we have that

$$\begin{aligned}
\xi_b^{(k)} \le 1 \Leftrightarrow \frac{\sum_n \gamma_b^n}{N + (\xi_b^{(k-1)} - 1)\beta} \le 1 &\Leftrightarrow \beta \le \frac{1}{1 - \xi_b^{(k-1)}} \sum_n (1 - \gamma_b^n) \\
&\Leftrightarrow \beta \le \frac{1}{1 - \xi_b^{(k-1)}} \sum_n \frac{p(x_b^j \mid \theta_b^{1:2})(1 - \xi_b^{(k-1)})}{p(x_b^j)} \\
&\Leftrightarrow \beta \le \sum_n \frac{p(x_b^j \mid \theta_b^{1:2})}{p(x_b^j)}
\end{aligned}$$

We observe that the upper bound can be written as:

$$\begin{aligned}
\sum_n \frac{p(x_b^j \mid \theta_b^{1:2})}{p(x_b^j)} &= \sum_n \frac{p(x_b^j \mid \theta_b^{1:2})}{p(x_b^j \mid \theta_b^{1:2})(1-\xi_b^{(k-1)}) + p(x_b^j \mid \theta_b^0)\xi_b^{(k-1)}} \\
&= \sum_n \frac{p(x_b^j \mid \theta_b^{1:2})}{p(x_b^j \mid \theta_b^{1:2}) + \left(-p(x_b^j \mid \theta_b^{1:2}) + p(x_b^j \mid \theta_b^0)\right)\xi_b^{(k-1)}} \\
&= \sum_n \frac{p(x_b^j \mid \theta_b^{1:2}) \pm \left(p(x_b^j \mid \theta_b^0) - p(x_b^j \mid \theta_b^{1:2})\right)\xi_b}{p(x_b^j \mid \theta_b^{1:2}) + \left(p(x_b^j \mid \theta_b^0) - p(x_b^j \mid \theta_b^{1:2})\right)\xi_b^{(k-1)}} \\
&\geq \frac{\sum_n p(x_b^j \mid \theta_b^{1:2}) \pm \left(p(x_b^j \mid \theta_b^0) - p(x_b^j \mid \theta_b^{1:2})\right)\xi_b^{(k-1)}}{\sum_n p(x_b^j \mid \theta_b^{1:2}) + \left(p(x_b^j \mid \theta_b^0) - p(x_b^j \mid \theta_b^{1:2})\right)\xi_b^{(k-1)}} \\
&= N + \frac{\sum_n p(x_b^j \mid \theta_b^{1:2}) - \left(p(x_b^j \mid \theta_b^0) - p(x_b^j \mid \theta_b^{1:2})\right)\xi_b^{(k-1)}}{\sum_n p(x_b^j \mid \theta_b^{1:2}) + \left(p(x_b^j \mid \theta_b^0) - p(x_b^j \mid \theta_b^{1:2})\right)\xi_b^{(k-1)}} \\
&> N
\end{aligned}$$

where the inequalities come from the previous Lemma 1 and Lemma 2.

This means that the bound for β holds if the model with two Sigmoids is in general better than the one with one Sigmoid. This requirements is reasonable.

Therefore, if the ratio of sums on the right hand side is positive, it is true that $\beta < N$ is a good bound.

2 PPMI analysis

Parkinson's disease (PD) is a progressive neurodegenerative disorder affecting approximately 1% of the global population, making it the second most common neurodegenerative disease after Alzheimer's disease (AD) [**?**]. Externally, PD manifests with a wide variety of symptoms that can differ significantly from person to person; common symptoms include tremors, slowed movement (bradykinesia), rigid muscles, impaired posture and balance, loss of automatic movements, changes in speech, and writing difficulties.

Analysing DP-MoSt's solution on PPMI dataset, we provide the probability of a split across clinical scores. We also show the estimated trajectories with solid coloured lines, and show subjects based on their estimated sub-population. We observe that the solution strongly depends on the choice of the prior parameters:

- Low prior parameters: when low prior parameters are considered, the method tends towards overfitting, resulting in a high probability for all biomarkers to present a split. Despite the similarities in the distribution of the sub-populations to that of SuStaIn, with one sub-population being significantly

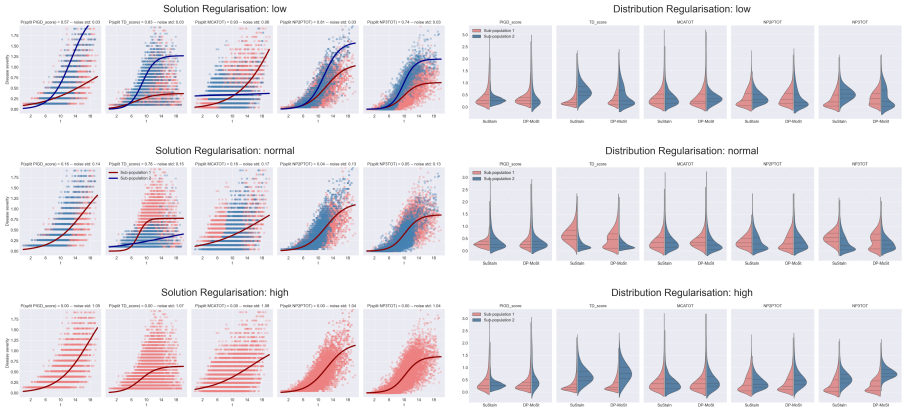

Fig. 2. The figure illustrates by each row the results obtained with different order of magnitude for the prior parameters of DP-MoSt. The first column shows biomarrker trajectories with estimated sub-populations, while the second columnshows the comparison between the biomarkers values between DP-MoSt and SuStaIn.

more populated than the other, we cannot appreciate any similarity in trajectories. The DP-MoSt's population associated with lower *TD_score* is the same that is related with lower *PIGD_score*, in clear contrast with the trajectory provided by SuStaIn. The differences in biomarker trajectories can be explained by the fact that our method includes longitudinal information, making the trajectory evaluation more consistent.

– Normal prior parameters: when normal prior parameters are considered, DP-MoSt associates high split probability to *TD_score*. DP-MoSt identifies a sub-population with a high percentage of TD subjects (76%) and another with a higher percentage of PIGD subjects (62%), providing a clinically meaningful partitioning of the subjects.
– High prior parameters: When high prior parameters are used, our method is less inclined to stratify the data into sub-populations, leading to no split across biomarkers and resulting in a solution that essentially performs like a logistic regression.

	DP-MoSt low		**DP-MoSt normal**		**SuStaIn**	
Condition	Sub-pop 1	Sub-pop 2	Sub-pop 1	Sub-pop 2	Sub-pop 1	Sub-pop 2
Intermediate	0.64	0.36	0.56	0.44	0.84	0.16
PIGD	0.83	0.17	0.38	**0.62**	0.78	0.22
TD	0.69	0.31	**0.76**	0.24	0.48	0.52
N data	73%	27%	56%	44%	72%	28%

Table 1. The Table shows the subdivision between different sub-populations considering two different regularisation parameters for DP-MoSt (low and normal) and SuStaIn.