Multiple orderings of events in disease progression

No Author Given

No Institute Given

Abstract. The event-based model allows a discrete picture of disease progression to be constructed from cross-sectional data sets, with each event corresponding to a new biomarker becoming abnormal. However, it relies on the assumption that all subjects follow a single event sequence. This is a major simplification for sporadic disease data sets, which are highly heterogeneous, include distinct subgroups, and contain significant proportions of outliers. In this work we relax this assumption by considering two extensions to the event-based model: a generalised Mallows model, which allows subjects to deviate from the main event sequence, and a Dirichlet process mixture of generalised Mallows models, which models clusters of subjects that follow different event sequences, each of which has a corresponding variance. We develop a Gibbs sampling technique to infer the parameters of the two models from multimodal biomarker data sets. We apply our technique to data from the Alzheimer's Disease Neuroimaging Initiative to determine the sequence in which brain regions become abnormal in sporadic Alzheimer's disease, as well as the heterogeneity of that sequence in the cohort. Our results suggest that fitting a generalised Mallows model provides a more realistic estimation of the variance in the event sequence across subjects. Fitting a Dirichlet process mixture reduces this variance by detecting an outlier cluster. The Gibbs samples additionally provide an estimate of the uncertainty in each of the model parameters, for example an individual's latent disease stage and cluster assignment. The distributions and mixtures of sequences that this new family of models introduces offer better characterisation of disease progression of heterogeneous populations, new insight into disease mechanisms, and have the potential for enhanced disease stratification and differential diagnosis.

1 Introduction

The sequence in which biomarkers become abnormal provides a simple, intuitive description of disease progression, providing insights into the underlying disease biology and a potential mechanism for disease staging and differential diagnosis. The sequence of biomaker abnormality in sporadic neurodegenerative diseases, e.g. Alzheimer's disease, has been a topic of intense debate amongst neurologists (Jack et al.). Reconstructing this sequence for sporadic neurodegenerative diseases is difficult because the position of subjects with respect to the full disease time course is unknown. Typically clinical diagnoses are used as a time proxy,

but this limits the temporal resolution of the sequence, e.g in Alzheimer's disease there are only three stages: cognitively normal, mild cognitive impairment and Alzheimer's disease. Additional complications arise due to the long disease time course, thought to span several decades, and the inherent heterogeneity of sporadic disease datasets. Many different factors contribute to this heterogeneity, for example genetic disease subtypes, mixed pathology, environmental factors, and misdiagnosed subjects.

The event-based model (Fonteijn et al.) considers disease progression as a series of events, where each event corresponds to a new biomarker becoming abnormal. By considering cross-sectional patient data as snapshots of a single common biomarker abnormality event sequence, the event-based model is able to probabilistically reconstruct the ordering of events across subjects, without relying on a-priori disease staging. Taking samples of the posterior probability of this sequence provides insight into the uncertainty in this single event ordering across the population. The application of this model has been demonstrated in familial Alzheimer's disease and Huntington's disease (Fonteijn et al.) to determine the sequence in which regional brain volumes become abnormal, and in sporadic Alzheimer's disease to determine the sequence in which cerebrospinal fluid (CSF) markers, cognitive test scores, and a limited set of regional atrophy and brain volume biomakers become abnormal (Young et al.). Young et al. found that this biomarker abnormality sequence is different in APOE4 positive individuals, who have an increased genetic risk of sporadic Alzheimer's disease, compared to the whole population, suggesting that the whole population contains a proportion of subjects who do not follow the single ordering of events encoded by the event-based model.

The assumption made by the event-based model of a single ordering of events in all subjects is a major simplification for heterogeneous sporadic disease datasets. In this work we relax this assumption by considering a family of models that allow for multiple orderings of events. The first is a generalised Mallows model, which parameterises the variance in the single ordering, allowing subjects to deviate from the central event sequence. The second is a Dirichlet process mixture model, which allows for subject subgroups that follow different event sequences. We build on the fitting techniques of Huang et al. for generalised Mallows models and Meila et al. for Dirichlet process mixtures of generalised Mallows models to develop tractable inference algorithms for large sets of events. We apply these models to determine the sequence in which FDG-PET, CSF markers, cognitive test scores and regional brain volumes become abnormal. We consider a much more extensive set of regional volumes than have been seen previously for sporadic Alzheimer's disease.

2 Models

2.1 The event-based model

The event-based model of disease progression consists of a set of events $\{e_1, \ldots, e_N\}$ and an ordering $\sigma = (\sigma(1), \ldots, \sigma(N))$, where $\sigma(k) = i$ means that event e_i occurs in position k with respect to σ. The occurence of event e_i in subject j is informed by a corresponding biomarker measurement x_{ij}. In practise we only observe a snapshot of the event sequence for each subject. This corresponds to a partition of the event set, or partial ranking, $\gamma_k = e_{\sigma(1)}, \ldots, e_{\sigma(k)} | e_{\sigma(k+1)}, \ldots, e_{\sigma(N)}$, where the vertical bar indicates that the first set of events precedes the second. If a subject is at stage k in the sequence σ the events $e_{\sigma(1)} \ldots e_{\sigma(k)}$ have occurred and events $e_{\sigma(k+1)} \ldots e_{\sigma(N)}$ have yet to occur. The generative model of the biomarker data is

$$k_j \sim P(k)$$

$$x_{\sigma(i),j} \sim p(x_{\sigma(i),j} | e_{\sigma(i)}) \text{ if } i \leq k_j$$

$$x_{\sigma(i),j} \sim p(x_{\sigma(i),j} | \neg e_{\sigma(i)}) \text{ otherwise}$$

where $p(x|e)$ and $p(x|\neg e)$ correspond to the probability of observing biomarker measurement x given that event e has or has not occurred respectively, and $P(k)$ is a prior on the disease stage k.

2.2 The generalised Mallows event-based model

We formulate the generalised Mallows event-based model by using a generalised Mallows prior to parameterise the variance in a central event sequence π through the spread parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{N-1})$. For this model each subject has their own latent ordering σ_j, which is assumed to be a sample from a generalised Mallows model. The generative model of the biomarker data in the event-based model is therefore preceded by

$$\pi, \boldsymbol{\theta} \sim P(\pi, \boldsymbol{\theta} | \nu, \boldsymbol{r})$$

$$\sigma_j \sim GM(\pi, \boldsymbol{\theta})$$

where $GM(\pi, \boldsymbol{\theta})$ is a generalised Mallows distribution, and $P(\pi, \boldsymbol{\theta} | \nu, \boldsymbol{r})$ is a conjugate prior over the generalised Mallows distribution parameters of the form $P(\pi, \boldsymbol{\theta} | \nu, \boldsymbol{r}) \propto e^{-\nu \sum_j [\theta_j r_j + \ln \psi_{n-j}(\theta_j)]}$, with $\psi_{n-j}(\theta_j) = \frac{1 - e^{-(n-j+1)\theta_j}}{1 - e^{-\theta_j}}$.

2.3 Dirichlet process mixtures of generalised Mallows event-based models

Dirichlet process mixtures of generalised Mallows models assume that each subject has their own central ordering π_j and spread parameters $\boldsymbol{\theta}_j$, which are

sampled from a distribution G that is drawn from a Dirichlet process. The generative model of the biomarker data in the event-based model is now preceded by the process

$$G \sim DP(\alpha, P(\pi, \boldsymbol{\theta}|\nu, \boldsymbol{r}))$$

$$\pi_j, \boldsymbol{\theta}_j \sim G$$

$$\sigma_j \sim GM(\pi_j, \boldsymbol{\theta}_j)$$

where $DP(\alpha, P(\pi, \boldsymbol{\theta}|\nu, \boldsymbol{r}))$ is a Dirichlet process. Each data point π_j can be characterised by an association with a cluster label $c_j \in 1, \ldots, C$ and each cluster c with a set of generalised Mallows parameters σ_c and $\boldsymbol{\theta}_c$.

3 Inference

3.1 The event-based model

Inference in the event-based model can be performed by taking Markov Chain Monte Carlo (MCMC) samples of $P(\sigma|X) = \frac{P(X|\sigma)P(\sigma)}{P(X)}$ where

$$P(X|\sigma) = \prod_{j=1}^{J} \left[\sum_{k=0}^{K} P(k) \left(\prod_{i=1}^{k} p(x_{\sigma(i),j}|e_{\sigma(i)}) \prod_{i=k+1}^{N} p(x_{\sigma(i),j}|\neg e_{\sigma(i)}) \right) \right] \quad (1)$$

3.2 The generalised Mallows event-based model

We use Gibbs sampling to infer the parameters of the generalised Mallows event-based model. This consists of two stages. First, generating a set of sample event sequences $\sigma_{1:J}$. We sample from an augmented model (Huang et al.), by alternating between sampling a subject's ordering σ_j and disease stage k_j, which are used to deterministically reconstruct their partial ranking γ_j. The Gibbs sampling updates are therefore

$$\sigma^{(j)} \sim P(\sigma|\boldsymbol{\gamma} = \gamma_j, \pi, \boldsymbol{\theta})$$

$$k^{(j)} \sim P(k|\boldsymbol{\sigma} = \sigma_j, X_j)$$

Second, sampling the model parameters given the set of sample orderings $\sigma_{1:J}$ using the updates

$$\pi \sim P(\pi|\boldsymbol{\theta}, \nu, \boldsymbol{r}, \sigma_{1:J})$$

$$\theta_k \sim P(\theta_k|\pi, \nu, \boldsymbol{r}, \sigma_{1:J})$$

3.3 Dirichlet process mixtures of generalised Mallows event-based models

We formulate another Gibbs sampler to infer the parameters of Dirichlet process mixtures of generalised Mallows event-based models. We generate a set of candidate sample orderings $\sigma_{1:J,1:C}$, disease stages $k_{1:J,1:C}$, and partial rankings $\gamma_{1:J,1:C}$, which are conditioned on the parameters for each cluster via the updates

$$\sigma^{(j,c)} \sim P(\sigma|\boldsymbol{\gamma} = \gamma_{jc}, \pi_c, \boldsymbol{\theta}_c)$$

$$k^{(j,c)} \sim P(k|\boldsymbol{\sigma} = \sigma_{jc}, X_j)$$

From these samples we sample the cluster assignment c_j of each subject conditioned on the cluster assignments of the other subjects c_{-j}, where c_{-j} is the set of cluster assignments for all subjects except subject j, the subject's sample ordering for each cluster $\sigma_{j,1:C}$, disease stage $k_{j,1:C}$ and their biomarker data X_j. We then update the generalised Mallows model parameters for each cluster, π_c and $\boldsymbol{\theta}_c$, from the set of subject orderings assigned to each cluster, $\boldsymbol{\sigma}_c$. So we have the updates

$$c^{(j)} \sim P(c|c_{-j}, \sigma_{j,1:C}, \boldsymbol{\theta}, \alpha, \nu, \boldsymbol{r}, X_j, k_{j,1:C})$$

$$\pi^{(c)} \sim P(\pi|\boldsymbol{\theta} = \boldsymbol{\theta}_c, \nu, \boldsymbol{r}, \boldsymbol{\sigma}_c)$$

$$\theta_k^{(c)} \sim P(\theta_k|\boldsymbol{\pi} = \pi_c, \nu, \boldsymbol{r}, \boldsymbol{\sigma}_c)$$

4 Implementation

4.1 ADNI dataset

We calculated regional brain volumes for 42 regions for the set of 382 subjects (135 cognitively normal, 149 mild cognitive impairment, 98 Alzheimer's disease) that had a 1.5T T1 scan at baseline. Stuff about neuromorphometrics atlas. Summed left and right brain regions. Normalised by regressing against TIV. Selected subset of regions with significant differences between cognitively normal and Alzheimer's disease subjects. Write about CSF and FDG.

4.2 Model fitting

We compare the result of fitting the event-based model, generalised Mallows event-based model and Dirichlet process mixtures of generalised Mallows event-based models to the ADNI data set. Follwing previous work (Fonteijn et al.) we model the probability that a biomarker is normal, $p(x|\neg e)$, as a Gaussian distribution, and the probability that a biomarker is abnormal, $p(x|e)$, as a uniform distribution to reflect the range of severity that corresponds to an abnormal biomarker, and to allow for a small proportion of subjects whose regional brain volumes are abnormal but look normal on a population-wide level. We use a

mixture model to fit these distributions to the data to account for a proportion of outliers in the control population. We fix the uniform component to cover the range of the biomarker values. In subjects that had missing data points we imputed the biomarker values such that $p(x|e) = p(x|\neg e)$, i.e. it is equally probable that the event e has or has not occurred. The prior probability that a subject is at a particular disease stage $P(k)$ is assumed to be uniform. To fit the generalised Mallows model we need to sample σ from $P(\sigma|\gamma, \pi, \boldsymbol{\theta})$. We approximate this by sampling from a generalised Mallows model for each of the event sets in the partial ranking γ separately; the set of events γ_e that have occurred and the set of events $\gamma_{\neg e}$ that have yet to occur. We sample

$$\sigma_e \sim GM(\pi_{\gamma_e}, \boldsymbol{\theta}_{\gamma_e})$$

$$\sigma_{\neg e} \sim GM(\pi_{\gamma_{\neg e}}, \boldsymbol{\theta}_{\gamma_{\neg e}})$$

This means that the precedence of events specified by the partial ranking is preserved, and that the central ordering of the generalised Mallows model for each event set, π_{γ_e} and $\pi_{\gamma_{\neg e}}$, has the minimal Kendalls tau distance from the central ordering π of the full generalised Mallows model. We sample k from $P(k|\sigma, X_j)$ using equation 1, i.e.

$$P(k|\sigma, X_j) \propto \prod_{i=1}^{k} p(x_{\sigma(i),j}|e_{\sigma(i)}) \prod_{i=k+1}^{N} p(x_{\sigma(i),j}|\neg e_{\sigma(i)})$$

The remaining sampling updates follow the work of Meila et al.. We sample π exactly using a stage-wise algorithm, and $\boldsymbol{\theta}$ using a beta function approximation. We update the Dirichlet process mixture model cluster assignments c_j and generalised Mallows model parameters π_c and $\boldsymbol{\theta}_c$ for each cluster using the Beta-Gibbs algorithm (Meila et al.). When updating the cluster assignments we calculate the probability that a subject belongs to each cluster given their sampled ordering for that cluster $\sigma_{j,c}$ as in Meila et al.. We weight this probability by $P(X_j|\sigma_{j,c}, k_{j,c})$. We fix the priors to be $\nu = 1$, $\boldsymbol{r} = \mathbf{1}$, $\alpha = 1$. We initialise the central ordering π of each sampler randomly. We initialise each subject's partial ranking γ such that γ_e is the set of events with $p(x|e) > p(x|\neg e)$ and $\gamma_{\neg e}$ is the set of events with $p(x|e) \leq p(x|\neg e)$. We initialise the Dirichlet process mixture of generalised Mallows event-based models to have 25 clusters.

5 Results and Discussion

5.1 The event-based model

We first estimated the central ordering of events from the event-based model. Figure 1 shows a positional variance diagram of the MCMC samples of the single event sequence returned by the event-based model. We find that CSF markers are the first to become abnormal, followed by cognitive test scores, then memory-related brain regions, then FDG-PET, and then other Alzheimer's

disease-related brain regions. This sequence complements the findings of other studies, but provides a much more detailed picture of the regional progression of volume changes than has been seen previously in sporadic Alzheimer's disease, and a direct comparison of the sequence of regional changes relative to a multimodal set of biomarkers.

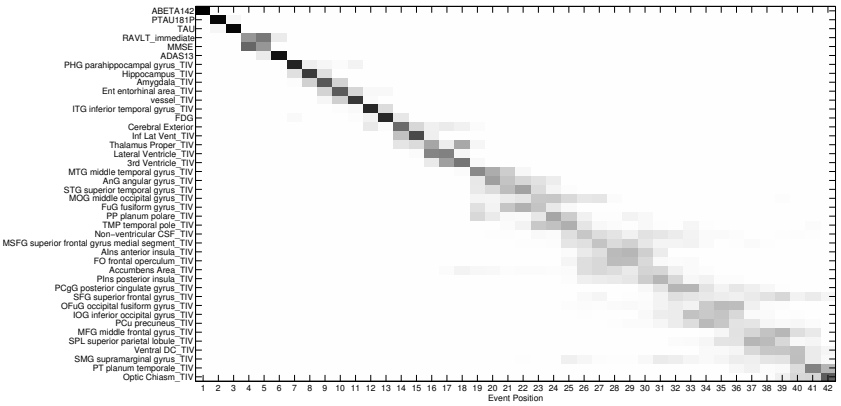

Fig. 1: Central ordering estimated by the event-based model: Positional variance diagram of the MCMC samples of the maximum likelihood event sequence σ. The events on the y-axis are ordered by the maximum likelihood sequence estimated by the model. Each entry of the positional variance diagram represents the proportion of samples in which a particular event appears in a particular position in the central ordering.

5.2 The generalised Mallows event-based model

The generalised Mallows event-based model estimates both the central ordering of the events across the population (Figure 2 - *figures 2 and 4 to be replaced with picture made by Raz), as well as the variance in this single event ordering (Figure 3). Figure 2 shows a positional variance diagram of the Gibbs samples of the central ordering π of the generalised Mallows event-based model, i.e. the uncertainty in the average ordering of events across the population. As you would expect, the central event sequence has a similar ordering and positional variance to the event-based model, the main difference being an increase in the positional variance of later Alzheimer's disease-related brain region events. However, the spread parameter $\boldsymbol{\theta}$ (Figure 3) estimated by the generalised Mallows model reveals that this central ordering has high variance (Figure 3), reflecting the uncertainty in the sequence of events followed by individuals given the het-

erogeneity in the data, and that only a cross-sectional snapshot of the progression in each subject is available.

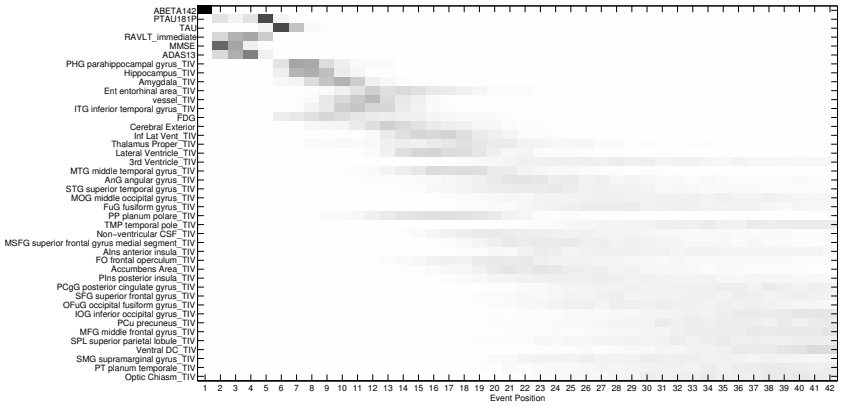

Fig. 2: Central ordering estimated by the generalised Mallows event-based model: Positional variance diagram of the Gibbs samples of the central ordering parameter π. The events on the y-axis are ordered by the maximum likelihood sequence in the event-based model to allow direct comparison with Figure 1. Each entry of the positional variance diagram represents the proportion of samples in which a particular event appears in a particular position in the central ordering.

5.3 Dirichlet process mixtures of generalised Mallows event-based models

We fitted a Dirichlet process mixture of generalised Mallows event-based models to allow for clusters of subjects that follow different sequences of events, of which each cluster has its own central ordering π_c and variance $\boldsymbol{\theta}_c$. The Dirichlet process mixture model identifies three main clusters in the data, with an average size of 189 ($\pm$ 5), 110 ($\pm$ 45), and 83 ($\pm$ 46) subjects over the Gibbs samples. Figure 4 (*figures 2 and 4 to be replaced with picture made by Raz) shows a positional variance diagram of the Gibbs samples of the central ordering of each cluster. The first two clusters look more Alzheimer's disease-like, with CSF Aβ_{1-42} being an early biomarker, whereas the third cluster likely captures outliers that do not fit the Alzheimer's disease sequence of events. The assignment of subjects to the first cluster is much more stable, with the second and third clusters seemingly representing more of an ad-mixture between outliers and Alzheimer's disease subjects. Fitting a mixture of sequences reduces the spread $\boldsymbol{\theta}$ (Figure 3), suggesting that the inclusion of outliers with different progression patterns contributes to the uncertainty in the event sequence. Our Gibbs sampling technique

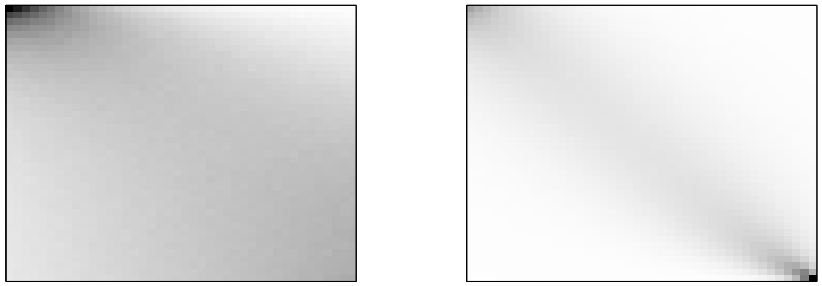

Fig. 3: Comparison of the spread parameter $\boldsymbol{\theta}$ for the generalised Mallows event-based model (shown on the left) and the Dirichlet process mixtures of generalised Mallows event-based models (shown on the right). An entirely black diagonal indicates zero variance, i.e. a single event sequence for all subjects. An entirely grey plot indicates maximal variance, i.e. total disagreement in the event sequence amongst subjects. $\boldsymbol{\theta}$ is only shown for one of the Dirichlet process mixture clusters as each cluster had a similar variance.

returns samples of all the model parameters for both the generalised Mallows event-based model and Dirichlet process mixture. For example, we are able to estimate the uncertainty in the disease stage of each subject for both models, and the cluster assignment of each subject from the Dirichlet process mixture.

6 Conclusions

We have fitted a family of disease progresssion models to regional imaging data from ADNI to determine the sequence of regional volume loss in sporadic AD. This sequence incorporates a much more extensive, multi-modal set of biomarkers than has been seen previously. We have formulated a Gibbs sampling algorithm that allows these progression models to be inferred for large event sets, as well as providing an estimate of the uncertainty on each model parameter. The generalised Mallows model shows that the variation in the central event sequence across the population is high. Fitting a Dirichlet process mixture reduces this variance by finding subject subgroups that follow more homogeneous event sequences.

The family of disease progression models we describe have wide potential further application to any disease or developmental process. The sampling techniques described naturally extend to incorporate multiple time points within an individual (Huang et al.). The multiple orderings of events described by these models have potential use for outlier detection, differential diagnosis and to characterise disease subtypes, e.g. genetic subtypes, for improved patient stratification.

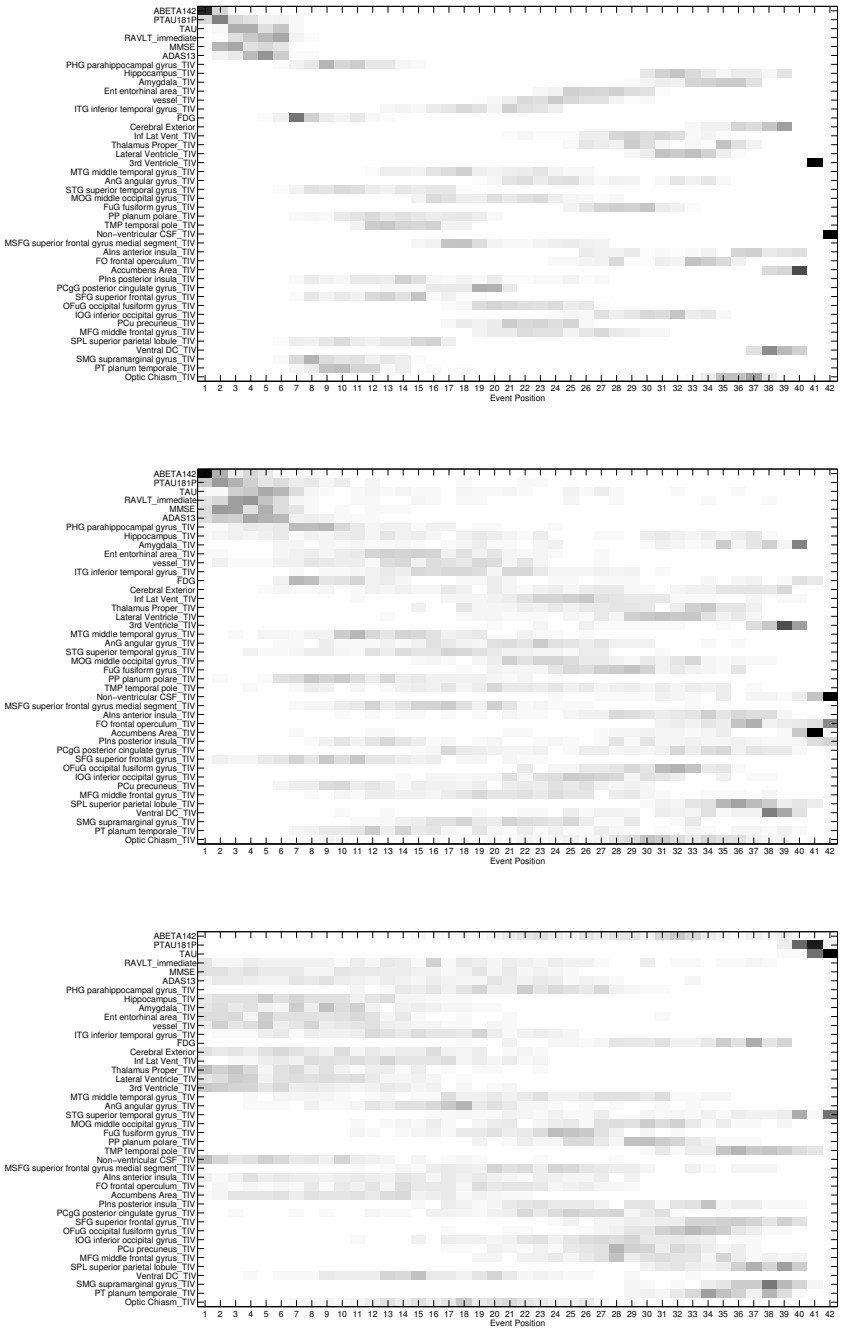

Fig. 4: Central ordering for each of the three clusters of the Dirichlet process mixtures of generalised Mallows event-based models: Positional variance diagram of the Gibbs samples of the central ordering parameter π_c for each cluster. The events on the y-axis are ordered by the maximum likelihood sequence in the event-based model to allow direct comparison with Figure 1. Each entry of the positional variance diagram represents the proportion of samples in which a particular event appears in a particular position in the central ordering.