Methods S1

In this supplementary material we will outline a probabilistic model-based clustering approach to functional parcellation as well as a way to translate group level results to individual subjects. The basis of this approach is a mixture model, i.e. we assume that the data is comprised of K clusters and each voxel belongs to one such cluster.

Likelihood

For each cluster k and time point t, the cluster mean μ_{kt} and precision τ_{kt} are drawn from a normal-gamma distribution

$$\mu_{kt}, \tau_{kt} \sim \mathcal{NG}(\mu_0, \kappa_0, a_0, b_0).$$

For every voxel i belonging to cluster k and time t, data point x_{it} is drawn from a normal distribution with mean μ_{kt} and precision τ_{kt}

$$x_{it} \sim \mathcal{N}(\mu_{kt}, \tau_{kt}^{-1}).$$

For notational convenience, let $\boldsymbol{\theta} = [\mu_0, \kappa_0, a_0, b_0]$ denote the vector of hyperparameters of the normal-gamma prior. Let

$$\begin{aligned} \kappa_k &= \kappa_0 + n_k \\ a_k &= a_0 + n_k/2 \\ b_{kt} &= b_0 + \frac{\sum_{i=1}^{n_k}(x_{it} - \overline{x}_{kt})^2}{2} + \frac{\kappa_0 n_k (\overline{x}_{kt} - \mu_0)^2}{2(\kappa_0 + n_k)}, \end{aligned}$$

where n_k is the number of elements in cluster k, and $\overline{x}_{kt}$ is the empirical mean of cluster k at time t. The marginal likelihood of the data in a cluster k (integrating out μ_{kt} and τ_{kt}) is given by:

$$\begin{aligned} P(\mathbf{X}_k | k, \boldsymbol{\theta}) &= \prod_{t=1}^{T} P(\mathbf{x}_k | k, \boldsymbol{\theta}) \\ &= \prod_{t=1}^{T} \frac{\Gamma(a_0) b_0^{a_0}}{\Gamma(a_k) b_{kt}^{a_k}} \left(\frac{\kappa_0}{\kappa_k (2\pi)^{n_k}} \right)^{\frac{1}{2}}, \end{aligned} \quad (1)$$

where $\mathbf{X}_k$ and $\mathbf{x}_t$ represent data from voxels assigned to cluster k for all and individual timepoints respectively. The full data likelihood is then simply the product of cluster likelihoods.

Because each timepoint has its own parameters c_{kt} and τ_{kt}, this model is hierarchical over timepoints. This allows the model to adapt to noise that is specific in time and/or space. It also means that group analyses can be performed simply by temporally

concatenating data, without any assumptions on a global noise level over subjects. It should be noted that the c_{kt}'s are assumed to be independent; that is, they are coupled only in the sense that the same parcellation must hold for all of them. Although functional signals contain autocorrelations, this is a useful approximation to maintain computational feasibility, as modelling these temporal dependencies require the estimation of $T \times T$ covariance matrices.

Prior

In this section we introduce the prior on clustering. Let $\boldsymbol{\pi}$ denote a vector to encode this parcellation such that $\pi_i = k$ if voxel i is part of cluster k. Every π_i can then be drawn from a K-dimensional multinomial distribution, but this would require specifying the number of clusters K. This involves either an arbitrary choice or some form of model selection to motivate the number of clusters. Bayesian non-parametric models steer clear of this issue by letting the model complexity, the number of clusters in this case, be determined by the data. This is a key advantage of our approach over clustering methods commonly applied in neuroscience. For an introduction to this type of model see (Gershman and Blei 2012).

The standard prior on $\boldsymbol{\pi}$ for a non-parametric clustering model would be the Chinese Restaurant Process (CRP). The analogy for the CRP is that each cluster is one of an infinite number of tables in a restaurant and customers (voxels, in our case) sit down at a random table. The probability to join any non-empty table is proportional to the number of customers already sitting at that table and proportional to the concentration parameter α for an empty table.

The CRP prior assumes exchangeability, that is, it does not matter in what order customers enter the restaurant. We would like to incorporate a spatial constraint such that clusters are contiguous, however, which violates this assumption. The distance-dependent Chinese Restaurant Process (dd-CRP; Blei and Frazier 2011) is a generalization of the CRP that allows non-exchangeable elements. In the analogy for the dd-CRP, each customer picks another customer and joins their table with probability inversely proportional to the distance between them. Customers can also join themselves with probability equal to the concentration parameter of the dd-CRP, in which case they start a new table.

The probability of a set of customer assignments $\boldsymbol{\lambda}$, where λ_i identifies with whom customer i sits, is given by:

$$P(\boldsymbol{\lambda}|\mathbf{A}) = \prod_{i=1}^{N} P\left(\lambda_i = j \middle| a_{ij}\right) = \prod_{i=1}^{N} \frac{a_{ij}}{\sum_{n=1}^{N} a_{in}}, \quad (2)$$

where, for notational convenience, $\mathbf{A}$ denotes a matrix incorporating both the inverse distances between customers (off-diagonal elements), as well as the concentration parameter (diagonal elements).

Finally, let $\boldsymbol{\pi}(\boldsymbol{\lambda})$ denote a partition that is determined by $\boldsymbol{\lambda}$, i.e. $\pi_i = \pi_j$ if node i can reach j by traversing customer links or vice versa. Combining likelihood and prior gives a posterior of the form:

$$P(\boldsymbol{\lambda}|\mathbf{X}, \boldsymbol{\theta}, \mathbf{A}) \propto P\big(\mathbf{X}\big|\boldsymbol{\theta}, \boldsymbol{\pi}(\boldsymbol{\lambda})\big)P(\boldsymbol{\lambda}|\mathbf{A}). \quad (3)$$

Single subject parcellations

To obtain single subject parcellations that represent the $K \times K$ group level covariance matrix $\boldsymbol{\Sigma}$, we fix the number of clusters to K and combine the Gaussian observation model from Eq. (1) with a multivariate normal prior for the subject-specific cluster time courses as

$$P(\mathbf{X}, \boldsymbol{\mu}|\boldsymbol{\Sigma}, \tau) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{x}_t|\mathbf{Z}\boldsymbol{\mu}_t, \tau^{-1}\mathbf{I}_N)\mathcal{N}(\boldsymbol{\mu}_t|0, \boldsymbol{\Sigma}),$$

where $\mathbf{x}_t$ is a $N \times 1$ vector containing the measured data from time instant t, $\boldsymbol{\mu}_t$ is a $K \times 1$ vector containing the cluster means at time t, and each row of the binary $N \times K$ matrix $\mathbf{Z}$ has only one non-zero entry indicating the cluster assignment of the corresponding voxel. To facilitate the inference on the cluster assignments $\mathbf{Z}$ we integrate $\boldsymbol{\mu}_t$ to obtain the marginal likelihood

$$P(\mathbf{X}|\mathbf{Z}, \boldsymbol{\Sigma}, \tau) = \prod_{t=1}^{T} \left(\frac{2\pi}{\tau}\right)^{\frac{-N}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}}|\mathbf{S}_t|^{-\frac{1}{2}} \exp\left(\frac{\tau^2}{2}\mathbf{x}_t^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{S}_t\mathbf{Z}\mathbf{x}_t - \frac{\tau}{2}\mathbf{x}_t^{\mathrm{T}}\mathbf{x}_t\right), \quad (4)$$

where the conditional covariance of $\boldsymbol{\mu}_t$ is given by $S_t = (\tau\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \boldsymbol{\Sigma}^{-1})^{-1}$. The key difference to the marginal likelihood used in dd-CRP is that now the cluster mean time courses are spatially coupled according to the group level covariance.

Since K is fixed, a standard Dirichlet-multinomial prior can be assigned to the cluster assignments:

$$P(\mathbf{Z}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(N + \sum_{k=1}^{K}\alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)} \quad (5)$$

where $\alpha_1, \ldots, \alpha_K$ are the concentration parameters and n_k is the number of voxels assigned to cluster k, i.e., the sum of the elements in the k:th column of $\mathbf{Z}$. The concentration paremeters were set to $\alpha_k = 1$. A gamma prior $\tau \sim \mathcal{G}(a_0, b_0)$ with a shape parameter $a_0 = 2$ and a scale parameter $b_0 = 1$ was assigned to the overall noise level τ, in keeping with the parameter settings of the full model.

Probabilistic inference

In order to estimate the posterior (Eq. (3)), we made use of Gibbs sampling (Geman and Geman 1993). The Gibbs sampler is a Markov chain Monte Carlo approach which works by cycling through all elements and reassigning them according to the full conditionals.

Full model

Having integrated out the cluster timecourses and precisions, all that remains to be sampled is the list of voxel assignments. In each step we sweep over all voxels in random order and resample their connections (customer assignments) by removing the current assignment and choosing a new assignment conditioned on the resulting partitioning as outlined in (Blei and Frazier 2011). The sampling scheme of the dd-CRP has the added benefit that resampling the link allows large moves to be made, which should benefit convergence. For the prior, we can use Eq. (2) and for the likelihood we can use Eq. (1) for clusters affected by the resampling. The conditional probability of a new link is then

$$P(\lambda_i = j|\boldsymbol{\lambda}_{-i}, \mathbf{X}, \boldsymbol{\theta}, \mathbf{A}) \propto P(\mathbf{X}|\boldsymbol{\pi}(\boldsymbol{\lambda}'), \boldsymbol{\theta}) P(\lambda_i = j|\mathbf{A}),$$

where $\boldsymbol{\lambda}_{-i}$ is the vector of customer assignments, disregarding the ith voxel, $\boldsymbol{\pi}(\boldsymbol{\lambda}')$ is the parcellation that follows from resampling λ_i. Given our specification of $\mathbf{A}$ and the factorisation of our likelihood, this reduces to

$$P(\lambda_i = j|\boldsymbol{\lambda}_{-i}, \mathbf{X}, \boldsymbol{\theta}, \mathbf{A}) \propto \begin{cases} \dfrac{P(\mathbf{X}_{k\cup l}|\boldsymbol{\theta})}{P(\mathbf{X}_k|\boldsymbol{\theta})P(\mathbf{X}_l|\boldsymbol{\theta})} & \text{If the link joins clusters } k \text{ and } l \\ 1 & \text{otherwise.} \end{cases}$$

In order to speed up convergence we adopted a three step scheme inspired by the evidence accumulation clustering framework (Fred and Jain 2005). First, we ran 1000 parallel sampling chains, initialised with every voxel in its own cluster, for 30 steps each and saved the final sample. These chains quickly converge to some local minimum in the tails of the posterior distribution and the second step is to get some estimate of the centre of this distribution. For this second step, we split the samples into 20 sets of $S = 50$ samples and applied average linkage agglomerative hierarchical clustering to the voxels where, for each set, the distance q_{ij} between voxels i and j was

$$q_{ij} = 1 - S^{-1} \sum_{s=1}^{S} \mathbf{1}_{\pi_i^{(s)} = \pi_j^{(s)}}.$$

The resulting dendrograms were cut at an average distance of 0.5. In order to finetune the posterior estimates we used these parcellations to initialize a new set of sampling runs. Each run was intitialised to a parcellation by drawing, conditioned on that parcellation, a random customer assignment from the prior and was subsequently run for 100 steps. These samples were pooled across chains and our maximum a posteriori (MAP) estimate was the sample from this pool with the highest posterior probability.

Single subject model

The Gibbs sampling was conducted by drawing the cluster assignments $\mathbf{z}_i$ one by one for $i = 1, \ldots, N$ conditioned on the remaining assignments $\mathbf{Z}_{-i}$ and the noise level τ using the conditional density

$$P(\mathbf{z}_i|\mathbf{Z}_{-i}) \propto P(\mathbf{X}|\mathbf{Z}, \boldsymbol{\Sigma}, \tau) P(\mathbf{Z}|\boldsymbol{\alpha}), \qquad (6)$$

where the likelihood is given by (4) and the prior by (5). After each sweep over the cluster assignments $i = 1, \ldots, N$, the noise level was sampled by first drawing the cluster means for $t = 1, \ldots, T$ from

$$P(\boldsymbol{\mu}_t | \mathbf{Z}, \tau) = \mathcal{N}(\mathbf{m}_t, \mathbf{S}_t),$$

where $\mathbf{S}_t = (\tau \mathbf{Z}^{\mathrm{T}} \mathbf{Z} + \boldsymbol{\Sigma}^{-1})^{-1}$ and $\mathbf{m}_t = \tau \mathbf{S}_t \mathbf{Z}^{\mathrm{T}} \mathbf{x}_t$. Subsequently, τ was drawn from

$$P(\tau | \boldsymbol{\mu}_t, \mathbf{Z}) = \mathcal{G}\left(a^0 + \frac{1}{2} NT, b^0 + \frac{1}{2} \sum_{t=1}^{T} \| \mathbf{x}_t - \mathbf{Z} \boldsymbol{\mu}_t \|^2 \right)$$

The marginal likelihood term (4) that is required for evaluating (6) can be computed efficiently by pre-computing $\mathbf{S}_t$ and $|\mathbf{S}_t|$ before each sweep over the voxels and subsequently updating them using rank-1 computations that result from the single-row adjustments of $\mathbf{Z}$.

References

Blei DM, Frazier PI. 2011. Distance dependent Chinese restaurant processes. J Mach Learn Res. 12:2461–2488.

Fred ALN, Jain AK. 2005. Combining Multiple Clusterings Using Evidence Accumulation. IEEE Trans Pattern Anal Mach Intell. 27:835–850.

Geman S, Geman D. 1993. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. J Appl Stat. 20:25–62.

Gershman SJ, Blei DM. 2012. A tutorial on Bayesian nonparametric models. J Math Psychol. 56:1–12.